

ESTUDIOS SOBRE DENGUE: EXPERIENCIAS Y PERSPECTIVAS

Grupo de modelación en salud pública-GMSP



Editores:
Miguel G. Valdés
Carlos N. Bouza
José M. Sautto
Sira Allende
Agustín Santiago



Universidad Autónoma de Guerrero
Universidad de la Habana

Grupo de Modelación en Salud Pública-GMSP.

Estudios sobre dengue: Experiencias y
perspectivas

Editores:

Miguel G. Valdés

Carlos N. Bouza

José M. Sautto

Sira Allende

Agustín Santiago

Modelación de Dengue

Acapulco, Guerrero, México 2012

Primera edición. Marzo del 2012.
Universidad Autónoma de Guerrero y Universidad de la Habana.
Derechos Reservados

Autores: A. Santiago, C. N. Bouza, J. Legorreta, J.M. Sautto, J.G. Villavicencio, M.G. Valdés, O. Juárez,
S. Allende, S. López. Grupo de Modelación en Salud Pública - GMSP

ISBN: 978-607-7760-61-0 (Versión impresa)

ISBN: 978-959-16-1559-6 Edición Digital Conjunta Universidad de La Habana y Universidad Autónoma de Guerrero.

Editorial Universitaria.

Impreso en La Habana y Mexico/Printed in Habana and México

El contenido de este libro es responsabilidad de los autores.

La presentación y disposición en conjunto de este libro son propiedad de los editores.

Ninguna parte de esta obra puede ser reproducida o transmitida, mediante ningún sistema o método, electrónico o mecánico, sin consentimiento de los autores.

Comité científico:

Dr. Carlos N. Bouza Herrera, departamento de Matemáticas, Universidad de la Habana, Cuba.

Dr. José Legorreta Soberanis, Unidad Académica Centro de Investigación de Enfermedades Tropicales (CIET), Universidad Autónoma de Guerrero.

Índice

Presentación Editorial	1
1. Importancia clínica y epidemiológica del dengue.	4
1.1. El Dengue y la salud pública.	4
1.2. El Dengue como enfermedad	6
1.3. El virus del Dengue	9
1.4. La vigilancia epidemiológica del Dengue	10
2. Incidencia de dengue y su asociación con índices larvarios y pupales en Acapulco, Guerrero.	19
2.1. Introducción.	19
2.1.1. Antecedentes	20
2.2. Material y métodos.	25
2.2.1. Definición operacional de las principales variables resultado	26
2.2.2. Instrumentos utilizados	27
2.2.3. Prueba piloto	27
2.2.4. Trabajo de campo	27
2.2.5. Análisis de los datos	28
2.2.6. Consideraciones éticas	28
2.3. Resultados	29
2.3.1. Información de los hogares	29
2.3.2. Incidencia de la enfermedad	29
2.3.3. Análisis bivariado	29
2.3.4. Análisis multivariado	30
2.3.5. Índices entomológicos	30
2.3.6. Costos	31
2.4. Discusión	36
2.4.1. Limitaciones del estudio	36
2.4.2. Incidencia.	36
2.4.3. Factores relacionados a padecer dengue	37

2.4.4. Costos	38
2.5. Conclusiones	39
Bibliografía	40
3. Consenso y evaluación de clasificadores	45
3.1. Análisis de la efectividad de la diagnóstico del dengue en el Estado de Guerrero, México 2007: hechos estadísticos y consenso.	45
3.1.1. Introducción	46
3.1.2. Teoría general	48
3.1.3. Análisis del consenso entre clasificadores.	52
3.1.4. Conclusiones.	59
Bibliografía	61
3.2. Medidas de exactitud de la clasificación y curvas ROC.	64
3.2.1. Introducción	64
3.2.2. Metodología	65
3.2.3. Medidas de exactitud de un clasificador	68
3.2.4. Resultados	74
3.2.5. Conclusiones	77
Bibliografía	81
4. Modelos de diagnóstico del dengue	82
4.1. Diagnostico mediante modelos de árbol de clasificación.	82
4.1.1. Introducción.	83
4.1.2. Materiales y métodos.	85
4.1.3. Modelo de árbol de clasificación bajo un enfoque clásico.	86
4.1.4. Modelos de árbol de clasificación bajo el enfoque Bayesiano.	88
4.1.5. Resultados y discusión.	92
4.1.6. Modelación bajo el enfoque de Breiman et al. (1984) y S-Plus (2000).	93
4.1.7. Modelación bajo el enfoque Bayesiano (Chipman et al., 1998).	96
4.1.8. Discusión de los modelos.	99
Bibliografía	101
4.2. Diagnóstico mediante regresión logística binaria.	102
4.2.1. Introducción	102
4.2.2. Metodología.	103
4.2.3. Aplicación a datos de dengue.	108
4.2.4. Conclusiones	112
Bibliografía	113
4.3. Un modelo para la transmisión vertical de dengue	115

4.3.1.	Introducción	115
4.3.2.	Formulación de un modelo del Dengue con dos cepas y transmisión vertical en las poblaciones de humanos y vectores	119
4.3.3.	Conclusiones	131
	Bibliografía	133
4.4.	Construcción óptima de estratos para estudios asociados a epidemias de dengue.	135
4.4.1.	Introducción	135
4.4.2.	El muestreo de poblaciones finitas.	137
4.4.3.	Muestreo estratificado	145
4.4.4.	El enfoque de la programación dinámica.	154
4.4.5.	Estratos óptimos para estudios relacionados con el dengue.	157
4.4.6.	Conclusiones.	158
	Bibliografía	160
5.	Estudios perspectivas sobre dengue	162
5.1.	Estimación de proporciones en áreas pequeñas	163
5.1.1.	Introducción	163
5.1.2.	Estimación con datos de dengue	164
	Bibliografía	168
5.2.	Modelos para mortalidad y tasas enfermedad en el estado de Guerrero.	169
5.2.1.	Introducción.	169
5.2.2.	Modelos de la familia exponencial	170
5.2.3.	Modelos semiparamétricos	170
5.2.4.	Modelos para datos binarios	171
5.2.5.	Posibles proyectos de investigación.	176
	Bibliografía	178

Índice de cuadros

2.1. Análisis bivariado entre padecer dengue y las variables incluidas en el estudio. IC95 % ajustado por clusters.	33
2.2. Análisis multivariado del riesgo de padecer dengue con las variables que se mantuvieron en el modelo final.	34
2.3. Promedio de gastos en pesos de los pacientes que padecieron dengue y se atendieron en su hogar.	34
2.4. Promedio de gastos en pesos de los pacientes que refirieron haber recibido atención médica al enfermar por dengue en Acapulco, Guerrero durante el 2009, sin tomar en cuenta los que se atendieron en su hogar.	34
2.5. Promedio de días perdidos de trabajo y al cuidado del enfermo estratificado por sexo en Acapulco, Guerrero, 2009.	35
3.1. Clasificación de n objetos por dos expertos.	49
3.2. Clasificación por ExpUT y ExpLab.	53
3.3. Clasificación por ExpUT y ResultLab.	53
3.4. Clasificación por ExpLab y ResultLab.	54
3.5. Clasificación general y Resultado IgM.	54
3.6. Clasificación general y Resultado IgM.	55
3.7. Clasificación general de dengue clásico y hemorrágico.	55
3.8. Clasificador de dengue clásico.	56
3.9. Clasificador de dengue hemorrágico.	56
3.10. Clasificación de DC y DH por jurisdicción sanitaria.	57
3.11. Clasificación de DC y DH por municipio.	58
3.12. Comparación de índices para la clasificación de DC Vs DH e IC al 95 %.	59
3.13. Comparación de índices para la clasificación de DC Vs No DC e IC al 95 %.	59
3.14. Comparación de índices para la clasificación de DH Vs No DH e IC al 95 %.	59
3.15. Matriz de confusión.	66
3.16. Matriz de confusión.	67

3.17. Categorización de la razón de verosimilitud.	71
3.18. Frecuencias absolutas para la clasificación de dengue.	75
3.19. Frecuencias condicionadas para la clasificación de dengue.	75
3.20. Medidas de exactitud adicionales para la clasificación de DC y DH.	75
3.21. Medidas de exactitud por regiones de clasificaciones de DC.	76
3.22. Medidas de exactitud por regiones de clasificaciones de DH.	77
3.23. Área bajo la curva ROC con IC asintótico al 95 %. La variable (o variables) de resultado de contraste: DC, DH tiene al menos un empate entre el grupo de estado real positivo y el grupo de estado real negativo. Los estadísticos pueden estar sesgados. (a) Bajo el supuesto no paramétrico, (b) Hipótesis nula: área verdadera = 0.5.	79
4.1. Prueba de independencia de la variable dengue con el resto de variables, utilizando la prueba Ji-cuadrada.	93
4.2. Comportamiento de la edad según la variable dengue.	93
4.3. Comportamiento de la temperatura según la variable dengue.	94
4.4. Comportamiento de la temperatura según la variable dengue.	95
4.5. Características de los modelos después del proceso de podado del “mejor” modelo.	96
4.6. Características de los nodos terminales del mejor modelo de árbol de clasificación.	96
4.7. Características de los nodos terminales del mejor modelo de árbol de clasificación.	97
4.8. Características de los diez modelos de árbol de clasificación con el valor del logaritmo de verosimilitud más grande, en cada corrida de 5,000 iteraciones.	98
4.9. Características de los diez modelos de árbol de clasificación finales de cada cadena de Markov de 5,000 iteraciones.	98
4.10. Características de los nodos terminales del modelo de clasificación	100
4.11. Características de los nodos terminales del modelo de clasificación 10.	100
4.12. Estimación de los parámetros del modelo e IC al 95 %.	110
4.13. Interpretación de las exponenciales de los parámetros.	111
4.14. Tabla de clasificaciones(a).	112
4.15. Probabilidades predichas de sufrir DC para diferentes combinaciones de síntomas significativos y sexo.	112
4.16. Estudio en Guerrero, México. Ganancia en precisión relativa de las diferentes estratificaciones óptimas.	157
4.17. Estudio en West Bengal, India. Ganancia en precisión relativa de las diferentes estratificaciones óptimas.	158

5.1.	Tamaños poblacionales de la población “dengue” en las regiones.	164
5.2.	Eficiencia relativa estimada de los estimadores para la población “dengue” en el dominio U_6 (Costa chica).	166
5.3.	Estimación de proporciones y varianzas por regiones.	167

Presentación Editorial

Dr. C. Miguel Gerardo Valdés Pérez,

Colegio Universitario San Gerónimo de La Habana, Universidad de La Habana, Cuba.

El concepto de Responsabilidad Social y la importancia de su aplicación real y práctica en todos los contextos sociales han ido cobrando relevancia creciente en los últimos años.

Existe consenso entre los investigadores que la Responsabilidad Social entraña una construcción simbólica no sólo inherente a los procesos de producción y servicios que se generan en el mundo empresarial. Dicha construcción simbólica también está presente en otros ámbitos cuya razón de ser es la investigación científica y la difusión del conocimiento, los que a su vez, se reconocen como dos pilares básicos de las universidades contemporáneas.

Es a partir de la Baja Edad Media, como un resultado del tránsito feudal a la incipiente urbanización y del equilibrio sociopolítico entre la iglesia y las instituciones civiles, que comienza el surgimiento de las principales universidades europeas, inicialmente, como congregaciones donde se ponían de manifiesto las relaciones discípulo-maestro que habían caracterizado a los gremios europeos (*Studium y Studium generale*) cuya principal razón era la agrupación, la relación enseñanza-aprendizaje; así como la protección de sus miembros.

Entre las razones socio-económicas e históricas que permitieron el posterior desarrollo de las universidades suele mencionarse a la aparición de nuevas materias de estudio, nuevos métodos de enseñanza, así como a la organización que trajo consigo el nacimiento y consolidación de los estados europeos. Sin dudas, estas circunstancias estuvieron notoriamente influidas por los grandes movimientos renacentistas y humanistas, y por el desarrollo y evolución del pensamiento filosófico posterior, todo lo que contribuyó a la consolidación de la universidad moderna en los comienzos del siglo XIX.

La UNESCO ha identificado las funciones esenciales de la universidad con-

temporánea, funciones que rebasan el limitado criterio de mera institución educativa o formadora de profesionales y técnicos de aquellas primeras etapas y que suponen la generación, la transmisión, la aplicación y la crítica del conocimiento, desde la perspectiva integradora del saber aprender, saber hacer y saber ser, todo lo que consolida la dimensión humanista que toma como principal sustrato el enfoque histórico cultural que más allá del exclusivo compromiso docente educativo tiende al compromiso social derivado hacia la creciente lucha por la transformación de la sociedad y el mejoramiento de su calidad de vida y de índices satisfactorios de salud.

Estos indicadores serían impensables sin la confluencia de la ciencia y la tecnología como soportes que permitan incentivar el desarrollo sostenible a partir de estrategias de acción participativa y de transformación social.

Práctica creciente es la integración de especialistas de universidades de disímiles polos geográficos. Hoy en día se habla de laboratorios y de redes sociales universitarias en franco desafío al aislamiento creciente que la globalización ha ido imponiendo en las últimas décadas.

La monografía “**Estudios sobre Dengue: Experiencias y perspectivas**”, que sometemos a valoración y consulta de los interesados en el tema, es un ejemplo elocuente de la integración de intenciones y saberes universitarios en aras de contribuir a la producción de la investigación científica, a su necesaria divulgación y socialización, y al mejoramiento y preservación de los índices de salud social como resultado de las estrategias de bien público.

Sin lugar a la duda, el presente estudio y valoración estadística de una enfermedad de alta incidencia en todo el mundo y en particular, en diferentes países de Latinoamérica, que ha contado, además, con el concurso y el trabajo colaborativo de profesionales de distintos centros de investigación y de enseñanza superior, es un importante resultado y un meritorio aporte científico.

Con casi doscientos años de presencia, el Dengue, ha sido reportado por catorce países en América tras su primer brote importante, ocurrido en Cuba en 1981. Las estadísticas de la Organización Mundial de la Salud refieren que después de la década de los noventa, esta enfermedad, transmitida por el mosquito *Aedes aegypti*, y producido por el virus *Togaviridae*, subgénero *Flavivirus*, ha emergido y evolucionado con fuerza, con el consiguiente riesgo para aquellas poblaciones susceptibles de adquirirla. Poblaciones que actualmente se cuentan en más de cien países con reportes estadísticos de cincuenta millones de casos aproximadamente en todo el mundo.

Por estas consideraciones, actualmente, el Dengue es evaluado como uno de los problemas de salud y retos a enfrentar más importantes a escala mundial, pues se estima que más del 40% de la población mundial se encuentra en riesgo de contraer la enfermedad.

Estructurada en cinco capítulos, la monografía de referencia, cuyo tema es el estudio de esta enfermedad en Acapulco, Estado de Guerrero, México, muestra los resultados de una enjundiosa investigación en la que los análisis matemáticos operacionales permiten una cabal comprensión del comportamiento -refrendado por tablas y gráficos- del Dengue. Todo lo que de hecho convierte a la investigación en una obra de obligada consulta y referencia sobre el tema.

La confluencia de especialistas de alto relieve de la Universidad Autónoma de Guerrero, de la Universidad de La Habana, del Laboratorio Estatal de Salud Pública “Dr. Galo Soberón y Parra” de Guerrero, de la Universidad de Granada y de la Universidad de Jaén, España, deviene aporte fundamental para la producción científica al estudio y comportamiento del Dengue.

En medio del creciente agravamiento de las tensiones económicas internacionales, del deterioro y agresión al medio ambiente con los consiguientes efectos negativos sobre los cambios climáticos -desertificación e inundaciones incluidas- cobra especial connotación la integración del conocimiento científico como alternativa que contribuya a la solución de los graves problemas que cada vez más deberá enfrentar la humanidad y como expresión tangible de la Responsabilidad Social que le corresponde asumir a las universidades.

Capítulo 1

Importancia clínica y epidemiológica del dengue.

Dr. Saúl López Silva, Unidad Académica de Medicina, UAG.

1.1. El Dengue y la salud pública.

Con la globalización, el Dengue ha reemergido como uno de los problemas de salud pública más importantes en todo el mundo. Si bien es considerada una enfermedad “tropical”, el turismo y la movilidad de las personas derivada del comercio y la actividad económica internacional, hace que sea un riesgo biológico para cualquier ciudadano del mundo que viva o visite las áreas endémicas.

El número de casos de dengue ha aumentado de manera inusitada en los últimos años. Se estima que el 40 % de la población mundial se encuentran en riesgo directo de adquirir la enfermedad. La enfermedad está presente, es decir es endémica, en más de 100 países. Las regiones del planeta más afectadas son África, América (desde México hasta Argentina), el Mediterráneo Oriental, Asia Sudoriental y el Pacífico Occidental.

En el Reglamento Sanitario Internacional (RSI) se incluye el dengue como ejemplo de una enfermedad que puede constituir una emergencia de salud pública de interés internacional con implicaciones para la seguridad sanitaria, debido a la necesidad de interrumpir la infección y la rápida propagación de la epidemia más allá de las fronteras nacionales.

Las estimaciones de la Organización Mundial de la Salud (OMS) señalan que cada año puede haber al menos 50 millones de casos de dengue en todo el mundo. De ellos, cada año se producen unas 500 mil hospitalizaciones por la forma más grave de la enfermedad, de los cuales un alto porcentaje

corresponde a niños. La tasa de letalidad en estos casos puede alcanzar hasta el 20 % cuando no se cuenta con el manejo y tratamiento adecuado. Los países y sistemas sanitarios que garantizan un diagnóstico oportuno y la atención médica apropiada han disminuido las tasas de letalidad a menos del 1 %. A pesar de ello, número anual de muertes por Dengue puede alcanzar las 25,000 defunciones.

Después de los exitosos resultados de las campañas regionales y multinacionales de erradicación de enfermedades transmitidas por vector, como la Fiebre Amarilla y el Paludismo, en la década de los 50's y 60's, el Dengue se consideraba bajo control epidemiológico. Para la década de los 70's se incrementa el número de países con casos y se inicia un explosivo y sostenido incremento de la morbilidad y la mortalidad por esta patología. Su incidencia ha aumentado 30 veces en los últimos 50 años. Así, el Dengue hoy es catalogado como una enfermedad "reemergente", cuyo acelerado aumento de los casos se asocia a diversos factores antropogénicos.

El cambio climático derivado del calentamiento global, ha originado significativas variaciones en las precipitaciones pluviales de las áreas endémicas. Las modificaciones en el ciclo del agua generalmente favorecen la sobrevivencia del vector, dificultando además los esfuerzos de control que emprenden los sistemas de salud. Como las consecuencias de este fenómeno son progresivas, el efecto sobre el dengue se presume que se mantendrá en los próximos años. Al ser una enfermedad predominantemente urbana, el crecimiento demográfico y el surgimiento de megaciudades en los países endémicos, favorece la transmisión del padecimiento. Solo en América Latina, existe un aumento creciente de ciudades con más 3 millones de habitantes, donde las condiciones de hacinamiento y sobrepoblación permiten que la tasa de infección se incrementen hasta en el 80 % de los individuos. Con el crecimiento demográfico, la urbanización no controlada ni planificada de los centros habitados, origina problemas de desabasto de agua. La respuesta de la sociedad para enfrentar la disponibilidad de agua, incluye su almacenamiento temporal en recipientes que permiten la proliferación de los criaderos el control del vector. De manera agregada, el inadecuado ordenamiento ambiental en estas grandes ciudades, se traduce en agudos conflictos para el manejo de los desechos líquidos y sólidos, que generan ambientes propicios para el desarrollo de criaderos del vector difíciles de identificar y controlar. Finalmente, aunque los países endémicos son principalmente países de bajo y mediano ingreso, los países de alto ingreso no están exentos del riesgo de Dengue, debido a la exportación de casos derivados del turismo y los fenómenos migratorios.

Ante el panorama epidemiológico que parece empeorar globalmente, la respuesta esperada es el desarrollo de una vacuna que permita la prevención de la enfermedad y que reduzca la severidad y la letalidad de la infección

por Dengue. Actualmente hay diversos grupos de investigadores desarrollando vacunas contra el Dengue, que si bien usan inmunógenos y tecnologías de producción diferentes, se encuentran en fases muy avanzadas de validación de su seguridad y eficacia. Se espera que para el 2015, se pueda contar con este tipo de vacunas para usarse masivamente.

1.2. El Dengue como enfermedad

“el dengue es una sola enfermedad con presentaciones clínicas diferentes y a menudo con evolución clínica y resultados impredecibles”.

El dengue es un síndrome febril agudo, infeccioso de origen viral, transmitido por un vector, de curso autolimitado, incapacitante y con riesgo de complicaciones letales. También conocido como fiebre rompe huesos, es una enfermedad de inicio agudo, caracterizada por fiebre bifásica, mialgias, artralgias, exantema, leucopenia y linfadenopatía. Causada por un virus que pertenece al grupo de los arbovirus (virus transmitidos por artrópodos, “arthropod-borne viruses”) que se transmite por la picadura de las hembras de mosquitos del género *Aedes* (A), (del latín Aedes = Casa), un mosquito de hábitos domésticos. La especie principal es *A. aegypti*, aunque otras especies de *Aedes* como son *A. polynesiensis*, *A. scutellaris* y *A. albopictus* también pueden transmitir la enfermedad en circunstancias específicas.

El termino Dengue es una adaptación del vocablo “suahili dinga”, “dyenga” o “ki denga” pepo, que significa “ser presa repentina de un espíritu maligno” (utilizado para calambre súbito). En el siglo XVII se denominó “knokkelkoorts” en Indonesia, y “breakbone fever” o “Dandy fever” en Filadelfia. La referencia más antigua sobre la enfermedad, se encuentra en la enciclopedia China de enfermedades, síntomas y remedios, que data de la dinastía china Tang, donde fue llamada agua venenosa por los chinos, en asociación de los insectos voladores y el agua.

Como una enfermedad viral aguda, sus manifestaciones clínicas son variables dependiendo de la edad en la que se presente. En la literatura médica, existen diversas formas de describir y clasificar la infección por Dengue, desde las clasificaciones históricas y hoy obsoletas como Dengue Clásico y Dengue Hemorrágico; hasta la más actual que lo separa en el cuadro de Fiebre por Dengue y Fiebre Hemorrágica por Dengue. El periodo de incubación de esta enfermedad es de tres a quince días y en promedio siete días. Después del período de incubación, la enfermedad comienza abruptamente y le siguen tres fases de evolución: la febril, la crítica y la de convalecencia.

De acuerdo a la guía de la Organización Panamericana de la Salud, para el Diagnostico, Tratamiento, Prevención y Control del Dengue, publicada en el

2010, para una enfermedad tan compleja en sus manifestaciones, el manejo debe ser relativamente sencillo, económico y muy efectivo para salvar vidas, siempre y cuando se instauren intervenciones correctas y oportunas. La clave es el reconocimiento temprano y la comprensión de los problemas clínicos que se presentan en las diferentes fases de la enfermedad, lo cual lleva a un abordaje racional del manejo de los casos y a buen resultado clínico. Por ello, el manejo clínico inicial para identificar y evaluar a los pacientes, incluyendo los posibles diagnósticos diferenciales, representa el momento crítico para la evolución clínica del dengue. Una respuesta directa y bien manejada no solo reduce el número de hospitalizaciones innecesarias, sino que salva la vida de los pacientes con dengue. A ello hay que sumar la notificación temprana de los casos de dengue para la identificación oportuna de brotes y la iniciación de una respuesta organizada temprana.

La fase febril aguda inicia en forma abrupta y dura de 2 a 7 días y a menudo está acompañada de rubor facial, eritema de la piel, dolor corporal generalizado, mialgias, artralgias y cefalea. Algunos pacientes pueden tener dolor de garganta, faringe inyectada e inyección conjuntival. También son comunes la anorexia, las náuseas y el vómito. En la primera fase febril temprana, puede ser difícil el distinguir clínicamente el dengue de otras enfermedades febriles que no tienen relación alguna con el dengue. Si la prueba del torniquete resulta positiva en esta fase aumenta las probabilidades de que sea dengue. Se pueden observar manifestaciones hemorrágicas leves, como petequias y sangrado de mucosas (por ejemplo, nasal y de las encías). La anomalía más temprana en el cuadro hemático es una reducción progresiva del número total de glóbulos blancos.

Los expertos de la OMS/OPS coinciden en que la fase crítica se presenta al momento de la disminución de la fiebre, cuando la temperatura cae a 37.5°C o 38°C o menos y permanece por debajo de este valor, usualmente en los días 3 a 7 de la enfermedad, se puede presentar un aumento en la permeabilidad capilar junto con mayores valores del hematocrito. El período de extravasación de plasma dura generalmente entre 24 y 48 horas. La leucopenia progresiva seguida de una rápida disminución del número de plaquetas precede usualmente la extravasación de plasma. En este momento, los pacientes que no presentan aumento de la permeabilidad capilar mejoran, mientras que los que tienen un aumento de la permeabilidad capilar pueden empeorar como resultado de la pérdida del volumen plasmático. El grado de extravasación varía. El derrame pleural y la ascitis se pueden detectar clínicamente dependiendo del grado de extravasación de plasma y del volumen de reemplazo de líquidos. Por tanto, la placa de tórax y el ultrasonido abdominal pueden ser herramientas útiles para el diagnóstico. Un aumento superior al valor de referencia del hematocrito a menudo refleja la gravedad de la extravasación

de plasma. El choque se presenta cuando hay una pérdida crítica del volumen plasmático debida a la extravasación. A menudo está precedido por signos de alerta. La temperatura corporal puede estar por debajo de lo normal cuando ocurre el choque. Con un choque prolongado, la hipoperfusión que se presenta resulta en deterioro orgánico progresivo, acidosis metabólica y coagulación intravascular diseminada. Esto, a su vez, lleva a una hemorragia seria que hace que el hematocrito disminuya en el choque grave. En lugar de la leucopenia que se observa generalmente durante esta fase del dengue, el número total de glóbulos blancos puede aumentar en los pacientes con sangrado grave. Además, también se puede desarrollar un deterioro orgánico importante, con hepatitis, encefalitis o miocarditis, y, también sangrado grave, sin extravasación plasmática evidente o choque. Se dice que los pacientes que mejoran después de la caída de la temperatura tienen dengue no grave. Algunos pacientes progresan a la fase crítica de extravasación de plasma sin que haya disminución de la temperatura y, en estos pacientes, se deben usar los cambios en el cuadro hemático completo para determinar la aparición de la fase crítica y la extravasación de plasma. Los que empeoran, presentan signos de alerta. Los casos de dengue con signos de alerta probablemente se recuperarán con rehidratación intravenosa temprana. Algunos casos pueden agravarse hasta llegar a dengue grave.

Si el paciente sobrevive a la fase crítica de 24 a 48 horas, en las siguientes 48 a 72 horas tiene lugar una reabsorción gradual de los líquidos del compartimiento extravascular. Mejora el bienestar general, regresa el apetito, disminuyen los síntomas gastrointestinales, se estabiliza el estado hemodinámico y se presenta diuresis. Algunos pacientes pueden tener una erupción parecida a “islas blancas en un mar rojo”. Algunos pueden presentar prurito generalizado. Son comunes en esta etapa la bradicardia y los cambios en el electrocardiograma. El hematocrito se estabiliza o puede ser menor debido al efecto de dilución de los líquidos reabsorbidos. El conteo de leucocitos generalmente comienza a subir inmediatamente después de la disminución de la fiebre, aunque la recuperación del número de plaquetas generalmente es posterior al del número de leucocitos. La insuficiencia respiratoria producida por el derrame pleural masivo y la ascitis puede ocurrir en cualquier momento si se han administrado líquidos intravenosos en exceso. Durante la fase crítica y la fase de recuperación, el reemplazo excesivo de líquidos se relaciona con edema pulmonar o insuficiencia cardíaca congestiva.

1.3. El virus del Dengue

El virus del dengue es un virus de Acido Ribonucleico (ARN), del género *Flavivirus*, familia *Flaviviridae*. La partícula madura del virus del dengue es esférica, con un diámetro de 50 nm, y contiene múltiples copias de las tres proteínas estructurales, una membrana de doble capa derivada del huésped y una copia única de un genoma de ARN compuesto por una sola molécula de RNA de cadena sencilla lineal, de sentido positivo, de 10 703 nucleótidos y de alta variabilidad genómica, que codifica para tres proteínas estructurales (cápside, C, prM, el precursor de membrana, M, proteína y envoltura, E) y siete proteínas no estructurales.

De acuerdo a sus características antigénicas, el virus del dengue está agrupado en cuatro serotipos: Den-1, Den-2, Den-3 y Den-4. Cada serotipo crea inmunidad específica a largo plazo contra el mismo serotipo (homólogo), así como una inmunidad cruzada de corto plazo contra los otros tres serotipos, la cual puede durar varios meses. Los cuatro serotipos son capaces de producir infección asintomática, enfermedad febril y cuadros severos que pueden conducir hasta la muerte.

La relación genética de los serotipos Den-1 y Den-4, Den-3 y Den-4 y la escasa homología de la secuencia entre el serotipo Den-2 sugiere que estos pudieron tener orígenes evolutivos diferentes. Algunas variantes genéticas parecen ser más virulentas o tener mayor potencial epidémico. Las frecuentes mutaciones del virus explican la existencia en cada serotipo, de un gran número de cepas antigénicamente diferentes circulando en áreas distintas del mundo. Existen métodos que permiten clasificar esas cepas en grupos llamados «topotipos», “genotipos” o “subtipos” según el método utilizado. El serotipo Den-1 está compuesto por 8 topotipos o 5 genotipos; el serotipo Den-2 por 10 topotipos o 5 subtipos o 5 genotipos; el serotipo Den-3 por 5 topotipos o 4 genotipos o 3 subtipos y el serotipo Den-4 por 5 topotipos o 1 (ó 2) genotipos. No se han podido establecer correlaciones específicas entre las diferencias genómicas y antigénicas y la virulencia de las cepa.

Las hembras del mosquito *Aedes* adquieren el virus del dengue al picar a un hospedero vertebrado virémico. El virus infecta las células epiteliales del intestino medio del mosquito, se disemina a través de la lámina basal hacia la circulación y termina por infectar las glándulas salivales. El virus establece una infección persistente y se replica en grandes cantidades en estas células. Posteriormente las glándulas lo secretan a la saliva. Tras picar al hospedero, la hembra del mosquito regurgita su saliva llena de virus hacia la sangre de su víctima. Una vez inoculado en el tejido subcutáneo, se disemina en forma rápida hacia los tejidos adyacentes; es fagocitado por los macrófagos tisulares y transportados por éstos a través de los vasos linfáticos hasta los ganglios

linfáticos regionales, en donde se efectúa una primera replicación. Posteriormente, por medio de los vasos linfáticos eferentes pasa al conducto torácico y llega al torrente circulatorio, en donde se efectúa la primera viremia; a continuación los virus son captados por el sistema reticuloendotelial y la piel, en donde se efectúa la segunda replicación y una segunda viremia. La presencia de viriones en la sangre del infectado, permite que cuando es picado por un vector libre de virus, el ciclo se complete y se mantenga la transmisión de la enfermedad.

La inmunidad a un serotipo se considera de larga duración y efectiva frente a una segunda infección por el mismo serotipo. La existencia de los 4 serotipos virales posibilita que se produzcan incluso infecciones terciarias y cuaternarias. En los individuos que sufren su primoinfección, los anticuerpos IgG anti-dengue comienzan a incrementarse lentamente a partir del tercer día del comienzo de los síntomas, los cuales son máximos hacia los 15-21 días. Después declinan y permanecen detectables poco más o menos durante toda la vida. En el transcurso de una infección secundaria, los anticuerpos IgG se elevan casi al mismo tiempo del comienzo de los síntomas; permanecen elevados durante varias semanas y luego van declinando. Esta elevación significativa permite el diagnóstico presuntivo en monoseros tomados durante la fase aguda de la enfermedad. Los anticuerpos IgM anti-dengue que se producen en respuesta a la infección se desarrollan rápidamente y hacia el quinto día de la enfermedad la mayoría de los individuos presentan cantidades detectables. En general, estos anticuerpos IgM declinan hacia niveles no detectables entre los 30 y 90 días del comienzo de los síntomas. Se ha observado que un pequeño número de individuos que sufren una infección secundaria por dengue, pueden no presentar niveles de anticuerpos IgM 28% aproximadamente de los casos.

1.4. La vigilancia epidemiológica del Dengue

Un elemento central en la prevención y control del Dengue, lo representa la vigilancia epidemiológica del padecimiento. La vigilancia epidemiológica se entiende como el estudio permanente y dinámico de las condiciones de salud de una población, para tomar decisiones tendientes a reducir los riesgos de enfermedad y muerte. Este proceso incluye la recopilación, procesamiento y análisis de información relativa a los daños y riesgos en salud. En nuestro país, el derecho a la protección de la salud, obliga al Estado Mexicano a operar un Sistema Nacional de Vigilancia Epidemiológica (SINAVE) capta, registra y analiza los datos de morbilidad, mortalidad, y daños y riesgos en salud a través del Sistema Único de Información para la Vigilancia Epide-

miológica (SUIVE). En el caso del Dengue, la información proviene de la notificación semanal de casos de enfermedad, en una red nacional digitalizada. Para ello se cuenta con definiciones operacionales de caso que unifican los criterios de detección, notificación y clasificación de los casos de dengue. Estas definiciones deben contar con una elevada sensibilidad para detectar la mayor proporción posible de los casos a través de los signos y síntomas más frecuentes de la enfermedad y de las pruebas de tamizaje. Mientras que la especificidad del diagnóstico depende de las pruebas de laboratorio que se utilizan para el seguimiento y clasificación final de cada caso.

De acuerdo a los “Lineamientos para la Vigilancia Epidemiológica de Fiebre por Dengue y Fiebre Hemorrágica por Dengue” emitidos en México por la autoridad sanitaria en Septiembre del 2008, las definiciones operacionales son las siguientes:

Caso Sospechoso de Fiebre por Dengue (FD): Toda persona de cualquier edad que resida o proceda de una región en la que haya transmisión de la enfermedad y que presente cuadro febril inespecífico o compatible con infección viral.

Caso Probable de Fiebre por Dengue: Todo caso sospechoso que presente fiebre y dos o más de las siguientes características: cefalea, mialgias, artralgias, exantema o dolor retroocular. En menores de 5 años, el único signo a considerar puede ser la fiebre.

Caso Confirmado de Fiebre por Dengue: Todo caso probable en el que se confirme infección reciente por denguevirus mediante técnicas de laboratorio, esté asociado epidemiológicamente a otro caso confirmado o no se disponga de resultado de laboratorio.

Caso Probable de Fiebre Hemorrágica por Dengue (FHD): Toda persona que, además de un cuadro probable de Fiebre por Dengue, desarrolle fiebre y una o más de las siguientes características: datos de fuga de plasma (ascitis, derrame pleural, edema, hipoalbuminemia); o datos de fragilidad capilar (petequias, equimosis, hematomas); o hemorragias a cualquier nivel (gingivorragia, hematemesis, metrorragia); o trombocitopenia menor a 100 mil plaquetas por ml³ o hemoconcentración con uno o más de los siguientes datos: incremento del hematocrito (Hto) 20 % o más en la fase aguda; decremento del Hto en 20 % después del tratamiento; tendencia del Hto en muestras secuenciales (por ejemplo, 40, 43, 45, etc.); relación hema-

tocrito/hemoglobina (Hto/Hb): sugestivo 3.2 a 3.4, indicativo 3.5 o mayor; o hipoalbuminemia.

Caso Confirmado de Fiebre Hemorrágica por Dengue: Toda persona con un cuadro probable de Fiebre Hemorrágica por Dengue confirmado por laboratorio que, además, presente lo siguiente:

1. Datos de fuga de plasma evidenciada por cualquiera de los siguientes datos:
 - A) CLINICA: Edema, piel moteada, ascitis o derrame pleural;
 - B) LABORATORIO: Medición de la Hb, Hto - elevación en 20 % en etapa aguda, o disminución de 20 % en etapa de convalecencia, o elevación de Hto o Hb en forma secuencial (a partir del tercer día) o hipoalbuminemia;
 - C) GABINETE: Ultrasonido (líquido perivisceral y en cavidad abdominal o torácica) y radiología (derrame pleural o ascitis).
2. Más uno de los siguientes datos:
 - A) Datos de fragilidad capilar: prueba de torniquete positiva (a partir del tercer día); (petequias, equimosis, hematomas, etc.) o
 - B) Trombocitopenia menor de 100 mil plaquetas por microlitro.

Caso Probable de Síndrome de Choque por Dengue (SCHD): Toda persona con cuadro probable de FD o FHD y que presente súbitamente datos de insuficiencia circulatoria (pulso rápido y débil, extremidades frías); alteraciones en el estado de conciencia (confusión mental); tensión arterial disminuida o reducción en la tensión diferencial sistólica-diastrólica menor a 20 mm/Hg, ejemplo 90/80 o 80/70, etc.; o bien, estado de choque profundo.

Caso Confirmado de Síndrome de Choque por Dengue: Todo caso probable de SCHD en el que se confirme infección reciente por dengue mediante técnicas de laboratorio.

Para la detección, notificación, estudio, seguimiento y clasificación de casos y defunciones, se siguen los procedimientos específicos siguientes:

DetECCIÓN

Acciones ante un caso sospechoso de Fiebre por Dengue: La definición de caso sospechoso sólo se utiliza en brotes o estudios especiales. Los

casos sospechosos de FD no deben ser notificados al área administrativa inmediata superior a menos que pasen a otra clasificación. Así mismo, no debe ser tomada muestra para diagnóstico de laboratorio.

Acciones ante un caso probable de Fiebre por Dengue: Ante la presencia de casos probables de FD, deberán efectuarse las siguientes acciones:

1. Notificación de la totalidad de los casos probables a través del “Informe Semanal de Casos Nuevos de Enfermedades”, a la Jurisdicción Sanitaria, al estado y a la DGAE.
2. Notificación, en caso de brote, a través del formato “Estudio de Brote de Dengue”.
3. La toma de muestras deberá realizarse de acuerdo a los siguientes criterios:
 - a. En ausencia de brotes o en periodos de baja transmisión, deberá tomarse muestra al 100 % de los casos probables a efecto de identificar la ocurrencia oportuna de casos.
 - b. Una vez demostrada la existencia de brotes, deberá tomarse muestra al 30 % de los casos probables.
4. Una vez confirmada la presencia de la enfermedad mediante diagnóstico de laboratorio en un área determinada, la clasificación de los casos podrá realizarse sobre bases clínico-epidemiológicas, de acuerdo con la definición operacional de caso.
5. Deberá llenarse el “Estudio Epidemiológico de Fiebre por Dengue y Fiebre Hemorrágica por Dengue”, a todos los casos que cumplan definición operacional de caso probable y enviarse de manera inmediata a la Jurisdicción Sanitaria o Delegación correspondiente, ***se tome o no muestra para diagnóstico.***
6. La Jurisdicción Sanitaria o Delegación deberá realizar la captura de todos los casos probables en el Módulo de Dengue de la Plataforma Única de Información, incluyendo la georreferencia de los mismos.
7. Los casos a quienes se les toma muestra para diagnóstico se imprimirá el estudio epidemiológico y se enviará junto con la muestra al Laboratorio Estatal de Salud Pública (LESP). Toda muestra deberá contar con el nombre completo del paciente y el folio de captura para ser recibida por el laboratorio.

8. Una vez procesada la muestra, el laboratorio ingresará el resultado en el Módulo de Dengue de la Plataforma Única de Información.
9. Las muestras de los casos con resultados positivos a dengue mediante NS1 deberán ser enviadas al InDRE o laboratorio avalado por este Instituto para la identificación del denguevirus causal de acuerdo a lo establecido en el Procedimiento para la Aplicación del Nuevo Algoritmo para Diagnóstico por Laboratorio de Fiebre por Dengue y Fiebre Hemorrágica por Dengue.
10. En todo caso probable de FD, deberán darse indicaciones al paciente o responsable del mismo sobre los signos y síntomas de FHD, los signos de alarma de la enfermedad y la conducta a seguir, incluyendo la unidad hospitalaria a las que deberá dirigirse en caso de presentarse alguno de ellos.
11. Los casos probables de FD deben clasificarse en un plazo no mayor a diez días hábiles posteriores a la detección del caso.
12. Realizar la vigilancia clínica, entomológica, virológica y de factores de riesgo para dengue, canalizando la información correspondiente a los Comités Jurisdiccional y Estatal para la Vigilancia y Control de Dengue u homólogo.

Acciones ante un caso probable de Fiebre Hemorrágica por Dengue:

1. Referir los casos a unidades de segundo o tercer nivel para valorar su hospitalización.
2. Llenado y envío inmediato del estudio epidemiológico (preliminar) de dengue a la Jurisdicción Sanitaria o Delegación correspondiente.
3. Captura inmediata de la información preliminar del estudio epidemiológico de caso (en las primeras 24 horas de conocido el caso por los Servicios de Salud) en la Plataforma Única de Información.
4. Toma y envío de muestra para diagnóstico de laboratorio al 100 % de los casos probables. Al igual que con FD, las muestras deberán ir acompañadas del estudio de caso previamente capturado en la Plataforma Única de Información.

5. Seguimiento del caso en unidad hospitalaria mediante la toma de muestras a todos los casos que cumplan con definición operacional para la determinación de Hto, Hb, plaquetas y otros estudios disponibles en la unidad, de acuerdo a lo establecido en el apartado de manejo de casos probables de Fiebre Hemorrágica por Dengue, enfatizando que:
 - La primera muestra deberá *tomarse* el día que inicien las manifestaciones de fuga de líquidos o en el primer contacto con los Servicios de Salud.
 - Se tomarán muestras de monitoreo hemodinámico dependiendo de la evolución del paciente cada 8 a 24 horas durante la fase aguda de la enfermedad.

6. Buscar:
 - Presencia o ausencia de escape de líquidos.
 - Trombocitopenia menor a 100 mil por ml³.
 - Hemoconcentración, con base en los resultados del monitoreo hemodinámico y de acuerdo con la presencia de uno o más de los siguientes criterios:
 - Incremento del Hto en 20 % o más (por ejemplo de 40 a 48) durante la fase crítica de la enfermedad.
 - Decremento del Hto en 20 % o más (por ejemplo de 48 a 40) entre la fase crítica y la convalecencia.
 - Tendencia ascendente del Hto (por ejemplo 40, 43 y 45 en muestras secuenciales).
 - Valor de la relación Hto / Hb: el valor normal es 3.0; si la relación es de 3.2 a 3.4 se considera sugestiva; si es igual o mayor de 3.5, es indicativa de hemoconcentración.
 - Presencia de derrames cavitarios o intersticiales.
 - Hipoalbuminemia.

7. Todos los casos deberán contar con sus estudios de monitoreo hemodinámico para establecer los criterios de trombocitopenia o hemoconcentración. ***Los casos a los cuales no se les realice estas pruebas de acuerdo a lo establecido en el apartado correspondiente serán considerados como positivos al criterio de hemoconcentración y plaquetopenia.***

8. Desde el punto de vista epidemiológico y para la clasificación de los casos, se considerará como un caso con adecuado seguimiento de Hto,

Hb y conteo plaquetario, todo aquel con cifra igual o mayor de 50 mil plaquetas y tendencia ascendente en su conteo plaquetario.

9. Obtención de copia del expediente clínico del paciente.
10. Actualización permanente en el apartado de seguimiento del Módulo de Dengue de la Plataforma Única de Información, asegurando incorporar toda la información disponible.
11. Una vez resuelto el cuadro clínico, el epidemiólogo jurisdiccional deberá realizar la clasificación final del caso con toda la información clínico-epidemiológica debidamente requisitada en el Módulo de Dengue en la Plataforma Única de Información en un plazo no mayor a 10 días hábiles posteriores a la detección del caso. Posteriormente y de considerarse necesario, dicha clasificación será ratificada o rectificada en el seno de los Comités de Vigilancia u homólogos.
12. Los casos a los que se les tomó muestra, no clasificados bajo los criterios enunciados en los párrafos anteriores serán clasificados por los Comités Estatales y Nacional de Vigilancia Epidemiológica, a efecto de garantizar la obtención de información oportuna y veraz que oriente la toma de decisiones.
13. Los casos sin toma de muestra para diagnóstico de laboratorio serán clasificados como FD y se tomarán como casos mal estudiados en la elaboración del indicador de seguimiento de casos de FHD.

Acciones ante defunciones por Probable Fiebre Hemorrágica por Dengue:

1. Notificación inmediata de la defunción en las primeras 24 horas de conocida por los Servicios de Salud. La notificación deberá incluir los signos y síntomas que confirme definición operacional de probable FHD.
2. Realizar y enviar de manera inmediata el Estudio Epidemiológico de Caso a la Jurisdicción Sanitaria o Delegación para su ingreso inmediato en el Módulo de Dengue de la Plataforma Única de Información.
3. Personal de la unidad médica tratante deberá obtener copia del expediente clínico y enviarlo a la Jurisdicción Sanitaria correspondiente, que a su vez enviará copia al estado y a la DGAE en un periodo no mayor a 5 días hábiles posteriores a la detección de la defunción. En caso de que la unidad médica no cuente con personal de vigilancia epidemiológica,

el epidemiólogo jurisdiccional se hará responsable de la obtención del expediente clínico.

4. Garantizar la toma de muestra para diagnóstico, así como su envío al LESP.
5. Dictaminación por el Comité Estatal u homólogo de la defunción en un periodo no mayor de 10 días posteriores al deceso y envío del acta correspondiente a la DGAE. Deberá contar con toda la información clínica-epidemiológica, así como los resultados del LESP. En las defunciones donde la clasificación final dependa del resultado virológico, el periodo de tiempo para la dictaminación de la defunción se extenderá a un máximo de seis semanas.
6. Revisión de la documentación (estudio de caso, expediente clínico y acta) por el CONAVE para avalar o no el dictamen.
7. Envío del certificado de defunción y el formato de “Reporte de causas de Muerte Sujetas a Vigilancia Epidemiológica” de INEGI en periodo no mayor de cuatro semanas a la Dirección General de Epidemiología, donde se realiza la ratificación o rectificación de la defunción dictaminada por el Comité Estatal de Vigilancia Epidemiológica u homólogo.
8. Las defunciones en las cuales no exista acuerdo en la dictaminación realizada por los comités estatales, serán dictaminadas por el CONAVE con presencia del epidemiólogo estatal y el médico tratante, se informarán a los servicios de salud, con la especificación de los desacuerdos atendidos, el dictamen final que se ingresará a los registros oficiales.
9. Toda defunción en la cual no haya sido tomada muestra para diagnóstico de laboratorio deberá ser dictaminada mediante criterios clínico-epidemiológicos por el Comité Estatal y el CONAVE.
10. Realizar la vigilancia clínica, entomológica, virológica y de factores de riesgo para dengue, canalizando la información correspondiente a los Comités Jurisdiccionales y Estatales de Vigilancia.

Acciones ante un caso probable de Síndrome de Choque por Dengue: Los procedimientos generales de vigilancia en los casos probables de SCHD son los mismos señalados para los casos probables de FHD. Los casos probables de SCHD serán tratados preferentemente en las unidades de cuidados intensivos en los hospitales que cuenten con ellas.

Acciones ante un caso confirmado de Fiebre por Dengue, Fiebre Hemorrágica por Dengue o Síndrome de Choque por Dengue:

1. Ingreso del resultado en el Módulo de Dengue de la Plataforma Única de Información.
2. Inclusión del caso en los registros, publicaciones e informes oficiales.
3. En brotes realizar permanentemente la estimación de casos de dengue, la cual se realizará aplicando el porcentaje de positividad a los casos probables no muestreados más los casos confirmados por laboratorio.
4. Intensificar las acciones integrales de vigilancia mediante el análisis de la situación epidemiológica de dengue en el seno de los Comités Jurisdiccionales y Estatales de Vigilancia Epidemiológica a efecto de orientar la toma de decisiones.

Capítulo 2

Incidencia de dengue y su asociación con índices larvarios y pupales en Acapulco, Guerrero.

José Legorreta Soberanis¹, María de los Ángeles Salgado Jiménez², Arcadio Morales Pérez¹, Elizabeth Nava Aguilera¹, Miguel Flores Moreno¹, Alejandro Balanzar Martínez¹, Neil Andersson¹

1. Unidad Académica Centro de Investigación de Enfermedades Tropicales, Universidad Autónoma de Guerrero.

2. Alumna de la maestría en Ciencias Médicas vertiente Epidemiología Aplicada, Unidad Académica Centro de Investigación de Enfermedades Tropicales, Universidad Autónoma de Guerrero.

2.1. Introducción.

El dengue es una arbovirosis ampliamente diseminada en, prácticamente, la mitad del mundo. En México la población de los estados de Veracruz, Colima y Guerrero ha sido afectada por la enfermedad. El municipio de Acapulco, en Guerrero, ha presentado incremento de dengue en los últimos años de manera alarmante, por esta razón, los objetivos del presente trabajo consistieron en generar información que permita proponer estrategias para la disminución del dengue en Acapulco, Guerrero; estimar la incidencia de dengue en la población de Acapulco; identificar factores de riesgo asociados a padecer dengue; estimar los índices larvarios y pupales de *Aedes aegypti* (índice de

recipiente, casa, Breteau y pupa-persona) y estimar costos de atención de la enfermedad de dengue en el municipio de Acapulco. El estudio se realizó bajo la siguiente metodología: en diciembre del 2009 se llevó a cabo la línea de base de un ensayo por conglomerados aleatorizado para la reducción de criaderos *A. aegypti*. En un primer paso se seleccionaron al azar 30 colonias de Acapulco con al menos 100 hogares contiguos, en un segundo paso, posterior a la realización de la línea de base, se aleatorizó la intervención integrando 15 colonias al brazo de intervención y 15 al brazo de control. Se aplicó un cuestionario que incluyó preguntas sobre aspectos socio-demográficos, haber padecido dengue en los 12 meses previos a la encuesta, factores asociados a padecer dengue, el costo económico para la familia del enfermo. Se revisaron todos los potenciales criaderos con presencia de agua en el momento de la encuesta y se obtuvieron, en caso de ser positivos, las larvas y pupas las que se trasladaron a laboratorio para ser clasificadas. Se encuestaron 4 389 hogares abarcando información de 19 707 personas. En este estudio presentamos los resultados de la línea de base. Como resultados del estudio se obtuvo que la incidencia de dengue fue del 3.5 % (663/19 707). Se encontró que el haber referido tener peligro de enfermarse por dengue, tener un jefe de familia con un nivel de estudios superior al sexto año de primaria, contar con un trabajo remunerado, habitar en sitios urbanos y no tener larvas en sus depósitos en donde se almacena agua son factores asociados a padecer dengue. El índice de casa fue de 15.5 %, el índice de recipiente fue de 5.1 %, índice de Breteau de 20.3, el índice de pupa-persona fue de 8.4 %. El gasto promedio por concepto de medicina, consulta y transporte fue de \$785. Los pacientes que se atendieron en instituciones públicas por concepto de medicina, consulta, transporte gastaron \$527 pesos aproximadamente y aquellos que se atendieron en instituciones privadas gastaron \$1 276. ($P < 0.0001$). Concluimos que la incidencia encontrada en el 2009 en el presente estudio es media. Tomando en cuenta los niveles de riesgo de epidemias de dengue de acuerdo a los tres índices médicos el índice de casa, de recipiente y de Breteau se encuentran en nivel medio.

2.1.1. Antecedentes

A finales del siglo XV llega a América el transmisor del dengue y la fiebre amarilla *Aedes aegypti*, se cree que fue traído por los primeros navegantes, la enfermedad se propagó por el Golfo de México y el Pacífico. En 1648 se registraron en Campeche y Mérida epidemias de fiebre amarilla, posteriormente se propagó al puerto de Veracruz, Ciudad del Carmen, Tlacotalpan y Alvarado. Para 1830 ya estaba afectada la península yucateca y para 1858, el transmisor de la fiebre amarilla afectaba desde Tamaulipas hasta Yucatán[1].

Entre 1867 y 1877 se registraron epidemias con un alto índice de mortalidad en Veracruz, Tamaulipas y Tabasco, y en 1880 se identificaron casos de dengue en Acapulco y Mazatlán, pero no es hasta 1903 cuando se realizan los primeros trabajos cuya finalidad fue erradicar al mosquito transmisor del dengue en su fase acuática, en 1957 se inició el programa de erradicación del vector con el método de rociado intradomiciliario con dicloro difenil tricloroetano (DDT)[2].

En México, en 1978 se registró un brote de dengue en Chiapas, desde entonces el serotipo 1 es endémico, en 1984 el serotipo 2 y 4 iniciaron su circulación y para 1985 el serotipo 3 se introdujo en la costa del pacífico[3].

Para 1997 prácticamente todos los países de América se encontraban reinfestados y los ministros de salud de los países del continente aprobaron una resolución, durante una reunión del consejo directivo de la Organización Panamericana de la Salud (OPS) en septiembre de 1996, la cual solicitaba a los gobiernos la colaboración en la elaboración de un plan continental de ampliación e intensificación para combatir al *A. aegypti*. El plan, elaborado entre el 22 y el 24 de abril de 1997, tenía como objetivo “el incremento de las acciones para combatir a la hembra hematófaga para alcanzar niveles de infestación cercanos a cero y lograr su erradicación y así conseguir la eliminación de la circulación de los virus del dengue”[4].

La incidencia del dengue ha aumentado en todo el mundo en los dos últimos decenios, aproximadamente 2.5 mil millones de personas están en riesgo de adquirir esta enfermedad. Según cifras reportadas por la OMS cada año puede haber 50 millones de casos de dengue[5].

La enfermedad es endémica en más de 100 países de África, las Américas, el Mediterráneo Oriental, Asia Sudoriental y el Pacífico Occidental. Del 2001 al 2007 se registraron en más de 30 países del continente Americano 4 332 731 casos de dengue[6], dentro de esta cifra, 106 037 fueron de dengue hemorrágico, contabilizándose un total de 1 299 defunciones. En Barbados, Colombia, República Dominicana, Guyana Francesa, México, Perú, Venezuela y Puerto Rico, se identificaron de forma simultánea en un año los cuatro serotipos[7]. Para llevar a cabo un buen control del dengue se requieren un mejoramiento del medio ambiente, educación y la participación comunitaria activa[8]. Toledo Romaní y colaboradores encontraron falta de motivación de los líderes comunitarios y, en general, que la “participación” es interpretada como “colaboración”[9].

La Organización Mundial de la Salud (OMS) estima que cada año 500 000 casos de dengue requieren hospitalización y se registran 20 000 defunciones[4]. El dengue es transmitido de una persona enferma a una susceptible a través de la picadura de un mosquito, hembra, hematófago conocido como *Aedes aegypti*. El mosquito *A. aegypti* pertenece al Phylum: *Artropoda*, cla-

se: *Insecta*, orden: *Diptera*, suborden: *Nematocera*, familia: *Culicidae*, tribu o subfamilia: *Culicini*, género: *Aedes*, subgénero: *Stegomyia*, grupo: "A", especie: *aegypti*. El ciclo de vida del *A. aegypti* tiene cuatro etapas: huevo, larva, pupa y adulto o imago[10].

La dispersión del vuelo del *A. aegypti* es de 100 metros aproximadamente, aunque en hembras grávidas se ha demostrado que son capaces de volar hasta tres kilómetros si no encuentran algún lugar apropiado para llevar a cabo su ovoposición. La mayor dispersión del *A. aegypti* ocurre cuando huevos o larvas son transportadas a través de recipientes por medios de transporte de carga[11].

Existen dos tipos de dengue, fiebre por dengue o dengue clásico (FD) y fiebre hemorrágica por dengue (FDH), la complicación es el síndrome de choque por dengue. Los principales síntomas son fiebre de 2 a 7 días de duración, cefalea, dolor ocular, mialgias, artralgias, dolor abdominal, erupción maculopapular, manifestaciones hemorrágicas (leves de piel y mucosas), diaforesis[12]. Se han registrado casos de coinfección en Delhi, India con el virus Chikungunya el cual es transmitido por el mismo vector y el cuadro clínico es similar[13].

Las variaciones climáticas ejercen una influencia positiva en la aparición de casos de dengue según refiere Rifakis y colaboradores[14]. En un estudio se logró asociar el aumento de formas larvarias de *A. aegypti* con las épocas de lluvia, pero no se correlacionó con la incidencia del dengue[15]. En Mérida, México, Manrique y colaboradores demostraron que existe un incremento en el número de contenedores con productividad pupal en épocas de lluvia de 6.5 % a 7.7 %, este último dato es de importancia ya que el 99 % de las pupas llegan a convertirse en moscos adultos[16].

El dengue es un problema multifactorial, tiene relación directa con el clima como la temperatura, precipitación pluvial, vegetación y altitud, pero la influencia de las condiciones sociales es más decisiva que el descongelamiento de los polos[17].

Pozo y colaboradores encontraron factores asociados a la infestación de viviendas por *A. aegypti*, como son la presencia de botellas, como potenciales criaderos, dentro de la vivienda, que la casa esté ubicada a menos de 200 metros de una llantería y tener un jardín en el interior[18]. En cuanto a educación, el nivel bajo de educación de los jefes de las familias está relacionado con un aumento de criaderos de larvas de mosquitos[19].

Por otra parte Cáceres y colaboradores encontraron como factor protector de la enfermedad el adecuado suministro de agua (OR: 0.39, IC, 95 %: 0.24-0.63) y como factores predisponentes: los bebederos (OR: 1.99 IC95 %: 1.33-2.98), las botellas (OR: 1.97 IC: 95 %: 1.23-3.17) y los recipientes con larvas (OR: 1.96 IC: 95 % 1.11-3.46)[20]. Navarrete Espinosa y colaboradores encontraron como principales factores vinculados con la infección del virus del dengue ser

mayor de 25 años, la presencia de larvas en el domicilio y el bajo conocimiento de la enfermedad[21]. Vezzani y colaboradores demostraron un incremento en el índice de contenedor en aquellos recipientes ubicados en área sombrosa comparado con aquellos ubicados en áreas expuestas al sol, 12.8 % y 6.9 %, ($P < 0.001$). La productividad pupal es más alta en el mes de marzo[22] y el efecto sombra juega un papel importante en la productividad pupal[15]. La mayoría de los criaderos se encuentran en el peridomicilio según refiere Calderón Arguedas y colaboradores[23].

Una revisión sistemática sobre los efectos de las intervenciones para el control del vector del dengue basados en parámetros entomológicos (índice de Breteau, índice de contenedor o recipiente e índice de casa), evaluó los tipos de control biológicos, químicos y físicos, concluyó que la mejor intervención para lograr dicho control es aquella en la cual participa la comunidad con programas de educación, ya que incrementa el conocimiento y se reducen los índices larvarios y pupales[24].

La interrogante continúa siendo el sitio de infección, Mammen y colaboradores realizaron un estudio prospectivo para resolver algunas incertidumbres sobre el lugar de transmisión del dengue en niños de poblaciones rurales de Tailandia, se concluyó que los casos positivos pudieron haber adquirido la infección cerca de su casa[25].

El costo en casos de epidemias es elevado, como ejemplo podemos citar la de Cuba en 1997, catalogada como la mayor epidemia de dengue clásico de las Américas, en la que se registró más de medio millón de personas afectadas y aproximadamente 4.5 millones de infectados por el serotipo 2. Se estimó la afectación económica incurrida tomando en cuenta gastos de hospitalización, control clínico terapéutico, lucha antivectorial y vigilancia de laboratorio, el costo total fue de 10 251 539 USD. Se realizó un análisis comparando gastos de la epidemia con campañas de prevención del dengue y se concluyó que si se desea reducir costos a largo plazo se debe invertir en programas de control de la enfermedad[26].

A pesar de la vigilancia epidemiológica y las medidas de prevención, los casos de dengue van en incremento. En Tailandia la transmisión endémica epidémica muestra un patrón cíclico cada tres años, de acuerdo a la hipótesis ADE (Antibody dependent enhancement) o potenciación inmunitaria, una vez que se conoce el serotipo circulante y el estado inmunitario de la población se establece un modelo de transmisión aunque América, a excepción de Cuba, parece no apegarse a ese modelo[27].

En México se presenta un modelo gravitacional, y se explica cuando en los focos gravitacionales primarios, en los que se presentan casos, se observa una dispersión a un número limitado de localidades secundarias. La diferencia entre las primarias y las secundarias es la densidad de población; cuanto

mayor es ésta el número de casos tiende a mantenerse constante y la velocidad de dispersión del brote depende de la densidad de población y las vías de comunicación entre las poblaciones primarias y secundarias[28].

También las cifras de dengue han ido ascendiendo de manera alarmante, en el 2005 los casos de fiebre por dengue (FD) aumentaron en un 170 % y la fiebre hemorrágica por dengue (FDH) en un 120 % respecto al año anterior[29], en el 2006 aumentaron en un 40 % y 7 % respectivamente[30], en el 2007 en un 85 % y 110 % llama la atención un aumento más notorio en los casos FDH[31], pero en el 2008 se registró un descenso en los casos de dengue en un 34 % y 20 %[32], para el 2009 las cifras se encontraban de nuevamente en ascenso en un 60 % y 50 % respectivamente[33].

En la República Mexicana desde el año 2008 no ha habido modificaciones en los grupos etarios más afectados por FD y FDH, en ambos casos han sido afectados 15-19 años, 10-14 años y 20-24 años[29-33].

En el año 2008, el estado de Morelos presentó la incidencia más alta, 440.8 por 100 mil habitantes[32], para el 2009 fue Colima quién obtuvo la incidencia más alta con 727.3 por 100 mil habitantes[33].

Los serotipos del dengue que circularon en Guerrero en el año 2008 fueron 2 y 3[32], en el año 2009 fueron el 1 y 2[33]. La incidencia en Guerrero para fiebre por dengue y fiebre hemorrágica por dengue en el 2008 fue de 89.5-178.9[32]; en el 2009 fue de 125.8-250.6 por 100 000 habitantes[33].

Acapulco es el municipio del Estado con una incidencia superior al resto de los municipios, por tanto es un problema económico no sólo para la Secretaría de Salud del Estado (programa de vectores, personal, abatización, fumigaciones, hospitalización), sino también para la población (días de ausencia laboral, atención en clínicas privadas, medicamentos, etc.).

En México, se utiliza con mayor frecuencia el control químico, el larvicida más utilizado es el temefós, se ubica dentro del grupo de los organofosforados, cuyo mecanismo de acción consiste en interferir en el sistema de comunicación de las células nerviosas, bloquea la acción de una enzima que inactiva al neurotransmisor acetilcolina, ocasionando convulsiones, parálisis y por consecuencia la muerte[34]. Dentro del control físico, que es el más favorable para el medio ambiente, se encuentran el sellar correctamente los recipientes en donde se almacene agua, con ello se evita la ovipostura, también se recomienda la instalación de mosquiteros en ventanas del hogar y el de mosquiteros en las camas.

El control biológico consiste en introducir a los recipientes, donde se almacena agua, organismos depredadores de larvas de mosquitos como son los peces larvivoros, sin embargo el olor que despide el agua limita su uso doméstico, otro ejemplo de control biológico lo constituyen las tortugas miniatura, copépodos depredadores, y larvas de otras especies (*Toxorhynchites spp*)[10].

La participación comunitaria aunada a un buen control físico, biológico y químico reduce los índices larvarios, pero se desconoce si reduce la transmisión del dengue[35]. Cordero Conejo y colaboradores proponen la participación comunitaria bajo seguimiento de las autoridades de salud, educación y aplicación de control biológico como alternativa ya que la fumigación no permite el control ni la erradicación del *Aedes aegypti*[36].

El objetivo del presente estudio fue generar información que permita proponer estrategias para la disminución del dengue en Acapulco, Guerrero así como la incidencia, identificar factores asociados a padecer dengue, estimar los índices larvarios y pupales de *Aedes aegypti* (índice de recipiente, casa, Breteau, pupal y pupa-persona) y estimar costos de atención de la enfermedad de dengue en el municipio de Acapulco.

2.2. Material y métodos.

Diseño del estudio. Se llevó a cabo la línea basal de un ensayo por conglomerados aleatorizado (ISRCT 27581154) para la reducción de criaderos *A. aegypti* en el mes de diciembre del 2009. Se aplicó un cuestionario que incluyó preguntas sobre información socio demográfica, conocimientos, actitudes y prácticas que tiene la población para proteger sus recipientes con agua y el haber padecido dengue en los últimos doce meses. Se realizó revisión entomológica en los hogares encuestados, inspeccionándose 17 331 recipientes que en ese momento contenían agua.

Se estimaron costos de hospitalización, de medicina, transporte y atención médica con base en información proporcionada por el respondiente en el momento de la encuesta.

Universo de estudio: municipio de Acapulco. Primero se obtuvo un listado de todas las colonias pertenecientes al municipio de Acapulco, posteriormente se ubicaron en un mapa del INEGI, como regla en cada uno de los conglomerados debía existir una escuela primaria o jardín de niños, cada colonia debió tener, al menos, 100 casas contiguas.

Selección de la muestra. Con los datos proporcionados por la Secretaría de Salud de Guerrero (SSA), en un primer paso el municipio de Acapulco se dividió en seis áreas administrativas de acuerdo con la misma secretaria, utilizando la población por área como ponderador, se seleccionaron al azar 30 colonias; se llevó a cabo una visita a las colonias con la finalidad de poder localizar la escuela primaria o jardín de niños dentro o lo más cercano al conglomerado. En un segundo paso, posterior a la realización de la línea de base,

se aleatorizó la intervención integrando 15 colonias al brazo de intervención y 15 el brazo de control, por personal externo del Centro de Investigación de Enfermedades Tropicales (CIET).

Criterios de inclusión. Todos aquellos de hogares en los que una persona con edad igual o superior a 18 años, autorizó, por medio del consentimiento informado verbal, su participación y permitió la revisión entomológica dentro del hogar.

Criterios de exclusión. Domicilios de colonias seleccionadas que se encontraron deshabitados, o que la persona encontrada en el momento de la encuesta no tuviera la capacidad mental que le permitiera otorgar el consentimiento informado.

2.2.1. Definición operacional de las principales variables resultado

Caso de dengue: Toda persona en la cual el respondiente manifestó que padeció dengue en los 12 meses previos a la entrevista.

Larva: Ejemplar de larva identificada como *Aedes aegypti* en el laboratorio.

Pupa: Ejemplar de pupa identificada como *Aedes aegypti* en el laboratorio.

Criadero positivo: Es un recipiente natural o artificial donde se encuentra al menos una larva o pupa de *Aedes aegypti*.

Casa positiva: Casa inspeccionada por personal capacitado del CIET, donde se encontró al menos un criadero positivo a larva o pupa de *Aedes aegypti*.

Peridomicilio: Espacio ubicado en el perímetro de la vivienda hasta 50 metros de ella, incluyendo sus paredes exteriores.

Índice de casa: Es el número de casas con al menos un criadero positivo a larvas entre número de casas inspeccionadas multiplicado por cien.

Índice de recipiente: Es el número de recipiente con al menos una larva o pupa entre número de recipientes inspeccionados multiplicado por cien.

Índice de Breteau: número de recipientes con al menos una larva o pupa encontrados en 100 casas inspeccionadas.

Índice pupal: Es el número de pupas entre el número de casas inspeccionadas multiplicado por cien.

Índice pupa-persona: Es el número de pupas encontradas dividido entre el número de personas.

2.2.2. Instrumentos utilizados

Se utilizaron los siguientes instrumentos: Cuestionario con 49 preguntas, carta de consentimiento informado y un formato de revisión entomológica con la finalidad de recolectar información sobre número y tipo de contenedores, su utilización, cuántos estuvieron infestados con larvas de *Aedes aegypti* y cantidad de ellas.

2.2.3. Prueba piloto

Todos los instrumentos que se utilizaron para realizar el presente trabajo fueron sometidos a pruebas piloto, de tal manera que tanto el cuestionario como el formato de revisión entomológica fueron modificados de acuerdo al resultado de la prueba, con la intención de mejorar la estructura, lógica y coherencia de las preguntas.

Las pruebas pilotos se realizaron en colonias de Acapulco con características similares a las incluidas en la muestra. Previa capacitación en el aula y en campo del personal del CIET, se conformaron brigadas de trabajo constituidas por cinco parejas (un encuestador y un revisor entomológico) y un coordinador de brigada. Una vez probados los instrumentos se procedió a seleccionar el personal para el trabajo de campo, el cual debió tener mínimamente un nivel de estudios de bachillerato.

2.2.4. Trabajo de campo

Los coordinadores de cada brigada visitaron los 30 conglomerados para corroborar la existencia de al menos 100 casas, se estableció contacto con los representantes de cada colonia o con personas reconocidas por la población y se explicaron los objetivos de la investigación. Cada pareja estuvo acompañada por un miembro de la comunidad para generar confianza en los miembros de las casas visitadas. Se visitó casa por casa del conglomerado sin excluir ninguna, se aplicó el cuestionario el cual fue anónimo y confidencial, sólo se preguntó el nombre de los niños entre 3 y 9 años que dieron muestra de saliva, se realizó la revisión entomológica, en caso de encontrarse recipientes con larvas o pupas se identificó la muestra con el folio de la casa en una etiqueta,

se anexó también la fecha, el número del conglomerado, tipo de recipiente positivo a larvas o pupas, las que se trasladaron al los laboratorios del CIET, se mantuvieron en congeladores a -20 grados centígrados y fueron clasificadas por personal de la escuela de Ciencias Biológicas de la Universidad Autónoma de Guerrero. Dos meses después se obtuvo una segunda muestra de saliva de los niños que habían proporcionado muestra durante la encuesta, lo cual no es tema del presente artículo.

2.2.5. Análisis de los datos

Se realizó doble captación de datos para minimizar los errores, los que fueron codificados, capturados, validados y analizados con los paquetes estadístico-epidemiológico de dominio público Epi-data[37] y CIET map[38]. Las variables de exposición se analizaron a través de de tablas de contingencia 2 x 2, se realizaron contrastes para poder estimar la relación entre haber padecido dengue y las variables de estudio. La estimación del riesgo (Razón de momios), el nivel de significancia y el análisis multivariado para identificación de factores de confusión o modificadores de efecto se hizo con el proceso de Mantel-Haenszel[39], los intervalos de confianza se calcularon con la propuesta de Miettinen[40].

La prueba de significancia estadística que se utilizó para evaluar la diferencia de los promedios de los costos por hospitalización y por atención médica de pacientes que padecieron dengue fue la prueba de la t de Student (GraphPad Quick Calcs: t test calculator)[41].

2.2.6. Consideraciones éticas

El protocolo fue aprobado por el Comité de Ética del Centro de Investigación de Enfermedades Tropicales conformado por miembros que no estuvieran participando en el proyecto para evitar conflicto de interés. Se les entregó a los representantes de las colonias un oficio en el cual se les daba a conocer la finalidad del proyecto de investigación.

Durante el trabajo de campo se pidió autorización a los jefes de los hogares por medio del consentimiento informado verbal para llevar a cabo la entrevista y la revisión entomológica en su hogar, ésta se realizó, de preferencia, acompañados por algún habitante del hogar.

Se entregaron resultados preliminares en las casas encuestadas en el mes de marzo de 2010. Los resultados finales se darán a conocer a los participantes del estudio de cada comunidad, mientras que con las autoridades de salud serán discutidos para que sean usados oportunamente para elaborar o reforzar programas de prevención.

2.3. Resultados

2.3.1. Información de los hogares

Se encuestaron 4 389 hogares de 30 colonias del municipio de Acapulco en el mes de diciembre del 2009, recojiéndose información de 19 707 personas. En el 93 % (4 078/4 379) de las viviendas el uso era exclusivamente hogar, en el 6 % (258/ 4389) el uso era hogar más negocio, 0.9 % (43 / 4 389) el uso era exclusivamente negocio, en el 0.2 % (10/4389) no se obtuvo información.

El tipo de vivienda predominante fue la permanente 63 % (2779/4389), seguida de semipermanente, 25 % (1075/4389); y endeble, 12 % (508/4389). El 85 % (3731/4389) obtienen el agua para beber de garrafón; llave, 5.8 % (256/4389); manantial, 5.1 % (223/4389), el resto se distribuye entre pozo, filtro, río y pipas.

El 47 % (2052/4389) refirió poco peligro de enfermar por dengue, mucho peligro, 26.4 % (1158/4389); ningún peligro 26.1 % (1147/4389) y peligro regular, 0.2 % (8/4389).

El 97 % (4249/4389) de los respondientes identificó físicamente a la larva, 2.8 % (124/4389) no la identificó.

2.3.2. Incidencia de la enfermedad

La tasa de casos de dengue notificada por los respondientes fue del 3.5 % (663/19 707), en niños de 3 a 9 años fue del 0.6 %, la relación hombre-mujer fue 1:1. De los 663 casos, el 60 % (390/663) se atendieron en instituciones públicas, el 31 % (208/663) en clínicas y farmacias privadas, el 5.9 % (39/663) en su hogar y el 4 % (26/663) no proporcionaron información al respecto.

En la distribución de los casos por grupo de edad, podemos observar que los grupos etarios más afectados son 10-14 años, 15-19 años y de 5-9 años (Figura 2.1).

2.3.3. Análisis bivariado

Dentro de los factores asociados a padecer dengue se encontró el utilizar productos antimosquitos (RM 1.3, IC95 % ajustado por clusters 1.07-1.72), el tener un grado de estudios superior al sexto año de primaria (RM 1.5, IC ajustado por clusters 1.28-1.96), el tener un trabajo remunerado (RM 1.5, IC ajustado por clusters 1.17-2.08), el no tener larvas en sus recipientes (RM 1.2, IC95 % IC ajustado por clusters 1.0-1.67), el habitar en sitios urbanos (RM 1.5, IC ajustado por clusters 1.13-2.10). Ver cuadro (2.1).

Figura 1. Distribución de casos de dengue en los últimos 12 meses por grupos etarios

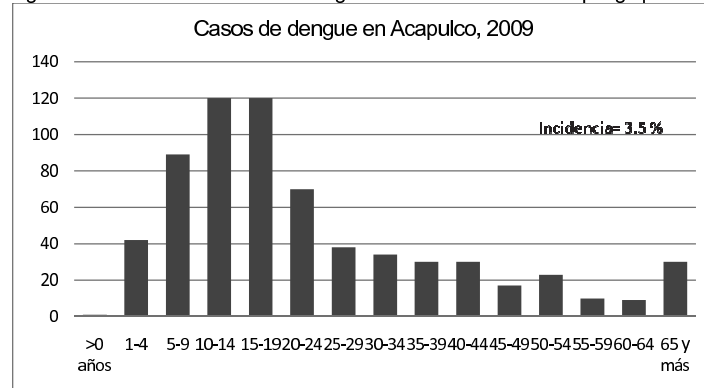


Figura 2.1: Distribución de casos de dengue en los últimos 12 meses por grupos etarios.

2.3.4. Análisis multivariado

Se realizó el análisis multivariado manteniéndose, en el modelo final, como factores asociados a padecer dengue el haber referido tener peligro de enfermar por dengue (RM 2.17, IC 95% ajustado por clusters 1.69-2.79), el tener un jefe de familia con un nivel de estudios superior al sexto año de primaria (RM 1.32, IC95% ajustado por clusters 1.06-1.63), el tener un trabajo remunerado (RM 1.48, IC95% ajustado por clusters 1.09-2.02), el habitar en sitios urbanos (RM 1.53, IC95% ajustado por clusters 1.53) y el no tener larvas en sus depósitos en donde se almacena agua (RM 1.24, IC95% ajustado por clusters 1.00-1.54). Los resultados se indican en el cuadro (2.2)

2.3.5. Índices entomológicos

Se estimó un índice de casa positivo de 15.5%, índice de recipientes positivos de 5.1%, índice pupa- persona de 8.4%, un índice de Breteau de 20.3% y un índice pupal 37.8%.

El 26% (4 412/16 797) de los recipientes se encontraron dentro del domicilio, el 30% (5 021/16 797) de los recipientes se encontraron con abate, el 44% (7 401/16 797) se encontraron tapados, sin embargo el 5% (812/16 797) de las tapas funcionan como criaderos, el 33% (5 597/16 797) de los recipientes se encontraron sellados, de estos el 4% (243/5 597) de los recipientes se encontraron positivos a larvas y el 1% (72/5 597) se encontraron positivos a pupas, las ollas, 7%; pilas, 4%; tambos, 2%; fueron los recipientes con mayor positividad.

El 2% (85/5 021) de los recipientes donde convencionalmente se almacena agua que se encontraron con abate en el momento en que se realizó la revisión entomológica tenían pupas y el 4% (208/5 021) tenían larvas.

2.3.6. Costos

Costos por concepto de consulta, medicina y transporte Se calculó el costo en pesos de la atención médica por concepto de consulta, medicinas, transporte de los pacientes que refirieron haber padecido dengue en la encuesta, en promedio un enfermo de dengue gastó 785 pesos, con un rango de 10 a 20 000, la moda fue de 500, mediana de 400 y la suma de los gastos de los 588 pacientes fue de \$461 981.

Atención en su hogar. De los 663 pacientes padecieron dengue según información proporcionada por el respondiente, 39 se atendieron en su hogar pero solo 29 personas refirieron haber gastado en su atención médica, medicinas y transporte. Ver cuadro (2.3)

Pacientes atendidos en instituciones públicas y privada De los 663 pacientes que refirieron haber padecido dengue en el momento en que se les entrevistó, 559 refirieron haber gastado en su paciente, en el cuadro (2.4) se pueden observar las diferencias respecto a costos.

Días que perdieron al cuidado del enfermo De los 663 pacientes que respondieron haber padecido dengue el último año, 633 mencionaron haber perdido días de trabajo y al cuidado del enfermo, esta información permitió estimar la diferencia de días perdidos entre ambos sexos alcanzando confianza estadística, cuadro (2.5).

Estimación de costos por concepto de hospitalización Se estimaron costos por concepto de hospitalización, el costo estimado fue de 1 288 pesos, se procedió a estratificar por sexo, pero las diferencias encontradas en el gasto promedio entre hombres y mujeres, fácilmente se explica al azar, ($p = 0,2$, $t = 1,09$, $df = 123$, $EE D = 411,1$).

Estimación de costos en pacientes que no requirieron hospitalización estratificado por sexo. De los 440 pacientes que no requirieron hospitalización se procedió a estratificar por sexo, pero las diferencias encontradas en el gasto promedio entre hombres y mujeres, fácilmente se explica al azar, ($p = 0,8$, $t = 0,2$, $df = 438$, $EE D = 89,3$).

Costos generados por la compra de insecticidas De los 4 389 hogares encuestados sólo en 2 195 los respondientes afirmaron haber gastado mensualmente en la compra de productos antimosquitos, en promedio cada hogar gastó 66 pesos mensuales en la compra de insecticidas, con un rango de (1-700), y una media de 50, la suma del gasto mensual generado por la compra de productos antimosquitos en los hogares encuestados fue de \$144 941.

Cuadro 2.1: Análisis bivariado entre padecer dengue y las variables incluidas en el estudio. IC95 % ajustado por clusters.

Variable	Dengue		RM	IC95 %	χ^2 M-H
	Sí	No			
1= Hogar más negocio	45	615	1.10	0.70 - 1.71	0.34
2= Exclusivamente hogar	1 154	17 289			
1= Vivienda permanente	433	220	1.18	0.91 - 1.55	3.88
2= Vivienda semipermanente	11 470	6 878			
1= No percibir peligro de padecer	581	81	2.34	1.74 - 3.15	53.46
2= Percibir algún grado	13 871	4 531			
1= Tiene abate en sus recipientes	474	170	1.30	1.0 - 1.68	8.35
2= No tiene abate	12 221	5 695			
1= Menos de dos meses con abate	523	120	1.16	0.82 - 1.63	2.01
2= Más de dos meses con abate	14 131	3 752			
1= Utiliza productos antimosquitos	392	265	1.35	1.06 - 1.72	13.97
2= No utiliza productos antimosquitos	9 582	8 755			
1= Si conoce las larvas <i>A. aegypti</i>	649	10	1.68	0.92 - 3.07	2.68
2= No conoce las larvas	17 952	466			
1= Identifica por dist. nombres la larva <i>A. aegypti</i> .	594	54	1.15	0.79 - 1.67	0.96
2= No la identifica	16 257	1 702			
1= Conocimiento correcto del origen de las larvas	170	476	1.32	1.06 - 1.64	9.43
2= Sin conocimiento	3 824	14 156			
1= No estar de acuerdo que los zancudos que transmiten el dengue vienen de charcos o arroyos	118	538	1.26	0.97 - 1.65	5.12
2= Estar de acuerdo	2 695	15 530			
1= No piensa eliminar los criaderos en su casa	7	655	1.20	0.56 - 2.58	0.23
2= Si tiene pensado eliminarlos	162	18 265			
1= Depende de la gente el control del dengue	505	155	1.04	0.81 - 1.34	0.17
2= Depende del sector salud o de ambos	13 892	4 431			
1=Platica con sus vecinos sobre cómo evitar el dengue	303	357	1.24	1.04 - 1.48	7.44
2= No platica	7 483	10 948			
1= Sexto grado o más de educación	522	136	1.60	1.29 - 1.98	23.52
2= Quinto grado o menos	12 920	5 383			
1= Trabajo remunerado	584	72	1.34	1.01 - 1.77	5.37
2= No tener trabajo remunerado	15 786	2 608			
1= No tener larvas	581	81	1.29	1.0 - 1.67	4.46
2= Tener larvas	15 672	2 810			
1= No tener pupas	34	628	0.76	0.57 - 1.03	2.29
2= Tener pupas	1 223	17 259			
1= Sitios urbanos	534	128	1.54	1.13 - 2.10	19.14
2= Sitios rurales	13 493	4 989			
1= Sexo femenino	303	357	1.24	1.04 - 1.48	7.44
2= Sexo masculino	7 483	10 948			
1= De cero a 24 años de edad	441	221	2.08	1.68 - 2.58	79.98
2= De 25 o más	9 042	9 438			

Cuadro 2.2: Análisis multivariado del riesgo de padecer dengue con las variables que se mantuvieron en el modelo final.

Variable	RMns	IC95 %		χ^2 het	p
		wt RM	cl adj wt RM		
No tener larvas	1.24	0.97 - 1.58	1.00-1.54	9.1	0.86
Peligro de padecer dengue	2.22	1.77 - 2.79	1.73 - 2.86	12.1	0.66
Grado escolar \geq 6 años	1.41	1.15 - 1.72	1.12 - 1.77	11.6	0.7
Tener trabajo remunerado	1.13	0.87 - 1.48	0.83 - 1.56	7.84	0.92
Conglomerados urbanos	1.5	1.23 - 1.82	1.05 - 2.14	10.05	0.81

Cuadro 2.3: Promedio de gastos en pesos de los pacientes que padecieron dengue y se atendieron en su hogar.

Estadístico	Atención en su hogar $n = 29^*$
Promedio	190
Desviación estándar	168
Rango de gastos	10-600
Moda	200
Mediana	125

*Solo se incluyen casos de dengue que refirieron haber gastado.

Cuadro 2.4: Promedio de gastos en pesos de los pacientes que refirieron haber recibido atención médica al enfermarse por dengue en Acapulco, Guerrero durante el 2009, sin tomar en cuenta los que se atendieron en su hogar.

$n = 559$	Instituciones públicas	Instituciones privadas
	$n_1 = 356^*$	$n_2 = 203^*$
Promedio	554	1 276
Desviación estándar	867	1 912
Rango de gastos	10 - 7 000	50- 20 000
Moda	200	500
Mediana	300	800

$p < 0,001, t = 6,10, df = 557, EED = 118,20$

*Solo se incluyen casos de dengue que refirieron haber gastado.

Cuadro 2.5: Promedio de días perdidos de trabajo y al cuidado del enfermo estratificado por sexo en Acapulco, Guerrero, 2009.

	Masculino	Femenino
$n = 633$	$n_1 = 317^*$	$n_2 = 316^*$
Promedio	12.8	11.5
Desviación estándar	8.6	7.6
Rango de gastos	1-31	1-31
Moda	7	7
Mediana	10	8

$p = 0,046, t = 1,996, df = 631, EED = 0,650$

*Solo se incluyen casos de dengue que refirieron haber perdido días.

2.4. Discusión

2.4.1. Limitaciones del estudio

La temporalidad del estudio no está perfectamente definida, por tanto no se puede concluir si la causa precede al efecto, en la asociación encontrada entre padecer dengue y tener larvas en sus depósitos con agua.

Se estimaron los días de ausencia laboral y los destinados al cuidado del enfermo en general, pero no fue posible estimar cuantos días perdieron la población estudiantil y la trabajadora debido a que no preguntamos si la persona que padeció dengue trabajaba, estudiaba o realizaba las dos actividades.

2.4.2. Incidencia.

La incidencia de dengue en la población general fue del 3.5 %, en niños y adolescentes fue del 1.6 %, sin predominio de sexo, en 1997 en Santiago de Cuba, Orozco Hechaverría y colaboradores encontraron una incidencia superior, reportaron un 3.5 % en niños y adolescentes con predominio en el sexo masculino[42]. En el 2002 en Colima, Espinoza Gómez y colaboradores encontraron una incidencia del 1.7 % en la población general[43]. En el 2007 en Maracay, Venezuela, Espino y colaboradores encontraron en población mayor a cinco años, tasas de incidencia de 5.7 y 18.6 por 100 000 personas por día para infecciones sintomáticas y asintomáticas, respectivamente[44]. En el tercer y cuarto estudio se realizaron las pruebas serológicas con ensayo inmunoenzimático de captura de IgM anti-dengue y el presente estudio se basó en lo notificado por el respondiente.

Conocimiento del agente transmisor de la enfermedad. En el presente estudio en el 79 % (3 376/4 281) de los hogares no sabían o tenían un conocimiento erróneo sobre el origen de las larvas, mientras que Hoyos Rivera y colaboradores refirieron el 90 % (482/535) de la población[45]. Ambas cifras se asemejan, y denotan carencia de conocimiento sobre el ciclo biológico del *Aedes*, el cual es importante para poder entender el mecanismo de transmisión y control de la enfermedad. El instrumento utilizado por Hoyos Rivera fue una encuesta escrita en la cual los respondientes tuvieron ocho opciones mientras que en el primer estudio fue una pregunta abierta.

Sitio de reproducción. Acosta Cabrera y colaboradores encontraron que el 93.7 % (558/596) de la población encuestada sabía que los zancudos se reproducen en el agua depositada[46], mientras que en el presente estudio en

el 84 % (3 698/4 389) de los hogares piensan que los zancudos que transmiten el dengue vienen de charcos o arroyos, es decir que se originan de agua que está en constante movimiento, existe una diferencia notable entre los conocimientos de la población de la Habana y la de Acapulco.

Empoderamiento de la prevención de la enfermedad. Acosta Cabrera y colaboradores encontraron que el 81 % de la población encuestada opina que la responsabilidad de disminuir los focos de mosquitos es labor de “organismos, organizaciones y comunidad” [46] esta cifra difiere con la encontrada en nuestro estudio ya que únicamente el 16.7 % (732/4 389) de los hogares encuestados opinaron que el control del dengue dependía de ambos, es decir de ellos mismos y del personal de salud. Por tanto es necesario lograr una participación comunitaria, supervisada por la secretaría de salud.

2.4.3. Factores relacionados a padecer dengue

Se encontró asociación entre padecer dengue y los siguientes factores:

1. Habitar en sitios urbanos. El aumento de la población es mayor en estas áreas, por tanto hay mayor consumo de agua entre otras cosas, la escasez de la misma en sitios urbanos obliga a la población a almacenarla y si no se lleva a cabo un sellado correcto de los recipientes pueden funcionar como criaderos de larvas y pupas.
2. El no tener larvas. Una vez que la Secretaría de Salud detecta un caso de dengue, le siguen acciones como el rociado espacial en la manzana donde se encuentra ubicada la vivienda con el caso positivo a dengue, así como colocar abate en depósitos de agua, realizar revisión entomológica y una capacitación a los habitantes del domicilio lo que trae como consecuencia que, cuando se visitan esos hogares, no encontremos larvas en los potenciales criaderos.
3. El tener un jefe de familia con nivel de estudios superior al sexto año de primaria. Lo anterior probablemente se refleje en prestar mayor atención a los distintos mensajes que otorga promoción a la salud a la población, de tal manera que este factor puede influir en entender más la problemática del dengue.
4. El haber referido tener peligro de enfermarse por dengue. El percibir peligro puede estar relacionado con tener conocimiento sobre la enfermedad y sobre todo con el mecanismo de transmisión pero también

puede estar relacionado al haber padecido dengue por parte de algún miembro del hogar.

2.4.4. Costos

Días de ausencia laboral. La población perdió en promedio 12 días por la enfermedad y el cuidado del enfermo, cifra similar hallada por Suaya y colaboradores quienes encontraron que un paciente con diagnóstico de dengue, tratado de manera ambulatoria, perdió en promedio 11.9 días, mientras que un hospitalizado perdió 12 días de trabajo, entre los hospitalizados un estudiante perdió 5.6 días de clases mientras que un trabajador se ausentó 9.9 días[47]. En nuestro estudio no obtuvimos información específica sobre ausentismo de población estudiantil y trabajadora, por tanto no podemos comparar entre estos dos tipos de poblaciones. En general este ausentismo laboral representa un gran impacto en la economía de los países, sobre todo aquellos que abarcan áreas endémicas.

Costos de atención médica y hospitalización. En el presente estudio una persona que enfermó por dengue gastó en promedio \$785 y en caso de haber sido hospitalizado gastó \$1 288, Añez y colaboradores estimaron \$1 056 por concepto de hospitalización[48], se realizó la conversión a pesos tomando como referencia el valor del dólar en el año en que se realizó el estudio[49], son cifras inferiores si se comparan con las encontradas por Suaya y colaboradores, ellos reportaron \$6 682 por atención ambulatoria y \$18 122 por paciente hospitalizado[47], Airmen y colaboradores reportaron \$3 695 y \$11 853 respectivamente[50], el segundo estudio fue efectuado en Zulia Venezuela, el tercer estudio fue multicéntrico el cual abarcó cinco países de América y tres de Asia, el cuarto estudio se llevó a cabo en Panamá. Los estudios referidos se realizaron en diferentes años, el primero se efectuó en el 2009, el segundo en el 2004, el tercero y el cuarto en el 2005.

Índices entomológicos. El valor de el índice de casa, de recipiente y de Breteau encontrados en el presente estudio se asemejan a los del rango mayor encontrados en Costa Rica en 2007. Allá, en 15 localidades que pertenecen al Cantón de Limón, registraron un índice de casa que varió de 3.9 a 19.7, el índice de recipiente varió de 1.9 a 9.4 y el índice de Breteau de 1.6 a 20.3[51]; Hayes y colaboradores encontraron un índice de Breteau de 62 y un índice de casa de 36[52]. Las diferencias encontradas se podrían explicar ya que se efectuaron en diferentes años y situaciones, el primero fue en Acapulco en el

2009; el segundo en el 2007 en Costa Rica y el tercero en Las Pampitas en El Salvador durante el año 2000 mismo periodo en que se registró la epidemia.

2.5. Conclusiones

La incidencia de dengue encontrada en el presente estudio es media. Tomando en cuenta a los niveles de riesgo de epidemias de dengue de acuerdo a los tres índices médicos (PAHO 1994) el índice de casa, de recipiente y de Breteau se encuentran en nivel medio.

Existe falta de interés, conocimiento y organización de la población para desempeñar labores en las colonias para el control del vector del dengue que finalmente son de gran beneficio para todos en diferentes ámbitos sean económicos, salud.

El costo de un paciente que enfermó por dengue es alto comparado con el salario mínimo que percibe un alto porcentaje de la población de la población y aumenta un 40 % en caso de requerir hospitalización.

Bibliografía

- [1] Torres-Muñoz A. La Fiebre Amarilla en México. Erradicación del Aedes Aegypti. Salud pública de Mexico 1995; 37(Supl 1): s103-s110.
- [2] Góngora-Biachi R. La erradicación de la fiebre amarilla en Mérida, Yucatán: una historia de tenacidad y éxito. Rev Biomed 2004; 15(4):251- 258.
- [3] Montesano-Castellanos R, Briseño-García B, Rodríguez-Solís E, Jaimes-Domínguez F, Solís-Santamaría I, Araujo-Carranza F. Manual Simplificado para la Vigilancia Epidemiológica del Dengue en México, Dirección General de Epidemiología, segunda edición enero 1997. Disponible en: <http://www.dgepi.salud.gob.mx/infoepi/manuales/Man19-Dengue/Man19.htm> [Consultado el 15 de agosto del 2010].
- [4] Rodríguez-Cruz R. Estrategias para el control del dengue y del Aedes aegypti en las Américas. Rev Cubana de Med Trop 2002;54(3):189-201.
- [5] Organización Mundial de la Salud. Dengue y dengue hemorrágico. Nota descriptiva 117, marzo de 2009. <http://www.who.int/mediacentre/factsheets/fs117/es/> [Consultado el 18 de agosto del 2010].
- [6] PAHO. Number of reported cases of dengue and dengue hemorrhagic fever (DHF), Region of the Americas (by country and subregion). Washington, DC, Pan American Health Organization, 2008. (<http://www.paho.org/english/ad/dpc/cd/dengue.htm>).
- [7] Dengue –Guidelines for diagnosis, Treatment, prevention and control. New edition 2009.For research on diseases of poverty. World Heart Organization. Disponible en: <http://whqlibdoc.who.int/publications/2009/9789241547871-eng.pdf> (consultado el 25 de octubre del 2010).
- [8] Erlanger TE, Kerser J, Utilinger J. Effect of dengue vector control intervenciones on entomological parameters in developing countries: a systematic review and meta-analysis. The Royal Entomological Society, Medical and Veterinary Entomology(2008) 22, 203-221.
- [9] Toledo-Romaní ME, Baly-Gil A, Ceballos-Ursula E, Boelaert M, van der Stuyft P.Participación comunitaria en la prevención del dengue: un abordaje desde la perspectiva de los diferentes actores sociales. Salud Publica Mex 2006; 48(1):39-44.

- [10] Fernández- Salas I. *Biología y Control del Aedes Aegypti*. Manual de Operaciones. Universidad Autónoma de Nuevo León 1999. Primera edición. Páginas 14-19, 27-30.
- [11] Araña Oliver O, Méndez Martínez J, Portillo Hernández R, Suárez Hernández R. Comportamiento del *Aedes aegypti* en la provincia de Matanzas durante los años 2004-2008. *Rev méd electrón [Revista en la internet]* 2010;32(1). Disponible en: <http://scielo.sld.cu/scielo.php?script=sci-arttext&pid=S1684-18242010000100005&lng=es>
- [12] Faingezicht I, Ávila M. Diagnóstico clínico y de laboratorio del paciente con dengue. *Rev. Méd. Hosp. Nac. Niños (Costa Rica)* 1999; 34(Supl):s33-s41. [revista en la Internet]. [citado 2011 Ene 03]. Disponible en: <http://www.scielo.sa.cr/scielo.php?script=sci-arttext&pid=S1017-85461999000100004&lng=es>.
- [13] Chahar H, Bharaj P, Dar L, Guleria R, Kabra S, Broor S. Co-infections with Chikungunya Virus and Dengue Virus in Delhi, India. *Emerg Infect Dis.* 2009; 15(7): 1077-1080.
- [14] Rifakis P; Gonçalves N; Omaña W; Manso M, Espidel A, Intingaro A, et al. Asociación entre las variaciones climáticas y los casos de dengue en un hospital de Caracas, Venezuela, 1998-2004. *Rev Peru Med Exp Salud Pública* 2005; 22 (3): 183-190.
- [15] Dias-Zeiler J; Amezaga-Acosta PO; Pereira-Barreto P; Da Silva-Cordeiro J. Vírus dengue em larvas de *Aedes aegypti* e sua dinâmica de infestação, Roraima, Brasil. *Rev Saú de Pública* 2008;42(6):986-991.
- [16] Manrique-Saide P, Davies C, Coleman P, Rebollar-Tellez E, Che-Mendoza A, Dzul-Manzanilla F et al. Pupal surveys for *Aedes Aegypti* surveillance and potencial targeted control in residential areas of Mérida, México. *Journal of the American Mosquito Control Association*, 2008; 24(2):289-298.
- [17] Gómez-Dántes H. Elementos económicos y políticos que impactan en el control del dengue en México. *Salud pública de México* 2007, vol. 49, edición especial, XII congreso de investigación en salud pública.
- [18] Pozo EJ, Neyra M.C, Vílchez E, Meléndez M. Factores Asociados a la Infestación Intradomiciliaria por *Aedes aegypti* en el distrito de Tambogrande, Piura 2004. *Rev Peru Med Exp Salud Pública* 2007; 24(2): 144-51.
- [19] Danis-Lozano R, Rodríguez MH, Hernández-Ávila M. Escolaridad con relación al género de los jefes de familia y el riesgo de cría de *Aedes aegypti* en el sur de México. *Salud Publica Mex* 2002; 44(3):237-242.
- [20] Cáceres JL., Giovinco L, Rubén M, Rondón R, Díaz MJ, Rojas V. Factores asociados a infección por virus dengue en menores de 15 años de edad. Maracay Venezuela 1997. *Boletín de Malariología y saneamiento ambiental.* 2002; 42(1,2):13-16.
- [21] Navarrete-Espinosa J, Acebedo-Vales J, Huerta-Hernández E, Torres-Barranca, Gavaldón-Rosas D. Prevalencia de anticuerpos contra dengue y leptospira en la población de Jáltipan, Veracruz. *Salud Publica Mex* 2006;48(3):220-228.

- [22] Vezzani D, Albicócco A. The effect of shade on the container index and pupal productivity of the mosquitoes *Aedes aegypti* and *Culex pipiens* breeding in artificial containers. *Medical and Veterinary Entomology* 2009; 23(1):78-84.
- [23] Calderón-Arguedas O, Troyo A, Solano M. Caracterización de los sitios de multiplicación de *Aedes aegypti* (Diptera: Culicidae) en el caserío "La Carpio", San José, Costa Rica durante la estación seca del año 2003. *Rev Biomed* 2004; 15(2):73-79.
- [24] Erlanger TE, Kerser J, Utlinger J. Effect of dengue vector control intervenciones on entomological parameters in developing countries: a systematic review and meta-analysis. *The Royal Entomological Society, Medical and Veterinary Entomology*(2008) 22, 203-221.
- [25] Mammen MP Jr, Pimgate C, Koenraadt CJM, Rothman AL, Aldstadt J, Nisalak A et al. Spatial and temporal clustering of dengue virus transmission in Thai villages. *Plos Med* 2008;5 (11): e 205.
- [26] Valdés L, Vila Mizrahi J, Guzmán M. Impacto económico de la epidemia de dengue 2 en Santiago de Cuba, 1997. *Rev Cubana Med Trop* 2002;54(3):220-27.
- [27] Nisalak A. Serotype-specific dengue virus circulation and dengue disease in Bangkok, Thailand from 1973 to 1999. *Am J Trop Med Hyg* 2003;68(2):191-202.
- [28] Viboud C, Bjornstad O, Smith D, Simonsen L, Miller M, Grenfell B et al. Synchrony, Waves, and Spatial Hierarchies in the Spread of Influenza. *Science* 2006; 312 (5772):447-451.
- [29] Panorama epidemiológico del dengue y dengue hemorrágico en entidades federativas. Información publicada a la semana 52 (del 25 al 31 de diciembre del 2005). Disponible en: <http://www.cenave.gob.mx/dengue/panorama/panoramasemana52.pdf>. [Consultado el 18 de agosto del 2010].
- [30] Panorama epidemiológico del dengue y dengue hemorrágico en entidades federativas. Secretaría de salud. Información publicada a la semana 52 (del 24 al 30 de diciembre del 2006). Disponible en: <http://www.cenave.gob.mx/dengue/panorama/Panoramasemana522006.pdf>. [Consultado el 18 de agosto del 2010].
- [31] Panorama epidemiológico del dengue y dengue hemorrágico en entidades federativas. Secretaría de salud. Información publicada a la semana 52 (Del 23 al 29 de diciembre del 2007.) Disponible en: http://www.cenave.gob.mx/dengue/panorama/Panoramasemana_522007.pdf. [Consultado el 18 de agosto del 2010].
- [32] Panorama epidemiológico de fiebre y fiebre hemorrágica por dengue. Secretaría de Salud. (Del 28 de diciembre del 2008 al 3 de enero del 2009, semana 53) Disponible en; [http://www.dgepi.salud.gob.mx/denguepano/2008 -PANORAMAS/ PANORAMA](http://www.dgepi.salud.gob.mx/denguepano/2008-PANORAMAS/PANORAMA)

- [33] Panorama epidemiológico de fiebre y fiebre hemorrágica por dengue. Secretaría de Salud. (Del 27 de diciembre 2009 al 02 de enero del 2010, semana 52) Disponible en: <http://www.dgepi.salud.gob.mx/denguepano/2009- PANORAMAS/ PANO- RA- MA>
- [34] Dirección del Programa de Enfermedades Transmitidas por Vector. Métodos de control de *Aedes aegypti* mosquito vector del virus del dengue en México. <HTTP://WWW.CENAVECE.GOB.MX/DENGUE/ APROBADOS2009.PDF> [Consultado el 25 de mayo del 2010].
- [35] Heintze C, Velasco-Garrido M, Kroeger A. What do community-based dengue control programmes achieve? A systematic review of published evaluations. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 2007; 101(4):317-325.
- [36] Cordero-Conejo A, Sandí- Nogueral F, Mejías- Fonseca G, Silva- Silva R, Campos- Rodríguez M, Soto L, et al . Reseña de una experiencia de promoción de la salud y lucha contra el dengue utilizando el control biológico, en escuelas de Chacarita, Puntarenas, Costa Rica. *Rev. costarric. salud pública* 2000; 9(17): 20-25.
- [37] Lauritsen JM, Bruus M. EpiData (version 3). A comprehensive tool for validated entry and documentation of data. Odense, Denmark: The EpiData Association; 2003.
- [38] Andersson N, Mitchell S. CIETmap: Free GIS and epidemiology software from the CIETgroup, helping to build the community voice into planning. World Congress of Epidemiology, Montreal, Canada, August 2002.
- [39] Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease *J Natl Cancer Inst* 1959;22:719-748.
- [40] Miettinen, O., and M. Nurminen. 1985. Comparative analysis of two rates. *Statistics in Medicine* 4: 213-226.
- [41] GraphPad Quick Calcs: t test calculator. (2002). Retrieved, Enero 2004, from <http://graphpad.com/quickcalcs/ttest1.cfm?Format=C>.
- [42] Orozco- Hechavarria N, Díaz-Portuondo M, Cañete A, Martínez-Delgado Y. Incidencia de dengue en niños y adolescentes. *Rev Cubana Med Trop* 2001;53(1): 16-9.
- [43] Espinoza-Gómez F, Hernández-Suárez CM, Rendón-Ramírez R, Carrillo-Alvarez ML, Flores-González JC. Transmisión interepidémica del dengue en la ciudad de Colima, México. *Salud Publica Mex* 2003;45:365-370.
- [44] Espino C, Comach G, Sierra G, Guzmán D, Camacho D, Cabello de Quintana M et al. Incidencia de infecciones sintomáticas y asintomáticas por virus dengue en Maracay, Venezuela: 2006 - 2007. *Bol. Mal. Salud Amb* 2010;50(1):65-74.
- [45] Hoyos Rivera A, Pérez Rodríguez A. Nivel de conocimientos sobre el dengue en San Mateo, Anzoátegui, Venezuela. *Revista Cubana de Salud Pública* 2009; 35(4):161-172.

- [46] Acosta Cabrera O, Chaljub Moreno A, Baydes González R, Abraham Maciel E. Modificación de los conocimientos, actitudes y prácticas de la población sobre la prevención de los mosquitos. *Rev Cubana Hig Epidemiol* 1999; 37(1):6-12.
- [47] Suaya J, Shepard D, Siqueira J, Martelli C, Lum L, Tan L, et al. Cost of Dengue Cases in Eight Countries in the Americas and Asia: A Prospective Study *Am J Trop Med Hyg* 2009; 80(5): 846-855.
- [48] Añez G, Balza R, Valero N, Larreal Y. Impacto económico del dengue y del dengue hemorrágico en el Estado de Zulia, Venezuela, 1997-2003. *Rev Panam Salud Publica*. 2006;19(5):314-20.
- [49] Banco Central de la República Bolivariana de Venezuela. Tipos de cambio (anual) [sitio en Internet]. Disponible en: <http://www.bcv.org.ve/cuadros/2/212a.asp?id=64>. Acceso el 03 de marzo del 2011.
- [50] Armien B, Suaya J, Quiroz E, Sah B, Bayard V, Marchena L, et al. Clinical Characteristics and National Economic Cost of the 2005 Dengue Epidemic in Panama. *Am J Trop Med Hyg* 2008;79(3): 364-371.
- [51] Marín Rodríguez R, Marquetti Fernández M, Díaz Ríos M. Índices larvales de *Aedes aegypti* antes y después de intervenciones de control en Limón, Costa Rica. *Rev Cubana Med Trop* [revista en la internet] 2009; 61(2). Disponible en: <http://scielo.sld.cu/scielo.php?script=sci-arttext&pid=S0375-07602009000200008&lng=es>.
- [52] Hayes J, García Rivera E, Flores Reyna R, Suarez Rangel G, Rodríguez Mata T, Coto Portillo R. Risk factors for infection during a severe dengue outbreak in el Salvador en el 2000. *Am J Trop Med Hyg* 2003;69(6): 629-633.

Capítulo 3

Consenso y evaluación de clasificadores

3.1. Análisis de la efectividad de la diagnóstico del dengue en el Estado de Guerrero, México 2007: hechos estadísticos y consenso.

Sira Allende¹, Carlos Bouza¹; Saúl López Silva² y Agustín S. Moreno³

1. Universidad de La Habana, Cuba

2. Laboratorio Estatal de Salud Pública "Dr. Galo Soberón y Parra", Guerrero, México

3. Universidad Autónoma de Guerrero, México

Resumen: En este trabajo se presenta un estudio de los criterios emitidos por la clínica respecto a las personas que acudieron a consulta, considerándose como infectados por el dengue, en el estado de Guerrero, México, durante el 2006. La clasificación de la clínica fue evaluada tomando los resultados de los exámenes de laboratorio como regla de oro. Para hacerlo se llevó a cabo un estudio con todas las personas consultadas que tenían los síntomas. El consenso de los médicos y el laboratorio fue evaluado al estudiar datos recogidos aplicando los índices de consenso más populares; Se contrastó una baja consensualidad en la predicción del dengue clásico mientras que los niveles para el hemorrágico fueron más adecuados. Este estudio se llevó a cabo analizando cada jurisdicción del estado. Este estudio del acuerdo entre clínica y laboratorio fue desarrollado por vez primera. Este junto con el análisis de datos brindaron una panorámica del comportamiento de esa epidemia.

Key words: Consensus index, kappa, golden rule, efficiency, dengue and dengue hemorrhagic fever.

3.1.1. Introducción

El dengue es un síndrome febril agudo de origen viral, considerada como una enfermedad emergente, representa un problema de salud pública en todo el mundo y una enfermedad endémica en América, Sudeste de Asia, Oeste del Pacífico, África y el este del Mediterráneo. Es causada por un virus que pertenece al grupo de los arbovirus que se transmite por la picadura de las hembras de mosquitos de la especie *Aedes aegypti*. La enfermedad está caracterizada por fiebre, mialgias, artralgias, exantema, leucopenia y linfadenopatía. Se han descrito tres cuadros clínicos causados por este virus, el Dengue Clásico (DC), el Dengue Hemorrágico (DH) y el Síndrome de Choque por Dengue (SCD). La vigilancia epidemiológica del Dengue constituye un elemento esencial para la prevención y control de esta enfermedad. En México se han establecido criterios para la toma, manejo y recepción de muestras para su determinación en laboratorio del Dengue, entre los que se incluye, el análisis serológico para la identificación de anticuerpos. Durante el diagnóstico de los casos se incluye la valoración clínica de los afectados en la unidad de salud y su clasificación como casos probables de DC o DH (experto clínico). Una vez hecha esta clasificación se obtiene una muestra de suero que se envía al Laboratorio Estatal de Salud Pública (LESP) para su confirmación serológica. Esta muestra se acompaña de un resumen clínico del caso, que eventualmente permite al personal especializado en el laboratorio reclasificar el caso probable como DC o DH (experto de laboratorio). Finalmente el resultado de la determinación de anticuerpos IgM contra el virus Dengue (gold standard), confirma o descarta el caso. Durante 2006, el LESP del estado de Guerrero recibió 16,604 muestras para confirmación serológica de casos de dengue, con un índice de positividad general de 28.8%. La certeza en la realización del diagnóstico clínico presenta diferencias importantes. Mientras que el índice de positividad para casos de DC es del 23.4%, para los casos de DH es de 77.7%. Esta disparidad se acentúa en algunas regiones de la entidad, sugiriendo que existen diferencias en las habilidades diagnósticas de los expertos clínicos que realizan el primer contacto con los pacientes que presentan fiebre. De ahí la importancia de explorar algunos modelos que pudieran ayudar a tomar decisiones en torno a la capacitación de los expertos clínicos y ulteriormente generar una herramienta matemática que permita estimar con mayor certidumbre la presencia de la enfermedad. En el presente trabajo hacemos un resumen de las aportaciones de un colectivo de autores sobre algunas técnicas para la clasificación de los pacientes

de dengue, estas incluyen índices tipo kappa para la evaluación del consenso entre clasificadores de primer contacto con los pacientes (médico tratante), evaluación de clasificadores a partir de las frecuencias absolutas en tablas de clasificación doble y la clasificación a partir de las probabilidades condicionales, dados datos auxiliares, mediante un modelo de regresión logística binaria. Los clasificadores, en este caso, médicos de primer nivel, son los responsables de discriminar en dos grupos o sub-poblaciones de una población general, a partir de una o más características medidas. La evaluación de los clasificadores se realiza mediante técnicas estadísticas de decisión que permiten evaluar una decisión como positiva o negativa, sobre la presencia de la variable de interés. Se usan varias medidas para evaluar la decisión de un clasificador con objeto de establecer si se han tomado buenas decisiones en el diagnóstico. Las medidas que se usan con frecuencia en la literatura estadística son la sensibilidad, la especificidad, la fracción de falsos positivos, fracción de falsos negativos, exactitud de la clasificación, razones de verosimilitud positiva y negativa, la razón de ventajas u Odds ratio, los valores predictivos o probabilidad a posteriori de las clasificaciones, curvas ROC como una representación visual de la sensibilidad y el complemento de la especificidad y diferentes medidas de consenso en la clasificación con objeto de valorar el rendimiento de uno a más clasificadores con respecto a un gold standard. Existen otros procedimientos basados en los modelos lineales generalizados que nos permiten hacer clasificaciones de las unidades poblacionales a partir de la estimación de la probabilidad condicional de que clasifique en uno u otro grupo, dada información auxiliar relacionada con la variable de interés. En este tipo de técnicas consideraremos a la regresión logística, que como se verá en el presente trabajo mejora significativamente la clasificación y permite disponer de una herramienta probabilística eficiente para el diagnóstico de enfermedades a partir de los síntomas, en nuestro caso, permite diferenciar entre enfermos de dengue clásico y dengue hemorrágico. Otros procedimientos para el diagnóstico de dengue, incluyen lo que se conoce en la literatura estadística como árboles de clasificación y regresión (C&RT) y un segundo método que se ubica en el enfoque bayesiano. Se cuenta con una base de datos con una variable de respuesta con dos categorías: “dengue Sí” y “dengue No” y 35 posibles variables regresoras. Se realiza una selección de variables regresoras y se construyen modelos de árbol de clasificación bajo los dos enfoques, con objeto de realizar el diagnóstico de la enfermedad con sólo algunos síntomas sin llegar a la fase de análisis de laboratorio. El problema de medir el acuerdo es sustancialmente diferente del de la asociación dada su naturaleza subjetiva. Este se asocia al establecer en qué medida un experto o criterio de clasificación concuerda con otro u otros. Esto es muchas veces conocido como “fiabilidad entre los clasificadores”. Otro problema es

el de establecer la concordancia entre la clasificación para los mismos objetos bajo diferentes condiciones. Entonces se busca medir la “fiabilidad de la clasificación-reclasificación”. El uso de una medida de consenso es necesario en estudios en los que es de interés establecer cuan reproducibles son ciertos resultados al utilizar un método alternativo de evaluación. Si uno de los clasificadores es “perfecto” podremos confiar en las valoraciones de otro que tiene un alto nivel de concordancia con él. Tal es el caso de la clasificación de los pacientes con dengue, a partir de la valoración, que hace un médico “principiante”, si tienen un alto nivel de concordancia con las realizadas por un médico muy experimentado. Si el médico novel está bien capacitado, se espera un alto grado de concordancia entre su valoración de los pacientes y las valoraciones de un experto. La noción asociada al concepto de concordancia o consenso está dada por el que se obtenga un mismo resultado usando diferentes expertos, métodos o criterios. La existencia de una asociación no implica que necesariamente exista un consenso. Por lo anterior podemos concluir que la existencia de consenso está dada en función de las coincidencias en la clasificación. Las discrepancias no son aceptadas aunque su causa sea un sesgo, una escala o un criterio de selección. La búsqueda del consenso en la clasificación se basa en el supuesto de que hay k clases disjuntas en las que el objeto i puede ser clasificado y J clasificadores (criterios, expertos, etc.). Cada i se asocia a m variables $Y_j^i, \dots, Y_j^i, i = 1, \dots, n$. Los objetos conforman una población o muestra y se supone que la distribución es invariante para cualquier permutación $\{1, \dots, m\}$. Si cada clasificador da respuesta a una misma pregunta, las discrepancias, bajo la hipótesis de acuerdo, no es sino un error y esto lo modelamos al fijar como respuesta del individuo j a $Y_j^i = X^i + \varepsilon_{ij}, j = 1, \dots, J, i = 1, \dots, n$. donde ε_{ij} es un error que puede ser considerado aleatorio o sistemático.

3.1.2. Teoría general

Asumiendo que la clasificación es realizada usando el método de “doble a ciegas”, lo que implica independencia, se trabaja con una variable que solo toma dos posibles valores

$$Y_t = \begin{cases} +, & \text{si } t \in A; \\ -, & \text{si } t \notin A. \end{cases}$$

Se valora cada objeto t , usando dos criterios o dos expertos, y se obtienen dos valores de Y_t . Cada objeto se clasifica como “positivo” o “negativo”. Al analizar los reportes de los dos expertos sobre n objetos los resultados se pueden reflejar en una tabla como la siguiente:

Cuadro 3.1: Clasificación de n objetos por dos expertos.

Experto 1	Experto 2		Total
	Positivo	Negativo	
Positivo	n_{11}	n_{12}	n_{1+}
Negativo	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

Note que este sería el mismo enfoque cuando cruzamos dos preguntas dicotómicas. Si por ejemplo indagamos sobre el precio y la calidad de un producto identificamos positivo con bueno y negativo con malo. Por tanto n_{11} identificaría a aquellos entrevistados que dicen que el producto tiene un buen precio y una calidad adecuada. Tenemos que no solo n_{11} es un valor importante pues también lo son los valores de las otras casillas y sobre todo funciones de ellas. Diferentes índices han sido propuestos para evaluar este tipo de problemas. Ellos hacen uso de las proporciones

$$p_{ij} = \frac{n_{ij}}{n}; q = \frac{1}{2}(p_{2+} + p_{+2}); p = \frac{1}{2}(p_{+1} + p_{1+}) \quad \text{donde} \quad p_{i+} = \frac{n_{i+}}{n}, p_{+i} = \frac{n_{+i}}{n}$$

de una u otra forma. Es razonable que el interés se centre en un criterio asociado a las concordancias existentes entre aquellos considerados como positivos por al menos uno de los expertos. Este enfoque es el usual en la taxonomía numérica en la que n_{22} no se toma en cuenta. Uno de esos índices es el propuesto por Dice (1945)

$$I_D = \frac{2p_{11}}{p_{1+} + p_{+1}} = \frac{p_{11}}{p} \quad (3.1.1)$$

que es la probabilidad empírica condicionada de concordancia entre los expertos dada la positividad del criterio asociado a la casilla “positivo-positivo”. El criterio opuesto establece el uso de

$$I'_D = \frac{2p_{22}}{p_{2+} + p_{+2}} = \frac{p_{22}}{p}. \quad (3.1.2)$$

Para estos índices el valor cero determina un perfecto desacuerdo de los expertos y el valor 1 un perfecto acuerdo. Cuando el acuerdo observado es equivalente al de un experimento aleatorio es igual a 0,5.

Por su parte Rogot-Goldberg (1966) propusieron usar la suma de estos dos índices

$$I_{RG} = \frac{p_{11}}{p} + \frac{p_{22}}{q} = I_D + I'_D \quad (3.1.3)$$

entonces este se mueve entre 0 y 2. El óptimo consenso determina un valor de 2 y la aleatoriedad en el acuerdo 1. Por otra parte usando

$$I'_{RG} = \frac{I_{RG}}{4}. \quad (3.1.4)$$

En este índice 0,5 es el valor para el consenso perfecto y 0,25 el de uno aleatorio.

En estos índices se ha tomado como principio el uso de medidas de posición- si se usaran criterios de dispersión se puede usar el índice de Armitage

$$I_A^2 = \frac{n}{n-1} [p_{11} + p_{12} - (p_{11} - p_{22})^2] \quad (3.1.5)$$

En el caso de desacuerdo perfecto este es cero y lo fuese el acuerdo toma el valor $n/n - 1$. Armitage-Blendys-Smyl (1966) reescalán este índice proponiendo

$$I'_A = \frac{p_{11} + p_{12} - (p_{11} - p_{22})^2}{1 - (p - q)^2}. \quad (3.1.6)$$

Uno de los más populares índices es de Kruskal (1954)

$$I_K = \lambda_r = \frac{(p_{11} + p_{22}) - q}{p} = \frac{2p_{11} - (p_{12} + p_{21})}{2p_{11} + (p_{12} + p_{21})} = 2I_D - 1. \quad (3.1.7)$$

Este índice se mueve en $[-1,1]$. Si el acuerdo es perfecto $\lambda_r = 1$. Si $n_{11} = 0$ el índice toma el valor -1 lo que es un severo problema al hacer ciertos análisis cuando el objetivo del evaluador esta en los objetos calificados de positivos por al menos un experto. El índice Scott (1955) es

$$I_S = \frac{4(p_{11}p_{22} - p_{12}p_{21}) - (p_{12} - p_{21})^2}{(p_{1+} + p_{+1})(p_{2+} + p_{+2})}. \quad (3.1.8)$$

Podemos considerar que los expertos hacen su valoración usando un mecanismo aleatorio independiente. Esta hipótesis fija que

$$Prob\{t \in C_i \text{ para experto 1 y } t \in C_j \text{ para experto e}\} = \pi_{ij} = \pi_{i+}\pi_{+j}$$

donde π_{i+} es la probabilidad marginal de que el experto 1 ubique un objeto en C_i y π_{+j} es la probabilidad de que el experto 2 ubique un objeto en C_j . Al analizar un índice I^* sin usamos este modelo al calcular su esperanza condicional:

$$E(I^* | \pi_{ij} = \pi_{i+}\pi_{+j}, \forall i, j = 1, \dots, k) = I_0$$

lo que establece que

$$p_{ij} = p_{i+}p_{+j}.$$

Entonces como $1 - I_0$ es el máximo de las diferencias obtenibles respecto a una clasificación aleatoria, una medida del acuerdo es el índice tipificado

$$M(I^*) = \frac{I^* - I_0}{1 - I_0}. \quad (3.1.9)$$

Cuando el acuerdo observado es mayor que el atribuible a un fenómeno aleatorio $M(I^*) > 0$, si es menor $M(I^*) < 0$ y será igual a cero si son equivalentes. Es conveniente acotar el hecho de que si $I_0 = 0,5$ entonces $\text{Min}\{M(I^*)\} = -1$. Por tanto este es un índice que es deducido al aceptar la existencia de un modelo superpoblacional que fija que

$$E(n_{ij}) = E(n_{i+})E(n_{+j}).$$

Esta esperanza no tiene nada que ver con el diseño muestral utilizado. Un índice popular el llamado "kappa"

$$\kappa = \frac{\sum_{i=1}^k p_{ii} - \sum_{i=1}^k p_{i+}p_{+i}}{1 - \sum_{i=1}^k p_{i+}p_{+i}} \quad (3.1.10)$$

debido a Cohen (1960). Este compara las frecuencias relativas observadas en la muestra. Utilizando este marco podemos decir que el κ -índice es un estimador insesgado de

$$\kappa_{pob} = \frac{\sum_{i=1}^k \pi_{ii} - \sum_{i=1}^k \pi_{i+}\pi_{+i}}{1 - \sum_{i=1}^k \pi_{i+}\pi_{+i}}. \quad (3.1.11)$$

En general el número de posibles calificaciones de un objeto pueden ser $k \geq 2$. En particular no es recomendable hacer preguntas con más de 5 alternativas. Pero esta es una recomendación asociada a la práctica pues para valores relativamente "grandes" de k los entrevistados tienden a verse desestimulados a ser "exactos". Podemos analizar la esperanza de los índices más populares bajo este modelo de aleatoriedad.

Para (3.1.1)

$$E(I_D | \pi_{ij} = \pi_{i+}\pi_{+i}, \forall i, j = 1, \dots, k) = \frac{p_{1+}p_{+1}}{p}$$

y el índice tipificado es

$$M(I_D) = \frac{2(p_{11}p_{22} - p_{12}p_{21})}{p_{1+} + p_{+2} + p_{+1} + p_{+2}}$$

que es igual a (3.1.3). Para I_{RG} tenemos

$$E(I_{RG} | \pi_{ij} = \pi_{i+}\pi_{+i}, \forall i, j = 1, \dots, k) = \frac{p_{1+}p_{+1}}{p} + \frac{p_{2+}p_{+2}}{q}$$

por lo que nuevamente obtenemos que $M(I_{RG}) = \kappa$.
 Similarmente a partir de que

$$E(I_A | \pi_{ij} = \pi_{i+} \pi_{+j}, \forall i, j = 1, \dots, k) = \frac{p_{1+} p_{2+}}{1 - (p - q)^2} + \frac{p_{1+} p_{+2}}{1 - (p - q)^2}$$

el índice tipificado correspondiente

$$M(I_A) = \kappa.$$

Fleiss (1975) destacó el hecho de que el índice κ caracteriza la clase de índices

$$\mathfrak{S} = \left\{ I \mid I = \frac{I * - I_{\Omega}}{\text{Max}(I * - I_{\Omega})} \right\}$$

donde $I_{\Omega} = E(I * | \Omega)$.

Estos hechos soportan la popularidad del índice propuesto por Cohen aunque en la práctica se utilice muchas veces uno de los introducidos al principio. Razones históricas, entre las que incluyen la costumbre, hacen que en ciertas áreas se utilicen uno de ellos en vez del κ -índice.

3.1.3. Análisis del consenso entre clasificadores.

Los datos de la población de sospechosos de dengue en el estado de Guerrero, durante el 2006, pueden ser organizados por municipio, Región o Jurisdicción Sanitaria y a nivel estatal sin considerar Municipios o regiones. El objetivo de esta sección es establecer cuan consensuadas fueron las clasificaciones de los especialistas respecto al estándar, aceptando como estándar el resultado del laboratorio (ResultLab), a partir del cual, un caso “sospechoso” pasa a la categoría de “confirmado”. El resultado del laboratorio se comparará con la clasificación realizada a nivel de municipio, región o jurisdicción sanitaria y a nivel estatal, por distintos especialistas a nivel de unidad tratante (ExpUT) y especialistas del laboratorio estatal de salud (ExpLab).

La clasificación de los pacientes hecha por el médico de atención primaria y el experto del laboratorio se muestra en la tabla (3.2). En esta clasificación se observa que ExpUT clasificó como DC 8741 casos, coincidiendo con el ExpLab en 8443 casos que representan el 96.59 % y 298 fueron reclasificados como DH, que representan el 3.41 % de los casos. De los 589 pacientes clasificados como DH por el ExpUT, existe coincidencia en la clasificación con el ExpLab en 477 casos, es decir, en el 80.98 % de los casos y solo 112 (19.02 %) son reclasificados como DC, entonces, el ExpLab clasifica 8555 casos (91.69 %) como DC y 775 (8.31 %) como DH.

Cuadro 3.2: Clasificación por ExpUT y ExpLab.

ExpUT	ExpLab		Total
	DC	DH	
DC	8443	298	8741
	96.59 %	3.41 %	100 %
	98.69 %	38.45 %	93.69 %
DH	112	477	589
	19.02 %	80.98 %	100 %
	1.31 %	61.55 %	6.31 %
Total	8555	775	9330
	91.69 %	8.31 %	100 %
	100 %	100 %	100 %

Cuadro 3.3: Clasificación por ExpUT y ResultLab.

ExpUT	ResultLab		Total
	DC	DH	
DC	1658	6627	8285
	20.01 %	79.99 %	100 %
	82.28 %	97.38 %	93.93 %
DH	357	178	535
	66.73 %	33.27 %	100 %
	17.72 %	2.62 %	6.07 %
Total	2015	6805	8820
	22.85 %	77.15 %	100 %
	100 %	100 %	100 %

En las tablas (3.3) y (3.4) se tienen las comparaciones de los ExpUT y el ExpLab con la regla de oro, es decir, el resultado del examen en el laboratorio y la interpretación es análoga a la realizada anteriormente.

En la tabla (3.2) se obtuvo que el ExpUT clasifica 8741 casos como DC y en la tabla (3.3) solo aparecen 8285 de ellos, es decir, 456 casos menos. La razón de esta disminución se debe a que la muestra de suero que se envió al laboratorio era inadecuada para hacer la clasificación a partir de ella. De acuerdo a estos resultados, de los 8285 casos válidos de DC, solo el 20.01 % reportados como DC, resultaron en DC y el 79.99 % resultaron negativos. De los 535 casos válidos de DH clasificados por el ExpUT, el 66.73 % resultaron positivos y el 33.27 % negativos. Se tiene una positividad general de solo el 22.85 %, es decir, 77.15 % de los casos fueron mal clasificados por el ExpUT. En la tabla (3.4) se observa que el experto del laboratorio clasifica correctamente 19.65 % de los casos de DC y 80.35 % es mal clasificado. De los

Cuadro 3.4: Clasificación por ExpLab y ResultLab.

ExpLab	ResultLab		Total
	DC	DH	
DC	1681	6874	8555
	19.65 %	80.35 %	100 %
	87.42 %	92.80 %	91.69 %
DH	242	533	775
	31.23 %	68.77 %	100 %
	12.58 %	7.20 %	8.31 %
Total	1923	7407	9330
	20.61 %	79.39 %	100 %
	100 %	100 %	100 %

775 casos que clasifica como DH, el 31.23% resultó en positivo, mientras que el 79.39% resultó en negativo, en general se tiene una positividad en la clasificación del ExpLab de 20.61%.

Pasemos a revisar el caso general de clasificación y el resultado IgM. Esto se muestra en la tabla (5.1) en la que se observa que existe una positividad general en la clasificación de dengue, sin distinguir entre clásico y hemorrágico, de 21.55%, 17.64% para DC y 71.83% para DH, es decir, se clasifica mejor el dengue tipo hemorrágico.

Cuadro 3.5: Clasificación general y Resultado IgM.

Clasificación Gral.	Resultado IgM		Total
	Negativo	Positivo	
DC	7850	1681	9531
	82.36 %	17.64 %	100 %
	97.41 %	75.93 %	92.78 %
DH	209	533	742
	28.17 %	71.83 %	100 %
	2.59 %	24.07 %	7.22 %
Total	8059	2214	10273
	78.45 %	21.55 %	100 %
	100 %	100 %	100 %

Resultados generales y discusión para el Estado de Guerrero.

Los especialistas del laboratorio consideran como los índices más significativos para su estudio los de Dice y Rogot-Goldberg . La tabla (5.2) brinda los resultados obtenidos en el estado para los diversos índices de consenso. Se observa que el consenso en la clasificación de dengue es muy bueno en el

dengue tipo hemorrágico para los índices más notables. κ , I_D , I'_D , I_{RG} y I'_{RG} están cerca del óptimo que es el valor 1 para los tres primeros, de 2 para el segundo y de 0.5 para el ultimo. Al hacer un análisis similar para el dengue tipo clásico resulta que el consenso es muy malo.

Cuadro 3.6: Clasificación general y Resultado IgM.

Clasificación	κ	I_D	I'_D	I_{RG}	I'_{RG}	I_A^2	I'_A	I_K	I_S
General	-0.098	0.286	0.047	0.334	0.083	0.907	0.926	-0.428	-0.666
DC	0.030	0.161	0.300	0.461	0.115	0.065	0.066	-0.678	-0.539
DH	0.826	0.989	0.836	1.825	0.456	0.160	0.690	0.978	0.825

Dada la consistencia entre los resultados desarrollamos los análisis a otros niveles usando solo los de Dice y el kappa.

Los resultados en la Tabla (3.7) permiten estudiar el consenso entre las clasificaciones de dengue clásico y hemorrágico hechas por los especialistas en las unidades tratantes, expertos del laboratorio y resultado de laboratorio.

Cuadro 3.7: Clasificación general de dengue clásico y hemorrágico.

Clasificadores	I_D	I'_D	κ
ExpUT-ExpLab	0.9763(B)	0.6994(R)	0.6762(R)
ExpUT-ResultLab	0.3219(N)	0.0485(N)	-0.0721(N)
ExpUT-ResultLabGral.	0.9198(B)	0.7107(R)	0.6341(R)
ExpLab-ResultLab	0.2862(N)	0.0475(N)	-0.0977(N)
ExpLab-ResultLabGral.	1.0000(B)	1.0000(B)	1.0000(B)

Como se ve para el caso general de clasificación como DC o DH es necesario fijar las características de ese índice respecto al modelo considerado para la comparación con una regla de oro. En este caso el resultado del laboratorio representa esta regla. El valor del Índice Kappa es de aproximadamente -0.098, este debe variar entre 0 y 1 si el azar (la casualidad) modela el consenso entre la valoración del médico y la prueba de laboratorio. Si el valor es próximo a cero pero positivo se considera que fue el azar lo que caracteriza los acuerdos observados. En el caso de ser cero esto indica que no hay un método para predecir el acuerdo. Si es menor que cero lleva a afirmar que la falta de consenso observada, como es este caso, entre las clasificaciones a nivel estatal, que es negativa, es debida a una causa no explicada por la aleatoriedad. Esto sugiere que son mal aplicadas sistemáticamente la metodología de clasificación de los síntomas.

Estos resultados indican que es alto el consenso entre unidad tratante y laboratorio estatal en clasificar positivos como positivos (I_D), pero que es

regular para detectar los negativos (I'_D) y en general (κ). Para el caso general de clasificación como DC o DH es necesario fijar las características de ese índice respecto al modelo considerado para la comparación con una regla de oro. En este caso el resultado del laboratorio representa esta regla. El valor del índice Kappa es de -0.098 para el experto del laboratorio y de -0.0721 para el experto de unidad tratante, este debe variar entre 0 y 1 si el azar (la casualidad) modela el consenso entre la valoración del médico y la prueba de laboratorio. Si el valor es próximo a cero pero positivo se considera que fue el azar lo que caracteriza los acuerdos observados. En el caso de ser cero esto indica que no hay un método para predecir el acuerdo. Este análisis nos lleva a la valoración cualitativa que aparece entre paréntesis en las tablas (3.7) y (3.8). Es de notar el bajo nivel de consenso en los pares ExpUT-ResultLab y ExLab-ResultLab en los cuales se puede afirmar que no hay consenso bajo ninguno de los criterios analizados. Los resultados por tipo de dengue se muestran en las tablas (3.8) y (3.9).

Cuadro 3.8: Clasificador de dengue clásico.

Clasificadores	I_D	I'_D	κ
ExpUT-ExpLab	0.9757(B)	0.6943(R)	0.6706(R)
ExpUT-ResultLab	0.2976(N)	0.1248(N)	0.0132(N)
ExpLab-ResultLab	0.2999(N)	0.1604(N)	0.0303(N)

En el dengue tipo clásico, tabla (3.8), el consenso en la clasificación es malo tanto para el experto de laboratorio como el experto en la unidad tratante ($\kappa = 0.0303$ y 0.0132 , respectivamente) pero, debemos observar que el consenso en la clasificación de ambos expertos sigue siendo de 0.6706 , es decir, no existen diferencias significativas en la clasificación que hacen ambos expertos. Los resultados para ExpUT-ResultLab y ExpLab-ResultLab son malos para todos los criterios de consenso.

Cuadro 3.9: Clasificador de dengue hemorrágico.

Clasificadores	I_D	I'_D	κ
ExpUT-ExpLab	0.6990(R)	0.9772(B)	0.6761(R)
ExpUT-ResultLab	0.5955(R)	0.9744(B)	0.5700(R)
ExpLab-ResultLab	0.8362(A)	0.9891(B)	0.8260(A)

Se observa en la tabla (3.9) que el consenso en la clasificación del tipo de dengue es bueno en el dengue tipo hemorrágico para ExpUT-ExpLab y ExpLab-ResultLab, dado que en ningún caso los índices están por debajo de 0.5. Sin

embargo es malo para ExpUt-ResultLab para el consenso en detectar lo positivos y en general. Observemos, además que se tiene un consenso moderado en la clasificación de dengue entre los expertos de unidad tratante y experto de laboratorio ($\kappa = 0,6761$).

Resultados y discusión por jurisdicción sanitaria y municipio.

Llevamos a cabo también un análisis a nivel de Jurisdicción Sanitaria o región, los resultados obtenidos fueron los que se muestran a continuación. Cada jurisdicción sanitaria coincide con cada una de las siete regiones del estado, de manera que los resultados pueden presentarse para cualquiera de los dos casos. La tabla (3.10) brinda los resultados de los índices de consenso.

Cuadro 3.10: Clasificación de DC y DH por jurisdicción sanitaria.

Jurisdicción	Clasificadores	I_D (DC)	I'_D (DC)	κ (DC)	I_D (DH)	I'_D (DH)	κ (DH)
T. Caliente	EUT-ELab	0.9891(B)	0.0000(N)	-0.0043(N)	0.0000(N)	0.9988(B)	0.0000(N)
	EUT-RLab	0.1651(M)	0.0055(N)	0.0005(N)	0.0000(N)	0.9988(B)	0.0000(N)
	ELab-RLab	0.1447(M)	0.0376(N)	0.0030(N)	0.0000(N)	0.6667(R)	0.0000(N)
Norte	EUT-ELab	0.9918(B)	0.7188(R)	0.7107(R)	0.7188(R)	0.9918(B)	0.7107(R)
	EUT-RLab	0.2457(M)	0.0495(N)	0.0017(N)	0.6122(R)	0.9911(B)	0.6035(R)
	ELab-RLab	0.2500(M)	0.0746(N)	0.0110(N)	0.7945(R)	0.9941(B)	0.7889(R)
Centro	EUT-ELab	0.9908(B)	0.5313(R)	0.5222(R)	0.5313(R)	0.9908(B)	0.5222(R)
	EUT-RLab	0.1518(M)	0.0364(N)	-0.0058(N)	0.3529(M)	0.9894(B)	0.3436(M)
	ELab-RLab	0.1545(M)	0.0360(N)	0.0031(N)	0.7660(R)	0.9968(B)	0.7629(R)
Montaña	EUT-ELab	0.9904(B)	0.1429(M)	0.1335(M)	0.1481(M)	0.9908(B)	0.1391(M)
	EUT-RLab	0.4810(M)	0.0120(N)	-0.0061(N)	0.0870(N)	0.9912(B)	0.0782(N)
	ELab-RLab	0.4793(M)	0.0334(N)	0.0106(N)	0.9231(B)	0.9992(B)	0.9223(B)
C. Grande	EUT-ELab	0.9879(B)	0.3214(M)	0.3106(M)	0.3214(M)	0.9879(B)	0.3106(M)
	EUT-RLab	0.2989(M)	0.0217(N)	-0.0013(N)	0.1463(M)	0.9885(B)	0.1350(M)
	ELab-RLab	0.2704(M)	0.0431(N)	0.0069(N)	0.7692(R)	0.9964(B)	0.7659(R)
C. Chica	EUT-ELab	0.9907(B)	0.1818(M)	0.1732(M)	0.1818(M)	0.9907(B)	0.1732(M)
	EUT-RLab	0.1264(M)	0.0250(N)	-0.0047(N)	0.1111(M)	0.9913(B)	0.1051(M)
	ELab-RLab	0.1270(M)	0.0166(N)	0.0011(N)	0.7692(R)	0.9985(B)	0.7678(R)
Acapulco	EUT-ELab	0.9177(B)	0.7498(R)	0.6700(R)	0.7504(R)	0.9180(B)	0.6710(R)
	EUT-RLab	0.4140(M)	0.4008(M)	0.1064(M)	0.6543(R)	0.9080(B)	0.5624(R)
	ELab-RLab	0.4749(M)	0.5135(R)	0.1943(M)	0.8449(A)	0.9534(B)	0.7998(R)

Se observa que el consenso en la clasificación de dengue que hacen los expertos del laboratorio, dentro de jurisdicciones sanitarias o regiones, fue muy bueno en el dengue tipo hemorrágico. Los valores de Kappa fluctúan entre 0.763 y 0.922, si eliminamos Tierra Caliente, donde no hubo casos registrados. La mejor clasificación fue realizada por la Jurisdicción sanitaria 4, de la montaña de Guerrero. El consenso para los expertos de unidad tratante, es inferior a la clasificación que hacen los expertos del laboratorio. Note que los únicos

valores del índice kappa altos son los que corresponden a la región norte y la región Acapulco, con valores de 0.603 y 0.562, respectivamente. Para el dengue tipo clásico el consenso en la clasificación fue muy malo. Tomando el índice Kappa se tiene que este varió entre 0.001 y 0.194. Esto es soportado por el hecho de que un valor cercano a cero es evidencia de falta de consenso en la clasificación. Podemos observar, además, que otros índices nos indican una buena clasificación dentro de cada jurisdicción para el dengue hemorrágico, y muy mala, para el dengue clásico.

Cuadro 3.11: Clasificación de DC y DH por municipio.

Municipio	Clasificadores	I_D (DC)	I'_D (DC)	κ (DC)	I_D (DH)	I'_D (DH)	κ (DH)
Acapulco	EUT-ELab	0.9166(B)	0.7437(R)	0.6630(R)	0.7444(R)	0.9169(B)	0.6640(R)
	EUT-RLab	0.4092(M)	0.3941(M)	0.1021(M)	0.6542(R)	0.9094(B)	0.5636(R)
	ELab-RLab	0.4679(M)	0.5080(M)	0.1884(M)	0.8450(A)	0.9541(B)	0.8005(A)
Apaxtla	EUT-ELab	0.9792(B)	0.6667(R)	0.6458(R)	0.6667(R)	0.9792(B)	0.6458(R)
	EUT-RLab	0.6849(R)	0.1481(M)	0.0464(N)	0.4000(M)	0.9684(B)	0.3697(M)
	ELab-RLab	0.7123(R)	0.2759(M)	0.1626(M)	0.8571(A)	0.9895(B)	0.8468(A)
Zihuatanejo	EUT-ELab	0.9528(B)	0.3478(M)	0.3101(M)	0.3478(M)	0.9528(B)	0.3101(M)
	EUT-RLab	0.4965(M)	0.0928(N)	0.0203(N)	0.1714(M)	0.9539(B)	0.1285(M)
	ELab-RLab	0.5188(R)	0.2491(M)	0.0950(N)	0.8000(A)	0.9824(B)	0.7831(R)

El análisis a nivel de municipio permitió establecer que nuevamente el consenso en la clasificación de dengue clásico es malo dentro de todos los municipios. La mejor clasificación se hace en los municipios de Acapulco Juárez, Apaxtla de Castrejón y José Azueta, los que presentamos en la tabla (3.11).

Estudio de los errores e intervalos de confianza

Un procedimiento estadístico para validar los resultados, consiste en hacer un remuestreo para computar los errores estándar de la distribución de medias de los índices. Utilizamos el método Bootstrap para la estimación de medias de los índices y sus errores cuadráticos medios, ver por ejemplo Hole-Fleiss (1993). Esto nos permite estimar un intervalo de confianza. Se tomaron 50 muestras aleatorias de tamaño 500 y se estimaron las medias y varianzas de cada uno de los índices, además del error estándar y un intervalo de confianza para la media de los índices. Los resultados estadísticos fueron los dados en las tablas (3.12) - (3.14).

En la Tabla (3.12) se confirma que el índice kappa para el caso general fue negativo, lo que indica que la clasificación general de dengue es mala y el signo negativo es debido a una causa no explicada por la aleatoriedad. Esto

Cuadro 3.12: Comparación de índices para la clasificación de DC Vs DH e IC al 95 %.

Índices	Media	D.T.	E.T.	LI	LS
Kappa	-0.0927	0.0252	0.0036	-0.0998	-0.0855
I_D	0.2937	0.0274	0.0039	0.2860	0.3015
I'_D	0.0486	0.0141	0.0020	0.0446	0.0526

confirma la hipótesis de que es mal aplicada la metodología de clasificación de los síntomas.

Cuadro 3.13: Comparación de índices para la clasificación de DC Vs No DC e IC al 95 %.

Índices	Media	D.T.	E.T.	LI	LS
Kappa	0.03039	0.00561	0.00079	0.02900	0.03217
I_D	0.30696	0.02759	0.00390	0.29916	0.31476
I'_D	0.15753	0.02461	0.00348	0.15057	0.16449

En el caso de la tabla (3.13) con el dengue clásico, se confirma que es mala la clasificación hecha por los expertos, toda vez que el índice kappa es muy próximo a cero (0.030), lo cual indica falta de consenso en la clasificación. Explicación similar se hace para los otros índices.

Cuadro 3.14: Comparación de índices para la clasificación de DH Vs No DH e IC al 95 %.

Índices	Media	D.T.	E.T.	LI	LS
Kappa	0.8168	0.0627	0.0089	0.7990	0.8345
I_D	0.8431	0.1472	0.0208	0.8015	0.8848
I'_D	0.9781	0.0767	0.0108	0.9564	0.9998

Los resultado de la Tabla (3.14) confirman que la clasificación hecha por los expertos para el dengue tipo hemorrágico es buena (kappa = 0.82) con un error típico muy cercano a cero (0.0089).

3.1.4. Conclusiones.

La existencia de pocos casos confirmados con relación a los casos sospechosos de dengue, se traduce en un alto costo para identificar los casos reales. Por ello podemos concluir que los médicos tratantes no clasifican adecuadamente

los casos de dengue clásico, aunque sí clasifican mejor el hemorrágico. Esto puede ser motivado por el temor a equivocarse en la clasificación, toda vez que lo hacen fundamentalmente mediante el reconocimiento clínico de la enfermedad. Para ello se basan en su experiencia con la enfermedad y la aplicación del manual de dengue, el cual obliga a que hagan un reporte escrito, tomen una muestra de suero y sea enviada al Laboratorio Estatal de Salud para su estudio y en su caso, confirmación del diagnóstico.

Bibliografía

- [1] AGRESTI, A. (1988): A model for agreement between ratings on the ordinal scale. *Biometrics*, 44, 539-548.
- [2] AGRESTI, A. and J. LANG (1993): Quasi-symmetric L class models with application to rater agreement. *Biometrics*, 49, 131-139.
- [3] ARMITAGE, P.; L.M. BLENDYS and H.C SMYLLIE (1966): The measurement of observer disagreement in the recording of signs. *J. Royal Sta. Soc. Series A*, 129, 96-109.
- [4] BLECH, D.A. and H.C. KRAEMER (1989): 2x2 kappa coefficient: measures of agreement and assessment. *Biometrics*, 45, 269-287.
- [5] BOUZA, C.N. (1987): A ratio estimator of the kappa index of agreement between two observers. *Biometrical J.*, 29, 1011-1015.
- [6] CICHETTI, D. V. and A.R. FEINSEIN (1990): High agreement but low kappa: II. Resolving he paradoxes. *J. of Clinical Epidemiology*. 43, 551-558.
- [7] COHEN, J. (1960): A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 20, 37-46.
- [8] COHEN, J. (1968): Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bull.*, 70, 213-220.
- [9] DONNER, A.; M. ELIASZUO and N. KLAR (1996): Testing the homogeneity of kappa stratified. *Biometrics*, 52, 176-183.
- [10] DONNER,A.; M. ELIASZUO and N. KLAR (1992): A goodness of fit approach to inference procedures for the kappa stratified: confidence interval construction, significance tests and sample size estimates. *Statistics in Medicine*, 11, 1511-1519.
- [11] FLEISS, J. (1975): Measuring agreement between two judges on the presence in the absence of a trait. *Biometrics*, 31, 651-659.
- [12] FLEISS, J. (1971): Measuring a nominal scale agreement among many raters. *Psychological Bull.* 76, 378-382.
- [13] FLEISS, J. (1965): Estimating the accuracy of dichotomous judgements. *Psychometrika*, 30, 469-479.

- [14] FORMAN, A. (1993): Fixed latent class models for the analysis of sets of 2-way contingency tables. *Biometrics*, 49, 511-521
- [15] GÓMEZ H., RAMOS B. TAPIA R. .(1995): El riesgo de transmisión del dengue: un espacio para la estratificación. *Salud. Pública. Mex.* ; 37:supl:88-97.
- [16] HOLE, C.A. and J.L. FLEISS (1993): Interval estimates under 2 study designs for kappa with binary classification. *Biometrics*, 49, 523-524.
- [17] KRAEMER, H.C. (1992): Measurement of reliability for categorical data in medical research. *Statistical Methods in Medical Res.* 1, 183-200.
- [18] KRAUTH, J. (1984): "Multivariate Behandlungstabilitat. Klinischer Skalen", *Psychol. Beitr.* 23, 438-457.
- [19] LAU, T.S. (1993): "Higher order kappa type statistics for the dichotomous attribute in multiple ratings", *Biometrics*, 49, 535-542.
- [20] LEE, J.J. and Z.N. TU (1994): A better confidence of interval for the kappa on measuring agreement between two raters with binary outcome. *J. of Computational and Graphical Stat.* 3,301-321.
- [21] LIPSITZ, S.R.; N.M. LAIRD and T.A. BREMEN (1994): Simple moments estimators of the kappa coefficient and its variance. *Applied Stat.* 43, 309-323.
- [22] OMS. (2006): Dengue and Dengue Haemorrhagic fever. FACT Sheet No. 117. Revised April 2006. <http://rhone.b3e.jussieu.fr/DegueNet/pages/f-info/html>
- [23] OPS.(1995): Dengue y Dengue Hemorrágico en las Américas: guías para su prevención y control . Organización Panamericana de la Salud. Publicación Científicas 548, Washington , DC.
- [24] ROGOT, E. and I.D. GOLBERG (1966): A proposed index for measuring agreement in test-retest studies. *J. Chronic Dis.*19, 991-1006.
- [25] ROSNER, B. (1982): Statistical methods in ophthalmology: an adjustment for the intraclass correlation between eyes. *Biometrics*, 38, 899-905.
- [26] SCHOUTEN, H.J.A. (1986): The kappa coefficient of agreement among observers. *Psychometrika*, 51,453- 466.
- [27] SCOTT, W.A. (1955): Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly*.
- [28] SHOUKRI, M.M.; S.W. MARTIN and I.U.H. MIAAN (1995): Maximum likelihood estimation of the kappa coefficient from models of matched binary responses. *Statistics in Medicine*, 124, 83-99.
- [29] SSM: (2006): Manual para la Vigilancia, Diagnóstico, Prevención y Control del Dengue. Secretaría de Salud. México. , México, DF.

- [30] TANNER, M.A. and M.A. YOUNG (1985): Modeling agreement among raters. *J. Amer. Stat. Ass.* 80, 175-180.
- [31] UBERSAX, J.C. (1993): Statistical modeling of expert ratings in medical treatment appropriateness. *J. Amer. Stat. Ass.* 88, 321-427.
- [32] VON EYE, A. and S. SORENSEN (1991): Model chance when measuring interrater agreement with kappa. *Biometrical J.*, 33,781-787.

3.2. Medidas de exactitud de la clasificación y curvas ROC.

Agustín Santiago¹, José Maclovio Sautto¹ y Melina Fernando²

1. Universidad Autónoma de Guerrero, México

2. Maestría en Estadística Aplicada, UAM-UAG.

Resumen: Un clasificador, en este caso un medico de primer nivel, discrimina a una población de dengue en dos grupos o sub poblaciones. En este estudio se discrimina entre dengue clásico o hemorrágico, de pacientes sospechosos de dicha enfermedad. El clasificador toma su decisión a partir de la observación de un conjunto de características asociadas a la enfermedad. La evaluación de la clasificación se hace mediante la aplicación de técnicas estadísticas que nos permiten valorar una decisión positiva o negativa, sobre la presencia de dengue clásico o dengue hemorrágico, en la población de pacientes sospechosos de dicha enfermedad. Las medidas que hemos usado son la sensibilidad, la especificidad, la fracción de falsos positivos, fracción de falsos negativos, exactitud de la clasificación, razones de verosimilitud positiva y negativa, la razón de ventajas u Odds ratio, los valores predictivos o probabilidad a posteriori de las clasificaciones y curvas ROC, como una representación visual de la sensibilidad y especificidad, con objeto de valorar el rendimiento de uno a más clasificadores con respecto a un gold standard.

Key words: Dengue fever, measures of accuracy of classifiers, evaluation of accuracy of classifiers, Receiver Operating Characteristic (ROC).

3.2.1. Introducción

Uno de los periodos más críticos del control epidemiológico del dengue se presentó en el año 2006, en el estado de Guerrero, México. Aunado a este hecho, surgió otro problema que por sus implicaciones económicas y de salud agravó la situación. La clasificación que los médicos de primer contacto hicieron con los pacientes sospechosos de la enfermedad, propició que se enviaran al laboratorio estatal de salud pública 16604 muestras de suero para confirmación serológica, de los cuales, resultaron positivas 3399 en dengue clásico y 1270 en dengue hemorrágico, es decir, la efectividad en el diagnóstico fue de solo el 28.12% del total de casos. En un estado como el de Guerrero, en el cual el presupuesto destinado a salud es muy bajo, un gasto en recursos humanos, reactivos y materiales para hacer las pruebas de laboratorio, resulta muy oneroso e innecesario. Por esta razón, las autoridades sanitarias

pensaron en buscar algunas alternativas para ayudar a mejorar el diagnóstico y así evitar que a causa de un mal diagnóstico se aplique un tratamiento innecesario, se retrase o prive al paciente del tratamiento adecuado y, como en este caso, se ocasione un gasto innecesario para la confirmación serológica de la enfermedad. Existen criterios clínicos de la enfermedad que han sido definidos en el manual para la vigilancia, diagnóstico, prevención y control del dengue, editado por la Secretaría de Salud, México (2006), elaborado en base a los criterios establecidos por la OMS e incluye mecanismos de vigilancia, diagnóstico, prevención y control del vector transmisor del dengue. Díaz, F.A., et al. (2006), realizaron un estudio con objeto de formular una escala diagnóstica de mayor exactitud que la formulada por la OMS para la identificación clínica temprana del dengue en áreas endémicas; Allende, Sira, et al. (2007), realizaron un estudio sobre la efectividad de la diagnosis de la enfermedad a partir de índices tipo kappa y calcularon medidas de exactitud. La metodología propuesta por Allende, Sira, et al. (2007) no es complicada para los expertos en estadística, pero el personal de laboratorio ha tenido dificultades para aplicarla. Es por esta razón que proponemos una metodología que hace uso de las clasificaciones hechas por los médicos de primer contacto, al cual llamaremos clasificador, y el resultado de laboratorio como gold standard. Pretendemos que las medidas de clasificación y de exactitud de clasificadores, así como de curvas ROC, constituyan una herramienta estadística asequible al personal encargado de la evaluación de clasificadores, no solo en el caso del dengue, sino en otras áreas del laboratorio.

3.2.2. Metodología

La clasificación de los individuos de una población se realiza a partir de una característica (o varias) de la misma, que se denomina clasificador. Puede entenderse como clasificador cualquier mecanismo que permite clasificar o discriminar entre los estados alternativos de los individuos de una población, a partir de la información proporcionada por ellos, e indicar sobre cual de estos estados mutuamente excluyentes presenta cada individuo en ese momento. Como es sabido, un clasificador tiene dos propósitos fundamentales (ver, por ejemplo, Sox, et al. 1989): proporcionar información fiable sobre el estado o condición de un individuo e influir en la acción apropiada para el estado pronosticado de un individuo. Aunque puede haber más de dos estados alternativos para los individuos de la población, los problemas de mayor relevancia corresponden a estados extremos, es decir, clasificaciones dicotómicas. En el presente trabajo nos centraremos en el análisis de una clasificación binaria, para discriminar entre los dos posibles estados de un individuo de la

población. El verdadero estado de un individuo está determinado por un *gold standard* o patrón de oro, y a este estado se le denomina:

- Presencia: la condición de interés está presente en el individuo.
- Ausencia: la condición de interés está ausente en el individuo.

De este modo, la respuesta o predicción del clasificador para un individuo puede codificarse como:

- Positivo: la respuesta del clasificador indica la presencia de la condición de interés.
- Negativo: la respuesta del clasificador indica la ausencia de la condición de interés.

Designemos por C la variable dicotómica gold standard, donde $C = 1$ si la condición está presente y $C = 0$, en caso contrario. Designemos por X la respuesta del clasificador, donde $X = 1$ indica la presencia de la condición y $X = 0$ en caso contrario. El verdadero estado de la condición no tiene por qué coincidir con el resultado de la clasificación proporcionada por el clasificador, excepto en el caso utópico de un clasificador ideal o perfecto. Por tanto, el análisis de un clasificador incluye el rendimiento del mismo, es decir, la capacidad para diagnosticar correctamente entre estados alternativos, dado que la clasificación por sí misma no tiene mayor interés si no se conoce su exactitud. En general, el estudio del rendimiento o exactitud de un clasificador se suele iniciar representando una matriz de confusión o decisión formada por una tabla de doble entrada, con las dos categorías disponibles para la respuesta del clasificador, según el estado real de la condición como se muestra en la tabla (15).

Cuadro 3.15: Matriz de confusión.

	Positivo	Negativo	Total
Presencia	VP	FN	TCP
Ausencia	FP	VN	TCA
Total	TRP	TRN	N

Por supuesto que los datos en la clasificación se recogen de una muestra, es decir, el cómputo resulta de las frecuencias absolutas de las respuestas del clasificador según el estado real de la condición de interés sobre el conjunto de individuos observados. En este sentido, los valores VP y VN de la diagonal principal representan las decisiones tomadas correctamente, por lo que se

denomina diagonal de aciertos, siendo sus elementos el número de casos con presencia de la condición y predicción correcta, es decir, el número de casos clasificados como positivos correctamente o Verdaderos positivos (VP) y el número de casos con ausencia de la condición y predicción correcta, es decir, el número de casos clasificados con negativo correctamente o Verdaderos negativos (VN). Asimismo, los valores que aparecen fuera de la diagonal de aciertos representan los resultados incorrectos de la clasificación (errores o confusión), distinguiéndose entre el número de casos en los que está ausente la condición y con predicción incorrecta, es decir, el número de casos clasificados como positivos erróneamente o Falsos positivos (FP) y el número de casos en los que está presente la condición y predicción incorrecta, es decir, el número de casos clasificados como negativos erróneamente o Falsos negativos (FN). Además, las sumas por columnas o por filas de la matriz de confusión corresponden a:

- $TPC = VP + FN$: Total de respuestas con presencia de la condición de interés.
- $TCA = FP + VN$: Total de respuestas con ausencia de la condición de interés.
- $TRP = VP + FP$: Total de respuestas positivas.
- $TRN = FN + VN$: Total de respuestas negativas.

Las frecuencias de las observaciones suelen resumirse en una tabla de frecuencias condicionadas para las respuestas del clasificador, en las diferentes sub muestras de individuos con o sin la condición de interés, como se observa en la tabla (16).

Cuadro 3.16: Matriz de confusión.

	Positivo	Negativo	Total
Presencia	$FVP = VP/TCP$	$FFN = FN/TCP$	1
Ausencia	$FFP = FP/TCA$	$FVN = VN/TCA$	1

Estas tasas o fracciones constituyen la base de las diferentes medidas de exactitud de un clasificador que estimaremos aquí. Con objeto de evitar confusión en la terminología usada, respecto de la que se usa habitualmente en la literatura, haremos las siguientes observaciones. Aunque hemos introducido inicialmente el concepto matriz de confusión en términos de de sucesos sobre la clasificación o predicción en la población, la interpretación práctica de la

misma se realiza en términos de las frecuencias observadas de estos sucesos. En segundo lugar, la base del análisis de la exactitud de un clasificador está formada por estos sucesos condicionados por el verdadero estado de la condición y sus probabilidades. Por tanto, con el fin de mantener la notación usual en la literatura, en adelante identificaremos los términos de la tabla (16) con sus respectivos sucesos en el contexto de la población de individuos, o con sus frecuencias cuando se refiera a una muestra, en consecuencia, las distintas medidas definidas a partir de la matriz de confusión serán parámetros o sus estimadores. Otras probabilidades que se utilizarán en el presente trabajo son las probabilidades a priori y a posteriori del estado de la condición; entre ellas destacamos la probabilidad a posteriori de que la condición esté presente (ausente) cuando la respuesta del clasificador ha sido positiva (negativa), es decir, $P(C = 1|X = 1)$ y $P(C = 0|X = 0)$, denominadas valores predictivos positivos y negativos, respectivamente, y la prevalencia o probabilidad a priori de que la condición esté presente $P(C = 1)$, que está determinada por el *gold standard*.

3.2.3. Medidas de exactitud de un clasificador

Entenderemos como exactitud de un clasificador a la capacidad del mismo para detectar correctamente una condición cuando está presente y descartarla cuando está ausente. Así, resultará de interés que las medidas de rendimiento sean inherentes al propio clasificador. La exactitud intrínseca de un clasificador se mide comparando las respuestas del clasificador (positiva o negativa) con el verdadero estado de la condición (presente o ausente) mediante el *gold standard*.

Sensibilidad y especificidad.

El rendimiento de un clasificador se expresa, usualmente, en términos de dos probabilidades independientes, su sensibilidad y especificidad, es decir, destacando las consecuencias asociadas a los resultados de la clasificación.

La *sensibilidad* es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea la presencia de la condición de interés. Su estimador, que denotaremos por S , es la proporción de respuestas positivas que son correctamente clasificadas:

$$S = \frac{VP}{VP + FN} = FVP \quad (3.2.1)$$

por este motivo, también se denomina fracción de verdaderos positivos (FVP) o exactitud positiva. Si la *sensibilidad* del clasificador es máxima, $S = 1$, entonces $FN = 0$, o equivalentemente, $VP = TCP$, ya que $TCP = VP + FN$.

En situaciones como la que nos ocupa, si las consecuencias son graves en caso de no detectarse a tiempo, resulta fundamental llevar a cabo procedimientos con el mínimo FN posible.

La *especificidad* es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea la ausencia de la condición. El estimador de la *especificidad* es la proporción de respuestas negativas que son correctamente clasificadas:

$$E = \frac{VN}{FP + VN} = FVN \quad (3.2.2)$$

por lo que también se denomina fracción de falsos positivos (FFP) o exactitud negativa. Si la *especificidad* es máxima, $E = 1$, entonces $FP = 0$ o equivalentemente $VN = TCA$, puesto que $TCA = FP + VN$. En nuestro caso, si se clasifica al paciente como dengue clásico o hemorrágico erróneamente, será sometido a un tratamiento para dengue, cuando en realidad su padecimiento es otro. Entonces, se requiere máxima *especificidad* para minimizar los falsos positivos. Dado que la *sensibilidad* y *especificidad* son proporciones, sus intervalos de confianza se construyen a través de métodos clásicos para las proporciones, siendo sus errores estándar

$$SD(S) = \sqrt{\frac{VP(FN)}{(VP + FN)^3}} \quad \text{y} \quad SD(E) = \sqrt{\frac{(FP(VN))}{(FP + VN)^3}} \quad (3.2.3)$$

y por tanto, sus intervalos de confianza asintóticos vienen dados por:

$$S \pm z_{1-\alpha/2}SD(S) \text{ y } E \pm z_{1-\alpha/2}SD(E). \quad (3.2.4)$$

Por otra parte, de la tabla (16) también se derivan otras dos medidas relacionadas con la exactitud del clasificador, las dos fracciones que indican clasificaciones incorrectas, es decir, la fracción de falsos positivos y la fracción de falsos negativos, dadas por

$$FFP = \frac{FP}{FP + VN} \quad \text{y} \quad FFN = \frac{FN}{VP + FN} \quad (3.2.5)$$

respectivamente. Estas fracciones quedan totalmente determinadas por la *sensibilidad* y *especificidad*, toda vez que las sumas por filas de la tabla (16) valen 1, y por tanto, $FFP = 1 - E$ y $FFN = 1 - S$.

Entonces, es posible expresar el rendimiento de un clasificador por medio de dos de estas probabilidades de respuesta del clasificador según el estado real de la condición, las cuales, a diferencia de las probabilidades a priori, se asume que al menos parcialmente están bajo el control del observador.

Índices de exactitud.

Consideraremos ahora algunas medidas de exactitud que incorporan tanto la *sensibilidad* como la *especificidad*, aunque hay que decir que ninguna de ellas ha sido completamente satisfactoria.

Exactitud: La exactitud o accuracy (AC) de un clasificador está definida por su probabilidad de predecir correctamente,

$$AC = P(X = 1|C = 1)P(C = 1) + P(X = 0|C = 0)P(C = 0) = \\ S\Delta P(C = 1) + E\Delta P(C = 0)$$

es decir, una combinación de la *sensibilidad* y *especificidad* ponderadas por la *prevalencia* y su complementario, respectivamente.

Este índice se calcula de la tabla (15) como la proporción de verdaderos positivos y negativos en la muestra,

$$AC = \frac{VP + VN}{N}. \quad (3.2.6)$$

Así, un clasificador ideal o perfecto es aquel cuyas respuestas son correctas, y por tanto su exactitud es uno, dado que $VP + VN = N$ o equivalentemente, $FP = FN = 0$, es decir, no produce clasificaciones erróneas. Por otra parte, en las situaciones que no sean deseables ninguno de los dos tipos de errores, por ser ambos graves y costosos, se buscará que la exactitud sea máxima.

Predominancia: Se define como el porcentaje de individuos con presencia de la condición en la muestra, es decir, el estimador de la *prevalencia* de la condición de interés. La *predominancia* (PD) se deduce de la fórmula anterior como

$$PD = \frac{AC - E}{S - E} = \frac{TCP}{N}. \quad (3.2.7)$$

Índice de Youden: Se define como la diferencia de las proporciones de respuestas positivas correctas e incorrectas, es decir, la diferencia entre la proporción de respuestas positivas del clasificador cuando la condición está presente y la proporción de respuestas positivas bajo el estado de ausencia de la condición

$$\gamma = FVP - FFP = S + E - 1 \quad (3.2.8)$$

esto es, una medida basada en la *sensibilidad* y *especificidad* que no depende de la *prevalencia* de la condición.

Este índice alcanza su máximo valor, $\gamma = 1$, cuando $S = E = 1$, es decir, en la situación de una clasificación ideal. Además, en el caso de una clasificación aleatoria, es decir, $S = E = 0,5$, su valor es nulo, por lo que este índice permite evitar o descartar los casos de clasificación aleatoria.

Razones de verosimilitud.

La capacidad discriminatoria del clasificador puede expresarse en función de la razón de verosimilitud (RV). Se define, para cada tipo de respuesta del clasificador, como el cociente entre la probabilidad de la respuesta bajo presencia de la condición y la probabilidad de dicha respuesta bajo el estado de ausencia.

En general, la razón de verosimilitud viene expresada como

$$RV(X) = \frac{P(X|C = 1)}{P(X|C = 0)} \quad (3.2.9)$$

donde cada tipo de respuesta del clasificador tiene su propia razón de verosimilitud, y se interpreta como el porcentaje de que la respuesta considerada ocurra con mayor o menor proporción en individuos con presencia de la condición con respecto a los individuos con ausencia de la misma. Por ello, se dispone de dos tipos de razones de verosimilitud dependiendo de la respuesta del clasificador, la razón de verosimilitud positiva (RVP), dada por el cociente entre la sensibilidad y el complementario de la especificidad y la razón de verosimilitud negativa (RVN) es el cociente entre el complementario de la sensibilidad y la especificidad, siendo sus estimadores

$$RVP = \frac{S}{1 - E} = \frac{FVP}{FFP} \quad \text{y} \quad RVN = \frac{1 - S}{E} = \frac{FFN}{FVN}. \quad (3.2.10)$$

La razón de verosimilitud refleja el grado de evidencia de una respuesta del clasificador a favor de la presencia de la condición con respecto a la ausencia de la condición. Jaeschke, et al. (2002), categorizan la capacidad de discriminar de un clasificador como sigue:

Cuadro 3.17: Categorización de la razón de verosimilitud.

<0.1	0.1-0.2	0.2-0.5	0.5-2	2-5	5-10	> 10
Excelente	Muy Bueno	Bueno	Neutro	Bueno	Muy Bueno	Excelente

Así, una razón de verosimilitud (positiva o negativa) igual a uno, indica que tal respuesta del clasificador es igualmente probable en individuos con

o sin presencia de la condición. Una razón de verosimilitud mayor que uno indica que la respuesta del clasificador es más probable entre individuos con presencia de la condición que sin la condición. Una razón de verosimilitud menor que uno indica que la respuesta del clasificador es más probable entre individuos con ausencia que con presencia de dicha condición. En general, las razones de probabilidad por encima de 10 o por debajo de 0.1 corresponden a clasificadores con alta capacidad para detectar o descartar la presencia correctamente.

Los intervalos de confianza se obtienen a través de métodos clásicos para tasas de riesgo, siendo sus errores estándar

$$SD(RVP) = \sqrt{\frac{1-S}{VP} + \frac{E}{FP}} \quad \text{y} \quad SD(RVN) = \sqrt{\frac{S}{FN} + \frac{1-E}{VN}} \quad (3.2.11)$$

y por tanto, sus intervalos de confianza asintóticos vienen dados por

$$\frac{S}{1-E} \exp(\pm z_{1-\alpha/2} SD(RVP)) \quad \text{y} \quad \frac{1-S}{E} \exp(\pm z_{1-\alpha/2} SD(RVN)). \quad (3.2.12)$$

Odds ratio.

Otra medida de la exactitud de un clasificador se basa en las odds (ventajas o preferencias) de las respuestas. En general, una odds de un suceso se define como el cociente entre su probabilidad y la de su complementario, es decir,

$$Odds(suceso) = \frac{P(suceso)}{1 - P(suceso)} \quad (3.2.13)$$

de donde podemos derivar una expresión para la probabilidad del suceso como

$$P(suceso) = \frac{Odds(suceso)}{1 + Odds(suceso)}. \quad (3.2.14)$$

De acuerdo con esto, serán de interés dos tipos de odds para analizar el rendimiento de un clasificador a través de su preferencia por las respuestas positivas en cada estado de la condición, es decir, la ventaja del clasificador para una respuesta positiva frente a una negativa bajo la presencia de la condición y la ventaja del clasificador para una respuesta positiva frente a una negativa bajo la ausencia de la condición. Estos índices se calculan a partir de la tabla (16), en función de la sensibilidad y especificidad, respectivamente:

$$Odds_{presencia} = \frac{S}{1-S} \quad \text{y} \quad Odds_{ausencia} = \frac{1-E}{E}. \quad (3.2.15)$$

A partir de estas dos razones sobre las preferencias del clasificador para las respuestas positivas frente a las negativas, se define el odds ratio

$$Oddsratio = \frac{Odds_{presencia}}{Odds_{ausencia}} \quad (3.2.16)$$

cuyo estimador es

$$Oddsratio = \frac{S.E}{(1-S)(1-E)} = \frac{VP.VN}{FN.FP} \quad (3.2.17)$$

y su error estándar

$$SD(Oddsratio) = \sqrt{\frac{VP.VN}{FN.FP} \left(\frac{1}{VP} + \frac{1}{FN} + \frac{1}{FP} + \frac{1}{VN} \right)} \quad (3.2.18)$$

Nótese que si alguno de los elementos de la matriz de confusión es nulo, entonces el *Odds ratio* no puede estimarse. Por otra parte, un *Odds ratio* igual a uno indica que la ventaja de una respuesta positiva es la misma en ambos estados de la condición, es decir, $S = 1 - E$. Un *Odds ratio* mayor que uno indica que la preferencia por una respuesta positiva es mayor cuando la condición está presente y si el *Odds ratio* es menor que uno, indica que la preferencia por una respuesta positiva es mayor cuando la condición está ausente.

Valores predictivos.

Ahora, si un individuo ha sido clasificado como positivo, ¿cuál es la probabilidad de que realmente presente la condición? o si un individuo ha sido clasificado como negativo, ¿cuál es la probabilidad de que realmente no presente la condición? La sensibilidad y la especificidad no dan estas dos probabilidades, toda vez que requieren de las probabilidades a posteriori, probabilidades sobre el verdadero estado de la condición tras conocer el resultado del clasificador, es decir, $P(C = 1|X = 1)$ y $P(C = 0|X = 0)$ denominadas valores predictivos positivo y negativo, respectivamente.

El valor predictivo positivo (VPP) o precisión de un clasificador, es la probabilidad de identificar correctamente los casos positivos, cuyo estimador es

$$VPP = \frac{VP}{VP + FP} \quad (3.2.19)$$

El valor predictivo negativo (VPN) de un clasificador, es la probabilidad de identificar correctamente los casos negativos, siendo su estimador

$$VPN = \frac{VN}{VN + FN} \quad (3.2.20)$$

Estos valores se calculan fácilmente a partir de la matriz de confusión, sin embargo, a la hora de ser utilizados para comparar el rendimiento de clasificadores, presentan la desventaja de depender de la proporción de casos con presencia de la condición en la población, es decir, de la prevalencia.

Estos valores pueden estimarse mediante el teorema de Bayes, como se puede ver en Franco, M., et al. (2008), o teniendo en cuenta la relación entre las probabilidades de un suceso y sus *odds*, los valores predictivos pueden calcularse también en términos de *odds* en lugar de probabilidades, según puede verse en Sox et al. (1989).

Existen otras medidas de exactitud desarrolladas para situaciones no balanceadas, es decir, cuando la distribución de los tipos o clases de respuesta del clasificador no es equilibrada, pero no las consideraremos en el presente trabajo.

Curva ROC.

La curva ROC es una representación gráfica del rendimiento de un clasificador, proporcionando una herramienta visual para examinar la relación entre la capacidad del clasificador para detectar correctamente los individuos con presencia de la condición de interés y su incapacidad para identificar los individuos del grupo de ausencia.

El espacio ROC consiste en un sistema de coordenadas donde se representa la fracción de falsos positivos en el eje de las abscisas frente a la fracción de verdaderos positivos en el eje de ordenadas. Así, el espacio ROC permite visualizar el rendimiento de un clasificador (o varios) mediante representaciones bidimensionales, a partir de los puntos (1-especificidad, sensibilidad). En general, el espacio ROC muestra la correspondencia relativa entre beneficios (verdaderos positivos) y costos (falsos positivos) de una clasificación.

3.2.4. Resultados

Hemos realizado el análisis de las clasificaciones a nivel global para el estado de Guerrero y por Jurisdicciones sanitarias, con objeto de identificar las regiones por su eficiencia en la clasificación.

Clasificación global.

A partir del total de casos estudiados en el laboratorio, en este caso 10283, toda vez que 607 se clasificaron como indeterminados porque las muestras de suero no eran adecuadas, se obtuvieron resultados sobre la clasificación global de los médicos de primer contacto, obteniendo los siguientes resultados:

Cuadro 3.18: Frecuencias absolutas para la clasificación de dengue.

DC	Positivo	Negativo	Total
Presencia	1681	534	2215
Ausencia	7850	218	8068
Total	9531	752	10283
DH	Positivo	Negativo	Total
Presencia	533	1682	2215
Ausencia	209	7859	8068
Total	742	9541	10283

De acuerdo con la tabla (19) de frecuencias condicionadas, la clasificación global para dengue clásico tiene una sensibilidad $S = 0.759$, aproximadamente del 76

Cuadro 3.19: Frecuencias condicionadas para la clasificación de dengue.

DC	Positivo	Negativo	Total
Presencia	0.758916479	0.241083521	1
Ausencia	0.972979673	0.027020327	1
DH	Positivo	Negativo	Total
Presencia	0.240632054	0.759367946	1
Ausencia	0.025904809	0.974095191	1

A partir de la tabla (18) podemos calcular las restantes medidas de exactitud, obteniendo que para el caso de dengue clásico las medidas globales de exactitud son:

Cuadro 3.20: Medidas de exactitud adicionales para la clasificación de DC y DH.

	AC	PD	γ	RVP	RVN	OR	VPP	VPN
DC	0.185	0.215	-0.214	0.780	8.922	0.0874	0.176	0.2899
DH	0.816	0.215	0.215	9.289	0.7796	11.916	0.718	0.824

Clasificación por regiones.

Con objeto de identificar la zona que mejor (o peor) clasifica, se estimaron las medidas de exactitud para cada una de ellas y se resumen en la tabla (21). La clasificación para dengue clásico muestra que la sensibilidad toma valores en $0,53 < S < 0,98$, su especificidad en $0,003 < E < 0,12$, la fracción de falsos positivos en $0,87 < FFP < 0,997$, la fracción de falsos negativos en $0,02 < FFN < 0,46$, la acuracidad en $0,08 < AC < 0,3$, la predominancia

en $0,07 < PD < 0,42$, el índice de Youden en $-0,34 < \gamma < -0,007$, la razón de verosimilitud positiva en $0,61 < RVP < 0,992$, la razón de verosimilitud negativa en $1,46 < RVN < 16,006$, la razón de ventajas en $0,04 < OR < 0,68$, el valor predictivo positivo en $0,06 < VPP < 0,316$ y el valor predictivo negativo en $0,2 < VPN < 0,89$.

Cuadro 3.21: Medidas de exactitud por regiones de clasificaciones de DC.

Estimador	JS1	JS2	JS3	JS4	JS5	JS6	JS7
S	0.97500	0.86256	0.88750	0.97087	0.92938	0.93243	0.53942
SD(S)	0.02469	0.02370	0.02498	0.00828	0.01362	0.02918	0.01605
E	0.01706	0.01355	0.00703	0.00344	0.00838	0.00315	0.12414
SD[E]	0.00598	0.00347	0.00211	0.00198	0.00215	0.00182	0.00910
FFP	0.98294	0.98645	0.99297	0.99656	0.99162	0.99685	0.87586
FFN	0.02500	0.13744	0.11250	0.02913	0.07062	0.06757	0.46058
AC	0.09234	0.14947	0.08870	0.31386	0.16045	0.07018	0.29996
PD	0.07859	0.16009	0.09275	0.32087	0.16511	0.07212	0.42336
I-Youden	-0.0079	-0.1239	-0.10557	-0.0257	-0.0622	-0.0644	-0.3364
RVP	0.99192	0.87441	0.89378	0.97423	0.93723	0.93538	0.61588
RVN	1.46563	10.14313	16.00568	8.46602	8.42750	21.44144	3.71008
SD(RVP)	0.02604	0.02771	0.02823	0.00876	0.01481	0.03135	0.03152
SD(RVN)	1.04779	0.30904	0.37360	0.64272	0.32138	0.72026	0.08117
OR	0.67679	0.08621	0.05584	0.11507	0.11121	0.04362	0.16600
SD(OR)	0.88329	0.09629	0.09278	0.21993	0.11074	0.15474	0.04308
VPP	0.07800	0.14286	0.08373	0.31521	0.15637	0.06778	0.31138
VPN	0.88889	0.34091	0.37931	0.20000	0.37500	0.37500	0.26853

Para el caso de dengue hemorrágico no incluiremos en el análisis a la JS1, porque algunos valores de la tabla de clasificaciones son cero, ver tabla (22), pero el hecho que AC sea de aproximadamente del 92 %, indica que algunos individuos fueron mal clasificados, lo que se confirma con la prevalencia de DH, cuyo valor es de aproximadamente el 8 %, aunque los médicos solo reportaron casos de dengue clásico. Para las otras regiones se observa que la sensibilidad toma valores en $0,02 < S < 0,461$, su especificidad en $0,87 < E < 0,998$, la fracción de falsos positivos en $0,002 < FFP < 0,1242$, la fracción de falsos negativos en $0,53 < FFN < 0,971$, la acuracidad en $0,68 < AC < 0,93$, la predominancia en $0,07 < PD < 0,424$, el índice de Youden en $0,02 < \gamma < 0,34$, la razón de verosimilitud positiva en $3,7 < RVP < 21,442$, la razón de verosimilitud negativa en $0,61 < RVN < 0,974$, la razón de ventajas en $6,02 < OR < 22,93$, el valor predictivo positivo en $0,62 < VPP < 0,86$ y el valor predictivo negativo en $0,68 < VPN < 0,933$.

Cuadro 3.22: Medidas de exactitud por regiones de clasificaciones de DH.

Estimador	JS1	JS2	JS3	JS4	JS5	JS6	JS7
S	0.00000	0.13744	0.11250	0.02913	0.07062	0.06757	0.46058
SD(S)	0.00000	0.02370	0.02498	0.00828	0.01362	0.02918	0.01605
E	1.00000	0.98645	0.99297	0.99771	0.99162	0.99685	0.87586
SD[E]	0.00000	0.00347	0.00211	0.00162	0.00215	0.00182	0.00910
FFP	0.00000	0.01355	0.00703	0.00229	0.00838	0.00315	0.12414
FFN	1.00000	0.86256	0.88750	0.97087	0.92938	0.93243	0.53942
AC	0.92141	0.85053	0.91130	0.68692	0.83955	0.92982	0.70004
PD	0.07859	0.16009	0.09275	0.32087	0.16511	0.07212	0.42336
I-Youden	0.00000	0.12389	0.10547	0.02683	0.06224	0.06442	0.33644
RVP		10.14313	16.00568	12.69903	8.42750	21.44144	3.71008
RVN	1.00000	0.87441	0.89378	0.97311	0.93723	0.93538	0.61588
SD(RVP)		0.30904	0.37360	0.76142	0.32138	0.72026	0.08117
SD(RVN)	0.00000	0.02771	0.02823	0.00869	0.01481	0.03135	0.03152
OR		11.60000	17.90781	13.05000	8.99189	22.92271	6.02407
SD(OR)		1.11700	1.66147	2.76769	0.99576	3.54708	0.25952
VPP		0.65909	0.62069	0.85714	0.62500	0.62500	0.73147
VPN	0.92141	0.85714	0.91627	0.68504	0.84363	0.93222	0.68862

3.2.5. Conclusiones

Clasificación global.

Toda vez que el objetivo central consistió en evaluar la clasificación que los médicos de primer contacto hacen con los pacientes sospechosos de dengue, a nivel global y a nivel de jurisdicciones sanitarias, encontramos que la sensibilidad general para la clasificación de pacientes sospechosos de dengue clásico fue de 0.75892, lo que indica que la probabilidad de clasificar correctamente a un individuo cuyo estado real sea la presencia de dengue clásico es de aproximadamente el 76 %, con intervalo de confianza al 95 % (0.74110295, 0.776730007). La especificidad, que es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea la ausencia de dengue clásico es de 0.02702, o equivalentemente, la fracción de falsos positivos es de 0.97298. Entonces, los médicos de primer contacto clasifican correctamente a los pacientes de dengue clásico, aproximadamente en el 76 % de los casos, aunque en el 97 % de las veces clasifica erróneamente como dengue clásico. En una situación como esta, lo ideal sería que la FFN fuera mínima, pero en este caso tenemos un valor 0.24108, que representa el 24.11 % de los casos con presencia de dengue clásico que fueron clasificados como negativos. La exactitud en la clasificación está dada por $AC = 0.18467$, que es un valor pequeño, lo ideal es obtener valores cercanos a uno, pero en este caso estamos muy lejos de él. El porcentaje de individuos con presencia de dengue clásico en la muestra, predominancia o prevalencia, es de 0.21540 o 21.54 %. Un valor del índice de Youden -0.21406, evidencia también una mala clasificación, toda vez que

para que la clasificación sea buena este valor debería estar próximo a uno. La razón de verosimilitud positiva toma el valor $RVP = 0.77999$, con intervalo de confianza al 95 % (0.76168379; 0.798740522), que corresponde a la categoría más baja de la tabla (17), por lo que se puede afirmar que la respuesta del clasificador es igualmente probable en individuos con o sin presencia de dengue clásico. El $OR = 0.08742$ indica que la preferencia por una respuesta positiva en dengue clásico, es mayor cuando el paciente presenta síntomas de dengue clásico (aproximadamente $1/0,08742 \approx 11$ veces más que si no presentara la condición). Respecto de los valores predictivos encontramos que $VPP = 0.17637$, lo que indica que la probabilidad de identificar correctamente los casos positivos dada una clasificación positiva, es de apenas el 17.64 % y que $VPN = 0.28989$, indica una probabilidad de identificar correctamente los casos negativos dada una clasificación negativa es de aproximadamente el 29 %. Entonces, la clasificación de los pacientes de dengue clásico por los médicos de primer contacto, está lejos de considerarse como buena.

Respecto de la clasificación de dengue hemorrágico encontramos que la sensibilidad y la especificidad son $S = 0.240632054$ y $E = 0.974095191$, respectivamente, lo que significa que la probabilidad de clasificar a un individuo cuyo estado real sea la presencia de dengue hemorrágico es de aproximadamente el 24 % y la probabilidad de clasificar correctamente a un individuo cuyo estado real sea ausencia de dengue hemorrágico es del 97.4 %, o de forma equivalente, la fracción de falsos positivos es de $FFP = 0.025904809$, es muy pequeña, de donde se deduce que aproximadamente el 2.6 % de los casos son clasificados erróneamente como positivos. El dengue hemorrágico es mortal si no se atiende a tiempo, por lo que la $FFN = 0.759367946$, nos indica que se puso en riesgo a aproximadamente al 76 % de los paciente confirmados por la gold standard de DH, toda vez que presentando la condición DH fueron clasificados como negativos. La exactitud en la clasificación como DH está dada por $AC = 0.81610425$, lo que indica que, a pesar de haber puesto en riesgo clasificando un número alto de FN, la probabilidad de predecir correctamente DH, por los médicos de primer contacto, es del 81.6 %. La predominancia o prevalencia de DH en la muestra es del 21.54 %, el índice de Youden es de $\gamma = 0,214727245$, muy lejos de la clasificación ideal. La razón de verosimilitud positiva, evidencia que la respuesta del clasificador es más probable entre los individuos que presentan síntomas de DH que en aquellos que no los presentan; según la tabla (17) clasifica como muy bueno. Se obtuvo un $OR = 11.91577297$, lo que indica que la razón de ventajas por una respuesta positiva en DH, es aproximadamente 12 veces mayor si están presentes los síntomas que si no lo están. Por último, los valores predictivos indican que, dada una clasificación positiva, la probabilidad de identificar correctamente los casos positivos es de 71.83 % (VPP), aproximadamente y, dada una

clasificación negativa, la probabilidad de identificar correctamente los casos negativos es del 82.37% (VPN), aproximadamente. De acuerdo con estos resultados podemos concluir que los médicos clasifican mejor a los pacientes con dengue hemorrágico que a los pacientes con dengue clásico, aunque la fracción de falsos negativos alta obtenida aquí, nos indica que necesitan mejorar las clasificaciones.

Cuadro 3.23: Área bajo la curva ROC con IC asintótico al 95%. La variable (o variables) de resultado de contraste: DC, DH tiene al menos un empate entre el grupo de estado real positivo y el grupo de estado real negativo. Los estadísticos pueden estar sesgados. (a) Bajo el supuesto no paramétrico, (b) Hipótesis nula: área verdadera = 0.5.

Var. de contraste	Área	E. típ. (a)	Sig. asintótica(b)	Li	Ls
Dengue Clásico	0.393	0.007	0.000	0.37852	0.40741
Dengue Hemorrágico	0.603	0.007	0.000	0.59296	0.62181

Figura 3.1: Curvas ROC para DC y DH

Observe que los gráficos de la figura(3.2.5) se encuentran por arriba (DC) y Por abajo (DH) de la línea diagonal de referencia, por lo que se deduce que la clasificación en ambos casos no se debe a la aleatoriedad, sino a las habilidades diagnósticas de los clasificadores. Para el caso del DC se trata de una clasificación peor que la aleatoria y para DH se observa que es ligeramente superior (regular). Esto significa que un individuo seleccionado aleatoriamente del grupo de enfermos de DC tiene un valor de la prueba mayor que uno seleccionado aleatoriamente del grupo de sanos en el 39% de las veces, aproximadamente y que un individuo seleccionado aleatoriamente del grupo de enfermos de DH tiene un valor de la prueba mayor que uno seleccionado aleatoriamente del grupo de sanos en el 60% de las veces aproximadamente.

Clasificación por Jurisdicción Sanitaria.

Los resultados obtenidos a nivel de jurisdicción sanitaria para dengue clásico, indican que quien mejor clasifica, en términos de sensibilidad, es JS4 (0.97087) y quien peor lo hace es JS7 (0.53942), o de forma equivalente, en la JS4 se detecta correctamente el DC en el 97.087% de las veces, mientras que en la JS7 se detecta en aproximadamente el 54% de las veces. Ordenados a partir del que mejor clasifica hasta el peor, mediante el criterio de la

sensibilidad, es JS4, JS6, JS5, JS3, JS2 y JS7. Debemos observar que la especificidad es de 0.00344 para la JS4 y de 0.12414 para la JS7, lo que conduce a las fracciones de falsos positivos 0.99656 para la JS4 y de 0.87586 para la JS7, es decir, la FFP para la JS4 es del 99.66 % y para la JS7 es de 87.6 %, aproximadamente. Debemos destacar que el la JS6 quien tiene una mayor FFP, cuyo valor es de 0.99685, ligeramente mayor que el de la JS4. Si observamos la FFN, el orden de mayor a menor de las JS es JS7, JS2, JS3, JS5, JS6 y JS4, nuevamente es evidente que quien mejor clasifica es la JS4, toda vez que tiene la FFN más pequeña (0.02913), aproximadamente 2.9 % de los casos de DC son clasificados como negativos, mientras que para la JS7, con una FFN de (0.46058), aproximadamente el 46 % de los casos de DC son incorrectamente clasificados como negativos. La medida de acuracidad (AC) en las jurisdicciones sanitarias en orden decreciente es JS4, JS7, JS5, JS2, JS3, JS6. Nuevamente la JS4, según evidencia, clasifica mejor, aunque la medida de exactitud es baja, alcanza apenas un valor de 0.31386, muy similar al valor alcanzado por la JS7, que es de 0.29996; entonces, quien peor clasifica es la JS6, con una exactitud de 0.07018. De acuerdo al índice de Youden, aunque para todas las jurisdicciones sanitarias se obtuvieron valores cercanos a cero y negativos (lo ideal es obtener valores cercanos a uno), sigue siendo la JS4 quien mejor situada está. La RVP indica que la respuesta de los clasificadores es igualmente probable en individuos con y sin presencia de la condición DC para todas las JS, toda vez que de acuerdo a la tabla (17), sus valores caen en la categoría neutro. El OR = 0.11507 para la JS4 y el OR = 0.16600, debe interpretarse como se ha hecho antes. Para el caso dengue hemorrágico, como se puede constatar en la tabla (22), quien clasifica mejor es la JS7 y quien peor lo hace es la JS4 y la situación se agrava cuando observamos la FFN que para la JS4 es del 97 %, mientras que para la JS7 es del 54 %, aproximadamente. Ambas clasifican mal a los pacientes de dengue Hemorrágico.

Bibliografía

- [1] Aguilera del Pino, A.M. (2001). Tablas de Contingencia Bidimensionales. Colección Cuadernos de Estadística (15). La muralla.
- [2] Allende, Sira; Bouza, Carlos; López, Saúl y Santiago, Agustín. Análisis de la efectividad de la diagnosis del dengue en el estado de Guerrero, México 2007: Hechos estadísticos y consenso. Revista Investigación Operacional Vol., 29 No 1, 35-47, 2008.
- [3] Díaz, Fredi A.; Martínez, Ruth A.; Villar, Luis A. Criterios clínicos para diagnosticar el dengue en los primeros días de enfermedad. Redalyc. Biomédica 2006, 26 22-30. [4] Farr, B.M. y Shapiro, D.E., (2000). Diagnostic test: distinguishing good tests from bad even ugly ones. Infect. Cont. Hosp. Epidemiol., 21(4), 278-84.
- [4] Franco, M., Vivo, J.M., (2008). Análisis de curvas ROC, Principios Básicos y Aplicaciones. Editorial la Muralla, S.A. España.
- [5] Jaeschke, R., Guyatt, G. y Lijmer, J. (2002). Diagnostic test. En Guyatt, G. y Rennie, D. (eds.), Users' guides to the medical literature. AMA Press, Chicago, 121-40.
- [6] Manual para la Vigilancia, Diagnóstico, Prevención y Control del Dengue. Secretaría de Salud, México (2006).
- [7] Sox, H., Stern, S., Owens, D. y Abrams, H.L. (1989). Assessment of diagnostic technology in health care. Rationale, methods, problems and directions. National Academy Press, Washintong, DC.

Capítulo 4

Modelos de diagnóstico del dengue

4.1. Diagnóstico mediante modelos de árbol de clasificación.

Octaviano Juárez¹, Ernestina Castell² y Agustín Santiago¹

1. Universidad Autónoma de Guerrero

2. Universidad de La Habana

Resumen: El dengue es una enfermedad de tipo viral transmitida principalmente por el mosquito *Aedes aegypti*, se clasifica en dengue clásico y dengue hemorrágico. El diagnóstico oportuno de la enfermedad significa ahorro de recursos humanos y materiales. Se construyen modelos de árbol de clasificación bajo dos enfoques diferentes. El primer enfoque basado en los modelos de Árbol de Clasificación y Regresión (C & RT) y un segundo método que se ubica dentro del paradigma Bayesiano. Se cuenta con una base de datos con una variable de respuesta con dos categorías: “Si dengue” y “No dengue” y 35 posibles variables predictoras. Se realiza una selección de las variables regresoras y se construyen modelos de árbol de clasificación bajo los dos enfoques, para realizar el diagnóstico de la enfermedad del dengue con sólo algunos síntomas que no requieren análisis de laboratorio.

Abstract: Dengue is a viral disease transmitted mainly by the *Aedes aegypti* mosquito, is classified as classic dengue and dengue hemorrhagic fever. The early diagnosis of the disease means saving human and material resources. Tree classification models are built under two different approaches. The first approach based on the models of Classification and Regression Tree (C

& RT) and a second method that is located within the Bayesian paradigm. There is a database with a variable response to two categories: “Yes dengue” and “No dengue” and 35 potential predictor variables. We present a selection of predictor variables and are built tree classification models under two approaches, to make the diagnosis of the disease of dengue fever with only a few symptoms that do not require laboratory analysis.

Key words: Tree Classification Models, Monte Carlo Markov Chain, Metropolis-Hastings Algorithm.

4.1.1. Introducción.

El dengue es una enfermedad de tipo viral que tiene importantes repercusiones económicas y sociales. Esta enfermedad se transmite de una persona enferma a otra persona sana, por la picadura de mosquitos conocidos como *Aedes aegypti*, principalmente, identificándose hasta la fecha cuatro serotipos, pero por sus manifestaciones pueden clasificarse en dengue clásico y dengue hemorrágico. Los puertos con comercio internacional son seriamente afectados por esta enfermedad y sus consecuencias, pero de manera particular, aquellos que son también centros turísticos. Debido a esto, un diagnóstico oportuno de la enfermedad posibilita el ahorro de recursos materiales y humanos.

El dengue clásico presenta los síntomas conocido como “tracazo” o “fiebre quebrantahuesos”, se presentan después de un periodo de incubación de cuatro a siete días (mínimo tres, máximo 14) y se caracteriza por un cuadro febril, de inicio brusco, con cefalea intensa (generalmente frontal), mialgias, artralgias y dolor de ojos (retroocular) que se incrementa con los movimientos oculares. Puede presentarse exantema transitorio, fotofobia, insomnio, prurito, diarrea, náusea, vómito, dolor abdominal y anorexia, hiperestesia, dolores generalizados, congestión faríngea y conjuntivitis. Este cuadro dura de tres a siete días. En menores de cinco años puede presentarse sólo como síndrome febril. Entre el 10 y 20% de los casos pueden presentar manifestaciones de fragilidad capilar en piel o mucosas (petequias, equimosis o hematomas) e incluso hemorragias menores (epistaxis o gingivorragia).

El dengue hemorrágico se caracteriza por la presencia de hemoconcentración debida a la fuga de plasma al espacio extravascular por el aumento en la permeabilidad de los vasos sanguíneos, lo que determina la severidad del cuadro clínico y lo diferencia del dengue clásico. Dicha hemoconcentración se manifiesta por hematocrito elevado y con frecuencia por la presencia de hemorragias (epistaxis, gingivorragia, sangrado urogenital, sangrado en sitios

de punción, hemoptisis y sangrado del tubo digestivo) y extravasación de líquidos (equimosis, hematomas o petequias).

Otros síntomas que suelen acompañar al dengue hemorrágico son: dolor en área hepática, dolor abdominal, derrame pleural, ascitis, edema en diversos órganos, hepatomegalia o esplenomegalia, leucopenia inicial y leucocitosis posterior, hiponatremia, hipoalbumemia, hipotensión con tendencia al acortamiento en el intervalo sistólico/diastólico. (Manual del dengue, SS, Wikipedia)

Dado que existen múltiples signos y síntomas de la enfermedad que dificultan su diagnóstico seguro, contar con un método de diagnóstico sencillo y fácil de manejar significaría el ahorro de recursos materiales y humanos. En este sentido, el objetivo del trabajo fue construir un modelo de árbol de clasificación que sirva como una herramienta para diagnosticar la enfermedad del dengue, con mayor certidumbre.

Los modelos basados en árboles binarios por las facilidades que proporcionan en su implementación y explicación tienen aplicación en diversas ramas. Los modelos de árbol de regresión y clasificación (C&RT, classification & regression tree), fueron introducidos en la estadística por Breiman et al. (1984). La idea principal de dicho método consiste en particionar el espacio de las variables independientes, en forma tal que los valores de la variable de respuesta sean cada vez más homogéneos dentro de las clases de dicha partición. Los modelos C&RT son fáciles de aplicar e interpretar, por tal motivo se han hecho populares en diferentes áreas.

Bajo el enfoque bayesiano existen diversas propuestas, por ejemplo Chipman et al. (1998a) y Denison et al. (1998), proponen nuevas metodologías que utilizan los modelos C&RT en combinación con los métodos de Monte Carlo vía cadenas de Markov (MCMC, Markov Chain Monte Carlo). En estos trabajos utilizan los métodos MCMC para la exploración de la distribución a posteriori (Gamerman, 1997 y Guttorp, 1995).

La base de datos con la cual se trabajó, contiene 36 distintas variables que contienen los síntomas de las personas posiblemente afectadas por la enfermedad del dengue (Laboratorio Estatal de Salud Pública del Estado de Guerrero, 2006). Por lo cual se realizó un trabajo previo de seleccionar las variables que tienen mayor influencia en el diagnóstico de la enfermedad. Para determinar el conjunto de variables que tienen mayor influencia estadística en la determinación del dengue se realizan pruebas de independencia. Con la metodología de los modelos C&RT en el enfoque de Breiman et al. (1984), utilizando el paquete estadístico S-Plus (Venables y Ripley, 2002; Clark y Pregibon, 1992) y las pruebas de independencia se ensayaron distintos modelos para determinar aquel que mejor explique el comportamiento de los datos. Bajo este enfoque se obtuvo un modelo de árbol de clasificación con

16 nodos terminales, en su construcción interviene siete variables regresoras y la tasa de error de clasificación es del 19.6%. Una vez que se seleccionó el conjunto de variables independientes de mayor influencia sobre el dengue, se construye un modelo de árbol de clasificación bajo el enfoque bayesiano considerando la propuesta de Chipman et al. (1998). Para realizar la aplicación se desarrolló el software en el código de S-Plus. En este caso el modelo que mejor explica el comportamiento de los datos tiene 20 nodos terminales, con siete variables regresoras y una tasa de error de clasificación de 19.0%. El ejercicio fue realizado con una muestra n de tamaño 300. El conjunto de variables independientes que se seleccionaron para construir el modelo de árbol de clasificación son síntomas que presentan los enfermos que son observables, sin exámenes de laboratorio. Para un paciente dado, se obtienen los datos correspondientes a estas variables, y mediante el modelo se intenta diagnosticar si el paciente tiene o no dengue. El modelo puede ser de ayuda para los médicos que se encuentran en el área rural.

4.1.2. Materiales y métodos.

Introducción.

Modelo de árbol de clasificación: En forma similar al modelo de regresión logística, se debe tener una variable de respuesta Y cualitativa y las variables explicativas X_1, X_2, \dots, X_p . Se supone que el conjunto de categorías de la variable Y es $C = \{C_1, C_2, \dots, C_K\}$. Mediante los valores observados de Y y de X , se construye un modelo de árbol binario con el propósito de realizar predicciones o como un dispositivo que posibilite la clasificación.

La idea general de la construcción de un árbol de clasificación, es particionar el espacio de las variables X , en subconjuntos donde la variable de respuesta Y sea cada vez más homogénea, es decir, que los datos de X correspondan a cada vez a menos categorías de Y , en cada nodo terminal. La partición del espacio X , es un proceso iterativo hasta que se satisface un criterio de parada.

Los elementos que constituyen el árbol binario T son: el nodo inicial o nodo raíz, los nodos internos o de división y los nodos terminales. Cada nodo (padre) se divide en dos, estos llamados nodos hijos izquierdo y derecho.

Definición 1: Regla de división. Para dividir el nodo t en los nodos hijo izquierdo t_I e hijo derecho t_D se realizan los siguientes pasos:

- i) Se selecciona la variable X_i del conjunto $\{X_1, X_2, \dots, X_p\}$.

- ii) Cuando $X_i \in [a_i, b_i]$ se selecciona r tal que $r \in (a_i, b_i)$. Los valores x_i que satisfacen $x_i \in [a_i, r]$ se asigna al nodo t_I ; y aquellos que $x_i \notin [a_i, r]$ se asigna a t_D . Cuando $X_i \in \{C_1, C_2, \dots, C_m\}$ se selecciona $A \subseteq \{C_1, C_2, \dots, C_m\}$. Los valores x_i que satisfacen $x_i \in A$ se asignan al nodo t_I ; y aquellos que satisfacen $x_i \notin A$ se asignan al nodo t_D .

A esta forma de dividir los nodos es llamada regla de división.

Para construir un modelo de árbol binario se le aplica la regla de división a cada nodo interno o de división. Existen distintos criterios de selección de la variable X_i para ser utilizada en la división del nodo, así como también, el valor de r , en el caso de una variable cuantitativa o el subconjunto A cuando la variable X_i es cualitativa (Lohn y Shib (1997)).

4.1.3. Modelo de árbol de clasificación bajo un enfoque clásico.

Para la construcción del modelo de árbol de clasificación se adopta el enfoque de Breiman et al. (1984) y los métodos de cálculo realizados en el paquete estadístico S-Plus. Este enfoque fue utilizado por las facilidades de cómputo que proporciona el paquete estadístico S-Plus. El árbol binario se considera como un modelo de probabilidad, esto es, en cada uno de los nodos t del árbol se tiene una distribución de probabilidad de las k categorías, es decir, p_{tk} es la probabilidad de la categoría k en el nodo t , donde $\sum_k p_{tk} = 1$ para $i = 1, 2, \dots, b$, suponiendo que el árbol de clasificación tiene b nodos terminales. Se tienen n observaciones de las variables y n_{tk} denota el número de observaciones de la categoría k en el nodo t , en consecuencia $n = \sum_t \sum_k n_{tk}$. Para evaluar el valor de una división de un nodo, se toma una medida llamada de impureza. En el caso de la metodología utilizada en S-Plus, esta medida es la deviance y para medir la bondad de ajuste del modelo a los datos, se considera la deviance media residual. Se define la deviance D_t en el nodo t como $D_t = -2 \sum_k n_{tk} \log p_{tk}$ para $t = 1, 2, \dots, b$ y la deviance asignada al árbol de clasificación con b nodos terminales es $D(T) = \sum_{t=1}^b D_t$.

Para evaluar la regla de división S en el nodo t , se considera la reducción de la deviance en la siguiente forma. Sean $t_I = 2 * t$, el nodo izquierdo y $t_D = 2 * t + 1$, el nodo derecho generado por la aplicación de la regla de división S . La reducción de la deviance será:

$$D_t - D_{t_I} - D_{t_D} = -2 \sum n_{tk} \log p_{tk} - \left[-2 \sum n_{t_I k} \log p_{t_I k} - 2 \sum n_{t_D k} \log p_{t_D k} \right]$$

Considerando que $n_t = n_{t_I} + n_{t_D}$ y simplificando la expresión anterior se tiene

que

$$D_t - D_{t_I} - D_{t_D} = \sum_k \left[n_{t_I k} \log \left(\frac{p_{t_I k}}{p_{tk}} \right) + n_{t_D k} \log \left(\frac{p_{t_D k}}{p_{tk}} \right) \right]$$

Dado que las distribuciones de probabilidad de las categorías en los nodos terminales son desconocidas, para la aplicación del criterio, se estiman de la siguiente manera:

$$\hat{p}_{tk} = \frac{n_{tk}}{\sum_k n_{tk}} = \frac{n_{tk}}{n_t} \quad \hat{p}_{t_I k} = \frac{n_{t_I k}}{\sum_k n_{t_I k}} = \frac{n_{t_I k}}{n_{t_I}} \quad \hat{p}_{t_D k} = \frac{n_{t_D k}}{\sum_k n_{t_D k}} = \frac{n_{t_D k}}{n_{t_D}}$$

Por tanto la reducción en la deviance en el nodo t será:

$$D_t - D_{t_I} - D_{t_D} = \sum_k \left[n_{t_I k} \log \left(\frac{\hat{p}_{t_I k}}{\hat{p}_{tk}} \right) + n_{t_D k} \log \left(\frac{\hat{p}_{t_D k}}{\hat{p}_{tk}} \right) \right]$$

La elección de la regla de división S , será aquella que haga máxima la reducción de la deviance en el nodo t , esto es, $\Delta D(S^*) = \max_S \{D_t - D_{t_I} - D_{t_D}\}$. Además, con la reducción de la deviance en cada nodo, se logra la reducción en el árbol de clasificación completo T . La deviance asociada al árbol de clasificación T es la suma de la deviance obtenida en cada nodo terminal y la deviance media residual será la deviance total dividida por los grados de libertad ($gl = n - b$), esto es, la diferencia entre el total de observaciones y el número de nodos terminales (Venables y Ripley (1994)).

Regla de parada: El árbol final T está constituido por nodos internos o de división y nodos terminales. Para que el nodo t sea declarado como nodo terminal, debe satisfacer alguna de las siguientes tres condiciones.

- i) Antes de dividir el nodo t , que el número de observaciones n_t sea menor que 10.
- ii) Después de dividir el nodo t , que alguno de los números de observaciones n_{t_I} o n_{t_D} en los nodos hijos, sean menores de cinco.
- iii) Que la deviance D_t en el nodo t sea menor que el 1% de la deviance en el nodo raíz.

Aunque estos son los valores dados por omisión por parte de S-Plus, sin embargo, pueden ser modificados por el usuario.

Asignación de la categoría a cada nodo terminal. La categoría k asignada a cada nodo terminal t , será aquella categoría que tenga la mayor frecuencia de observaciones en el nodo t . En caso de empate, se elige arbitrariamente. También este valor se puede considerar como, el valor estimado de la variable de respuesta Y en el nodo terminal t .

Las observaciones de la variable de respuesta Y que no coinciden con la categoría asignada al nodo terminal se les llaman observaciones mal clasificadas o subclasificadas. La suma de las observaciones subclasificadas de cada uno de los nodos terminales dividida por el total de observaciones, n , es la tasa de error de subclasificación. Esta es también una medida de la bondad de ajuste.

Podado del árbol bajo el criterio de costo de complejidad. Dado que el proceso de crecimiento del árbol no está restringido, el modelo resultante estará sobreajustado. Para lograr un modelo parsimonioso se realiza un proceso de podado sin pérdida significativa de bondad en el ajuste.

La forma de medir la importancia de un subárbol T' , está dado por la llamada medida del costo de complejidad. La cual se define como:

$$D_k(T') = D(T') + \alpha * \text{tamaño}(T') \quad (4.1.1)$$

donde $D(T')$ denota la deviance del subárbol T' , α es el parámetro del costo de complejidad y $\text{tamaño}(T')$ representa el número de nodos terminales del árbol T' . El criterio a seguir es seleccionar aquel subárbol que minimiza la deviance, dado el parámetro de complejidad, α . Para valores grandes de α se obtienen subárboles T' , con un número menor de nodos terminales y viceversa.

En el paquete estadístico S-Plus existe una opción para ajustar un modelo de árbol de clasificación a los datos, proporcionando la gráfica del árbol y un reporte general donde se incluye las variables del modelo, el número de nodos terminales, deviance media residual y la tasa de error de subclasificación. También se reportan los datos correspondientes a cada uno de los nodos como son: el número del nodo, regla de división, el número de observaciones, la deviance, el valor estimado de la variable de respuesta y las probabilidades estimadas para cada categoría.

4.1.4. Modelos de árbol de clasificación bajo el enfoque Bayesiano.

Para realizar la generación y selección de modelos de árbol de clasificación bajo el paradigma Bayesiano, se adopta el enfoque de Chipman et al. (1998). Se inicia con definiciones básicas para explicar el método.

Definición 2: Probabilidad de división de un nodo. La probabilidad de división del nodo interno t está dada por:

$$p = \alpha(1 + d)^{-\beta} \quad (4.1.2)$$

donde d es la profundidad del nodo, esto es, es el número de nodos de división arriba del nodo actual, y $\alpha \in (0, 1)$ y $\beta \geq 0$. Los parámetros α y β son valores predeterminados que influyen en la forma y tamaño del árbol.

Definición 3: Verosimilitud integrada. Sea la distribución conjunta de la variable de respuesta Y y el vector de parámetros Θ , $p(Y, \Theta|T)$, entonces la verosimilitud integrada se define como la distribución marginal

$$p(Y|T) = \int p(Y, \Theta|T)d\theta = \int p(Y|\Theta, T)p(\Theta|T)d\theta \quad (4.1.3)$$

donde $p(Y|\Theta, T)$ es la distribución de las observaciones de la variable de respuesta y $p(\Theta|T)$ es la distribución a priori de los parámetros.

Especificación de la distribución a priori del árbol y de los parámetros.

Distribución a priori del árbol. Bajo el presente enfoque, se hace una especificación implícita de la distribución a priori $p(T)$ del modelo de árbol de clasificación, por medio de un proceso estocástico árbol-generador, que consiste en un algoritmo de tres pasos (Chipman et al., 1998):

- i) Se inicia con el árbol cuyo único nodo es el nodo raíz ($t = 1$).
- ii) El nodo t se divide con probabilidad $p = \alpha(1 + d)^{-\beta}$.
- iii) Cuando el nodo t se divide, se elige la variable X_i del conjunto de predictores, utilizando una distribución uniforme discreta, de igual manera el valor de r o el subconjunto A , necesarios para la división. Con esta regla de división obtenida, se asignan a los nodos hijos izquierdo, $2 * t$, y derecho, $2 * t + 1$, sus correspondientes valores de las variables regresoras. Cuando el nodo no es dividido se le llama nodo terminal.

Se asigna t al siguiente nodo hijo izquierdo o derecho, y se repiten los pasos ii) y iii). Termina el algoritmo cuando no hay más nodos por dividir.

Con este algoritmo se asocia a los nodos de división la probabilidad de dividirse $p = \alpha(1 + d)^{-\beta}$ y a los nodos terminales el complemento $q = 1 - p$. Por lo tanto, la probabilidad $p(T)$ a priori del árbol T , es el producto de las probabilidades asociadas a cada nodo.

Distribución a priori de los parámetros. Se supone que las observaciones y_{ij} pertenecen a una de las K categorías, esto es, C_1, C_2, \dots, C_K , además se supone que $f(y_{ij}|\theta_i)$ se comporta de acuerdo a una distribución multinomial simple, esto es,

$$f(y_{i1}, y_{i2}, \dots, y_{in_i}|\theta_i) = \prod_{j=1}^{n_i} \prod_{k=1}^K p_{ik}^{I(y_{ij} \in C_k)} \quad \text{para } i = 1, 2, \dots, b$$

donde $\theta_i = p_i \equiv (p_{i1}, \dots, p_{iK}), p_{ik} > 0$ y $\sum_k p_{ik} = 1$. Observe que $P(y_{ij} \in C_k)|p_i = p_{ik}$.

La distribución a priori conjugada para los parámetros $\theta = (p_1, p_2, \dots, p_b)$ de la multinomial es la distribución de Dirichlet estándar de dimensión $K-1$ con parámetros $\alpha = (\alpha_1, \dots, \alpha_K)$ con $\alpha_k > 0$, esto significa que:

$$p_1, p_2, \dots, p_b | T \stackrel{\text{iid}}{\sim} \text{Dirichlet}(p_i | \alpha) \propto p_{i1}^{\alpha_1-1} p_{i2}^{\alpha_2-1} \dots p_{iK}^{\alpha_K-1}$$

Con esta especificación de la densidad a priori conjugada, se obtiene la verosimilitud integrada

$$P(Y|X, T) = \left(\frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \right)^b \prod_{i=1}^b \frac{\prod_k \Gamma(n_{ik} + \alpha_k)}{\Gamma(n_i + \sum_k \alpha_k)} \quad (4.1.4)$$

Donde $n_{ik} = \sum_j I(y_{ij} \in C_k), n_i = \sum_k n_{ik}$ con $k = 1, \dots, K$.

Exploración de la distribución a posteriori por el algoritmo de Metropolis-Hastings.

Para la determinación de la distribución a posteriori del modelo de árbol de clasificación $p(T|X, Y)$, salvo la constante de integración, se sustituyen las expresiones de la verosimilitud integrada $p(Y|X, T)$ y la distribución a priori del árbol $p(T)$ en la expresión:

$$p(T|X, Y) \propto p(Y|X, T)p(T). \quad (4.1.5)$$

Aún cuando la distribución de probabilidad a posteriori es discreta y el número de modelos de árbol es finito, no es posible tener una idea del comportamiento de la distribución por los métodos tradicionales. Una dificultad adicional es el desconocimiento de la constante de integración de la densidad a posteriori. Estas son las razones por las cuales la exploración de la distribución a posteriori se trabaja con uno de los métodos de Monte Carlo vía cadenas de Markov (MCMC). Uno de los métodos MCMC, es el llamado algoritmo de Metropolis-Hastings, este tiene la característica de no requerir la constante de integración y trabaja bien, aún cuando el vector de parámetros es de dimensión variable (Gamerman, 1997 y Guttorp, 1995).

Densidad generadora de candidatos.

Para este problema particular, el algoritmo de Metropolis-Hastings, simula una cadena de Markov de modelos de árbol, $T^0, T^1, T^2 \dots$, a partir de la cual se puede obtener una muestra simulada de la distribución a posteriori. El algoritmo se inicia con un árbol T^0 , con base en éste, se generan los siguientes elementos de la cadena de Markov, en general, a partir del i -ésimo elemento T^i de la cadena, se genera un árbol-candidato T^* con lo que: se toma $T^{i+1} = T^*$ con probabilidad

$$\alpha(T^i, T^*) = \min \left\{ \frac{q(T^*, T^i)p(Y|X, T^*)p(T^*)}{q(T^i, T^*)p(Y|X, T^i)p(T^i)}, 1 \right\}$$

en caso contrario se considera $T^{i+1} = T^i$.

La tasa de aceptación es el porcentaje del número de veces que la cadena acepta un candidato como nuevo elemento de la cadena. Este número es indicativo del comportamiento de la cadena (Chib y Greenberg, 1995).

Para modificar el árbol T^i y obtener el árbol candidato T^* , se selecciona en forma aleatoria uno de los cuatro pasos siguientes, propuestos por Chipman et al. (1998):

- i) Crecimiento (CR): siguiendo una distribución uniforme se selecciona un nodo terminal, para dividirse de acuerdo a las reglas de selección de variables y valores de división.
- ii) Podado (PD): siguiendo una distribución uniforme se selecciona un nodo interno que tenga sus nodos hijos, izquierdo y derecho, como nodos terminales, se eliminan estos nodos hijos.
- iii) Cambio (CA) de la regla de división de un nodo interno: siguiendo una distribución uniforme se selecciona un nodo interno para cambiar la regla de división.
- iv) Intercambio (IC) de las reglas de división de dos nodos internos: siguiendo una distribución uniforme se selecciona un par de nodos internos, padre e hijos, y se intercambian sus reglas de división.

Para implementar, en esta aplicación particular, el algoritmo de Metropolis-Hastings, se define una densidad generadora de candidatos, $q(T^i, T^*)$, incorporando los cuatro movimientos anteriores y que satisfaga la condición de reversibilidad (Guttorp, 1995).

Criterios de selección de los modelos.

El número de elementos que debe tener la cadena de Markov, deben ser los suficientes para que la cadena se estabilice alrededor de un punto modal. En la aplicación presente, la cadena de Markov tiene 5,000 elementos y se realizaron diez cadenas, para posibilitar encontrar más puntos próximos a distintos puntos modales. Se seleccionaron de cada cadena dos modelos: el primero, aquel modelo de árbol de clasificación con el mayor valor del logaritmo de la verosimilitud; el segundo, aquel modelo que se obtiene en la última iteración de la cadena. Un criterio adicional que se utiliza para seleccionar el modelo más adecuado, es la tasa de error de subclasificación. El criterio de seleccionar el modelo con el mayor valor de la verosimilitud integrada $p(Y|X, T)$, en lugar del modelo con la mayor probabilidad a posteriori, es para evitar la influencia de la distribución a priori de árbol $p(T)$.

4.1.5. Resultados y discusión.

Análisis descriptivo de la base de datos sobre el dengue.

La modelación estadística se realizó sobre la base de datos del dengue correspondientes al año 2006 en distintos municipios del estado de Guerrero. La base de datos original contiene 10,282 registros de los cuales 8,068 corresponden a expedientes de personas que no tienen dengue y 2,214 que si lo tienen. Existen 533 con dengue hemorrágico y 1,681 con dengue clásico. Se realiza una depuración de la base de datos original para evitar los registros con datos faltantes, quedando la base con 3,272 registros. Las características de la base depurada se presentan en la Tabla No. 1 del Anexo.

Prueba de independencia.

Se utilizó la prueba de la Ji-Cuadrada para determinar el grado de dependencia de las distintas variables cualitativas con respecto a la variable del dengue. Para realizar la prueba se consideró que las categorías de la variable de respuesta son: dengue hemorrágico, dengue clásico y personas sin dengue; y como segunda variable, cada una de las 33 variables categóricas existentes en la base de datos. En la siguiente tabla se presenta el valor de la Ji-Cuadrada, calculada en el paquete estadístico S-Plus con la base de datos de 3,272 registros (ver Agresti, 1996). En la Tabla No. (4.1) se presentan los resultados de las pruebas.

Adicionalmente a las 33 variables cualitativas regresoras, la base de datos contempla las variables cuantitativas la edad y temperatura corporal de los

Cuadro 4.1: Prueba de independencia de la variable dengue con el resto de variables, utilizando la prueba Ji-cuadrada.

No.	Variable	Ji-Cuadrada	No.	Variable	Ji-Cuadrada
1	Escape líquidos	467.2249	18	Dolor retroocular	35.8263
2	Petequia	408.3843	19	Hematomas	35.3948
3	Exantema	366.1314	20	Fotofobia	34.6612
4	Prurito	333.965	21	Hematemesis	33.3377
5	Vomito	282.5316	22	Equimosis	33.3069
6	Hemorragias	203.7782	23	Rinitis	31.8548
7	Torniquete	176.5432	24	Melena	30.1176
8	Nauseas	173.2917	25	Ascitis	21.2505
9	Dolor abdominal	145.9509	26	Conjuntivitis	20.3121
10	Tos	103.4691	27	Hepatomegalia	17.9406
11	Diarrea	84.1348	28	Mialgia	17.7105
12	Congestión nasal	67.6205	29	Sexo	13.5224
13	Gingival	64.9367	30	Derrame pleural	11.3592
14	Epistaxis	51.8602	31	Fiebre	8.4153
15	Escalofrío	50.7935	32	Esplenomegalia	8.0232
16	Faringitis	44.5836	33	Cefalea	4.0916
17	Artralgia	37.9127			

pacientes. En las Tablas No. (4.2) y (4.3), se presentan sus valores mínimos, máximos, media, entre otras estadísticas de estas variables.

Cuadro 4.2: Comportamiento de la edad según la variable dengue.

Dengue	Mínimo	Máximo	Media	Mediana	1er cuartil	3er. cuartil	Desv. est.
No dengue	0.33	90	27.04	20	12	40	19.07
DC	1	85	20.7	15	10	28	15.12
DH	0.013	69	19.52	16	11	23	13.06

4.1.6. Modelación bajo el enfoque de Breiman et al. (1984) y S-Plus (2000).

Para realizar la modelación estadística, se trabajó con la base de datos que contiene 3,272 registros. Se consideraron sólo dos categorías de la variable de respuesta: “dengue Sí” y “dengue No”, en el trabajo de modelación.

Se utilizó el paquete estadístico S-Plus para la selección y construcción de los modelos de árbol de clasificación bajo el enfoque de Breiman et al. (1984). Con la base de datos, se ensayaron distintos subconjuntos de las 35 variables independientes, considerando aquellas con mayor índice de dependencia con la variable del dengue. El método seguido fue partir de un modelo con las cinco variables independientes con los mayores valores de la Ji-Cuadrada

Cuadro 4.3: Comportamiento de la temperatura según la variable dengue.

Dengue	Mínimo	Máximo	Media	Mediana	1er cuartil	3er. cuartil	Desv. est.
No dengue	36	41	38.32	38	38	38.9	0.66
DC	36	41	38.43	38	38	39	0.69
DH	37	40	38.91	39	38	39.5	0.74

(Tabla No. (4.1)) y agregar una a una las variables independientes hasta llegar a diez. Después de un proceso de agregar y quitar variables se llegó a obtener un modelo de árbol de clasificación, tentativamente el “mejor”. Este modelo tiene 9 variables explicativas, las cuales se muestran en la primera línea de la Tabla No. (4.4), con 157 nodos terminales, con una deviance media residual de 0.788 y una tasa de subclasificación de 17.2%. Los indicadores de bondad de ajuste de los modelos y las variables utilizadas en su construcción se muestran en la Tabla No. (4.4). Para evaluar los modelos se consideran el número de nodos terminales, la deviance media residual (DMR) y la tasa de error de subclasificación (TESC).

En la Tabla No. (4.4), se ilustra el proceso de selección de las variables, partiendo del mejor modelo e incorporando las variables con mayor índice de dependencia con respecto a la variable de respuesta. En la misma tabla a partir de la línea 8 se realizó otro ejercicio, quitando una variable al mejor modelo.

Tabla No. 7. Características de los modelos tentativos para los datos sobre dengue.

Dado que el proceso de crecimiento del árbol no está restringido, el modelo resultante está sobreajustado. Para lograr un modelo parsimonioso se realiza un proceso de podado sin pérdida considerable de bondad en el ajuste. La forma de medir la importancia de un subárbol T' , está dado por la llamada medida del costo de complejidad (1).

Para realizar el proceso de podado se toma el modelo 1 de la Tabla No. (4.4), considerando distintos valores del costo de complejidad. En la Tabla No. (4.5), se proporcionan los valores resultantes del proceso de podado para distintos valores del parámetro de costo de complejidad.

Evaluación estadística del modelo.

En la Tabla No. (4.5) se presentan sólo los indicadores de bondad de ajuste para cuatro valores del parámetro de complejidad, tomando en cuenta éstos, se seleccionó el modelo correspondiente a la tercera línea de dicha tabla, como el mejor. Este modelo de árbol de clasificación obtenido después de realizar

Cuadro 4.4: Comportamiento de la temperatura según la variable dengue.
 Modelo 1: Dengue \sim Edad + Temperatura + Doloretr + Exantema + Nauseas + Escalofrío + Dolorabdominal + Tos + Faringitis

No.	Variables independientes en el modelo	Nodos Terminales	DMR(*)	TESC(**)
1	(Modelo 1)	157	0.788	0.172
2	(Modelo 1)+Prurito	162	0.761	0.161
3	(Modelo 1)+Vomito	161	0.779	0.167
4	(Modelo 1)+Escape líquidos	156	0.773	0.168
5	(Modelo 1)+Petequia	157	0.769	0.164
6	(Modelo 1)+Torniquete	159	0.784	0.169
7	(Modelo 1)+Hemorragias	158	0.771	0.165
8	(Modelo 1)-Edad	135	0.873	0.189
9	(Modelo 1)-Temperatura	152	0.823	0.179
10	(Modelo 1)-Exantema	171	0.837	0.187
11	(Modelo 1)- Dolor retroocular	164	0.795	0.172
12	(Modelo 1)- Nauseas	157	0.790	0.169
13	(Modelo 1)- Escalofrío	169	0.795	0.169
14	(Modelo 1)-Dolor abdominal	156	0.789	0.173
15	(Modelo 1)- Tos	161	0.801	0.172
16	(Modelo 1)-Faringitis	164	0.793	0.172
17	(Modelo 1)-Escalofrío - Dolor abdominal	161	0.798	0.172
18	(Modelo 1)-Nauseas-Dolor abdominal	169	0.815	0.180
19	(Modelo 1)-Tos-Faringitis	157	0.839	0.178
20	(Modelo 1)-Exantema-Escalofrío	168	0.853	0.187

(*) Deviance media residual.

(**) Tasa de error de subclasificación.

el podado tiene 16 nodos terminales, la deviance media residual es de 0.887 y la tasa de error de subclasificación es 0.196. En el caso del modelo 1, que resulto ser el “mejor” tiene 157 nodos terminales, una deviance promedio de 0.788 y su tasa de subclasificación es 0.172. Una diferencia importante entre los dos modelos es la cantidad de nodos terminales y el número de variables utilizadas, aunque se sacrifica bondad de ajuste (Ver Figura No. 1)

En la Tabla No. (4.6) se muestran las características más relevantes de los 16 nodos terminales del mejor modelo de árbol de clasificación. Se observa que existen sólo seis nodos terminales donde la categoría asignada es “dengue Sí”, el otro hecho que hay que hacer notar es que se construye el modelo con el total de las observaciones existentes en la base de datos y que la tasa de error de subclasificación es del 19.6 %, esto significa que de los 3272 casos, 642 son mal clasificados.

Cuadro 4.5: Características de los modelos después del proceso de podado del “mejor” modelo.

Parámetro de costo de complejidad	No. de nodos terminales	RMD (*)	TESC (**)	Variables Excluidas.
	157	0.788	0.172	
6	24	0.872	0.193	Doloretro
8	16	0.887	0.196	Doloretro, Faringitis
10	12	0.897	0.198	Doloretro, Faringitis
15	8	0.909	0.201	Cinco

(*) Deviance media residual. (**) Tasa de error de subclasificación.

Cuadro 4.6: Características de los nodos terminales del mejor modelo de árbol de clasificación.

No.	No. de nodo Terminal	No. de Observaciones	No. de “No dengue”	No. de “Si Dengue”	Categoría asignada
1	5	37	30	7	“dengue No”
2	9	37	6	31	“Si Dengue”
3	13	176	174	2	“dengue No”
4	16	88	25	63	“Si Dengue”
5	17	285	129	156	“Si Dengue”
6	25	8	3	5	“Si Dengue”
7	29	175	138	37	“dengue No”
8	31	696	613	83	“dengue No”
9	48	307	266	41	“dengue No”
10	49	514	479	35	“dengue No”
11	57	19	5	14	“Si Dengue”
12	61	306	248	58	“dengue No”
13	112	209	117	92	“dengue No”
14	113	205	147	58	“dengue No”
15	120	203	143	60	“dengue No”
16	121	7	1	6	“Si Dengue”
Totales		3272	2524	748	

4.1.7. Modelación bajo el enfoque Bayesiano (Chipman et al., 1998).

Para la construcción de los modelos de clasificación bajo el enfoque Bayesiano, se toma como base la propuesta de Chipman et al. (1998). Para realizar el ejercicio se ha tenido que desarrollar el software en el código proporcionado por el paquete estadístico S-Plus, el cual tiene sus restricciones. Una de estas restricciones es la falta de flexibilidad para el manejo del número de variables, por esta razón se han retomado las variables más relevantes de la construcción del modelo bajo el primer enfoque trabajado. Una restricción adicional, es el número más pequeño que se puede manejar en el paquete, para el cálculo del logaritmo de verosimilitud integrada, por esta razón el

tamaño de la muestra aleatoria se restringe a 300 casos. Las variables que se trabajaron para la construcción de los modelos se muestran en la Tabla No. (4.7).

Cuadro 4.7: Características de los nodos terminales del mejor modelo de árbol de clasificación.

No.Var.	Variable	Ji-Cuadrada	Valor en la muestra
X1	Edad		Min=2, media=24.8,Max=80
X2	Temperatura		Min=36,media=38.35,Max=41
X3	Dolor retrocular	35.8263	Si=184 , No=116
X4	Exantema	366.1314	Si=38 , No=262
X5	Nauseas	173.2917	Si=117 , No=183
X6	Tos	103.4691	Si=100 , No=200
X7	Faringitis	44.5836	Si=86 , No=214
VarY	Dengue		Dengue=76, No dengue=224

En el caso de las variables cuantitativas que son la edad (X1) y la temperatura (X2), se define una retícula con el objetivo de reducir la cantidad de las particiones del espacio de las variables regresoras.

Resultados de la modelación.

Para la selección y construcción de los modelos de árbol de clasificación bajo el enfoque Bayesiano, se aplica el algoritmo de Metropolis-Hastings. Se genera una cadena de Markov de modelos de árbol, cuya longitud es de 5,000 modelos. El último modelo de cada cadena se selecciona y además el modelo con un valor del logaritmo de verosimilitud máximo en las 5,000 iteraciones. Se realizan 10 ejecuciones del programa con lo cual se logra tener 10 modelos finales y 10 modelos óptimos. En la Figura No. 1, se presenta la gráfica del logaritmo de la verosimilitud integrada en el eje de las ordenadas y el número de iteración en el eje de las abscisas, con los resultados de las 10 corridas del programa.

Figura 1

Evaluación estadística de los modelos.

En la Tabla No. (4.8) se presentan las características relevantes de los modelos de árbol de clasificación que obtuvieron el mayor valor del logaritmo de verosimilitud en una corrida del programa de 5,000 iteraciones y las características de los últimos modelos de la cadena de Markov se presentan en la Tabla No. (4.9).

Se realiza una comparación del conjunto de modelos óptimos con los finales, bajo distintos criterios. El logaritmo de la verosimilitud promedio de los

Cuadro 4.8: Características de los diez modelos de árbol de clasificación con el valor del logaritmo de verosimilitud más grande, en cada corrida de 5,000 iteraciones.

No.	Log de la verosimilitud	No. de nodos terminales	Variable de la 1ª. división	Tasa de error por subclasificación	Tasa de aceptación
1	-67.0988	19	X5	62/300=0.207	50.24 %
2	-66.6499	22	X2	63/300=0.210	51.80 %
3	-67.2567	18	X2	59/300=0.197	50.96 %
4	-64.2753	17	X4	55/300=0.183	53.02 %
5	-64.9854	17	X1	57/300=0.190	53:26 %
6	-65.1772	20	X3	59/300=0.197	45:06 %
7	-65.9276	20	X6	65/300=0.217	53:24 %
8	-63.7312	16	X4	59/300=0.197	52.42 %
9	-64.5207	18	X1	61/300=0.203	54.42 %
10	-64.1643	20	X1	57/300=0.190	52.04 %

Cuadro 4.9: Características de los diez modelos de árbol de clasificación finales de cada cadena de Markov de 5,000 iteraciones.

No.	Log de la verosimilitud	No. de nodos terminales	Variable de la 1ª. división	Tasa de error por subclasificación	Tasa de aceptación
1	-67.7850	23	X5	62/300=0.2066	50.24 %
2	-69.4710	22	X2	62/300=0.2066	51.80 %
3	-67.9884	22	X2	59/300=0.1966	50.96 %
4	-66.2029	16	X4	58/300=0.1933	53.02 %
5	-67.5929	20	X1	60/300=0.20	53:26 %
6	-66.1967	19	X3	58/300=0.1933	45:06 %
7	-68.6135	18	X6	68/300=0.2266	53:24 %
8	-64.8413	20	X4	60/300=0.20	52.42 %
9	-66.6554	22	X1	60/300=0.20	54.42 %
10	-68.1212	19	X1	59/300=0.1966	52.04 %

árboles óptimos es igual a -65.38 y para el caso de los finales es igual a -67.35. El número promedio de los nodos terminales de los óptimos es de 18.7 nodos terminales contra 20.2 de los árboles finales. En el caso de la tasa de error de subclasificación los óptimos tienen 19.9 % y los finales tienen 20.2 %. En consecuencia, los modelos de árbol de clasificación óptimos son mejores estadísticamente, en su conjunto, que los modelos finales de las cadenas de Markov.

Se realiza una comparación al interior de los modelos óptimos. El modelo número 8 de la Tabla No. (4.8), es el mejor considerando el criterio del logaritmo de la verosimilitud y el menor número de nodos terminales, sin embargo, considerando la menor tasa de error de subclasificación el mejor modelo es el número 4. Comparando estos dos modelos, se observa que ambos

tienen a la variable X4 como variable de división en el nodo raíz y tienen una estructura muy similar como árboles.

Un modelo alternativo pudiera ser el número 10. Con el criterio del mayor logaritmo de verosimilitud, después del mejor modelo, se encuentra el número 10 con un valor igual a -64.16. En cuanto a los nodos terminales el modelo 10 tiene una mayor cantidad (20). Si se considera la menor tasa de error de subclasificación, el modelo 10 es mejor que el número 8. El modelo 10 tiene como variable de división en el nodo raíz a X1, con esta variable existen 3 de los 10 modelos óptimos.

Como se señaló para realizar el ejercicio bajo el enfoque Bayesiano, se trabajó con una muestra de tamaño 300, por lo que, una forma de aprovechar toda la información contenida en la base de datos es aplicar la estructura de los dos modelos 8 y 10 a esta base de datos, para verificar su capacidad de clasificación. Se realizó el ejercicio y se obtuvo que el modelo 8 obtiene una tasa de error subclasificación del 25.06 %, esto es, de los 3272 casos, 820 casos los clasifica mal. En el caso del modelo 10, el porcentaje correspondiente es de 21.45 %, lo que significa que clasifica mal, 702 casos.

4.1.8. Discusión de los modelos.

Una vez realizada la modelación bajo los dos enfoques se puede señalar que, estadísticamente, los modelos obtenidos son muy similares considerando el número de nodos terminales y la tasa de error de subclasificación. Sin embargo, haciendo un análisis de los modelos del enfoque de Breiman et al. (1984) y el modelo 10 del enfoque Bayesiano, sobre los 748 casos de “dengue Sí” se puede señalar que: bajo el enfoque de Breiman, de los 748 casos de “dengue Sí”, clasifica bien 275, pero los 473 casos restantes los clasifica como “No Dengue”, esto significa que, del conjunto de los 748 casos el 36.76 % de los casos son bien clasificados; bajo el enfoque de Chipman et al. (1998), se obtuvo que de los 748 casos de “dengue Sí”, se clasifican bien el 39.03 %, lo que representa 292 casos.

Cuadro 4.10: Características de los nodos terminales del modelo de clasificación 8.

No.	No. de nodo Terminal	No. de Observaciones	No. de "NO"	No. de "SI"	Categoría asignada
1	4	32	8	24	"dengue Sí"
2	5	6	5	1	"No Dengue"
3	12	48	47	1	"No Dengue"
4	26	9	8	1	"No Dengue"
5	30	6	6	0	"No Dengue"
6	31	62	52	10	"No Dengue"
7	54	19	14	5	"No Dengue"
8	55	10	10	0	"No Dengue"
9	57	11	10	1	"No Dengue"
10	112	8	6	2	"No Dengue"
11	113	23	11	12	"dengue Sí"
12	116	17	9	8	"No Dengue"
13	117	11	10	1	"No Dengue"
14	118	9	9	0	"No Dengue"
15	238	14	7	7	"dengue Sí"
16	239	15	12	3	"No Dengue"
	Totales	300	224	76	

Cuadro 4.11: Características de los nodos terminales del modelo de clasificación 10.

No.	No. de nodo Terminal	No. de Observaciones	No. de "NO"	No. de "SI"	Categoría asignada
1	4	24	4	20	"dengue Sí"
2	14	21	21	0	"No Dengue"
3	15	11	9	2	"No Dengue"
4	21	10	6	4	"No Dengue"
5	24	6	2	4	"dengue Sí"
6	25	6	5	1	"No Dengue"
7	26	25	25	0	"No Dengue"
8	40	6	6	0	"No Dengue"
9	41	5	5	0	"No Dengue"
10	44	19	18	1	"No Dengue"
11	47	45	38	7	"No Dengue"
12	54	18	10	8	"No Dengue"
13	90	25	12	13	"dengue Sí"
14	92	8	6	2	"No Dengue"
15	93	13	10	3	"No Dengue"
16	111	5	5	0	"No Dengue"
17	182	23	16	7	"No Dengue"
18	183	9	9	0	"No Dengue"
19	220	15	14	1	"No Dengue"
20	221	6	3	3	"dengue Sí"
	Totales	300	224	76	

Bibliografía

- [1] Díaz-Quijano, F. A., Villar-Centeno, L. A. y Martínez-Vega, R. A. (2006). Complicaciones asociadas a la trombocitopenia profunda en pacientes con dengue”. *Revista Médica Chile*; 134: 167-173.
- [2] Secretaría de Salud, Gobierno Federal México. (2005): Manual para la vigilancia, diagnóstico, prevención y control del dengue.
- [3] Beiman, L.; Friedman J. H.; Olshen R. A. y Stone .J. (1984): *Classification and Regression Trees*, Waldsworth International Group: Belmont, California.
- [4] Chipman, H. A.; George, E. I. y Mcculloch, R. E. (1998): “Bayesian CART model search. (with discussion)”, *Journal of the American Statistical Association*, Vol. 93, No. 443, pp. 935-960.
- [5] Chipman, H. A.; George, E. I. And Mcculloch, R. E. (2002): “Bayesian treed models”, *Machine Learning* 48, pp. 299-320.
- [6] Denison, D. G. T.; Mallick, B. K. y Smith, A. F. M. (1998): “A Bayesian CART Algorithm”, *Biometrika* 85,2, pp. 363-377.
- [7] Denison, D. G. T.; Mallick, B. K. y Smith, A. F. M.(2002): *Bayesian Methods for Nonlinear Classification and Regression*, John Wiley & Sons, Ltd: USA.
- [8] Gamerman, D. (1997): *Markov chain Monte Carlo. Stochastic simulation for Bayesian inference*, Chapman & Hall: London.
- [9] Guttorp, P. (1995): *Stochastic Modeling of Scientific Data*. Chapman & Hall: Great Britain.
- [10] Laboratorio Estatal de Salud “Dr. Galo Soberón y Parra”, Gobierno del Estado de Guerrero (2006): “Base de datos sobre los reportes clínicos de los casos de dengue del año 2006”.
- [11] Chib, S. And Greenberg, E. (1995): “Understanding the Metropolis-Hastings algorithm”, *The American Statistician*, 49,327-335.
- [12] S-Plus 2000 Professional Release 1. (1999)Mathsoft, Inc.: Seattle, Washington.

4.2. Diagnóstico mediante regresión logística binaria.

Ana María Aguilera del Pino¹ , Agustín Santiago Moreno²

1. Universidad de Granada, España.

2. Universidad Autónoma de Guerrero, México. UAM.

Resumen: El dengue es una enfermedad febril aguda de origen viral que representa un problema de salud pública en diferentes regiones tropicales y subtropicales del mundo. La vigilancia epidemiológica del dengue basada en el diagnóstico temprano de la enfermedad a partir de los síntomas sufridos por los pacientes, es esencial para evitar la muerte de pacientes enfermos, prevenir la propagación de la enfermedad y reducir los costos derivados del diagnóstico basado en pruebas serológicas y el cuidado paciente. Este trabajo presenta una aplicación con datos reales, del estado mexicano de Guerrero, en que se considera que modelo de regresión logit con variables predictoras categóricas y proporciona a médicos una prueba de diagnóstico basado en los síntomas con alta sensibilidad para diferenciar entre el dengue la fiebre clásica y la forma más severa de fiebre hemorrágica. El modelo ajustado también permite identificar los síntomas significativos asociados a cada uno de estos dos formas de dengue.

Key words: logit models; Dengue

4.2.1. Introducción

Uno de los principales problemas en el control epidemiológico del dengue es la dificultad para la identificación temprana de esta enfermedad de otras causas de fiebre aguda. De hecho, el diagnóstico del dengue basado en los síntomas no es confiable en la actualidad y debe ser confirmado por estudios de laboratorio basados en pruebas moleculares y serológicas (Gubler DJ, 2002 y Beasley CM, 2008). Por lo tanto, existe una gran demanda para la detección y diferenciación rápida de la infección por virus de dengue en la fase aguda de la enfermedad para proporcionar el tratamiento clínico oportuno, investigación etiológica y el control de la enfermedad. Hasta ahora, se han realizado pocos estudios para determinar el nivel de síntomas que identifican la condición de dengue. Una escala para el diagnóstico clínico de dengue a partir de la más alta sensibilidad y especificidad fue propuesta por WHO y descrita en Arcari P, et al. (2007).

Como se ha indicado antes, la vigilancia epidemiológica del dengue es un elemento esencial para la prevención y control de la enfermedad. En nuestro caso, trabajamos con los 2215 casos confirmados de dengue, obtenidos de la población de 16604 muestras para confirmación serológica en el laboratorio estatal de salud pública en el estado de Guerrero (LESPEG).

El objetivo del presente trabajo consistió en proponer un modelo de diagnóstico de la enfermedad basado en los síntomas con una alta capacidad de diferenciar entre los pacientes con dengue clásico (DC) y dengue hemorrágico (DH). Teniendo en cuenta que la razón de positividad en el diagnóstico de DC es muy bajo, respecto del diagnóstico de DH (debido a la mayor claridad en la manifestación de síntomas de DH según los expertos clínicos), consideraremos como “respuesta” una variable aleatoria binaria Y que toma el valor 1 si el paciente es sufre de DC y cero si es está afectado por DH. Cada síntoma se representa en el modelo por una variable binaria X que toma valor uno si un paciente tiene el síntoma y valor cero en el caso contrario. Se trabajó con una base de datos generada en un estudio anterior sobre consenso (Allende S., et al., 2008) en que el interés principal se centró en evaluar la clasificación hecha por médicos en pacientes sospechosos de dengue usando índices tipo kappa, como herramienta de análisis. Teniendo en cuenta las características de las variables involucradas proponemos usar a un modelo logit con variables predictivas categóricas (Gibbons RV, et al. 2002, Gubler DJ, 1998). El modelo logit se ha usado en la literatura del dengue con propósitos diferentes: para hallar las variables asociadas con la trombocytopenia severa en pacientes con el dengue en Colombia (Díaz- Qijano FA, 2006), para estudiar la asociación entre la depresión y una segunda infección por dengue en el sur de Texas (Beasley CM, 2008), para identificar los síntomas, signos físicos y características de laboratorio que permiten diferenciar la fiebre de dengue de otras enfermedades febriles dentro de los primeros días de admisión hospitalaria en Singapur (Arcari P., 2007) o para mostrar que el riesgo relativo de tener la enfermedad clásica, después de una infección primaria de virus de dengue, aumenta con la edad (Beasley CM, 2008).

4.2.2. Metodología.

En primer lugar, daremos un breve resumen de la teoría básica de la regresión logística: la formulación del modelo, interpretación, estimación e inferencia estadística.

Sea $X = (x_{ij})_{n \times p}$ una matriz cuyas filas $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ son una muestra de n diferentes observaciones de un vector de variables (X_1, \dots, X_p) observadas sin error. Y sea $y = (y_1, \dots, y_n)^T$ una muestra aleatoria de una variable respuesta binaria Y asociada con las observaciones en X , es decir,

$y_i \in \{0, 1\}, i = 1, \dots, n$. Entonces, el modelo de regresión logística muestral se formula como

$$y_i = \pi_i + \varepsilon_i, i = 1, \dots, n, \quad (4.2.1)$$

donde ε_i son errores independientes con media cero y π_i son las esperanzas de Y dados los valores observados modelados por

$$\pi_i = P\{Y = 1 | X_1 = x_{i1}, X_2 = x_{i2}, \dots, X_p = x_{ip}\} = \frac{\exp(\beta_0 + \sum_{j=1}^p x_{ij}\beta_j)}{1 + \exp(\beta_0 + \sum_{j=1}^p x_{ij}\beta_j)}, \quad (4.2.2)$$

$i = 1, \dots, n$, siendo $\beta_0, \beta_1, \dots, \beta_p$ los parámetros del modelo.

Si definimos la transformación logit $l_i = \ln \frac{\pi_i}{1-\pi_i}, i = 1, \dots, n$, donde $\frac{\pi_i}{1-\pi_i}$ da los odds de respuesta $Y = 1$ para los valores observados x_i , entonces el modelo de regresión logística puede expresarse en forma matricial como un modelo lineal generalizado con la transformación logit como función de enlace o link (McCullagh y Nelder, 1983)

$$l = X\beta \quad (4.2.3)$$

con $l = (l_1, \dots, l_n)^T$ el vector de transformaciones logit previamente definidas, $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$ el vector de parámetros y $X = (\mathbf{1}|X)$ la matriz de diseño con $\mathbf{1} = (1, \dots, 1)^T$ un vector n -dimensional de unos.

Observe que cada parámetro del modelo logit puede ser interpretado como el cambio aditivo que es producido en la transformación logit por cada unidad de incremento en la variable explicativa asociada con el parámetro. Para una interpretación más intuitiva desde un punto de vista práctico, trabajaremos con la exponencial de los parámetros que pueden ser interpretados en términos de los odds ratios. Dada una única observación de las variables explicativas (x_1, \dots, x_p) , la exponencial del j -ésimo parámetro ($j = 1, \dots, p$) puede ser interpretada como una odds ratio. Eso significa que los odds de la respuesta $Y = 1$ se multiplican por su exponencial cuando los valores de la variable asociada X_j se incrementan en una unidad y las otras variables explicativas se mantienen fijas

$$\theta(\Delta X_j = 1 | X_k = x_k, \forall k \neq j) = \frac{\pi(x_1, \dots, x_{j+1}, \dots, x_p)}{1 - \pi(x_1, \dots, x_{j+1}, \dots, x_p)} \frac{\pi(x_1, \dots, x_j, \dots, x_p)}{1 - \pi(x_1, \dots, x_j, \dots, x_p)} = \exp\{\beta_j\}.$$

Observe que una variable categórica X_j con dos niveles A y B puede ser incluida en el modelo como una variable binaria que tiene valor 1 para individuos en A y valor 0 para individuos en B. En este caso la exponencial del parámetro β asociado a esta variable es la razón entre el odds de la respuesta $Y = 1$ para el sujeto en A y la misma odds para un sujeto en B.

El método más usado para estimar el modelo logístico es el método de máxima verosimilitud (ver por ejemplo, Ryan, 1992).

Sea $\mathcal{L} = \ln L(y; \beta)$ la log verosimilitud de los datos muestrales, donde la verosimilitud está dada por el producto de n funciones masa de probabilidad Bernoulli

$$L(y; \beta) = \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i}. \quad (4.2.4)$$

Diferenciando \mathcal{L} con respecto a β e igualando a cero obtenemos las ecuaciones de verosimilitud

$$X^T(y - \pi) = 0, \quad (4.2.5)$$

con $\pi = (\pi_1, \dots, \pi_n)^T$. Estas ecuaciones no son lineales en términos de los parámetros. Debido a esto, los parámetros del modelo son aproximados usando el método recursivo de Newton-Raphson.

Cuando las variables explicativas son continuas, puede fijarse una x_i diferente para cada sujeto en la muestra. Sin embargo, cuando las variables explicativas son categóricas (este es el caso con datos de dengue), hay más de una observación sobre Y para cada combinación diferente de categorías (una x_i fija). En estos casos y_i hace referencia al número total de respuestas $Y = 1$ para cada configuración diferente de las variables explicativas en lugar de respuestas binarias individuales. Por lo tanto, las variables que generan los valores observados y_i son variables aleatorias binomiales independientes con esperanza $m_i = n_i \pi_i$, donde n_i es el número total de observaciones de la respuesta binaria para cada x_i .

La estimación de los parámetros del modelo puede ser usada para calcular las probabilidades predichas y las frecuencias esperadas estimadas. Para hacer inferencia estadística sobre el modelo logístico usaremos las propiedades de insesgader y normalidad para muestras grandes de los estimadores de máxima verosimilitud (MV) de los parámetros proporcionados por el resultado asintótico general de Wald para estimadores MV. El test de bondad de ajuste para la regresión logística está dada por

$$\begin{aligned} H_0 : l_i &= x_i & i = 1, \dots, n \\ H_1 : l_i &= x_i \beta'_i \end{aligned}$$

Donde las frecuencias esperadas estimadas $\hat{m}_i = n_i \hat{\pi}_i$ son relativamente grandes, podemos hacer test de bondad de ajuste con un estadístico chi-cuadrado de Pearson o mediante razón de verosimilitud. El estadístico más frecuentemente usado es el estadístico de razón de verosimilitud de Wilks asociado al modelo (4.2.1) conocido como devianza y dado por

$$G^2(M) = 2 \sum_{i=1}^n \left[y_i \ln \frac{y_i}{\hat{\pi}_i} + (n_i - y_i) \ln \frac{n_i - y_i}{n_i - \hat{\pi}_i} \right] \underset{n \rightarrow \infty}{\overset{H_0}{\rightsquigarrow}} \chi_{n-p-1}^2. \quad (4.2.6)$$

Así, el modelo es rechazado con nivel de significación α cuando el p -valor asociado (mínimo nivel de significancia para rechazar H_0) es menor o igual que α .

Otro test importante en regresión logística es el test de significancia para cada una de las variables independientes en el modelo que está dada por

$$\begin{aligned} H_0 : \beta_j &= 0 \\ H_1 : \beta_j &\neq 0. \end{aligned} \quad (4.2.7)$$

Este test puede resolverse en diferentes formas. Una de estas está basada en el estadístico de Wilks asociado al estadístico de razón de verosimilitud condicional dado por

$$G^2(M_j/M) = 2 \sum_{i=1}^n \left[y_i \ln \frac{\hat{\pi}_{iM_j}}{\hat{\pi}_i} + (n_i - y_i) \ln \frac{n_i - \hat{\pi}_{iM_j}}{n_i - \hat{\pi}_i} \right] \underset{n \rightarrow \infty}{\overset{H_0}{\rightsquigarrow}} \chi_1^2. \quad (4.2.8)$$

donde M simboliza el modelo con todas las variables, M_j es un modelo simple que resulta cuando igualamos a cero el parámetro β_j en el modelo M y $\hat{\pi}_{iM_j}$ son las probabilidades estimadas por el modelo M_j . Este test estadístico del modelo M_j dado que M se conserva y puede descomponerse como la diferencia entre las devianzas de los dos modelos anidados $G^2(M_j|M) = G^2(M_j) - G^2(M)$.

Otra forma de resolver el test (4.2.7) está basado en el estadístico de Wald dado por

$$W_j = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)} \underset{n \rightarrow \infty}{\overset{H_0}{\rightsquigarrow}} \chi_1^2,$$

con $\hat{\sigma}(\hat{\beta}_j)$ el error estándar de los parámetros estimados $\hat{\beta}_j$.

Ambos estadísticos tienen una distribución asintótica chi-cuadrada con un grado de libertad. Así, la hipótesis nula se rechaza a un nivel de significación α (el parámetro β_j se dice significativo) cuando el p -valor asociado es menor o igual a α .

El test de significancia de los parámetros (test de razón de verosimilitud condicional o test de Wald) se usa en el método paso a paso (adelante y hacia atrás) para la selección de las mejores variables que explican la respuesta. El paso cero es para ajustar un modelo inicial que es usualmente la constante del modelo. En cada paso sucesivo, ajustamos todos aquellos modelos que resultan de la inclusión de cada una de las variables explicativas que no están

en el modelo seleccionado en el paso previo. Entonces, realizaremos la prueba de significancia que tiene como hipótesis nula el modelo seleccionado en el paso previo, y como hipótesis alternativa el modelo resultante de agregar cada nueva variable. De todas las variables para las cuales estos test son significativos (el p -valor es menor o igual que el nivel de significancia fijado previamente) se incluyen en el modelo aquellas variables asociadas con el mínimo p -valor. Se continúa la introducción de variables con este método hasta que ninguno de los test sea significativo.

Además, consideraremos en cada paso la posibilidad de remover algún parámetro del modelo seleccionado en el paso previo. Para quitar las variables, realizamos el test de significancia que tiene como hipótesis alternativa el modelo seleccionado en el paso previo, y como hipótesis nula el modelo resultante al quitar cada una de sus variables. De todas las variables para las cuales estos test no son significativos (el p -valor es mayor o igual que el nivel de significancia fijado para quitar la variable) se quitan del modelo aquellas variables asociadas con el máximo p -valor. La eliminación de variables continúa hasta que todos estos test sean significativos. Para evitar anular en un paso la variable que ha entrado justamente en el paso anterior, se fija para la eliminación de variables un nivel de significancia mayor que el fijado para la inclusión de variables.

Por otra parte, en base a la distribución normal para muestras grandes de los estimadores de los parámetros del modelo logit, podemos construir intervalos de confianza (IC) para cada parámetro β_j con nivel de confianza $(1 - \alpha) \times 100$ dado por $\widehat{\beta}_j \pm z_{\alpha/2} \widehat{\sigma}(\widehat{\beta}_j)$.

Por lo tanto, un IC para una muestra grande de los odds ratio dados por la exponencial de cada parámetro es

$$\left(\exp \widehat{\beta}_j - z_{\alpha/2} \widehat{\sigma}(\widehat{\beta}_j), \exp \widehat{\beta}_j + z_{\alpha/2} \widehat{\sigma}(\widehat{\beta}_j) \right).$$

Para medir la exactitud de la predicción del modelo logístico, se usa con frecuencia la razón de clasificaciones correctas (RCC) definida como la razón de observaciones clasificadas correctamente por el modelo logístico. El problema es que esta medida depende del punto de corte usado para clasificar a los individuos tomando en cuenta que predecimos valores $Y = 1$ para cualquier individuo con probabilidad predicha mayor o igual que el punto de corte y $Y = 0$ en otro caso. Debido a esto es mejor usar el área bajo la curva ROC (receiver operating characteristic) para medir la capacidad del modelo para clasificar correctamente. Esta es una representación gráfica de la sensibilidad (razón de verdaderos positivos) y 1-especificidad (razón de falsos positivos), para un clasificador binario (el modelo logit) cuando el umbral de la discriminación (punto de corte) es variado.

Finalmente, un análisis adicional del diagnóstico es necesario para describir la naturaleza de cualquier falta de ajuste que podría ser debido a una elección inapropiada de la función de enlace (transformación logit) o no linealidad en los efectos de las variables explicativas. La comparación de residuos observados y frecuencias ajustadas también se usan comúnmente. Hay dos tipos principales de residuos: residuos de Pearson y devianza calculados como la raíz cuadrada de las componentes de la bondad de ajuste estadístico asociado. Cada residuo de Pearson compara la diferencia entre una frecuencia observada y su valor fijado para la desviación estándar estimada de las frecuencias observadas. Para la i -ésima variable explicativa x_i fijada, el residuo de Pearson está dado por

$$r_i = \frac{y_i - n_i \hat{\pi}_i}{[n_i \hat{\pi}_i (1 - \hat{\pi}_i)]^{1/2}}.$$

Si n_i es grande r_i tomará una distribución aproximada normal estándar. Esto significa que valores absolutos más grandes que 2 indican posible falta de ajuste. En otro caso, podemos usar residuos ajustados que tengan una distribución normal estándar asintótica (ver Agresti, 2009, para una discusión detallada de estos residuos).

Los resultados asintóticos para modelos logit se aplican cuando el número n de diferentes niveles de las variables explicativas son fijos, y el número de observaciones n_i hechas para cada nivel es bastante grande. Estos resultados teóricos se aplican más naturalmente a modelos logit para respuesta categórica, como es el caso de datos de dengue analizados en este trabajo.

4.2.3. Aplicación a datos de dengue.

Como se discutió en la introducción, el objetivo del presente trabajo es discriminar entre DC y DH y determinar los síntomas asociados con mayor probabilidad a cada tipo de dengue.

La base de datos consiste de 2215 casos confirmados de fiebre por dengue en el estado Mexicano de Guerrero. De los casos considerados el 54.5 % son mujeres y el 75.9 % están afectados de DC. Este conjunto de datos fue analizado primeramente en un estudio de consenso donde se usaron índices tipo kappa para medir el acuerdo en la diagnosis de los médicos de atención primaria y el diagnóstico exacto dado por el laboratorio. Las conclusiones fueron que el acuerdo en el diagnóstico de DC era bajo, mientras que para el DH era alto (Allende et al. 2008). Para cada sujeto tenemos información adicional sobre diferentes variables observadas por los médicos de atención primaria en pacientes sospechosos de la enfermedad. La mayoría de estas variables corresponden a los síntomas que fueron observados en cuadro típico de dengue y son descritos en el manual de dengue.

Los síntomas se representan en el modelo estadístico por variables categóricas binarias que tienen valor 1 si los pacientes tienen el síntoma y valor 0 en otro caso. Las variables explicativas consideradas en este estudio fueron: temperatura corporal, edad, sexo, contacto del paciente con otros vectores, contacto con pacientes similares, transfusión sanguínea, fiebre, cefalea, mialgias, artralgias, dolor retrocular, exantema, prurito, vómito, náuseas, escalofrío, fotofobia, dolor abdominal, diarrea, conjuntivitis, congestión nasal, tos, faringitis, rinitis, hepatomegalia, esplenomegalia, escape de líquidos, hemorragias, petequias, equimosis, hematomas, torniquete positivo, ascitis, derrame pleural, gingival, epistaxis, hematemesis y melena.

Tomando en cuenta las características de las variables involucradas, consideramos un modelo de regresión logística para predecir la variable respuesta Y ($Y = 1$ para pacientes con DC y $Y = 0$ para pacientes con DH) de un conjunto de 38 variables predictoras observadas para cada paciente, de las cuales 36 son categóricas. Para seleccionar las variables significativas que explican el tipo de dengue (DC o DH) usamos la selección por pasos basada en el test de razón de verosimilitud condicional con niveles de significancia de entrada y salida de variables de 0.05 y 0.1, respectivamente. Así, seleccionamos un modelo logit con 13 variables. Las probabilidades de sufrir DC predichas por este modelo logit están dadas por

$$\widehat{p}(x) = \exp\left(\frac{\exp_{j=1}^{13} \widehat{\beta}_j x_j}{1 + \exp_{j=1}^{13} \widehat{\beta}_j x_j}\right), \quad (4.2.9)$$

con $x = (x_1, \dots, x_{13})$ los valores observados de la variable predictora seleccionada X_j que se muestra en la tabla (4.12). Todas las variables seleccionadas, excepto la primera, representan síntomas y están definidas por

$$x_j = \begin{cases} 1, & \text{si el paciente tiene el síntoma } X_j; \\ 0, & \text{si el paciente no tiene el síntoma } X_j. \end{cases}$$

La primera variable (sexo) está dada por

$$x_1 = \begin{cases} 1, & \text{si el paciente es hombre;} \\ 0, & \text{si el paciente es mujer.} \end{cases}$$

El p -valor asociado al test de bondad de ajuste de Hosmer-Lemeshow para el modelo ajustado es 0.151, mayor que el nivel de significancia $\alpha=0.05$. Esto significa que el modelo ajusta bien los datos. Con respecto al análisis de residuos, solamente el 3% son significativos y en todos los casos la medida de influencia asociada puede ser considerada nula.

Cuadro 4.12: Estimación de los parámetros del modelo e IC al 95 %.

Nombre	Variable	β	E.T.	Wald	df	p -valor	$Exp(B)$	LI	LS
Sexo(1)	X_1	-0.292	0.125	5.402	1	0.020	0.747	0.584	0.955
EnfSimil(1)	X_2	0.584	0.127	21.212	1	0.000	1.794	1.399	2.300
Cefalea(1)	X_3	0.963	0.273	12.423	1	0.000	2.620	1.533	4.475
Artralgias(1)	X_4	-0.663	0.206	10.344	1	0.001	0.515	0.344	0.772
Vómito(1)	X_5	-0.456	0.146	9.790	1	0.002	0.634	0.476	0.843
Nauseas(1)	X_6	-0.357	0.162	4.883	1	0.027	0.700	0.510	0.960
DolorAbdominal(1)	X_7	-0.628	0.140	20.143	1	0.000	0.534	0.406	0.702
Diarrea(1)	X_8	-0.380	0.151	6.344	1	0.012	0.684	0.509	0.919
Conjuntivitis(1)	X_9	0.452	0.172	6.902	1	0.009	1.572	1.122	2.203
EscapeLiquidos(1)	X_{10}	-1.195	0.261	20.943	1	0.000	0.303	0.182	0.505
Hemorragias(1)	X_{11}	-0.971	0.195	24.803	1	0.000	0.379	0.259	0.555
Petequias(1)	X_{12}	-0.926	0.254	13.301	1	0.000	0.396	0.241	0.652
Melena(1)	X_{13}	-1.400	0.578	5.874	1	0.015	0.247	0.080	0.765
Constante	X_0	1.778	0.283	39.513	1	0.000	5.917		

La estimación de los parámetros del modelo ajustado junto con sus errores estándar, el p -valor asociado al estadístico de Wald para probar la significancia de los parámetros, la exponencial de los parámetros y los intervalos de confianza al 95 % se muestran en la tabla (4.12). Del test de Wald podemos concluir que todos los parámetros son estadísticamente significativos al nivel $\alpha=0.05$. Esto es equivalente al hecho de que la unidad no está en el rango del IC del 95 % asociado con la exponencial de cada parámetro.

En relación a la interpretación de los parámetros, los síntomas con parámetro positivo están asociados con una mayor probabilidad de DC y aquellos con parámetro negativo al DH. Esto significa que hay una fuerte evidencia de asociación positiva entre DC y dolor de cabeza, contacto con pacientes similares y conjuntivitis. Por otra parte, el DH está positivamente con los otros síntomas significativos: escape de líquidos, melena, hemorragia, petequias, artralgias, dolor abdominal, vomito, diarrea y nausea. Los otros síntomas que no están en el modelo no permiten distinguir entre DC y DH debido a que son igualmente sufridos por los dos tipos de pacientes. Con respecto al sexo, la probabilidad de sufrir DC es mayor para mujeres que para hombres. Observe que cada odds ratio derivado de este modelo logit (exponencial de cada parámetro) se define como la razón de odds de DC para individuos que tienen el síntoma en cuestión y por aquellos que no lo tienen. La interpretación en términos de odds ratio de las exponenciales estimadas para cada parámetro puede verse en la tabla (4.13).

Para predecir el tipo de dengue que padece un paciente hemos designado un test de diagnóstico basado en un punto de corte óptimo. En la tabla (4.14) podemos ver la clasificación con la razón de verdaderos positivos, la razón de verdaderos negativos y la razón de clasificaciones correctas para dos diferentes

Cuadro 4.13: Interpretación de las exponenciales de los parámetros.

Variable	$\exp(\beta)$	Interpretación
Sexo(1)	0.747	El odds del DC es 1.3 veces mayor (1/0.747) para mujeres que para hombres.
EnfSimil(1)	1.794	El odds del DC es aproximadamente el doble para los pacientes que tuvieron contacto con otros pacientes similares.
Cefalea(1)	2.620	El odds del DC es aproximadamente tres veces mayor para los pacientes con dolor de cabeza.
Artralgias(1)	0.515	El odds del DH es de aproximadamente el doble (1/0.515) para pacientes con artralgias.
Vómito(1)	0.614	El odds del DH es 1.6 (1/0.614) para los pacientes con vómito.
Nauseas(1)	0.700	El odds del DH es 1.4 veces mayor (1/0.700) para los pacientes con nauseas.
DolorAbdominal(1)	0.534	El odds del DH es de aproximadamente el doble para pacientes con dolor abdominal.
Diarrea(1)	0.684	El odds del DH es 1.5 veces mayor para pacientes con diarrea.
Conjuntivitis(1)	1.572	El odds del DC es 1.5 veces mayor para pacientes con conjuntivitis.
EscapeLiquidos(1)	0.303	El odds del DH es aproximadamente tres veces mayor (1/0.303=3.3) para pacientes con escape de líquidos.
Hemorragias(1)	0.379	El odds del DH es aproximadamente tres veces mayor (1/0.379=2.6) para pacientes con hemorragia.
Petequias(1)	0.396	El odds del DH es 2.5 veces mayor (1/0.396=2.5) para pacientes con petequias.
Melena(1)	0.247	El odds del DH es aproximadamente cuatro veces mayor (1/0.247=4) para pacientes con melena.
Constante	5.917	El odds que una mujer que no sufra ningún síntoma de DC en lugar de DH es 5.917. Como un resultado, la probabilidad de que una mujer sufra DC sin padecer ningún síntoma es de 0.855

puntos de corte: 0.5 y 0.7. El punto de corte 0.7 es aproximadamente la proporción de pacientes con DC en la muestra dada la mejor clasificación con razones de sensibilidad y especificidad mas balanceadas.

Observe que el porcentaje de pacientes con DC correctamente clasificados es bastante alto (84.3%). Esto resalta la utilidad del modelo para diagnosticar correctamente pacientes con DC en términos de los síntomas y sexo.

Mediante un ejemplo, la tabla (4.15) presenta las probabilidades predichas de sufrir DC por pacientes afectados de diferentes combinaciones de síntomas. Observe que la probabilidad de DC para una paciente mujer que tiene dolor de cabeza, conjuntivitis y ha estado en contacto con pacientes similares es 0.9776. Por otra parte la probabilidad de DC para un paciente hombre con dolor abdominal, hemorragia y escape de líquidos es 0.2128, lo cual nos permite inferir que está infectado de DH.

Cuadro 4.14: Tabla de clasificaciones(a).

	Predicción con punto de corte 0.5			Predicción con punto de corte 0.7		
Observados	DC	DH	% correcto	DC	DH	% correcto
DC	1310	65	95.3	1159	216	84.3
DH	280	191	40.6	186	285	60.5
% global			81.3			78.2

Cuadro 4.15: Probabilidades predichas de sufrir DC para diferentes combinaciones de síntomas significativos y sexo.

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	$\hat{p}(DC)$
0	1	1	0	0	0	0	0	1	0	0	0	0	0.9776
1	1	1	0	0	0	0	0	1	0	0	0	0	0.9703
0	0	1	0	0	0	0	0	0	0	0	0	0	0.9394
1	0	1	0	0	0	0	0	0	0	0	0	0	0.9205
0	0	0	0	0	0	0	0	0	0	0	0	0	0.8554
1	0	0	0	0	0	0	0	0	0	0	0	0	0.8155
0	0	1	0	0	0	0	0	0	1	1	0	0	0.6399
1	0	1	0	0	0	0	0	0	1	1	0	0	0.5703
0	0	0	0	0	0	0	0	0	1	1	0	0	0.4042
1	0	0	0	0	0	0	0	0	1	1	0	0	0.3363
0	0	0	0	0	0	1	0	0	1	1	0	0	0.2658
1	0	0	0	0	0	1	0	0	1	1	0	0	0.2128
0	1	1	1	1	1	1	1	1	1	1	1	1	0.0392
1	1	1	1	1	1	1	1	1	1	1	1	1	0.0289
0	0	0	1	1	1	1	1	0	1	1	1	1	0.0055
1	0	0	1	1	1	1	1	0	1	1	1	1	0.0041

4.2.4. Conclusiones

Un modelo logit con variables categóricas ha sido aplicado en este estudio para predecir el tipo de dengue (DC o DH) a partir de los síntomas y otras variables relacionadas como sexo, temperatura corporal y edad. De las 38 variables inicialmente consideradas, se seleccionaron solo 13 como significativas para diferenciar entre DC y DH. Entre ellas, ser mujer, tener contacto con pacientes similares, sufrir dolor de cabeza y conjuntivitis están asociadas con alta probabilidad de DC, mientras que ser hombre y tener los síntomas escape de líquidos, melena, hemorragias, petequias, artralgias, dolor abdominal, vomito, diarrea y nauseas están positivamente asociadas a DH.

El modelo logit ajustado proporciona al médico un test de diagnóstico basado en los síntomas para predecir el tipo de dengue con una razón de verdaderos positivos (sensibilidad) mucho mayor que la basada en la habilidad diagnóstica del médico y las recomendaciones proporcionadas en el manual de dengue.

Bibliografía

- [1] Aburas H.M., Cetiner B.G., Sari M. 2010. Dengue confirmed-cases prediction: A neural network model. *Expert Systems with Applications* 37: 42564260.
- [2] Aguilera A.M. 2005. Modeling of Multidimensional Contingency Tables. Colección Cuadernos de Estadística, 33. Hespérides-La Muralla, Salamanca (in Spanish).
- [3] Agresti A. 1990. *Categorical Data Analysis*. Wiley, New York.
- [4] Allende S., Bouza C., López-Silva S., Moreno A. S., Soberón y Parra G. 2008. Analysis of the effectiveness of the diagnosis of dengue in the state of Guerrero , México 2007: statistical facts and consensus. *Revista Investigación Operacional* 29 (1): 35-47 (in Spanish).
- [5] Arcari P., Tapper N., Pfueller S. 2007. Regional variability in relationships between climate and dengue/DHF in Indonesia. *Singaporean Journal of Tropical Geography*. 28(3): 251-272.
- [6] Secretaría de Salud de México. 2006. Manual for the Surveillance, Diagnosis, Prevention and Control of Dengue (in Spanish).
- [7] Beasley C.M. 2008. Evidence of secondary dengue infection and potential for association with clinical depression in a subset of a randomly selected cohort in South Texas. *Texas Medical Center Dissertations* (via ProQuest). Paper AAI1454370.
- [8] Chadwick D., Arch B., Wilder-Smith A., Paton N. 2008. Distinguishing dengue fever from other infections on the basis of simple clinical and laboratory features: application of logistic regression analysis. *J. Clin. Virol.* 35(2): 147-53.
- [9] Díaz F.A., Martínez R.A., Villar L.A. 2006. Clinical criteria to diagnose dengue in the first days of illness. *Biomédica* 26: 22-30.
- [10] Díaz-Quijano F.A., Villar-Centeno L.A., Martínez-Vega R.A. 2006. Complications associated to severe thrombocytopenia in patients with dengue. *Rev Méd Chile* 134: 167-173.
- [11] Egger J.R., Coleman P.G. 2007. Age and Clinical Dengue Illness. *Emerg. Infect. Dis.* 13(6): 924927.
- [12] Gubler D. J. 1996. Serologic diagnosis of dengue/dengue haemorrhagic fever. *Dengue Bull.* 20.

- [13] Gubler D. J. 1998. Dengue and dengue hemorrhagic fever. *Clin. Microbiol. Rev.* 11: 480-496.
- [14] Gubler D.J. 2002. How effectively is epidemiological surveillance used for dengue programme planning and epidemic response. *Dengue Bulletin* 26.
- [15] Gibbons R. V., Vaughn D. W. 2002. Dengue: an escalating problem. *BMJ* 324:1563-1566.
- [16] Hosmer D., Lemeshow S 1989. *Applied Logistic Regression*. Wiley, New York.
- [17] Nor Azura H., Naomie S., Rahman A. 2006. Simulation of dengue outbreak prediction. *Proceedings of the Postgraduate Annual Research Seminar 2006 of the University Technology Malaysia*: 374-379.
- [18] Schreiber K.V. 2001. An investigation of relationships between climate and dengue using a water budgeting technique. *International Journal of Biometeorology* 45: 8189.
- [19] Shu P-Y, Huang, J-H. 2004 Current advances in dengue diagnosis. *Clin. Diagn. Lab. Immunol.* 11(4): 642650.
- [20] WHO. 1985. *Viral haemorrhagic fevers: general research needed and recommendations for viral haemorrhagic fevers*. World Health Organization.
- [21] Yang M.H., Ferreira C.D. 2008. Assessing the effects of vector control on dengue transmission. *Journal of Applied Mathematics and Computation* 198: 401413.

4.3. Un modelo para la transmisión vertical de dengue

Villavicencio-Pulido J. G., Castro-Salgado. V., Nila-Luévano C.
Unidad Académica de Matemáticas de la UAG-Acapulco

Resumen: Zhilan Feng, Jorge X. Velasco-Hernández [1] proponen un modelo de ecuaciones diferenciales ordinarias para la dinámica poblacional de la fiebre del Dengue. El modelo propuesto por Zhilan Feng, Jorge X. Velasco-Hernández [1] se ha tomado como base de nuestro artículo para analizar la dinámica del Dengue. Además de analizar el caso en el que se tienen dos cepas patógenas y dos poblaciones de individuos infectados con las dos cepas, se analiza el caso en los cuales aparecen infecciones verticales en las poblaciones de humanos y en las poblaciones de vectores [2, 3, 4, 5, 6, 7, 8]. Finalmente se argumentan los casos en los que se permite la coexistencia de las cepas estudiadas o la exclusión de alguna de las cepas que se analizan en el modelo.

Key words: Epidemiología Matemática, Dengue, Ecuaciones Diferenciales, Enfermedades transmitidas por vector, Dinámica Poblacional.

4.3.1. Introducción

La fiebre de dengue es una enfermedad viral endémica en muchas áreas del mundo que está invadiendo y recolonizando regiones donde ya sea que estuviera ausente o que haya sido erradicada. El virus de dengue tiene 4 diferentes serotipos. Construimos y analizamos un modelo matemático para esta dinámica de transmisión. El modelo es un sistema de ecuaciones diferenciales que incorpora tamaño poblacional variable tanto en el mosquito como en el huésped, dos cepas co-circulantes y tasas de picadura dependientes de la frecuencia. El modelo consituye un marco para la discusión de condiciones para la coexistencia o exclusión competitiva de cepas de patógenos relacionadas. En la siguiente sección damos una reseña de la enfermedad resumiendo su importancia epidemiológica así como las principales características incorporadas en nuestro modelo. Luego procedemos con la formulación del modelo y discutimos trabajo previo en el principio de exclusión competitiva. Esta sección está seguida de un análisis del modelo en ausencia de mortalidad inducida por la enfermedad (virulencia negligible). Luego, discutimos los resultados computacionales para el caso de virulencia no negligible y, finalmente, en la última sección, comparamos los resultados obtenidos con otros modelos de

superinfección, tamaño de población variable y tasas de infección dependientes de la frecuencia.

La fiebre de Dengue

En las últimas décadas se ha hecho énfasis en el análisis de las enfermedades comunicables, particularmente en la fiebre del dengue y sus formas severas de la enfermedad tal como la fiebre hemorrágica del dengue la cual puede llegar a ser muy severa y en algunos casos, fatales. La fiebre de dengue clásico es generalmente observada en niños mayores y en adultos y se caracteriza por un brote repentino de fiebre, dolor de cabeza frontal, náusea, vómito y otros síntomas. La enfermedad aguda dura de 3 a 7 días y es usualmente benigna. La forma hemorrágica de dengue y su síndrome de shock por dengue asociado (DHF-DSS) se observa principalmente en niños por debajo de los 15 años pero puede también ocurrir en adultos [16]. Se caracteriza por un brote agudo inicial de fiebre y una variedad de síntomas que duran de 2 a 7 días. Esta forma de dengue puede terminar en muerte del paciente.

Es importante mencionar que el dengue ha estado presente cerca de doscientos años durante los cuales esta enfermedad ha sido reconocida clínicamente, cabe mencionar que durante los siglos XVIII y XIX la propagación de la enfermedad se llevó a cabo de una manera muy lenta. En muchos casos la dinámica poblacional de individuos susceptibles y vectores tenía lugar en los barcos, donde se criaban cepas de *A. aegypti*, además de la población humana huésped, sin embargo, estos patrones cambiaron abruptamente después de la II guerra mundial, debido a que los virus del dengue fueron propagados por personal militar infectado a áreas donde no había dengue y los vectores por diferentes medios de transporte que servían como criaderos de dichas cepas. La diseminación de la enfermedad aumento después de la guerra por el rápido crecimiento de la población y la urbanización. Cambios en la epidemiología del dengue en la Posguerra ocurrieron en América algo más tarde que en Asia. El primer brote importante de fiebre de dengue en América ocurrió en Cuba en 1981 con 116 000 pacientes hospitalizados, 34 000 casos documentados de DHF y 158 defunciones. Durante la década que siguió al brote epidémico en Cuba, 14 países de América han reportado casos confirmados de DHF. Es importante notar que en América, la población urbana crecio casi al doble entre 1970 y 1990, y es en este periodo donde el dengue emergió con más fuerza.

En países en desarrollo, el crecimiento poblacional es un factor importante que contribuye al incremento en la incidencia de enfermedades comunicables que afectan principalmente a los habitantes de las zonas urbanas, con hijos y niños entre los grupos particularmente en riesgo. La urbanización y crecimien-

to poblacional incrementan la demanda de servicios básicos esenciales como casa, agua, etc., y al mismo tiempo inducen condiciones que incrementan la transmisión potencial de algunas enfermedades por vector. Insuficiencias en el abastecimiento de agua provocan el almacenamiento a gran escala de agua que genera un ambiente ideal para hábitats de crianza para mosquitos *Aedes*, los vectores de la fiebre de dengue, la fiebre hemorrágica de dengue y síndrome de shock por dengue así como la fiebre amarilla. Cambios en los hábitos alimenticios llevan al incremento del uso de comida enlatada y mayor uso de contenedores desechables que dan lugares de crianza para vectores de este tipo. En resumen, el crecimiento poblacional, la urbanización y la pobreza realzan la presencia y transmisión de las enfermedades infecciosas.

El dengue se produce por virus del género, *Togaviridae*, subgénero *Flavivirus*. Se han distinguido cuatro tipos distintos de virus de dengue, denotados por los tipos 1, 2, 3, 4. Los virus de dengue pueden infectar sólo un número restringido de vertebrados pero es esencialmente una enfermedad humana [17]. La infección por cualquier cepa de virus de dengue produce inmunidad a largo plazo pero sólo inmunidad temporal a los otros serotipos. Tres de los vectores son *Aedes aegypti* Linnaeus, *Aedes albopictus* Skuse, y *Aedes scutellaris* Walk. Los mosquitos *Aedes aegypti* adquieren la infección de individuos infectados de 6 a 18 horas antes del inicio de la fiebre y durante la duración de la fiebre. Se requiere un periodo mínimo de incubación extrínseca de 8 a 14 días luego de que consuman alimento de sangre infectada antes de que el mosquito se vuelva infeccioso. La infección en el vector es de por vida. El virus del dengue se transmite en dos ciclos: urbano y silvestre aunque, como se mencionó antes, es un virus predominantemente humano [17].

La transmisión de dengue ocurre a través del año en áreas endémicas tropicales pero existe, sin embargo, un patrón cíclico distinto asociado con la temporada de lluvias [16]. En particular, en Tailandia donde el ciclo de vida del vector es altamente domiciliario, la temperatura y las condiciones de humedad durante la temporada de lluvia favorecen la supervivencia de mosquitos infectados. En el continente Americano la situación es diferente ya que en estas áreas las larvas se desarrollan en el exterior. Aquí, los picos de transmisión toman lugar en los días de la estación con mayor lluvia y temperaturas más altas [16].

En regiones donde el mosquito y los humanos existen, una introducción del virus del dengue puede producir una epidemia dependiente de; a) la cepa de los virus (magnitud de influencia y duración de la viremia), b) la susceptibilidad de la población humana, c) la densidad, comportamiento, y la competencia de la población del mosquito vector, y d) la introducción del virus en un área donde tiene contacto con la población de mosquito local [16]. La severidad de la fiebre de dengue ha sido asociada con infecciones secundarias de dengue

aunque sus causas están lejos de ser explicadas. Los estudios epidemiológicos en Tailandia sugieren que un importante factor de riesgo para DHF-DSS es la presencia del anti-cuerpo de dengue preexistente al subneutralizar niveles. Además, la DHF-DSS endémica se encuentra en áreas donde las densidades de *Aedes aegypti* son altas y los virus de dengue de múltiples tipos son endémicos. Más aún, DHF-DSS está asociado con individuos mayores de un año de edad con una respuesta de anticuerpo secundario con infecciones primarias en bebés recién nacidos cuya madre era inmune al dengue [16,17]. Estos hechos llevan a la formulación de la hipótesis de infección secundaria o realce inmune para explicarla [16]. Esta hipótesis afirma que sólo aquellas personas que experimentan una segunda infección con serotipo de dengue heterólogo presentan DHF-DSS. En particular, se ha encontrado que sólo las infecciones secundarias de dengue-2-enfermedad está inmunológicamente realzada y la infección con este serotipo de virus causa la mayoría de casos de DSS [18]. Otros factores están también asociados con DHF-DSS. Estos son sexo (más frecuentemente infecciones en mujeres), estado nutricional (mayor incidencia en bebés bien alimentados o de clase socioeconómica media y alta), y el intervalo entre primera y segunda infección.

Por otro lado, se han realizado pocos estudios acerca de la infección por dengue durante el embarazo y la infección concomitante en el feto y el recién nacido.

En la literatura mundial encontramos pocos casos reportados, en 1981 en Cuba, de 59 mujeres estudiadas sólo en 4 RN se aisló la IgM específica, todos los niños fueron sanos y se les realizó seguimiento hasta los 5 años sin encontrar anomalías [4]. En 1988 en China, a 9 mujeres embarazadas infectadas por el virus del dengue, se les realizó amniocentesis y niveles de alfafetoproteína, sin encontrar anomalías, durante esta epidemia se demostró actividad anti-dengue en la leche materna [5]. Entre 1989-1990 en la Polinesia Francesa se encontraron 5 casos de RN con IgM (+), aislando el virus tanto en la madre como en el RN [6]. En 1991, en la epidemia ocurrida en Brasil, se estudiaron 10 mujeres embarazadas, todas con IgG (+), y 2 con IgM (+), en todos los RN se encontró IgG (+) con descenso progresivo de sus niveles entre los 3 y 8 meses de edad, todos fueron RN normales [7]. En 1993 en Tailandia, se reportó un caso de IgM (+) en el RN [3]. En 1996 en Malasia, de 2 casos reportados como transmisión vertical, uno presentó falla multi-orgánica y muerte a los 6 días de nacido, siendo el único caso reportado en la literatura de muerte por infección vertical por el virus del dengue, otros 2 casos son reportados en Malasia en 1996 (ver [2]) donde las madres que tuvieron dengue agudo 4 y 8 días antes del nacimiento de sus infantes.

4.3.2. Formulación de un modelo del Dengue con dos cepas y transmisión vertical en las poblaciones de humanos y vectores

En la literatura matemática se reportan modelos que analizan las distintas dinámicas de la fiebre del Dengue. En cada uno de ellos se describen estados o términos que describen distintos estados fenomenológicos, como pueden ser, los modelos costo-efectividad, modelos que incorporan un periodo de incubación o de latencia, otros describen la transmisión vertical en los vectores. Además existen modelos que discuten el problema de la coexistencia en interacciones huésped-patógeno [19, Levin y Pimentel], otros resultados postulan un principio de exclusión competitiva en una epidemia [10 Bremermann y Thieme], [20 Li, Castillo-Chavez, Huang], [22 Mena-Lorca, Velasco-Hernández y Castillo-Chavez].

El modelo presentado en este trabajo se basa en el modelo presentado por Zhilan Feng y Jorge X. Velasco-Hernández, en el cual se hace énfasis en el fenómeno de superinfección, además de que en este trabajo se analiza el efecto debido a las transmisiones verticales provenientes de las poblaciones de humanos y la de los vectores. En el modelo que analizamos aquí, consideramos tamaños poblacionales variables tanto de huésped como de vector, no incorporamos el compartimiento de expuestos, pero incluimos la existencia de una segunda cepa co-circulante que puede producir infecciones secundarias en aquellos individuos ya sean susceptibles o ya recuperados de una infección primaria con una cepa distinta.

En el dengue ciertas secuencias de infección parecen ser más dañinas para el paciente que otras. Modelamos este proceso usando un coeficiente de susceptibilidad (cf. [19]). Este coeficiente nos permite explorar grados variables de susceptibilidad a infecciones secundarias y su efecto en las dinámicas asintóticas de la enfermedad. A través de este modelo exploramos las consecuencias de emparejar dos poblaciones que difieren en el patrón de infección (*SIR* con superinfección en el huésped, *SI* en el vector), y el efecto de tasas de infección dependientes de la frecuencia sobre la coexistencia de cepas relacionadas de forma cercana.

La mayoría de las enfermedades se producen por un espectro de patógenos relacionados de cerca en lugar de una sola cepa y el dengue es claramente un ejemplo de esta afirmación. En el dengue ocurre un fenómeno análogo de superinfección ([23] Nowak y May, [21] May y Nowak). Una cepa invade la población de huéspedes, produce un breve periodo de inmunidad temporal a otras cepas pero cuando se pierde la inmunidad el huésped se vuelve susceptible de reinfección con otra cepa. En dengue antes de que pue-

da ocurrir la reinfección, hay un periodo donde el huésped es resistente, en grados variables, a todas las cepas [16]. En la fiebre de dengue estamos, entonces, confrontados con una enfermedad de transmisión por vector, cepas co-circulantes cierto grado de inmunidad cruzada o incluso susceptibilidad incrementada a la infección, y un tamaño de población del huésped variable. Bajo estas condiciones un problema teórico importante que analizamos aquí es aquel de la coexistencia de todas las cepas o la eventual extinción de alguna de ellas. Un problema similar se ha explorado teóricamente por varios autores [19, 10, 20, 5].

Ecuaciones del modelo

Considere una población humana denotada por $N(t)$, la cual es dividida en clases, asentada en una región donde una población de mosquito del género *Aedes* denotada por $T(t)$ está presente y es portadora del virus del dengue. Donde $N = S + I_1 + I_2 + Y_1 + Y_2 + R$ y $T = M + V_1 + V_2$ son los tamaños de población total del huésped y el vector respectivamente y están divididas en individuos infecciosos con dengue del tipo i , denotadas por I_i , individuos con infecciones secundarias denotados por Y_i , individuos recuperados, $R(t)$, vectores de provenientes de la cepa i , denotados por V_i y finalmente, los mosquitos susceptibles, denotados por $M(t)$, ($i = 1, 2$). Las tasas de las infecciones primarias en las poblaciones humanas huéspedes y en las poblaciones de vectores denotadas por $B_i = \frac{\beta_i V_i}{(c + \omega_h N)}$ y $A_i = \frac{\alpha_i (I_i + Y_i)}{(c + \omega_v N)}$ se describen a través de casos especiales de la respuesta funcional de Holling tipo II [14] y son también generalizaciones de las tasas de contacto del modelo de Ross-Macdonald para Malaria [9] y enfermedad de Chagas [25] (para $i = 1, 2$). (vea Tabla 1 para otras definiciones de parámetros y valores).

Las ecuaciones del modelo se plantean como sigue ($' = d/dt$):

$$\begin{aligned}
 S'(t) &= h - (B_1 + B_2)S - uS - (\eta I_1 + \eta I_2 + \eta Y_1 + \eta Y_2), \\
 I_1'(t) &= B_1 S - \sigma_2 B_2 I_1 - uI_1 + \eta I_1, \\
 I_2'(t) &= B_2 S - \sigma_1 B_1 I_2 - uI_2 + \eta I_2, \\
 Y_1'(t) &= \sigma_1 B_1 I_2 - (e_1 + u + r)Y_1 + \eta Y_1, \\
 Y_2'(t) &= \sigma_2 B_2 I_1 - (e_2 + u + r)Y_2 + \eta Y_2, \\
 R'(t) &= r(Y_1 + Y_2) - uR,
 \end{aligned} \tag{4.3.1}$$

y

$$\begin{aligned}
M'(t) &= q - (A_1 + A_2)M - \delta M - \theta(V_1 + V_2), \\
V_1'(t) &= A_1M - \delta V_1 + \theta V_1, \\
V_2'(t) &= A_2M - \delta V_2 + \theta V_2
\end{aligned}
\tag{4.3.2}$$

Tabla 2. Definiciones de parámetros y valores usados en las simulaciones usadas en las figuras.

Símbolo	Definición del parámetro	valor
h	Tasa de reclutamiento del huésped	variable
u^{-1}	Esperanza de vida del huésped	70 años
r^{-1}	Periodo de infecciosidad del huésped	14 días
α_i	Tasa per cápita de infecciosidad del vector (tasa de picadura \times probabilidad de infección del vector)	(0, 0.05)
β_i	Tasa per cápita de infección del huésped (tasa de picadura \times probabilidad de infección del huésped)	(0, 0.05)
q	Tasa de reclutamiento del vector	variable
δ^{-1}	Esperanza de vida del vector	14 días
c	Parámetro de reescalamiento (α_i/c y β_i/c tasa de infección cuando N es pequeña)	1
ω_i	Parámetro de saturación (α_i/ω_i and β_i/ω_i teniendo en cuenta la tasa infecciosa máxima)	0.5
e_i	Tasa per cápita de muerte inducida por la enfermedad	variable
σ_i	Índice de susceptibilidad a la cepa i	(0,5)
η	Tasa a la que nacen nuevos infecciosos en la población humana huésped i	variable
θ	Tasa a la que nacen nuevos vectores infecciosos	variable

Supóngase que una vez que el mosquito es infectado nunca se recupera y no puede ser re-infectado con una cepa distinta de virus. Las infecciones secundarias, entonces, pueden tener lugar sólo en el huésped. Pueden ocurrir dos casos: ya sea que previamente los individuos I_1 sean infectados por la cepa 2, a través de contacto con mosquitos infectados V_2 , convirtiéndose en huésped Y_2 o, previamente, I_2 individuos están infectados con la cepa 1, a través del contacto con mosquitos V_1 , para convertirse en Y_1 huéspedes infectados, a las tasas $\sigma_1 B_1 I_2$ y $\sigma_2 B_2 I_1$, respectivamente. Aquí σ_i es un número real positivo que puede imitar ya sea inmunidad cruzada ($\sigma_i < 1$) o susceptibilidad incrementada ($\sigma_i > 1$) por realce inmune. Este tipo de dinámica es análoga a la

super-infección (cf.[23]). En el dengue, la inmunidad desarrollada luego de la infección es un factor que no aparece en los modelos de super-infección. En el dengue, cualquiera de las poblaciones de infectados primarios puede ser re-infectada con otra cepa. Los resultados generales sobre los efectos de la inmunidad cruzada en modelos *SIS* y *SI* respectivamente [23, 22], indican que es posible, para ciertos valores de σ , la coexistencia de cepas competidoras. Como mostraremos, la existencia de inmunidad cruzada junto con la inducción de inmunidad permanente específica y tasas de contacto dependientes de la frecuencia, previenen la coexistencia. El resultado genérico de nuestro modelo es exclusión competitiva aunque, en algunos casos, en una escala de tiempo muy larga.

Para resumir, si $\sigma_i < 1$, las infecciones primarias otorgan inmunidad parcial a la cadena i ; si $\sigma_i = 1$ las infecciones secundarias con cepa i toman lugar como si fueran infecciones primarias, y si $\sigma_i > 1$ las infecciones primarias incrementan la susceptibilidad a la cadena i . Una vez que un individuo ha sufrido de ambas infecciones obtiene inmunidad a ambas cepas a la tasa r independientemente de la sucesión de infecciones. Obsérvese que la población humana huésped alcanza el equilibrio $N^* = \frac{h}{\mu}$ y la población total de vectores alcanza el equilibrio $T^* = \frac{q}{\delta}$ cuando $t \rightarrow \infty$.

El equilibrio libre de enfermedad y R_i , Equilibrios endémicos (exclusión competitiva)

Estamos interesados en las condiciones que garantizan la permanencia del dengue como una enfermedad endémica. En esta sección se dan condiciones sobre la existencia y estabilidad de puntos de equilibrios del sistema del modelo. Sin embargo es importante mencionar que solamente se encontraron de manera explícita el equilibrio libre de enfermedad, el cual existe para todos los valores de los parámetros, y equilibrios de frontera (donde sólo una cepa está presente), y se conjetura la existencia de un equilibrio de coexistencia.

El equilibrio libre de enfermedad y R_i : En primer lugar analizaremos las condiciones que se tienen que satisfacer para garantizar que no se propague una epidemia, además se darán condiciones sobre el parámetro umbral R_i que se definirá mas adelante.

Sea

$$\begin{aligned} \Omega = \{ & (S, I_1, I_2, Y_1, Y_2, R, V_1, V_2) : \\ & S + I_1 + I_2 + Y_1 + Y_2 + R \leq h/u, \\ & V_1 + V_2 \leq q/\delta \}, \end{aligned}$$

el conjunto acotado por el total de la población huésped y la población del vector en la ausencia de enfermedad. Podemos inmediatamente identificar que el equilibrio libre de enfermedad existe para todos los valores de los parámetros el cual está dado por $E_0^* = (S^*, 0, 0, 0, 0, 0, 0, 0)$, donde $S^* = \frac{h}{\mu}$.

Lema 4.1. *Para el modelo (1-2) el equilibrio libre de enfermedad es localmente asintóticamente estable si y sólo si $\mu > \eta$ y $\delta > \theta$.*

Demostración.

Los valores propios asociados al sistema están dados por; $\lambda_1 = -\mu + \eta$, $\lambda_2 = -\mu$, $\lambda_3 = \delta - \theta$, $\lambda_4 = -\mu + \eta - (r + e_1)$, $\lambda_5 = -\mu + \eta - (r + e_2)$.

De aquí se sigue que el equilibrio libre enfermedad es localmente asintóticamente estable si las tasas de nacimiento y muerte en las poblaciones de humanos y vectores es mayor que la tasa de nacimientos infecciosos en cada una de las clases.

Desafortunadamente el análisis del equilibrio libre de enfermedad del sistema no proporcionó información sobre el número reproductivo básico, el cual mide el número de infecciones secundarias que un individuo infeccioso es capaz de producir al ser introducido en una población totalmente susceptible durante su periodo de infecciosidad. Por esta razón es fundamental construir la matriz de nueva generación asociada al sistema y de esta manera obtener el parámetro de control R_i , el cual proporcionará un criterio de invasión para la dispersión inicial del virus en una población totalmente susceptible al patógeno,

Diekmann y Heesterbeek [15] presentan una definición precisa del número reproductivo básico para un modelo compartamental de transmisión de la enfermedad. En este caso se definirá la matriz de nueva generación que presenta Feng y Velasco Hernández [1].

Para encontrar el número reproductivo básico en nuestro modelo igualamos

(1-2) a cero y se re-escibe como

$$\begin{aligned}
S &= \frac{h - \eta(I_i + Y_i)}{u + B_1 + B_2}, \\
I_i &= \frac{B_i S}{u - \eta + \sigma_j B_j}, \quad i \neq j, \quad i, j = 1, 2, \\
Y_i &= \frac{\sigma_i B_i I_j}{u - \eta + e_j + r}, \\
R &= \frac{r(Y_1 + Y_2)}{u}, \\
V_1 &= \frac{qA_1}{\delta - \eta + A_1 + A_2}, \\
V_2 &= \frac{qA_2}{\delta - \eta + A_1 + A_2}.
\end{aligned}$$

Sea $K = (B_1, B_2, A_1, A_2)'$ (' denota transpuesta). Sustituyendo las expresiones de arriba en la definición de B_i y A_i obtenemos un sistema de ecuaciones para cuatro ecuaciones algebraicas no-lineales en términos de B_i y A_i . Denotamos este sistema por $\Phi(K)$. Por construcción, la matriz de nueva generación es el Jacobiano de Φ evaluado en el equilibrio libre de enfermedad (dado por $B_1 = B_2 = A_1 = A_2 = 0$);

$$D\Phi(0) = \begin{pmatrix} 0 & 0 & \frac{\beta_1 q}{\delta(\delta - \theta)(c + h\omega_h/u)} & 0 \\ 0 & 0 & 0 & \frac{\beta_2 q}{\delta(\delta - \theta)(c + h\omega_h/u)} \\ \frac{\alpha_1 h}{\mu(\mu - \eta)(c + h\omega_v/u)} & 0 & 0 & 0 \\ 0 & \frac{\alpha_2 h}{\mu(\mu - \eta)(c + h\omega_v/u)} & 0 & 0 \end{pmatrix}.$$

El número reproductivo básico está dado por

$$\mathcal{R}_0 = \max\{\sqrt{R_1}, \sqrt{R_2}\},$$

con

$$R_i = \frac{\alpha_i \beta_i h q}{\mu \delta (\mu - \eta) (\delta - \theta) (c + h\omega_h/\mu) (c + h\omega_v/\mu)}.$$

Es importante observar que para que R_i sea positiva y que implique la estabilidad asintótica del equilibrio libre de enfermedad se necesita que $(\mu - \eta)(\delta - \theta) > 0$, y particularmente que $\mu > \eta$ y $\delta > \theta$. Esta fórmula es una generalización del número reproductivo básico de Ross-Macdonald para el caso en el que existen diferentes cepas, tasas denso dependientes y tamaño poblacional variable en las poblaciones de los huéspedes y los vectores.

Se sigue entonces que si $\mathcal{R}_0 > 1$, entonces la enfermedad puede invadir la población de huéspedes. De otro modo, si $\mathcal{R}_0 \leq 1$ el virus eventualmente desaparece de la población de huéspedes (resultado local).

Equilibrios endémicos (exclusión competitiva): Es de interés fundamental encontrar condiciones sobre la existencia de equilibrios endémicos los cuales de existir garantizan la permanencia del dengue como una enfermedad endémica. Su existencia está determinada por la magnitud relativa del número reproductivo básico de cada cepa.

Las soluciones a $K = \Phi(K)$ dan, por construcción, todos los puntos de equilibrio de (1-2). De manera explícita se encuentran dos puntos fijos de $\Phi(K) = K$ los cuales son dos equilibrios de frontera y están dados por;

$$E_1^* = (S_1^*, I_1^*, 0, 0, 0, 0, V_1^*, 0),$$

donde sólo la cepa 1 sobrevive, y

$$E_2^* = (S_2^*, 0, I_2^*, 0, 0, 0, 0, V_2^*),$$

donde sólo la cepa 2 sobrevive.

Obsérvese que los equilibrios encontrados están en la frontera de Ω (donde sólo una cepa está presente).

Equilibrio de frontera

Siempre que $R_j < 1$ entonces B_j y A_j ambos se hacen cero. Entonces $R_i > 1, R_j < 1, i \neq j, i, j = 1, 2$, implica la existencia de E_i^* la cual es además un equilibrio endémico único (donde sólo una cepa está presente). Bajo estas condiciones la dinámica asintótica del sistema (1-2) está completamente contenida en una variedad invariante $\Omega_i \subset \Omega$ definida como

$$\Omega_i = \{(S_i, I_i, V_i) : 0 < S_i + I_i \leq h/u, 0 < V_i \leq q/\delta\}.$$

Para $V_j = I_j = Y_j = R = 0$ (en Ω_i) nuestro modelo está descrito por las ecuaciones

$$\begin{aligned} S'(t) &= h - B_i S - uS - (\eta I_i), \\ I_i'(t) &= B_i S - uI_i + \eta I_i, \\ V_i'(t) &= \frac{\alpha_i I_i}{c + \omega_v N} (q/\delta - V_i) - \delta V_i + \theta V_i. \end{aligned} \tag{4.3.3}$$

En ausencia de infecciosos la población huésped está dada por $N(t) = \frac{h}{\mu}$ (en Ω_i). En (3) el único equilibrio (no trivial) es la proyección de E_i^* en

el subespacio Ω_i . Mas aún, la virulencia (mortalidad extra inducida por el virus) no juega un rol en este sistema asintótico.

Ya que $N(t) \rightarrow h/u$ cuando $t \rightarrow \infty$, (3) puede reducirse más adelante a las ecuaciones

$$I_i'(t) = \frac{\beta_i V_i}{c + \omega_h h/u} (h/u - I_i) - u I_i + \eta I_i,$$

$$V_i'(t) = \frac{\alpha_i I_i}{c + \omega_v h/u} (q/\delta - V_i) - \delta V_i + \theta V_i,$$

que son equivalentes al modelo de Ross Macdonald para malaria (Aron y May [9]). Entonces, siempre que $R_i > 1$, este sistema límite es globalmente asintóticamente estable en Ω_i , es decir los valores propios asociados al sistema tiene parte real negativa si y sólo si $R_i > 1$.

Equilibrio de frontera

Se hará análisis sobre la existencia y las propiedades de estabilidad de los puntos de equilibrio, asumimos que el dengue produce mortalidad significativa ($e_j \neq 0$) esto es debido a que existen muchas regiones donde el dengue es una enfermedad endémica. Feng y Velasco-Hernández obtienen resultados sobre la estabilidad suponiendo que el dengue no produce mortalidad significativa ($e_j = 0$). Obsérvese que la suposición no es justificable en todas las regiones donde el dengue es endémico.

Para reducir el sistema en una dimensión se supondrá que no hay virulencia en la población huésped. De aquí se sigue que en ausencia de virulencia, la población huésped es asintóticamente constante y entonces nuestro modelo se puede reducir en una dimensión.

Para facilitar el análisis e interpretación de nuestros resultados calculamos la matriz Jacobiana de (1-2) reordenando como se indica en (4).

Considere el equilibrio de frontera

$$E_1^* = (V_1^*, I_1^*, V_2^*, I_2^*, Y_1^*, Y_2^*, R^*), \quad (4.3.4)$$

donde

$$Y_1^* = Y_2^* = R^* = 0,$$

y

$$V_i^* = \frac{\mu \delta (\mu - \eta) (\delta - \theta) (c + h\omega_h/\mu) (c + h\omega_v/\mu) (R_1 - 1)}{(N\omega_v(\mu - \eta)(\delta - \theta) + (\mu - \eta)(\delta - \theta) + h\alpha_1) \delta \beta_1},$$

$$I_i^* = \frac{\mu \delta (\mu - \eta) (\delta - \theta) (c + h\omega_h/\mu) (c + h\omega_v/\mu) (R_1 - 1)}{(\mu - \eta) \alpha_1 (\beta_1 q + \delta \mu c + \delta \mu \omega_h N)},$$

con

$$a_i = \frac{\alpha_i}{c + \omega \hat{N}}, \quad b_i = \frac{\beta_i}{c + \omega \hat{N}}.$$

Con este nuevo orden el jacobiano tiene la forma

$$J(E_i^*) = \begin{pmatrix} G_1 & G_2 \\ 0 & G_4 \end{pmatrix},$$

donde

$$G_1 = \begin{pmatrix} -\delta + \theta - a_i I_i^* & a_i(\hat{T} - V_i^*) \\ b_i(\hat{N} - I_i^*) & -\mu + \eta - b_i V_i^* \end{pmatrix},$$

$$G_4 = \begin{pmatrix} -\delta + \theta & a_j(\hat{T} - V_i^*) & 0 & a_j(\hat{T} - V_i^*) & 0 \\ b_j(\hat{N} - I_i^*) & -\mu + \eta - \sigma_i b_i V_i^* & 0 & 0 & 0 \\ 0 & \sigma_i b_i V_i^* & -(\mu + r + e_i) + \eta & 0 & 0 \\ \sigma_j b_j I_i^* & 0 & 0 & -(u + r + e_j) + \eta & 0 \\ 0 & 0 & r & r & -u \end{pmatrix}.$$

Los auto-valores de $J(E_i^*)$ están dados por los auto-valores de G_1 y G_4 . Cuando $R_i > 1$, G_1 tiene dos auto-valores con parte real negativa. Los auto-valores de G_4 están dados por $-\mu$, $-(\mu + r + e_1) + \eta$ y por las raíces del polinomio

$$\lambda^3 + A\lambda^2 + B\lambda + C = 0,$$

donde

$$A = \sigma_i b_i V_i^* + r + e_j + 2(\mu - \eta) + (\delta - \theta),$$

$$B = (\sigma_i b_i V_i^* + (\mu - \eta) + (\delta - \theta))(\mu - \eta + r + e_j) - \sigma_j a_j b_j \frac{\delta V_i^*}{a_i} - a_j b_j \frac{\hat{N}\hat{T}}{R_i} +$$

$$+ (\delta - \theta)(\sigma_i b_i V_i^* + \mu - \eta),$$

$$C = (\sigma_i b_i V_i^* + (\mu - \eta))(\mu - \eta + r + e_j)(\delta - \theta) - \sigma_j a_j b_j \frac{\delta V_i^*}{a_i} (\sigma_i b_i V_i^* + (\mu - \eta)) -$$

$$- a_j b_j (\mu + r + e_j - \eta) \frac{\hat{T}\hat{N}}{R_i}.$$

En lo anterior usamos las equivalencias

$$(\hat{N} - I_i^*)(\hat{T} - V_i^*) = \frac{\hat{N}\hat{T}}{R_i}, \quad \frac{u - \eta}{b_i V_i^*} = \frac{\delta + a_i \hat{N}}{\delta(R_i - 1)}, \quad I_i^*(\hat{T} - V_i^*) = \frac{\delta V_i^*}{a_i}.$$

Nuestros parámetros umbrales están dados en términos de cotas para los coeficientes de super-infección σ_1 y σ_2 . Entonces, ellos fijan un conjunto de

cotas para el nivel de la resistencia inducida o susceptibilidad incrementada que cada cepa produce en el huésped. Sea

$$\begin{aligned}\sigma_i^* &= \max \left\{ 0, \left(\frac{R_j}{R_i} - 1 \right) \frac{\delta + a_i \hat{N}}{\delta(R_i - 1)} \right\}, \\ \sigma_j^* = f(\sigma_i) &= \frac{(\delta - \theta)(\mu + r + e_j - \eta)}{a_j b_j I_i^*(\hat{T} - V_i^*)} \left(1 - \frac{(\mu - \eta)}{(u + \sigma_i b_i V_i^* - \eta)} \frac{R_j}{R_i} \right).\end{aligned}\quad (4.3.5)$$

Haciendo un análisis sobre G_4 se obtiene el siguiente resultado

- Lema 4.2.** 1. E_i^* es localmente asintóticamente estable si $\sigma_j < f(\sigma_i)$ para cada $\sigma_i > 0$, e inestable si $\sigma_j > f(\sigma_i)$.
2. E_j^* es localmente asintóticamente estable si $\sigma_j > g^{-1}(\sigma_i)$ para cada $\sigma_i > 0$, e inestable si $\sigma_j < g^{-1}(\sigma_i)$.
3. E_i^* y E_j^* son localmente asintóticamente estables si $f^{-1}(\sigma_i) < \sigma_j < f(\sigma_i)$.

Para demostrar el lema anterior se hará un análisis de estabilidad para los diferentes escenarios presentado en el modelo, el caso en el que se tiene exclusión competitiva entre ambas especies.

- Lema 4.3.** 1. $\sigma_i > \sigma_i^*$ implica $\sigma_j^* > 0$;
2. $\sigma_i > \sigma_i^*$ y $\sigma_j < \sigma_j^*$ implica $A > 0, B > 0, C > 0$ y $AB > C$;
3. $\sigma_i < \sigma_i^*$ o $\sigma_j > \sigma_j^*$ implica $C < 0$.

Las pruebas de 1 y 3 son directas.

Sea $\sigma_i > \sigma_i^*$ y $\sigma_j < \sigma_j^*$. Es fácil ver que $\sigma_j < \sigma_j^*$ implica $C > 0$. Note que

$$\delta(u + \sigma_i b_i V_i^*) > u \delta \frac{R_j}{R_i}.$$

Entonces

$$\begin{aligned}B(u + \sigma_i b_i V_i^*) &\geq (u + \delta)(u + \sigma_i b_i V_i^*)\delta - (u + \sigma_i b_i V_i^*)\sigma_j a_j b_j I_i^*(\hat{T} - V_i^*) \\ &\geq C,\end{aligned}$$

Pero, $C > 0$, entonces $B > 0$. Para mostrar que $AB > C$ notamos que $\sigma_j < f(\sigma_i)$, entonces obteniendo

$$\sigma_j a_j b_j I_i^*(\hat{T} - V_i^*) \leq \delta(u + r) - u \delta \frac{u + r}{u + \sigma_i b_i V_i^*} \frac{R_j}{R_i}.$$

Se sigue entonces que

$$\begin{aligned}
AB &\geq (2u + \sigma_i b_i V_i^* + r + \delta) \cdot \left((u + r)(u + \sigma_i b_i V_i^* + \delta) + \delta(u + \sigma_i b_i V_i^*) \right. \\
&\quad \left. - u\delta \frac{R_j}{R_i} - \delta(u + r) + u\delta \frac{u + r}{u + \sigma_i b_i V_i^*} \frac{R_j}{R_i} \right) \\
&> (u + r) \left(\delta(u + \sigma_i b_i V_i^*) - u\delta \frac{R_j}{R_i} \right).
\end{aligned}$$

Ya que

$$C < \delta(u + \sigma_i b_i V_i^*)(u + r) - (u + r)u\delta \frac{R_j}{R_i}.$$

Se sigue que

$$AB > C.$$

Usando el criterio de Routh-Hurwitz tenemos el siguiente resultado:

Corolario 4.1. *El equilibrio de frontera E_i^* es localmente asintóticamente estable si $\sigma_i > \sigma_i^*$, y $\sigma_j < \sigma_j^*$. En otro caso, es inestable.*

El Lema 1 y el Corolario dicen que siempre que el coeficiente de súper-infección de la primera cepa esté arriba del umbral, existen valores del segundo coeficiente de súper-infección que dan estabilidad asintótica al equilibrio de frontera E_1^* . Debido a que sólo se requiere que $\sigma_i^* > 0$ y $\sigma_j^* > 0$, la estabilidad asintótica de E_i^* está garantizada en principio ya sea cuando la primera cepa induce resistencia o cuando incrementa la susceptibilidad a la segunda cepa.

Usando la definición de f dada en (5) tenemos

Lema 4.4. *1. Si $R_j < R_i$, entonces $\sigma_i^* = 0$, $f(\sigma_i^*) = f(0) > 0$, $f'(\sigma_i) > 0$.
2. Si $R_j > R_i$, entonces $\sigma_i^* > 0$, $f(\sigma_i^*) = 0$, $f(0) < 0$, y $f'(\sigma_i) > 0$.*

Note que es posible tener un valores umbral para σ_i^* y para σ_j^* tales que ambos E_i^* y E_j^* sean localmente asintóticamente estables. Esta conclusión indica, por lo menos, que hay situaciones donde las dinámicas asintóticas de nuestro modelo dependen de las condiciones iniciales.

Note que la estabilidad del punto de equilibrio E_i^* se favorece cuando la virulencia (mortalidad extra del huésped inducida por el virus) de la cepa j es grande.

Más aún, en el caso límite donde $\sigma_i = \sigma_j = 0$ (no ocurre la super-infección), la condición $R_j < 1 < R_i$ hace Φ una contracción local alrededor de E_i^* . Entonces, la estabilidad asintótica local de E_1^* en Ω se sigue.

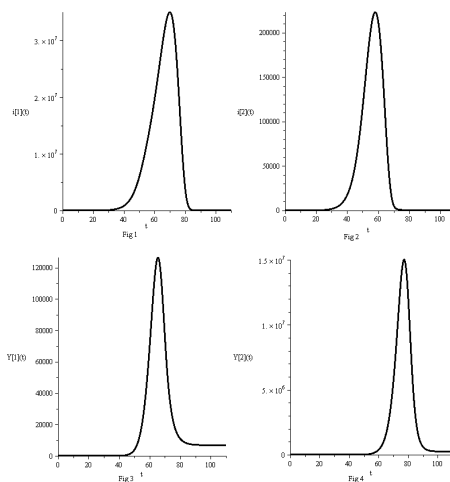


Figura 4.1: Gráfica para los huéspedes con infecciones primarias e infecciones secundarias

Sin embargo, si $0 < \sigma_i < \sigma_j$, E_i^* puede dejar de ser un atractor bajo Φ (con la condición $R_j < 1 < R_i$ manteniéndose). En este caso E_i^* es todavía un atractor global en Ω_i y (localmente) un punto silla en Ω .

Es importante mencionar que siempre que $R_i > 1$ y $R_j > 1$ simultáneamente, ambos equilibrios E_i^* y E_j^* existen.

Equilibrio endémico con coexistencia de las dos cepas de vectores

El caso en el que se tiene la presencia de las dos cepas de vectores se tiene que encontrar la solución de la ecuación $\Phi(K) = K$, donde las variables $I_1, I_2, Y_1, Y_2, V_1, V_2$ son diferentes de cero y este caso es algebraicamente intratable. Por esta razón solamente se presentan resultados numéricos que aportan evidencia fuerte sobre la existencia de un equilibrio endémico interior, esto es, un punto de equilibrio con densidades positivas para ambos tipos de huéspedes infectados. En estas simulaciones exploramos la interdependencia de tres parámetros clave: el número reproductivo básico, los coeficientes de super-infección y la tasa de muerte inducida por cada cepa.

Los tamaños de población variables pueden tener un efecto muy dramático en el resultado de interacción competitiva [22]. En particular puede “reducir” el área del espacio parametral en el cual es posible la coexistencia (comparada con el modelo equivalente con tamaño de población constante). Los siguientes resultados fueron encontrados a través de simulación numérica del modelo (1-2) donde $e_i > 0$.

1. Siempre que $R_i > 1$ para $i = 1, 2$ existe un punto de equilibrio en

el interior de Ω . Este punto tiene una variedad local inestable y una variedad local estable de dimensión positiva.

2. Si $R_i > 1 > R_j$ entonces el equilibrio de frontera E_j^* y el equilibrio endémico interior no existe y el equilibrio de frontera E_i^* es globalmente asintóticamente estable.
3. Cuando $R_i > R_j > 1$, los coeficientes de super-infección σ_1 y σ_2 pueden cambiar el comportamiento asintótico del sistema, imponiéndose la cepa j como la ganadora sobre la cepa i (la cual sería ganadora si $\sigma_1 = \sigma_2 = 1$).
4. Cuando $R_i > R_j > 1$ y ambos equilibrios de frontera son localmente asintóticamente estables, existe una separatriz que corta Ω en dos cuencas de atracción disjuntas.

Las simulaciones del modelo muestran que en la población huésped no hay persistencia a largo plazo para ambas cepas. Sin embargo, la naturaleza inusual del equilibrio endémico (o punto silla) produce un estado de cuasi-equilibrio estable relativamente prolongado (años de duración) cuando ambos R_i son mayores que uno. Dada la escala de tiempo inherente a la enfermedad (meses), este estado de cuasi-equilibrio luciría como un equilibrio endémico estable (vea Figs. 1,2,3,4). Bajo estas condiciones hay dos posibilidades dependiendo de cuántos equilibrios de frontera son localmente estables. Si sólo uno de ellos es asintóticamente localmente estable nuestras simulaciones de computadora indican que este equilibrio es también globalmente asintóticamente estable. Entonces, la exclusión competitiva de una de las cepas ocurre. Nuestras simulaciones computacionales también indican que aunque el brote inicial de infección primaria se conduce por la cepa con el mayor número reproductivo, es precisamente esta cepa la que puede ser competitivamente excluida. Esto ocurre si la infección primaria mejora (incrementa la susceptibilidad a ella) a las infecciones secundarias (vea Fig. 3). Entonces, la cepa con el menor número reproductivo puede terminar persistiendo en el huésped. La segunda posibilidad ocurre cuando ambos equilibrios de frontera existen y ambos son localmente asintóticamente estables. En este caso el resultado de la interacción-exclusión competitiva de la cepa, que depende de las condiciones iniciales.

4.3.3. Conclusiones

En este trabajo se incorpora la transmisión vertical de la infección en las poblaciones de humanos y la población de vectores en un modelo epidemiológico

propuesto por Feng y Velasco-Hernández (1999). La incorporación de dicha fenomenología ha sido analizada parcialmente en este trabajo. Se encuentra que la transmisión vertical afecta más la dinámica de la población de vectores que la población de humanos, además de que para tener la posibilidad de controlar el brote epidémico se necesita que siempre el número de recién nacidos infecciosos en cada clase sean menor que el número promedio de decesos en cada clase de humanos infecciosos y de vectores. Se demuestra que para un rango biológicamente factible en los índices de súper-infección σ_1 y σ_2 y bajo ciertas condiciones sobre los otros parámetros existen dos soluciones de equilibrio del sistema, los cuales representan estados endémicos de la fiebre del dengue, y se muestra numéricamente la existencia de un equilibrio endémico interior el cual es inestable (coexistencia entre las cepas) con variedades estables e inestables en la dimensión no-cero. La existencia de un punto de equilibrio endémico interior con esta caracterización produce un comportamiento de largo plazo potencialmente impredecible. Este rango cubre casos cuando ambas infecciones primarias inducen resistencia a infecciones secundarias o cuando se realiza la susceptibilidad es importante señalar que dicho rango cubre casos cuando ambas infecciones primarias inducen resistencia a infecciones secundarias o cuando se realiza la susceptibilidad. Las olas de infección nuevas luego del brote epidémico primario se adaptarán a un nivel de endemicidad bajo transitorio, aparentemente estable, donde las cepas co-circulen. Sin embargo, conforme el tiempo pase, la prevalencia de una de las cepas se incrementará eventualmente y a ritmo constante, mientras la otra cepa desaparece. Cual de las cepas será la ganadora depende de las condiciones iniciales preexistentes cuando la nueva cepa llegó y del nivel de susceptibilidad inducido por la infección primaria (Figs. 1,2,3, y 4). Un análisis más detallado muestra que para valores factibles de σ_i (e.g. $\sigma_i \in (0.1, 2)$) y siempre que $R_i > R_j$, el equilibrio interior existe a menos que σ_j sea muy pequeña o muy grande. En las figuras 1, 2, 3 y 4 ilustramos el caso cuando $R_1 > R_2$ (vea Tabla 1 para valores de los parámetros usados). Concluimos que la presencia de súper-infección fuerza la existencia del equilibrio interior. Sin embargo, parece que la super-infección no puede inducir estabilidad en este estado. El equilibrio endémico interior dejará de existir para un valor de $\sigma_2 \approx 0,001$ pero entonces sólo el equilibrio de frontera E_i^* será asintóticamente estable.

Bibliografía

- [1] Feng, Z. and Velasco-Her'andez, X. J., Competitive exclusion in a vector-host model for the dengue fever, *Journal of Mathematical Biology*. 35: 523-544, 1999.
- [2] Carles G, Peiffer H, Talarmin A. Effects of dengue fever during pregnancy in French Guiana. *Clin Infect Dis* 28:637-640. 1999
- [3] Chye, J. K, Lim C. T., Kwee B. Ng., Lim J.H.M, George R, Lam S. K. Vertical transmission of Dengue, *Clinical Infectious Diseases*, 25: 1374.1377, 1997.
- [4] Thaithumyanon P, Thisyakorn U, Deerojnawong J, Innis BL. Dengue infection complicated by severe hemorrhage and vertical transmission in a parturient woman. *Clin Infect Dis* 18:248-249, 1994
- [5] Fernandez R, Rodriguez T, Borbonet F, Vazquez S, Guzman MG, Kou G. Study of the relationship dengue- pregnancy in a group of cuba mothers. *Rev Cubana Med Trop* 46(2):76-78. 1994.
- [6] Figueiredo LT, Carlucci RH, Duarte G. Estudo prospectivo com lactantes cujas maes tiveram dengue durante a gravidez. *Rev Inst Med Trop. Sao Paulo*, 36(5):417-421, 1994.
- [7] Poli L, Chungue E, Soullignac O, Gestas P, Kuo P, Papouin-Rauzy M. Materno-fetal dengue. Apropos of 5 cases observed during the epidemic in Tahiti 1989. *Bull Soc Pathol Exot* 84(5):513-521, 1991.
- [8] Chong KY, Lin Kc. A preliminary report of the fetal effects of dengue infection in pregnancy. *Kao Hsiung I Hsueh Ko Hsueh Tsa Chic* 5(1):31-34. 1989.
- [9] Aron, J. and May, R. M., The population dynamics of Malaria, pages 139-179. *Population Dynamics of Infectious Diseases*. Chapman and Hall, London, 1982.
- [10] Bremermann, H. J. y Thieme, H. R., A competitive exclusion principle for pathogen virulence. *Journal of Mathematical Biology*, 27: 179-190, 1989.
- [11] Capasso, V. *Mathematical Structures of Epidemic System*. Springer Verlag, New York, 1993. *Lecture Notes in Biomathematics* 97.
- [12] Castillo-Chavez, C., Huang, w. and Li. J., Epidemiological models with age structure, proportionate mixing, and cross-immunity. *Journal of Mathematical Biology*, 27: 233-258, 1989.

- [13] Castillo-Chavez, C., Huang, W. and Li, J., Competitive exclusion in gonorrhea models and other sexually, transmitted diseases. Technical Report 93-36, Mathematical Science Institute, Cornell University, 1993.
- [14] Castillo-Chavez, C., Velasco-Hernandez, J.X. and Fritman, S., Modeling contact structures in biology. Lecture Notes in Biomathematics, 100: 454-492, 1995.
- [15] Diekmann, O., Heerterbeek, J. A. P. and Metz, J. A. J., On the definition and the computation of the basic reproduction ratio r_0 in models for infectious diseases in heterogeneous populations. Journal of Mathematical Biology, 28: 365-382, 1990.
- [16] Gubler, D. J., Dengue, volume 2 of The Arboviruses: Epidemiology and Ecology, pages 223-261. CRC Press, Inc., 1986.
- [17] Halstead, S. B. Dengue, Tropical and Geographical Medicine. Mac Graw Hill, New York, pp. 652-659, 1984.
- [18] Halstead, S. B., The XXth century dengue pandemic: need for surveillance and research. World Health Statistics Quarterly, 45: 292-298, 1992.
- [19] Levin, S. and Pimentel, D., Selection of intermediate rates of increase in parasite-host systems. American Naturalist, 117(3): 308-315, 1981.
- [20] Li, J., Castillo-Chavez, C. and Huang, W., Dynamics of multiple pathogen strains in heterosexual epidemiological models pages 289-298. Differential Equations and Applications to Biology and to Industry. Singapore, New Jersey, London, Hong Kong, 1996.
- [21] May, R. M., and Nowak, M. N., Super infection, metapopulation dynamics, and the evolution of virulence. Journal of Theoretical Biology, 170: 95-114, 1994.
- [22] Mena-Lorca, J., Velasco-Hernandez, J. X. and Castillo-Chavez, C., Super infection, virulence and density dependent mortality in an epidemic model. Technical Report BU 1299-M., Biometrics Unit, Cornell University, 1995.
- [23] Nowak, M. N., and May, R. M., Super infection and the evolution of parasite virulence. Proceedings of the Royal Society of London B, 255: 81-89, 1994.
- [24] World Health Organization. The urban crisis. World Health Statistics Quarterly, 44: 208-233, 1991.
- [25] Velasco-Hernandez, J. X., A model for Chagas disease involving transmission by vectors and blood transfusion. Theoretical Population Biology, 46(1):1-31, 1994.

4.4. Construcción óptima de estratos para estudios asociados a epidemias de dengue.

Carlos N. Bouza-Herrera¹, Sira M. Allende-Alonso¹, Gemayqzel Bouza-Allende¹, Dante Covarrubias- Melgar² y Lakshma Singh³

1. Departamento de Matemática Aplicada, Universidad de La Habana, Cuba.

2. Universidad Autónoma de Guerrero

3. Bhat & Sarkar Informatic Consultants, India

Resumen En este trabajo se presenta brevemente el problema general del muestreo antes de discutir el problema del muestreo estratificado. El problema de la estratificación óptima es discutido y se presenta el enfoque de la programación dinámica que utiliza las varianzas obtenidas con los diversos modelos como función objetivo. Datos de las tasas de mosquitos en un estudio desarrollado en West Bengala se utilizan para hacer una comparación entre los métodos de estratificación óptima.

Palabras clave: estratificación óptima, varianza generalizada, programación dinámica

Abstract This paper presents briefly the general sampling problem. Afterwards the optimal stratification problem is discussed and the dynamical programming approach to it is presented using the variances obtained by different models as objective functions. The rates of the population of mosquitoes obtained in a study developed in West Bengal are used for comparing the different optimal stratification methods.

Keywords: optimal stratification, generalized variance, dynamic programming

4.4.1. Introducción

El dengue es propagado por el mosquito *Aedes Aegypti*. Conocer las tasas del mismo es importante para planear y controlar el efecto de campañas para su erradicación. En el estudio del dengue es necesario por tanto el desarrollo de encuestas que permitan monitorear el comportamiento de la dinámica de la población de mosquitos. Así es necesario el estudio de zonas geográficas. Estas están divididas en áreas y dentro de ellas se muestrea con el fin de establecer un estimado de las tasas. Lo más recomendable es tomar información de cada una de las zonas lo que sugieren la recomendabilidad de usar el

concepto de estratificación. En la correspondiente teoría los estratos forman entidades disjuntas que cubren el área total. Dada la naturaleza de los estudios de este tipo es posible diseñarla construcción de los estratos. Por ello el determinarles en forma tal que esta estratificación minimice el error de estimación es una opción importante aunque no siempre utilizada. En este trabajo se presentan aspectos del muestreo como teoría lo que soporta el uso del muestreo estratificado. Se establecen los aspectos teórico más importante del mismo: estimación insesgada del promedio, afijación del tamaño de la muestra, ganancia en precisión debida a la estratificación. La determinación de estratos óptimos es abordada. Esta es presentada para el caso univariado para los casos de la selección simple aleatoria y con probabilidades desiguales. Se considera el caso general en que la información es deformada al ser generada por un modelo superpoblacional. Los modelos considerado son el producto de la variable de interés Y por una variable S cuya distribución es conocida y el del modelo de regresión simple $Y = b(X) + e$. Utilizándoles se presenta una metodología para determinar estratos óptimos aproximadamente. El uso de la programación dinámica es propuesto y se presenta el problema de optimización. Una base de datos de las tasas de mosquitos existentes en una zona de West Bengala es utilizada para ilustrar el comportamiento de los métodos estudiados. Este es valorado usando la ganancia en precisión debido a la estatificación. En la sección 2 se presenta el problema general del muestreo dentro de la teoría unificada. La sección 3 se dedica a discutir el problema del muestreo estratificado, donde se presentan las distintas afijaciones, cuando el submuestreo es llevado a cabo usando muestreo simple aleatorio con y sin reemplazo. Asimismo se presenta una medida de la ganancia en precisión debido al muestreo. La estratificación óptima es discutida y se justifica el uso de la aproximación de la distribución de Y por una distribución uniforme. En el capítulo 4 se presenta el uso del muestreo con probabilidades desiguales. En este caso estas son determinadas usando una variable auxiliar conocida X . Como caso particular se tendría el MSA cuando $X = 1$ para toda unidad poblacional. En este marco se considera el uso de un modelo superpoblacional $Y = SY$, Este modela el caso en que hay errores aleatorios que deforman la medición de Y aunque se conocen la media y varianza de S . El otro modelo es el clásico modelo superpoblacional de regresión simple. Si e es una constante nula y $b = 1$ se regresa al problema usual. Un estudio similar el llevado a cabo con el modelo producto se desarrolla. La estratificación óptima para el modelo superpoblacional es presentado usando una aproximación de la función de densidad del modelo. El capítulo 5 presenta el enfoque de la programación dinámica que utiliza las varianzas obtenidas con los diversos modelos como función objetivo. Datos de las tasas de mosquitos en un estudio desarrollado

en West Bengala se utilizan para hacer una comparación entre los métodos de estratificación óptima.

4.4.2. El muestreo de poblaciones finitas.

Algunos antecedentes

Las nociones de la inferencia en poblaciones finitas fueron expresadas hace más de 60 años en muchos libros clásicos como Cochran, Hansen, Hurwitz y Madow, Deming, Murthy, Des Raj y otros. La teoría de muestreo era aplicada desde la perspectiva misma de la selección aleatorizada de posibles muestras en la población finita. Dependiendo de las circunstancias prácticas, la selección se hacía de distintas maneras: muestreo aleatorio simple, muestreo aleatorio estratificado, muestreo de conglomerados, muestreo en dos etapas, etc. El muestreo era considerado como la actividad primaria y la estimación nunca fue considerada como una práctica separada sino como una consecuencia automática. Lo anterior se debía a que cada tipo de diseño de muestreo inducía un estimador cuyas propiedades estadísticas como el insesgamiento y la varianza eran establecidas de antemano con el diseño y así, la varianza era calculable y estimable. Así que, para la década de los 60's, muchos creyeron que la investigación en el campo del muestreo y de la inferencia en poblaciones finitas ya estaba muerta porque se deberían inventar nuevas formas de selección de muestras (tarea ardua y difícil), más allá de las que se cubrían en los libros clásicos del muestreo. Aunque el estimador de razón fue considerado en algún detalle por los textos de referencia, la inclusión de varias variables de información auxiliar no se vio como un tópico que prometiera rédito alguno para emprender el camino de la investigación en esa vía. En la década de los 70's, varios autores dieron un viraje en su perspectiva epistemológica de la inferencia en poblaciones finitas. Es así como Basu, Brewer, Godambe y Royall, entre otros, consideraron los modelos estadísticos (en sintonía con la estadística clásica Fisheriana) como los verdaderos fundamentos de la estimación e inferencia en poblaciones finitas. Su trabajo se cimentó alrededor de la posibilidad de tener una inferencia que dependiera estrictamente del modelo propuesto y no tuviera nada que ver con el diseño de muestreo utilizado en la recolección de los datos. Como consecuencia, la atención se tornó alrededor de la estimación y se dejó de lado el muestreo por la relación existente o propuesta entre la característica de interés y las variables de información auxiliar. El camino que tomó la historia del muestreo fue, precisamente, la incorporación de las dos corrientes de pensamiento bajo una sola sombrilla. Así que, fue posible combinar la Aleatorización clásica con un percepción más general de la relación de y con x . No hubo necesidad de sacrificar los

principios basados en la aleatorización. Así nació la inferencia asistida por modelos pero basada en la aleatorización (model assisted design-based inference por su original en inglés). Este nuevo tipo de inferencia se hizo muy atractiva porque la regresión y los modelos acompañan al estadístico desde sus primeros cursos y van tomando más fuerzas a medida que se avanza en el camino universitario. Así que, este pensamiento “asistido por modelos” es un matrimonio efectivo y tolerante que permite las ideas de la regresión junto con el paradigma de la aleatorización. Jan Wretttman opina que el ajuste de un modelo se ha convertido en parte integral de la teoría clásica del muestreo, aunque los principios de la misma deben permanecer intocables porque las propiedades de los estimadores son evaluadas con respecto al mecanismo de probabilidad que genera la muestra y no con respecto a cualquier modelo asumido.

Diseño muestral.

Tenemos una población finita $U = \{1, \dots, N\}$ y una variable de interés Y que es evaluada en cada elemento de U . Esto determina el vector $Y(U) = Y = (Y_1, \dots, Y_N)$ es un parámetro pero comúnmente denominado parámetro universal o vector paramétrico. $Y \in \mathfrak{R}^N$ y al fijar sus coordenadas se puede determinar un intervalo de definición para cada uno $a(i) < Y_i < b(i)$ por lo que se puede determinar una subespacio $\prod_{i=1}^N]a(i), b(i)[$ de $Y \in \mathfrak{R}^N$. Denotémosle $\Omega = \{Y = (Y_1, \dots, Y_N)^T | a(i) < Y_i < b(i), i = 1, \dots, N\}$ al que llamamos espacio paramétrico, $\Omega \subseteq \mathfrak{R}^N$. Por problemas teóricos es usualmente considerando que es un espacio propio o que contenga al menos algunos puntos notables como el 0, por ejemplo.

Definición. Una muestra $s \subset U$ es una sucesión de elementos de la población $\{i_1, \dots, i_n\}$. Cuando i_j es el elemento incluido en la muestra en la selección j -ésima esta está ordenada. Tomemos S como el conjunto de todas las muestras ordenadas. Es llamado espacio muestral.

Para formalizar el problema probabilístico consideremos la sigma álgebra dada por el conjunto potencia de S , $\mathbf{S} = 2^{|S|}$. Entonces $\{S, \mathbf{S}, d\}$ es un espacio de probabilidad definida sobre $\{S, \mathbf{S}\}$. Entonces d es una medida de probabilidad con una estructura particular que le asigna a cada muestra de S una probabilidad.

Definición. Un diseño muestral es una medida de probabilidad $d : \{S, \mathbf{S}\} \rightarrow [0, 1]$, $\sum_{s \in S} d(s) = 1$ con soporte $S(d) = \{s \in S | d(s) > 0\}$.

Note que no es necesario que los i_j 's sean diferentes.

Entonces podemos apuntar que:

- El punto de partida para afrontar un diseño muestral es definir cuál es el universo población de estudio $U = \{u(i), i = 1, \dots, N\}$.
- $u(i)$ es el individuo i -ésimo de la población y es perfectamente identificable.
- $S = \{U^\#\}$ es el espacio muestral (potencia de U) o sea que es una sigma álgebra.
- Por tanto tenemos un espacio de probabilidad U, S, d .

Definición. Un diseño es una medida de probabilidad

- $d : S \rightarrow [0, 1]$ que le asigna a cada muestra de $s \in S$ una probabilidad de ser seleccionada.

Al hacer un experimento aleatorio se selecciona una muestra $s = \{i_1, \dots, i_n\}$ donde i_j es el elemento seleccionado en el intento j . Se evalúa la variable de interés y se obtiene $y(s) = (y_1, \dots, y_n)$ donde la coordenada j -ésima es el valor de Y en el i_j . Entonces tenemos los datos.

Definición. La sucesión $D = \{s, y\} = \{(i, Y_i) | i \in s\} = \{(i, y_j), j = 1, \dots, n\}$ es llamada dato muestral. La colección $D^* = \{D\}$ de todos los datos es llamada espacio de los datos.

Antes de seleccionar la muestra cada $\{(i, Y_i) | i \in s\}$ es un dato observable.

Hay un relación biunívoca entre diseños y procedimiento de muestreo. Veamos esto describiendo como se implementa un diseño.

Implementación de la selección de una muestra usando un diseño muestral.

Paso 1. $Q(1) = Prob\{i = i_1\}$, $s(1)$ es la muestra provisional, ($Q(1)$ es una medida de probabilidad sobre U),

Paso 2. $Q(2) = Q(2, s(1)) = P(s(1))$ observada), ($Q(2, s(1))$ es una medida de probabilidad en el espacio de la muestras provisionales $\Omega[s(1)]$)

Paso 3. Se hace un experimento Bernoulli con probabilidad $Q(2, s(1))$. Si el resultado es un "éxito" seguir, sino parar.

Paso 4. $Q(3) = Q(3, i|s(1)) = P(i = i_j|s(1))$ (probabilidad condicional de observar a i en la selección j dada la muestra provisional $s(1)$).

Paso 5. Si $i = i_2$ es seleccionada la muestra provisional es $s(2) = \{s(1), i_2\}$ e ir al Paso 2.

Finalizar.

Note que este proceso termina cuando hay un fracaso.
Veamos el problema probabilístico. Tomemos

$$Q(1, i) = \alpha_i = \sum_{s \in S(i)} d(s), \quad \beta_i = d(\{i\})$$

$$\alpha_{ij} = \sum_{s \in S(i,j)} d(s), \quad \beta_{ij} = d(\{i, j\})$$

Entonces para la muestra provisional $s = \{i_1, \dots, i_k\}$

$$Q(2, s) = \begin{cases} 1 - \frac{\beta_{i_1, \dots, i_k}}{\alpha_{i_1, \dots, i_k}}, & \text{si } \alpha_{i_1, \dots, i_k} \neq 0 \\ 0, & \text{en otro caso.} \end{cases}$$

$$Q(3, i|s) = \begin{cases} 1 - \frac{\alpha_{i_1, \dots, i_k}}{\alpha_{i_1, \dots, i_k} - \beta_{i_1, \dots, i_k}}, & \text{si } Q(2, S) \neq 0 \\ 0, & \text{en otro caso.} \end{cases}$$

Hallando la probabilidad de observar un muestra $s = i_1, \dots, i_n$ esta es expresada por

$$P(s) = Q(1, i_1)Q(2, i_1)Q(3, i_2|i_1)Q(2, i_1, i_2) \cdots Q(3, i_{n-1}|i_1, \dots, i_{n-2}) \cdots Q(3, i_1 \cdots, i_n)$$

Sustituyendo

$$P(s) = \alpha_{i_1} \frac{\alpha_{i_1} - \beta_{i_1}}{\alpha_{i_1}} \frac{\alpha_{i_1 i_2}}{\alpha_{i_1} - \beta_{i_1}} \frac{\alpha_{i_1 i_2} - \beta_{i_1 i_2}}{\alpha_{i_1 i_2}} \cdots \frac{\alpha_{i_1 \cdots i_n}}{\alpha_{i_1 \cdots i_{n-1}} - \beta_{i_1 \cdots i_n}} = \alpha_{i_1 \cdots i_n} = d(s)$$

O sea que el procedimiento genera un solo diseño. El inverso es evidente. Por tanto tenemos

Teorema (Hanurav , 1968). Todo procedimiento de muestreo implementa un solo diseño muestral y viceversa.

La estimación

Se toma la muestra para hacer una estimación. De la teoría general de la estadística se deriva para el problema de la estimación en poblaciones finitas una definición particular . Esta es la siguiente:

Definición. Un estadístico es una función $t = t(D) = t(s, Y)$ definida sobre $\{S, \Omega\}$ en $\mathfrak{R}(t : \{S, \Omega\} \rightarrow \mathfrak{R})$. Si t es usado para estimar la función paramétrica $T(Y)$ es llamado estimador de T .

Al hacer la selección de una estrategia es preferida aquella con menor $EM(d)$. En la práctica es usual buscar un estimador con sesgo cero (insesgada respecto al diseño, d -insesgada) y tratar de hacer la varianza lo más pequeña posible dadas las restricciones de costo que se planteen.

Sea $D = \{(i, Y_i) | i \in s\}$ el dato observado, entonces conocemos las coordenadas de Y cuyo índice fue seleccionado e ignoramos el valor de las coordenadas $Y_j, j \notin s$. En términos de los datos podemos representar el estimador como $t(D) = t(s, Y)$. Por ejemplo, si queremos estimar el total de la variable de interés Y

$$T(Y) = \sum_{i \in s} Y_i + \sum_{i \notin s} Y_i.$$

Definición. Una estrategia muestral para el parámetro T está definida por $\xi(d, t)$ donde d es una medida de probabilidad definida de T . Su comportamiento es medido por el error de muestreo, $EM(d)$ dado por

$$M_d[t] = E_d(t - T)^2 = \sum_{s \in S} (t(s, Y) - T(Y))^2 d(s) = V_d(t) + B_d^2(t)$$

V_d es el operador varianza bajo el diseño y B_d el sesgo bajo del diseño.

Parece que el estimador debe establecer una relación con los datos observados $\{i, Y_i | i \in s\}$ y los no observados $\{i, Y_i | i \notin s\}$. Pero si la muestra s es fijada los datos lo son y la relación solo tendrá un sentido probabilístico. En realidad una vez tomada la muestra, D no tiene relación con las coordenadas no observadas, sino con Y a través de la distribución de probabilidad. Por ello lo que tiene sentido no es preocuparse por cuan cerca de T esta $t(D)$ sino como lo está en términos probabilísticos. De ahí, como en el resto de los problemas inferenciales basados en el criterio del muestreo repetido de la teoría frecuentista, lo que nos interesa es evaluar

$$E_d(t(s, Y)) = \sum_{s \in S} t(s, Y) d(s)$$

y

$$E_d(t(s, Y) - T)^2 = \sum_{s \in S} (t(s, Y) - T)^2 d(s) = \sum_{s \in S} (t(s, Y) - T(Y))^2 d(s) = V_d(t) + B_d^2(t)$$

No se puede pensar en hacer una estimación sin error a menos que $s = U$ por lo que $d(s) = 1$. En este caso hacemos un censo y no inferimos sino que se desarrolla un estudio descriptivo.

Definición. Si $t(U, Y) = T$ este es Cochran-consistente.

Existe al menos un estimador que no se asocie a un error. Si no podemos garantizar que Y_i es medido sin error y $U = s$ lo que se hace es un estudio analítico.

Definición. Un estimador t es d -insesgado para T si $E_d(t(s, Y)) = T$. Si su varianza toma el menor valor posible para todo Y se denomina uniformemente de mínima varianza (*UMV*).

El principio del muestro repetido sostiene entonces el uso de la desigualdad de Markov y tendremos que $P(|t - T|e) \geq 1 - M_d[t]/e^2$

Intervalos de confianza son obtenibles al fijar $e = \lambda$ pues podemos hacer uso, con precaución, de la identidad

$$P\{t - \lambda\sqrt{V_d(t)} < T < t + \lambda\sqrt{V_d(t)}\} = P\{T \in IC(\lambda, t, V_d)\} > 1 - \alpha$$

Este es construido sobre las repeticiones hechas usando $d(s)$. Este concepto no lleva de nuevo al hecho de que preferiremos aquel estimador que tenga la menor amplitud de los intervalos y que $1 - \alpha$ esté lo más cerca posible de 1.

Teorema Central del Límite (Erdős-Renyi-Hajek). Si t es una función lineal de los $Y_i, i \in s, n \rightarrow \infty$ y $N \rightarrow \infty$ entonces $z = (t - T)/V_d(t)$ sigue aproximadamente una distribución $N(0, 1)$.

Consideremos las clases de estimadores del total $T = ? \sum_{i \in U} Y_i$

$$T_1 = \left\{ \sum_{i=1}^n a_r y_r \mid y_r = Y_{i_r} \right\},$$

a_r se asigna a la variable aleatoria r -ésima, de acuerdo a su aparición en las selecciones.

$$T_2 = \left\{ \sum_{i \in s} Y_i = \sum_{i \in U} b(s, i) Y_i \right\},$$

$b(s, i)$ le asigna valores a las coordenadas incluidas en la muestra.

$$T_3 = \left\{ c(s) \sum_{i \in s} Y_i \right\},$$

$c(s)$ le asigna valores a la muestra seleccionada.

Esos coeficientes son conocidos al tomar la muestra o antes de ello y no dependen de información no conocida- Las propiedades de las clases son determinadas por las de sus coeficientes. En general los problemas inferenciales

de los cursos de estadística se trabaja con la primera clase. Por ejemplo la media aritmética es derivada por asignar $1/n$ a los resultados en cada selección. Podemos decir que Neyman la usó para hacer su trabajo usándola. Tomemos el vector paramétrico $Y = (Y_1, \dots, Y_N) \in \mathcal{Y} \subset \mathfrak{R}^N$. La no existencia de *EIMV* en el muestreo de poblaciones finitas es probado simplemente para estimadores de T_2 como sigue.

Teorema. No existe un *EUMV* en la inferencia en poblaciones finitas

Demostración

Sea $t(s, Y)$ un estimador insesgado del total de Y, T y $A \in \mathcal{Y} \subset \mathfrak{R}^N$ un punto arbitrario del espacio de definición de Y con $T(A) = \sum_{i=1}^N A_i$. Siempre existe un estimador insesgado

$$t(A) = t(s, Y) - t(s, A) - T(A) = \sum_{i \in s} \frac{Y_i}{\pi_i} - \sum_{i \in s} \frac{A_i}{\pi_i} + T(A) = \sum_{i \in s} \frac{Y_i - A_i}{\pi_i} + T(A)$$

Como A es arbitrario su varianza es cero en $Y = A$. Por tanto no existe un *EIMV*.

Dada la no existencia de estimadores óptimos en el sentido de minimizar la varianza unívocamente en la clase de los estimadores insesgados lineales se han generado otros criterios de óptimo.

Tomemos D como una clase diseños y T como una de estimadores. A partir de propiedades serán subindizados. Entonces denotemos una clase de estrategias como E . Entonces $\xi \in D \times T$. Consideremos un par de estrategias para $T, \xi = \{d, t\}$ y $\xi = \{d, t^*\}$ donde t y t^* pertenecen a la misma clase T .

Veamos algunas propiedades deseables de los estimadores.

Usando un punto de vista decisional diremos que si dos estimadores t y t^* pertenecen a la misma clase T , diremos que t es al menos tan bueno como t^* si $M_d(t) \leq M_d(t^*)$. Si la desigualdad es válida para algún $Y \in \mathcal{Y}$ se dice que t es mejor que t^* (t domina a t^*). Si t domina a t^* en T este es denominado el mejor estimador en T . Esto lleva a fijar una propiedad muy estudiada para estimadores del total $T(Y)$.

Definición. Si en una clase T ningún t^* domina a t se dice que este es admisible en T .

Esto es lo menos que se le puede pedir a un estimador. El estimador de Horvitz-Thompson aparece como admisible.

Fijando un diseño d y $T(o)$ una subclase de T . Si para todo $t \in T|T(o)$ existe un t_o de $T(o)$ que es mejor que t entonces esta subclase cumple con la definición de completitud. Si no hay una subclase de $T(o)$ completa esta es minimal y completa.

Si tomamos dos estrategias $\xi(1) = \{d(1), t(1)\}$ y $\xi(2) = \{d(2), t(2)\}$ de $\xi \in D \times T$. Se dice que una estrategia $\xi(1)$ es uniformemente mejor que $\xi(2)$ si $M_d(t(1)) \leq M_d(t(2))$ para todo $Y \in \mathcal{Y}$ cumpliéndose la desigualdad para l menos un Y^* .

Definición. Si no existe otra estrategia uniformemente mejor que se dice que es admisible.

La estrategia usando muestreo simple aleatorio con reemplazo (msacr) con la media aritmética no es uniformemente admisible pues

$$V(\bar{y}) = \frac{\sigma^2}{n}$$

Para el caso del MASSR

$$V(\bar{y}) = \frac{N - n}{N - 1} \frac{\sigma^2}{n}$$

es claro que el $N - n < N - 1$ a menos que $n = 1$.

Diversas variantes se han desarrollado pero estos son los razonamientos básicos.

De ahí que las propiedades de $\theta^* = f(\alpha_1^*, \dots, \alpha_m^*)$, donde ν^* estima a ν van a depender de un estudio de las propiedades estadísticas clásicas. Si f es una función lineal todo es sencillo. En otros casos una solución es desarrollar en Series de Taylor esa función. El ser insesgado es una propiedad popular pero todo va a depender de cuán grande es el sesgo del estimador. Vale analizar el valor de la razón entre los componentes del error cuadrático medio

$$BR(\theta^*) = B(\theta^*)/V(\theta^*)^{1/2}$$

Si es pequeño, la mayor parte de los procedimientos basados en la varianza, con los intervalos de confianza no se ven muy afectados pues un IC es

$$IC(\theta^*) = (\theta^* - z_{1-\alpha/2}V(\theta^*)^{1/2}, \theta^* + z_{1-\alpha/2}V(\theta^*)^{1/2})$$

En términos de la variable tipificada Z tenemos que si es una $N(0, 1)$

$$P\{Z(-z_{1-\alpha/2} - BR(\theta^*), z_{1-\alpha/2} + BR(\theta^*))\} \approx 1 - \alpha$$

si $BR(\theta^*) \approx 0$.

En el resto de la teoría estadística se habla de insesgadez asintótica. Está dada porque la sucesión $\theta_1^*, \dots, \theta_n^*, n \rightarrow \infty$ entonces $E(\theta_n^*)$ y consistente si $P\{|\theta_n^* - \theta| > \varepsilon\} \rightarrow 0$.

Entonces si θ_n^* es asintóticamente insesgado su estimador estará más cerca de θ para n grande. Si es consistente la distribución está muy concentrada

alrededor de θ . En la teoría del muestreo nada garantiza que $n \rightarrow N$ garantice nada de esto. Es necesario usar conceptos de procesos estocásticos y hacer n y N crecer hacia el infinito al mismo tiempo satisfaciendo que $n < N$. El marco es el siguiente:

Tenemos individuos i_1, \dots, i_k, \dots . A cada individuo se asocia un valor de Y . O sea para el k -ésimo individuo se tiene $Y(i_k)$. Se tiene una sucesión de poblaciones

$U(1) \subset U(2) \subset \dots \subset U(N_v) \subset \dots$ con $N_1 < N_2 < \dots < N_v < \dots$ y $v \rightarrow \infty$. El análisis es en términos de $E(\theta_{n(v)}^*) \rightarrow \theta$ y $P\{|\theta_{n(v)}^* - \theta| > \varepsilon\} \rightarrow 0$, cuando $v \rightarrow \infty$.

Ahora podemos darle una solución a la estimación de funciones no lineales. Si $\theta^* = f(\alpha_1^*, \dots, \alpha_m^*)$, lo fuera podemos hacer un desarrollo en series de Taylor, alrededor de $(\alpha_1, \dots, \alpha_m)$, y fijar como estimador a

$$\theta^* \approx \theta_0 = \theta + \sum_{j=1}^q \delta_j (\alpha_j^* - \alpha_j),$$

donde $\delta_j = \partial f / \partial \alpha_j^*$, $j = 1, \dots, q$ es evaluado en $(\alpha_1^*, \dots, \alpha_m^*) = (\alpha_1, \dots, \alpha_m)$. Si n es grande θ^* tiene un comportamiento similar al de θ_0 .

La exactitud de la solución depende solo de la muestra.

El ejemplo más popular es del estimador de la razón $R = T(y)/T(x)$. Este estimador es no lineal. $(\alpha_1^*, \dots, \alpha_m^*) = (y_1, \dots, y_n)$ y usando el método de la substitución

$$R^* = t(y)/t(x) = f(t(x), t(y))$$

$$\partial R^* / \partial t(y) = t(x)^{-1}, \quad \partial R^* / \partial t(x) = -t(y)t(x)^{-2}$$

Evaluándoles

$$\delta_1 = T(x)^{-1} \quad \delta_2 = -RT(x)^{-1}$$

Entonces

$$R^* \approx R_0 = R + (\sum y_i - Rx_i)/T(X) \Rightarrow B(R^*) \approx (\sum y_i - Rx_i)/T(X)$$

$$V(R^*) \approx [V(t(y) + R^2 V t(x) - 2R cov(t(x), t(y)))]/T(X)^2$$

4.4.3. Muestreo estratificado

Un muestreo aleatorio estratificado es aquel en el que se divide la población de N individuos, en k subpoblaciones o estratos, atendiendo a criterios que puedan ser importantes en el estudio, de tamaños respectivos N_1, \dots, N_K , $N = N_1 + \dots + N_K$. Seleccionamos en cada una de estas subpoblaciones una muestra de tamaño n_i con igual probabilidad (muestreo simple

aleatorio, MAS). Sea n el número de individuos a seleccionar, tamaño de la muestra, entonces $n = n_1 + \dots + n_k$. Sea Y la variable aleatoria que representa el carácter que intentamos estudiar. Sobre cada estrato puede definirse entonces la variable aleatoria. El parámetro (media) puede ser expresado mediante una combinación lineal de los estratos

$$\bar{Y} = \frac{\sum_{i \in U} Y_i}{N} = \frac{\sum_{j=1}^k \sum_{t \in U_j} Y_j}{N} = \frac{\sum_{j=1}^k N_j \bar{Y}_j}{N}$$

Tomando

$$W_i = \frac{N_i}{N} = Prob(U_i)$$

la podemos escribir como

$$\bar{Y} = \sum_{j=1}^k W_j \bar{Y}_j.$$

Si seleccionamos la muestra s_i del estrato i

$$\bar{y}_i = \frac{\sum_{j=1}^{n_i} \bar{y}_j}{n_i}, \quad i = 1, \dots, k, \quad n_i = |s_i|$$

$$\bar{y}_e = \sum_{j=1}^k W_j \bar{y}_j$$

es insesgado si usamos MAS en la sección de la muestra de cada estrato. Además

$$V(\bar{y}_e) = \sum_{j=1}^k W_j^2 V(\bar{y}_j)$$

es el error de muestreo.

Un problema raigal es determinar cuál es número de unidades a muestrear en cada estrato. Este es conocido como el problema de afijación a asignación del tamaño de la muestra. Hay cuatro criterios: asignar el mismo número a cada estrato, asignarlo arbitrariamente usar un criterio diferenciado como el proporcional y el óptimo.

La asignación es proporcional al tamaño del estrato correspondiente con respecto a la población total si:

$$n_i = \frac{N_i}{N}, i = 1, \dots, K$$

Consideremos que tenemos la función de costos que mide los gastos de la encuesta

$$C = C_0 + \sum_{h=1}^K c_h n_h.$$

El primer componente representa los gastos generales de la encuesta y c_h el de obtener la respuesta en un individuo del estrato h -ésimo. En general

$$V(\bar{y}_e) = \sum_{h=1}^K \frac{\varepsilon_h}{n_h}$$

Podemos plantear la desigualdad de Cauchy Schwarz

$$V(\bar{y}) \sum_{h=1}^K c_h n_h \geq \sum_{h=1}^K (\sqrt{\varepsilon_h n_h})^2 \Rightarrow \left(\sum_{h=1}^K \frac{\varepsilon_h}{n_h} \right) \left(\sum_{h=1}^K c_h n_h \right) = \sum_{h=1}^K (\sqrt{\varepsilon_h n_h})^2$$

Esto se cumple solo si

$$\forall h = 1, \dots, K : \frac{c_h n_h}{\left(\frac{\varepsilon_h}{n_h} \right)} = \text{constante} \Rightarrow n_h \propto \sqrt{\frac{\varepsilon_h}{c_h}}$$

Esto fija un criterio lógico pues esta fórmula establece que

1. El tamaño de la muestra debe ser pequeño si es barato hacer un muestreo en el estrato.
2. El tamaño de la muestra debe ser grande si la variabilidad del estrato es grande.

Resolviendo el problema de optimización tenemos que, dependiendo de si se hace reemplazo (CR) o no (SR) en la selección de la muestra del estrato h

$$\text{Si } n = \sum_h n_h \Rightarrow n_h = n \frac{W_h \sqrt{\varepsilon_h} / \sqrt{c_h}}{\sum_{h=1}^K W_h \sqrt{\varepsilon_h} / \sqrt{c_h}}$$

$$\varepsilon_h = \begin{cases} \sigma_h^2, & \text{si } CR \\ S_h^2, & \text{si } SR \end{cases}$$

La afijación es determinada en la práctica por

$$n_{h_0} = \begin{cases} n \frac{W_h \sigma_h / \sqrt{c_h}}{\sum_{h=1}^K W_h \sigma_h / \sqrt{c_h}}, & \text{si } CR \\ n \frac{W_h S_h / \sqrt{c_h}}{\sum_{h=1}^K W_h S_h / \sqrt{c_h}}, & \text{si } SR \end{cases}$$

Si los costos c_h son iguales tenemos la llamada fijación de Neyman.

$$n_{h_N} = \begin{cases} n \frac{W_h \sigma_h}{\sum_{h=1}^K W_h \sigma_h}, & \text{si } CR \\ n \frac{W_h S_h}{\sum_{h=1}^K W_h S_h}, & \text{si } SR \end{cases}$$

Ganancia en precisión debida a la estratificación

$$N\sigma^2 = \sum_{i=1}^K \sum_{j=1}^{N_i} (Y_{ij} - \bar{Y})^2 = \sum_{i=1}^K \sum_{j=1}^{N_i} [(Y_{ij} - \bar{Y}_i)^2 + (\bar{Y}_i - \bar{Y})^2 - 2(Y_{ij} - \bar{Y}_i)(\bar{Y}_i - \bar{Y})]$$

El primer término es

$$\sum_{i=1}^K \sum_{j=1}^{N_i} (Y_{ij} - \bar{Y}_i)^2 = \sum_{i=1}^K N_i \sigma_i^2$$

El segundo término mide la variación entre estratos

$$\sum_{i=1}^K \sum_{j=1}^{N_i} (\bar{Y}_j - \bar{Y})^2 = \sum_{i=1}^K N_i (\bar{Y}_j - \bar{Y})^2 = N\sigma_b^2$$

Entonces tenemos la ganancia debido a usar MAEAP

$$v(\bar{y}) = \frac{\sigma^2}{n} = \sum_{i=1}^K \frac{N_i}{nN} \sigma_i^2 + \frac{\sigma_b^2}{n} = V(\bar{y}_e|AP) + G(AP)$$

Por tanto el MAE con AP es más preciso que el uso del MASCR. A mayor diferencia entre las medias de los estratos, mayor es la ganancia.

Al usar AN hay una mayor precisión que la AP y esta crece con la diferencia entre las variaciones internas de los estratos

$$V(\bar{y}|AP) \pm V(\bar{y}|AN) = \frac{\left(\sum_{h=1}^K W_h S_h\right)^2}{n} + \left[\frac{\sum_{h=1}^K W_h S_h^2}{n} - \frac{\left(\sum_{h=1}^K W_h S_h\right)^2}{n} \right]$$

donde

$$\frac{\sum_{h=1}^K W_h S_h^2}{n} - \frac{\left(\sum_{h=1}^K W_h S_h\right)^2}{n} = \frac{\sum_{h=1}^K W_h (S_h - \bar{S})^2}{n}, \quad \bar{S} = \frac{\sum_{h=1}^K W_h S_h}{n}$$

Entonces

$$V(\bar{y}|AP) = V(\bar{y}|AN) + \frac{1}{n} \sum_{h=1}^K W_h (S_h - \bar{S})^2 = V(\bar{y}|AN) + G(AN)$$

Formación de estratos óptimos en el caso univariado

El caso del MAS.

Dalenius en la década del 50 del pasado siglo, se planteó el problema de construir estratos que garantizaran minimizar la varianza de la estimación de la media poblacional. Detalles de los resultados diversos derivados en esta línea pueden verse en Dalenius-Hodges (1959), Bethel (1989) y en libros como Cochran (1981).

Usando los criterios clásicos de la estadística paramétrica podemos decir que tenemos una fdp $f(t)$ tal que

$$W_i = Prob(U_i) = \int_{U_i} f(t)dt = \int_{y(i-1)}^{y(i)} f(t)dt$$

Consideramos que U_i está definido por cotas en la variable de interés Y . $Y \in [y(0), y(K)]$. Dada la continuidad $P(Y = y(i)) = 0$. Entonces tomando

$$M(i) = \bar{Y}_i/U_i \text{ es determinado por } [y(i-1), y(K)]$$

$$W_i M_i = \int_{U_i} t f(t) dt = \int_{y(i-1)}^{y(i)} t f(t) dt$$

$$W_i^2 \sigma_i^2 = \int_{y(i-1)}^{y(i)} t^2 f(t) dt - W_i M^2(i) \Rightarrow V(\bar{y}_e) = \sum_{i=1}^K \frac{\int_{y(i-1)}^{y(i)} t^2 f(t) dt - W_i M^2(i)}{n_i}$$

Por tanto, entonces al fijar los tamaños de muestra se puede plantear el problema de optimización

$$Min\{V(\bar{y}_e)\} = Min_{(y_0, \dots, y_K)} \left\{ \sum_{i=1}^K \left[\int_{y(i-1)}^{y(i)} t^2 f(t) dt - W_i M^2(i) \right] \right\}$$

Este es un problema de particionamiento en K subintervalos del intervalo $[y(0), y(K)]$.

Cuando usamos la AP

$$\frac{\partial V(\bar{y}_e)}{\partial y(i)} = -W_i M^2(i) - W_{i+1} M^2(i+1)$$

$$\frac{\partial W_i}{\partial y(i)} = f(y(i))$$

$$\frac{\partial W_{i+1}}{\partial y(i+1)} = -f(y(i))$$

$$\frac{\partial (W_i M_i)^2}{\partial y(i)} = 2W(i)M(i)y(i)f(y(i))$$

$$\frac{\partial (W_{i+1} M_{i+1})^2}{\partial y(i+1)} = -2W(i+1)M(i+1)y(i+1)f(y(i+1))$$

De ahí que

$$\frac{\partial \left[\int_{y(i-1)}^{y(i)} t^2 f(t) dt - W_i M^2(i) \right]}{\partial y(i)} = 0 \Rightarrow y(i) = \frac{M(i) - M(i-1)}{2}$$

El sistema de ecuaciones que se deriva es

$$\frac{(y(i) - M(i))^2 + \sigma_i^1}{\sigma_i} = \frac{(y(i+1) - M(i+1))^2 + \sigma_{i+1}^1}{\sigma_{i+1}}, \quad i = 1, \dots, K.$$

Actualmente su solución numérica parece no ser tan difícil pero aun continua siendo complicado obtener una solución adecuada, dada la dependencia de los parámetros de las cotas.

Una solución simple es la propuesta de Dalenius-Hodges (1959) de considerar que $f(y)$ es la densidad de la distribución uniforme en cada estrato. Entonces tendremos que si $f(i)$ es el valor de f , que es constante en $U_i : y$

$$W_i = \int_{y(i-1)}^{y(i)} f(t) dt \cong f(i)[y(i) - y(i-1)]$$

La varianza es

$$\sigma_i^2 \cong \left(\frac{y(i) - y(i-1)}{\sqrt{12}} \right)^2$$

Ahora podemos establecer que la integral de $\sqrt{f(x)}$ es aproximada por

$$\zeta(i) - \zeta(i-1) = \int_{y(i-1)}^{y(i)} \sqrt{f(t)} dt \cong f(i)[y(i) - y(i-1)]$$

Regresando a la expresión del error de muestreo podemos determinar que la relación a minimizar es la varianza del estimador, es decir

$$\sqrt{12} \sum_{i=1}^K W_i \sigma_i \cong \sum_{i=1}^K f(i)(y(i) - y(i-1))^2 = \sum_{i=1}^K (\zeta(i) - \zeta(i-1))^2$$

Esta suma es mínima si las diferencias son constantes. Entonces si hacemos la distribución acumulativa en $\sqrt{f(i)}$ y construimos intervalos de igual amplitud en función de Y tendremos unos estratos “aproximadamente óptimos”.

El Caso del Muestreo con probabilidades desiguales con reemplazo

Consideremos que tenemos una variable S independiente de Y con media \bar{S} y varianza σ_S^2 conocidas y que el reporte que se obtiene es $Z = SY$. Se usa la

variable X conocida para determinar diferentes probabilidades de selección dentro del estrato h

$$p_{i(h)} = \frac{X_i}{\sum_{i \in U_h} X_i}$$

El problema de la estimación correspondiente es caracterizado en el siguiente teorema

Teorema (estrategia PPXWR de Mahajan-Singh).

$$\bar{y}_{MS} = \frac{1}{N} \sum_{h=1}^K \frac{\sum_{i=1}^{n_h} \frac{Z_{i(h)}}{S_h p_{i(h)}}}{n_h}, \quad \bar{S}_h = \sum_{i \in U_h} \frac{S_i}{n_h}$$

es un estimador insesgado de \bar{Y} y su varianza es

$$V(\bar{y}_{MS}) = \frac{1}{N^2} \sum_{h=1}^K \frac{\vartheta_h^2 + \frac{\sigma_S^2}{\bar{S}^2} \sum_{i=1}^{n_h} \frac{Y_{i(h)}^2}{p_{i(h)}}}{n_h}, \quad \vartheta_h^2 = \sum_{i=1}^{n_h} \frac{Y_{i(h)}^2}{p_{i(h)}} - \bar{Y}^2$$

Demostración:

La prueba se sigue al utilizar el hecho de que tanto la varianza como la esperanza deben calcularse tomando en cuenta la dependencia de la distribución de S . Así

$$E(\bar{y}_{MS}) = E_1(E_2(y_{MS}/p)) = E_1 \left(\frac{1}{N} \sum_{h=1}^K E_2 \left(\frac{\sum_{i=1}^{n_h} \frac{Z_{i(h)}}{\bar{S}_h p_{i(h)}}}{n_h} / p \right) \right)$$

Como

$$E_2 \left(\frac{\sum_{i=1}^{n_h} \frac{Z_{i(h)}}{\bar{S}_h p_{i(h)}}}{n_h} / p \right) = \frac{\sum_{i=1}^{n_h} \frac{y_{i(h)}}{p_{i(h)}}}{n_h}$$

$$E(\bar{y}_{MS}) = E_1 \left(\frac{1}{N} \sum_{h=1}^K \frac{\sum_{i=1}^{n_h} \frac{y_{i(h)}}{p_{i(h)}}}{n_h} \right) = \bar{Y}$$

Para la varianza aplicamos la relación $V(\bar{y}_{MS}) = E_1(V_2(\bar{y}_{MS}/p)) + V_1(E(\bar{y}_{MS}/p))$. Es claro que

$$V_1(E_2(\bar{y}_{MS}/p)) = V_1 \left(\frac{1}{N} \sum_{h=1}^K \frac{\sum_{i=1}^{n_h} \frac{y_{i(h)}}{p_{i(h)}}}{n_h} \right) = \frac{1}{N^2} \sum_{h=1}^K \frac{\vartheta_h^2}{n_h}$$

Por otra parte

$$E_1(V_2(\bar{y}_{MS}/p)) = E_1 \left(\frac{1}{N^2} \sum_{h=1}^K \frac{\sum_{i=1}^{n_h} \frac{\sigma_{S_h}^2}{S_h^2} \times \frac{y_{i(h)}^2}{p_{i(h)}^2}}{n_h^2} \right) = \frac{1}{N^2} \sum_{h=1}^K \frac{\sigma_{S_h}^2}{n_h S_h^2} \sum_{i=1}^{n_h} \frac{y_{i(h)}^2}{p_{i(h)}}.$$

Consideremos el modelo superpoblacional $\xi : Y = b(X) + e$, donde e es una variable independiente con esperanza nula y varianza $q(x)$. Estratificando, usando la variable aleatoria X , definida en el intervalo $[x(0), x(K)]$ tendremos que

$$W_{X(i)} M(X(i)) = \int_{x(i-1)}^{x(i)} x f(x) dx, \quad W_{x(i)} = Prob(U_i) = \int_{U_i} f(x) dx = \int_{x(i-1)}^{x(i)} f(x) dx$$

Bajo este modelo se debe hallar el error esperado que es

$$E_\xi(V(\bar{y}_{MS})) \cong \sum_{h=1}^K W_{x(h)}^2 \frac{\mu_h \mu_{X(h)} \left(1 + \frac{\sigma_\xi^2}{S^2}\right) - \mu_{hr}^2}{n_h},$$

$$\mu_{(h)} = \frac{\int_{x(h-1)}^{x(h)} \frac{b^2(x)+q(x)}{x} f(x) dx}{W_{x(h)}}, \quad \mu_{X(h)} = \frac{\int_{x(h-1)}^{x(h)} x f(x) dx}{W_{x(h)}},$$

$$\mu_{Y(h)} = \mu_{b(h)} = \frac{\int_{x(h-1)}^{x(h)} b(x) f(x) dx}{W_{x(h)}}$$

pues

$$E_\xi(\vartheta_h^2) = E_1 E_2 \left(\sum_{i=1}^{n_h} \frac{(b(X_{i(h)}) + e_{i(h)})^2}{p_{i(h)}} - \left(\sum_{i=1}^{n_h} (b(X_{i(h)}) + e_{i(h)}) \right)^2 \right)$$

$$= E_1 \left(\sum_{i \neq j=1}^{N_h} \frac{(b(X_{i(h)})^2 + q(X_{i(h)}))}{X_{i(h)}} X_{j(h)} - \sum_{i \neq j=1}^{N_h} b(X_{i(h)}) b(X_{j(h)}) \right)$$

$$= N_h N_{h-1} (\mu_h \mu_{X(h)} - \mu_{hr}^2)$$

$$E_\xi \left(\sum_{i=1}^{n_h} \frac{y_{i(h)}^2}{p_{i(h)}} \right) = E_1 E_2 \left(\sum_{i=1}^{n_h} \frac{(b(X_{i(h)}) + e_{i(h)})^2}{p_{i(h)}} \right)$$

$$= E_1 \left(\sum_{i=1}^{n_h} \frac{b(X_{i(h)})^2 + q(X_{i(h)})}{X_{i(h)}} X_{j(h)} \right) = N_h N_{h-1} (\mu_h \mu_{X(h)})$$

Estos resultados soportan la demostración del siguiente teorema

Teorema (estrategia PPXWR de Mahajan-Singh Mahajan-Singh bajo ξ). Si se acepta que $\xi : Y_{i(h)} = b(X_{i(h)}) + e_{i(h)}$, $E(e_{i(h)}) = 0$, $V(e_{i(h)}) = q(x_{i(h)})$, $Cov(e_{i(h)}, e_{j(h)}) = 0$, si $i \neq j$, para todo $i = 1, \dots, N_h$, $N_h \cong N_{h-1}$, $h = 1, \dots, K$. Entonces

$$E_{\xi}(V(\bar{y}_{MS})) \cong \sum_{h=1}^K W_{X^{(h)}}^2 \frac{\mu_h \mu_{x(h)} \left(1 + \frac{\sigma_S^2}{S^2}\right) - \mu_{hr}^2}{n_h} = \sum_{h=1}^K W_{X^{(h)}}^2 \frac{\gamma_h^2}{n_h}$$

Para esta ξ -estrategia la afijación de Neyman es

$$n_{hN(MS)} = n \frac{W_h \gamma_h}{\sum_{h=1}^K W_h \gamma_h}$$

Por lo que se obtiene la expresión del error esperado bajo el modelo

$$E_{\xi}(V(\bar{y}_{MS})) \cong \frac{(\sum_{h=1}^K W_{X^{(h)}} \gamma_h)^2}{n}$$

Ahora el problema de la determinación óptima de las fronteras de los estratos se transforma en resolver el sistema

$$A(i) = A(h)$$

donde

$$A(Z) = \frac{\left(1 + \frac{\sigma_{S_Z}^2}{S_Z^2}\right) (\theta(x_z) \mu_{X(z)} + x(Z) \mu_{\theta(Z)}) - 2b(x(Z)) \mu_{b(Z)}}{\sqrt{\left(\mu_{X(Z)} \mu_{\theta(Z)} \left(1 + \frac{\sigma_{S_Z}^2}{S_Z^2}\right) - \mu_{b(Z)}\right)}}$$

$$Z = i, h; \quad i = h + 1, h = 1, \dots, K - 1.$$

$$\mu_{\theta(h)} = \frac{\int_{x^{(h-1)}}^{x^{(h)}} \frac{\theta(x) + q(x)}{x} f(x) dx}{W_{X^{(h)}}}$$

Una regla similar a la propuesta por Dalenius-Hodges (1959) para este caso es la derivada, ver Mahajan-Singh (2005), la que establece estratificar usando

$$\begin{aligned} \sqrt[3]{f_2(x)}, f_2(x) &= \frac{\tau(1) - 4\tau(2)\tau(3)}{3^{1/2}\sqrt{\tau(3)}} f(x), \\ \tau(1) &\left(q'(x) \left(1 + \left(1 + \frac{\sigma_{S_Z}^2}{S_Z^2} \right) \right) + b(x)b'(x) \left(1 + \frac{\sigma_{S_Z}^2}{S_Z^2} \right) \right), \\ \tau(2) &= \left(\theta'(x) \left(1 + \left(1 + \frac{\sigma_{S_Z}^2}{S_Z^2} \right) \right) - b^2(x) \right), \\ \tau(3) &= \left(q(x) \left(1 + \left(1 + \frac{\sigma_{S_Z}^2}{S_Z^2} \right) \right) + b^2(x) \left(1 + \frac{\sigma_{S_Z}^2}{S_Z^2} \right) \right). \end{aligned}$$

La ganancia en precios debido al uso de este método es

$$G(MS) \cong V(\bar{y}) - \frac{(\sum_{h=1}^K W_{X(h)} \gamma_h)^2}{n}.$$

4.4.4. El enfoque de la programación dinámica.

La programación dinámica (PD) es una de las técnicas de la programación matemática. Esta ha sido usada como método para construir estratos multivariados. Bühler y Deutler (1975) hicieron una propuesta, Bouza et al. (2009) la usaron en una propuesta de solución moderna. No hay una definición formal de ellas pero puede ser definida en forma general a partir de un problema secuencial de optimización. Una gran clase de estos problemas es el dado como

$$\min \sum_{t=1}^T \phi_t(Z_{t-1}, Z_t)$$

Sujeto a

$$Z_t = \tau_t(Z_{t-1}, \nu_t)$$

donde $Z_t \in \zeta_t, \nu_t \in D_t(Z_{t-1}), Z_0 = Z', t = 1, \dots, T$.

Este problema es resuelto en forma secuencial en T etapas. El estado en la etapa t es Z_t y ζ_t , es el espacio de los posibles estados. En dependencia del problema este estado será discreto o continuo. La variable ν_t es la variable de decisión que se asocia al espacio de decisiones $D_t(Z_{t-1})$. La descripción de la transición de los estados de una etapa $t - 1$ a una t es descrita por τ_t . Sea $\phi_t^*(Z_{t-1})$ el valor óptimo de la función objetivo $\phi_t(Z_{t-1})$ para la etapa t dado el estado Z_{t-1} . Estos son determinados usando del llamado Principio de

Optimalidad de Bellman (POB). Así se comienza en la etapa $t = T$ y se usa un método descendente hasta llegar a $t = 1$. El problema de optimización es resulta en forma recurrente a partir de la llamada Ecuación Funcional de Bellman (FEB):

$$\phi_t^*(Z_{t-1}) = \text{Min}_{\nu_t \in D_t(Z_{t-1})} \{ \phi_t(Z_{t-1}, Z_t) + \phi_{t+1}^*/Z_t = \tau_t(Z_{t-1}, \nu_t) \}$$

La solución óptima es el valor $\phi_1^*(Z_0)$ obtenido al final del proceso que es el valor optimal global de la función objetivo. Dado que se conocen los valores óptimos en cada etapa y para todos los estados al obtener la solución ν_1^* usando $\phi_1^*(Z_0)$ el vector optimal $\nu^{*T} = (\nu_1^*, \dots, \nu_T^*)$ es obtenido por inducción. En las aplicaciones que nos ocupan las áreas pueden estar definidas usando criterios administrativos que pueden relacionarse con las áreas de atención o con los niveles de infestación de mosquitos o con los de prevalencia del dengue. Digamos que haya M estratos definidos y que queramos subdividirle para obtener, estimaciones de los factores de interés. Sea K es el numero de tales particiones que se basan en la variable de interés Y un estrato es definido por un subintervalo del intervalo de definición de esta. Esto replantea el problema de la determinación óptima de estratos. Consideremos que en la subpoblación m -ésima, $P(m)$, de tamaño $N(m)$ se tiene una función de distribución, definida como $F_m(Y)$. La media de esta es denotada por $\mu(m)$ y su varianza por $\sigma^2(m)$. En general la distribución va a ser diferente para cada subpoblación. Sea $Q_m = N(m)/N$ el tamaño relativo de $P(m)$, $m = 1, \dots, M$. Entonces

$$\mu = \sum_{m=1}^M \frac{N(m)}{N} \mu(m)$$

Al hacer la partición de los $P(m)$, $m = 1, \dots, M$ en subpoblaciones U_i , $i = 1, \dots, K$ tenemos que

$$\mu = \sum_{m=1}^M Q_m \sum_{i=1}^K \frac{N_k(m)}{N(m)} \left(\frac{\sum_{j=1}^{N_k(m)} y_{mij}}{N_k(m)} \right) = \sum_{m=1}^M Q_m \sum_{i=1}^K W_k(m) \mu_k(m)$$

La estimación insesgada del mismo es obtenida, al usar MASCR por

$$\bar{x}_{2e} = \sum_{m=1}^M Q_m \sum_{i=1}^K W_k(m) \bar{y}_k(m), \quad \bar{y}_k(m) = \frac{\sum_{j=1}^{n_k(m)} Y_{mij}}{n_k(m)}$$

Tomando $W_{km} = Q_m W_k(m)$ podemos escribir

$$\bar{y}_{2e} = \sum_{m=1}^M \sum_{i=1}^K W_{km} \bar{y}_k(m)$$

La varianza asociada es

$$V(\bar{y}_{2e}) = \sum_{m=1}^M \sum_{i=1}^K W_{im}^2 V(\bar{y}_i(m)), \quad V(\bar{y}_i(m)) = \frac{\sum_{j=1}^{N_k(m)} (Y_{mij} - \mu_i(m))^2}{i(m)N_i(m)} = \frac{\sigma_{im}^2}{n_i(m)}$$

Ahora

$$\mu_{im} = \frac{1}{W_{im}} \int_{y(i-1)}^{y(i)} y dF_m(y)$$

$$\sigma_{im}^2 = \frac{1}{W_{im}} \int_{y(i-1)}^{y(i)} (y - \mu_i(m))^2 dF_m(y)$$

usando la afijación proporcional $n_i(m) = nW_{im}$.

El problema de PD para la estratificación óptima puede escribirse, ver Schneberger (1985), como

$$\min_{i=1}^K \left(\sum_{m=1}^M \phi_m(y(i-1), y(i)) \right)$$

Sujeto a

$$x(i) = x(i-1) + d(i), \quad i = 1, \dots, K,$$

$$Y \in [y(0), y(K)], d(i) \in (0, Y(K) - x(i-1)), \quad i = 1, \dots, K$$

Cada etapa se corresponde con el subintervalo i y el aspecto secuencial no es más que hacer la distribución del intervalo, en un cierto momento, de la porción $d(i)$ del dominio $(Y(0), Y(K))$. Después se debe determinar el FEB. Ahora $\phi_i^*(Z_{i-1})$ es el valor óptimo de la función objetivo para el estrato $U_i \cap P(m), i = 1, \dots, K, m = 1, \dots, M$. O sea

$$\phi_i^*(y(i-1)) = \min_{d(i) \in (0, b-y(i-1))} \left\{ \sum_{m=1}^M \phi_m(y(i-1), y(i)) \right. \\ \left. + \phi_{m+1}^*(y(m))/y(i) = y(i-1) + d(i) \right\}$$

Aunque el estado de decisión y el de la naturaleza son infinitos su solución es resuelta al moverse en forma discreta dentro de los intervalos, haciendo un compromiso entre el costo de computación y el nivel de precisión.

En trabajos posteriores consideraremos el problema estocástico siguiendo Krakenes y Moxnes (2005).

4.4.5. Estratos óptimos para estudios relacionados con el dengue.

El problema de la afijación en el caso multivariado ha recibido atención ocasional. Vea por ejemplo Khan y Ahsan (2003) y Kokan y Khan (1967), Skinne et al. (1994). Para evaluar el comportamiento de los estratos óptimos, construidos usando uno de los procedimientos aproximativos, se consideraron las dos afijaciones: la afijación proporcional (AP) o de Neyman (AN) con los datos correspondientes computamos

$$GP(AB) = \frac{G(AB)}{V(\bar{y})}$$

Se tomaron datos censales generados por dos investigaciones. Una sobre la infestación de mosquitos en West Bengal, India, en el 2001 y otra sobre la efectividad de los diagnósticos de dengue en el 2006 en el estado de Guerrero, México.

En el estado de Guerrero se analizaron 10891 casos de dengue por el Laboratorio Estatal de Salud Pública "Dr. Galo Soberón y Parra". Estos pertenecían a 80 municipios y su diagnóstico fue hecho por uno de los 7 sistemas que hicieron lo reportes. Ver Allende et al. (2008) y Bouza-Allende (2010). Estos definieron naturalmente 2 formas de agrupar administrativamente a los pacientes: por municipio y por sistema de salud. Se aplicó el método de la PD con cada uno de las divisiones correspondientes. Como variable se tomó la probabilidad predicha de diagnosticar correctamente al paciente. Estas se obtuvieron mediante un modelo logístico que fue ajustado con los datos. La variable auxiliar fue la temperatura.

Los resultados de la ganancia en precisión se brindan en la tabla (4.16).

Cuadro 4.16: Estudio en Guerrero, México. Ganancia en precisión relativa de las diferentes estratificaciones óptimas.

Ganancias en precisión respecto al MASCR

Casos	G(AP)	G(AN)	G(MS) con AP	G(MS) con AN	G(PD)
Dianosgósticos	0.193	0.199	0.257	0.272	–
Dianosgósticos (regiones)	0.197	0.201	0.204	0.215	0.275
Dianosgósticos Guerrero	0.074	0.066	0.159	0.190	–
Dianosgósticos Guerrero(Regiones)	0.185	0.08	0.163	0.167	0.199
Dianosgósticos Guerrero(sistemas)	0.069	0.065	0.037	0.071	0.163
Rango de consenso	4	5	3	2	1

Los resultados fijan claramente que el método basado en PD es el mejor de todos. Le sigue el uso del ajuste del modelo superpoblacional con afijación

de Neyman. Es de notar que la AN no es necesariamente mejor que la AP al usar la aproximación por la distribución uniforme.

En West Bengala se midió la prevalencia de mosquitos en 93 regiones construidas ad hoc para remedar la usada en Guerrero. En el estudio esta se midió en términos de $LC50(mgl/l) \times 10^4$. La variable auxiliar fue el área cubierta por embalses de agua. El uso de la estratificación óptima, dentro de las regiones, no arroja resultados significativamente diferentes del caso en que se derivan los estratos óptimos sin tomarlas en cuenta. Eso parece debido a lo no existencia de grandes diferencias regionales en la población de mosquitos. Sin embargo, en el estudio de las probabilidades de correcto diagnóstico del dengue si es notable la diferencia entre las regiones y los sistemas. Esto se valora como debido a que la precisión de estos depende mucho del sistema y de la zona donde se encuentran las unidades hospitalarias. Incluso la derivación de los EO's sin tomar en cuenta una subdivisión previa es mejor que considerar los sistemas.

Cuadro 4.17: Estudio en West Bengal, India. Ganancia en precisión relativa de las diferentes estratificaciones óptimas.

Ganancia en precisión relativa de las diferentes estratificaciones óptimas

Casos	G(AP)	G(AN)	G(MS) con AP	G(MS) con AN	G(PD)
Mosquitos	0.229	0.237	0.451	0.459	–
Mosquitos(regiones)	0.105	0.109	0.139	0.149	0.275
Mosquitos West Beangal	0.164	0.168	0.189	0.204	–
Mosquitos West Bengal(Regiones)	0.172	0.174	0.166	0.197	0.217
Mosquitos Guerrero(systems)	0.117	0.176	0.199	0.271	0.572
Rango de consenso	5	4	3	2	1

4.4.6. Conclusiones.

Estos resultados sugieren que:

1. En West Bengal el uso de la EO genera altas ganancias en precisión en la valoración de la población de mosquitos y por ende esto permitirá una mejor valoración de la infestación y con ello planear mejor las tares de saneamiento y prevención.
2. Es preferible usar la división por municipio en el estudio de los diagnósticos efectivos en Guerrero.
3. Es alta la ganancia en precisión debida al uso de los municipios en el estudio llevado a cabo en Guerrero.

4. En ambos casos el uso de la PD fue el mejor rankeado y le sigue el uso de la estratificaciones multivariadas con la AN.

Es de apuntar que lamentablemente en la mayor parte de las aplicaciones el uso de la PD es más costoso desde el punto de vista de la determinación de los EO.

Bibliografía

- [1] Allende, Sira, Carlos Bouza, Saúl López Silva y Agustín S. Moreno (2008): Análisis de la Efectividad de la diagnosis del dengue en el estado de Guerrero, México 2007: hechos estadísticos y Consenso. *Investigación Operacional* 29, 35-47
- [2] Bethel, J. (1989) Sample Allocation in Multivariate Surveys. *Survey Methodology*, 15, 47-57.
- [3] Bouza, Carlos N. y Sira Allende (2010): Aplicaciones Médicas de Índices del Tipo Kappa su Modelación. En *Investigación aplicada a la salud. Una mirada desde la Investigación de Operaciones*(J.F. Garcia y C.N. Bouza, Editores), . 69-85. UL-TRADIGITAL PRESS, S.A. México, D. F.
- [4] Bouza Herrera, Carlos N., Sira M. Allende Alonso, Eduardo Cairo Valcárcel y Luis Pedreira Andrade (2009): La afijación multicriterio de la muestra en el Modelo estratificado Multivariado. (EDS. J. C. LEYVA, E. AVILES Y J.J. ZEPEDA) PLAZA Y VALDÉS, SPAIN, 2009,383-404.
- [5] Bühler, A, y Deutler, M. (1975): Optimal stratification and grouping by Dynamic Programming. *Metrika*, 22, 161-175.
- [6] Cochran, W.G. (1981): *Técnicas del Muestreo*, CECSA, Mexico
- [7] Dalenius, T. y J. L. Hodges (1959): Minimum variance stratification. *J. Amer. Statist. Assoc.* 54, 88-101.
- [8] Khan M.G.M. and M. J. Ahsan (2003): A note on optimum allocation in multivariate stratified sampling. *S. Pac. J. Nat. Sci.*, 21, 91- 95.
- [9] Krakenes, A. and Moxnes, E. (2005): SOPS - A Tool to Find Optimal Policies in Stochastic Dynamic Systems. WPSD 2/05. System Dynamics Group, University of Bergen. (presented at The International System Dynamics Conference 2005, Boston)
- [10] Kokan, A.R., and Khan, S.U., (1967). Optimum Allocation in Multivariate Surveys: An Analytical Solution. *J.Roy. Stat. Soc., Ser. B*, 29, 115-125
- [11] Lavallée, P (1987): Some contributions to optimum stratification. Tesis de MSC., Carleton University, Ottawa.

- [12] Mahajan, P. K. and R. Singh (2005): Optimum stratification for scrambled response in pps sampling. *Metron*, LXIII, 103-114.
- [13] Salazar Gonzalez, Juan J- (2001): *Programación Matemática*. Ed. Diaz Santos. Madrid.
- [14] Scheneberger, H-(1985): Maxima, minima und Sattelpunkte Bei Optimaler Schichtung und Optimaler Aufteilung. *Allgemeines Statis. ARCHIV.*, 69, 286-297.
- [15] Singh, R. (1971): Approximately optimum stratification on the auxiliary variable. *J. Amer. Statist. Assoc.* 68, 829-833.
- [16] Skinner, C. J. D. J. Holmes and D. Holt (1994): Multiple Frame Sampling for Multivariate Stratification. *International Statistical Review*, 62, 333-347

Capítulo 5

Estudios prospectivos sobre dengue

Agustín Santiago¹, Antonio Arcos², María del Mar Rueda², Silvia González³

1. Universidad Autónoma de Guerrero, México.

2. Universidad de Granada, España.

3. Universidad de Jaén, España

Resumen: Los estimadores de áreas pequeñas para variables cuantitativas, han demostrado su eficiencia en distintas aplicaciones debido a la naturaleza de las variables de estudio y a las variables auxiliares involucradas. En este trabajo estudiamos estimadores de razón y regresión para la estimación de proporciones en dominios, incorporando información auxiliar en forma de totales o proporciones en la fase de estimación. La ventaja de los estimadores propuestos, frente a los métodos de estimación indirecta basada en modelos bajo el enfoque clásico o bayesiano, predominantes en la teoría estadística de las últimas décadas para la estimación en áreas pequeñas, es que se obtienen buenas estimaciones con la información auxiliar disponible en forma de totales o proporciones, provenientes de censos, estudios previos o registros administrativos, sin demandar el uso de información completa. Además, se ha comprobado la eficiencia de los estimadores propuestos mediante un estudio de simulación.

Abstract: The estimators for small areas and quantitative variables are very efficient in several applications due to the character of the interest and auxiliary variables implied. In this paper, we consider a ratio estimator and a regression estimator for the estimation of proportions in domains, adding auxiliary data in the form of totals or proportions in the estimation phase. The advantage of the proposed estimators in regards to the indirect estimation

methods based in classic or Bayesian models is that we obtain good estimations with the auxiliary data in the form of totals or proportions, coming from census, previous studies or administrative records, without the need of the full data. In addition to this, we have proof its efficiency in a simulation study.

Key words: información auxiliar, dominios, estimación basada en el diseño, estimación de proporciones, small area estimation.

5.1. Estimación de proporciones en áreas pequeñas

5.1.1. Introducción

Uno de los objetivos de la estimación en áreas pequeñas, consiste en proponer estimadores, que sin aumentar el tamaño de muestra, permitan obtener buenas estimaciones, es decir, el error de estimación sea menor que el de los estimadores conocidos. Hasta ahora, en el contexto de la estimación en áreas pequeñas, se han desarrollado distintos métodos de estimación, bajo el diseño y bajo el modelo, para abordar este problema. La estimación basada en modelos, que ligan la información auxiliar a la variable de interés con la cual se supone relacionada, ha demostrado ser más eficiente que la estimación basada en el diseño, si es que se cuenta con información auxiliar completa, es decir, se observan la variable respuesta y un vector de p variables auxiliares que son conocidas para todos los individuos de la población. Desgraciadamente, en la práctica este supuesto no es muy común, siendo más plausible que los datos asociados a las variables auxiliares se obtengan de censos y ficheros administrativos que proporcionan diferentes parámetros de estas variables auxiliares. Así, es común disponer de diversas medidas de posición (medias, medianas, momentos, etc.) pero es difícil tener acceso a los datos originales de cada individuo, fundamentalmente por motivos de privacidad o cuando se trata de información relacionada con políticas públicas. En este contexto es en el que proponemos la aplicación de un conjunto de estimadores de proporciones para áreas pequeñas, para la estimación de características de interés en los sectores de salud, educación y administración pública. En la siguiente sección desarrollamos una aplicación para datos de dengue, en el estado de Guerrero, México.

5.1.2. Estimación con datos de dengue

Los datos

La aplicación de los estimadores propuestos a un problema con datos reales, se hace con una base de datos de casos confirmados de dengue por el laboratorio estatal de salud pública del Estado de Guerrero (LESPEG), México (2006). La información contiene registros de los pacientes diagnosticados con dengue y características asociadas a los síntomas de un cuadro típico de dengue, organizados para las 7 regiones del estado. La tabla (5.1) muestra los tamaños poblacionales de esta regiones, que serán consideradas como dominios. Como variable principal consideraremos el tipo de dengue sufrido por el paciente (clásico o hemorrágico), donde $Y = 1$ si el paciente presenta dengue clásico y $Y = 0$, si presenta dengue hemorrágico. Para la selección de la variable auxiliar, tomamos en consideración la relación entre la variable principal y las distintas variables auxiliares, a partir de un modelo de regresión logística binaria. Las variables que mejor clasifican a los pacientes son sexo, enfermos similares en la localidad, cefalea, mialgias, artralgias, vómito, náuseas, dolor abdominal, diarrea, congestión nasal, escape de líquidos, hemorragias, petequias y melena. De entre las variables con mayor correlación, seleccionamos artralgias y realizamos las estimaciones con una muestra disponible de tamaño 300. La metodología, en este caso, al tratarse de una muestra real, consiste en estimar el error cuadrático medio de cada uno de los estimadores mediante Jackknife. Posteriormente, definimos la eficiencia relativa tomando como base el estimador directo.

Cuadro 5.1: Tamaños poblacionales de la población “dengue” en las regiones.

Región	Tamaño poblacional.
1. Tierra caliente	40
2. Zona norte	211
3. Zona centro	160
4. Montaña	412
5. Costa grande	354
6. Costa chica	74
7. Acapulco	964

Resultados

Seleccionada la muestra, encontramos que todas las submuestras correspondientes a las regiones eran no vacías. Los tamaños de las submuestras para cada dominio fueron $n_1 = 2$, $n_2 = 37$, $n_3 = 18$, $n_4 = 52$, $n_5 = 46$, $n_6 = 7$

y $n_7 = 138$. El dominio 1 solo tiene casos de dengue clásico y el dominio 6 tiene de los dos tipos, pero con solo 5 casos de dengue hemorrágico. Centramos nuestro estudio en el dominio 6, que es uno de los casos extremos que se pueden presentar en la selección de la muestra, toda vez que todos los valores correspondientes a la variable respuesta en la submuestra, resultaron ser unos, es decir, todos los pacientes en la submuestra están infectados de dengue clásico. Existen muchos casos en la práctica en los que podemos encontrar esta situación debido a la ocurrencia de factores relacionados a las características específicas del dominio, tales como una proporción a nivel dominio muy cercana a 1. En este caso, el uso de estimadores de regresión logística no es factible. Por otra parte, cabe esperar que el estimador base, que solo utiliza la información muestral de dominio, no ofrezca estimaciones precisas. La tabla (5.2) muestra las eficiencias relativas de los estimadores propuestos.

Como puede observarse la mayoría de los estimadores propuestos mejoran al estimador base con respecto a la eficiencia; en el caso de los estimadores $\widehat{P}_{regA_d}^{(c_4)}$ y $\widehat{P}_{regA_d}^{(c_5)}$, la ganancia en eficiencia alcanza el 56.34 %. De especial interés para nosotros es el estimador $\widehat{P}_{rA_d}^{(1)}$, toda vez que lo clasificamos como eficiente, y con esta muestra mejora la eficiencia en un 47.29 %, que sabemos usa información auxiliar a nivel de dominio. Los estimadores $\widehat{P}_{rA_d}^{(c_1)}$ y $\widehat{P}_{regA_d}^{(c_1)}$, aunque su eficiencia no es tan grande como en la población simulada, aquí habría que destacar que, aún con este caso extremo de muestra y con submuestras por dominio muy pequeñas, son ligeramente más eficientes que el estimador directo.

Esta situación nos hace deducir que precisamente la inclusión de la razón entre la proporción poblacional y muestral de la variable auxiliar en el dominio es lo que mejora la estimación. En resumen, podemos concluir, que el problema inicial de tener una submuestra en el dominio compuesta solo por unos, puede solventarse incluyendo en las estimaciones la información proporcionada por la variable auxiliar.

Cuadro 5.2: Eficiencia relativa estimada de los estimadores para la población “dengue” en el dominio U_6 (Costa chica).

Población U de tamaño $N = 2215$, con $n = 300$.

	Estimador	Eficiencia relativa
Directo	\widehat{P}_{A_d}	100.00
Razón	$\widehat{P}_{rA_d}^{(1)}$	144.29
	$\widehat{P}_{rA_d}^{(2)}$	147.16
	$\widehat{P}_{rA_d}^{(3)}$	8.41
Regresión	$\widehat{P}_{regA_d}^{(1)}$	108.06
	$\widehat{P}_{regA_d}^{(2)}$	108.48
	$\widehat{P}_{regA_d}^{(3)}$	12.45
Comb. Razón	$\widehat{P}_{rA_d}^{(c_1)}$	100.73
	$\widehat{P}_{rA_d}^{(c_2)}$	100.77
	$\widehat{P}_{rA_d}^{(c_3)}$	5.03
	$\widehat{P}_{rA_d}^{(c_4)}$	105.61
	$\widehat{P}_{rA_d}^{(c_5)}$	105.61
Comb. Regresión	$\widehat{P}_{regA_d}^{(c_1)}$	100.17
	$\widehat{P}_{regA_d}^{(c_2)}$	100.17
	$\widehat{P}_{regA_d}^{(c_3)}$	5.03
	$\widehat{P}_{regA_d}^{(c_4)}$	156.34
	$\widehat{P}_{regA_d}^{(c_5)}$	156.34

Estimación de proporciones en dominios

Se usaron los estimadores con mejores propiedades en la estimación de proporciones de pacientes con dengue clásico por regiones. La tabla 5.3 permite observar las proporciones estimadas para cada dominio. La estimación de proporciones en el dominio 1, que corresponde a la región Tierra caliente, no se ha realizado, toda vez que la muestra de dominio es demasiado pequeña ($n_1 = 2$) y todos los casos fueron confirmados con dengue clásico.

Cuadro 5.3: Estimación de proporciones y varianzas por regiones.

	Estimador	p	$\hat{\sigma}^2$
Tierra caliente	$\hat{P}_{rA_d}^{(1)}$	NA	NA
	$\hat{P}_{regA_d}^{(1)}$	NA	NA
	$\hat{P}_{rA_d}^{(c1)}$	NA	NA
	$\hat{P}_{regA_d}^{(c1)}$	NA	NA
Zona norte	$\hat{P}_{rA_d}^{(1)}$	0.800	0.007
	$\hat{P}_{regA_d}^{(1)}$	0.802	0.003
	$\hat{P}_{rA_d}^{(c1)}$	0.841	0.002
	$\hat{P}_{regA_d}^{(c1)}$	0.832	0.002
Zona centro	$\hat{P}_{rA_d}^{(1)}$	0.791	0.009
	$\hat{P}_{regA_d}^{(1)}$	0.792	0.005
	$\hat{P}_{rA_d}^{(c1)}$	0.838	0.003
	$\hat{P}_{regA_d}^{(c1)}$	0.827	0.002
Montaña	$\hat{P}_{rA_d}^{(1)}$	0.994	0.004
	$\hat{P}_{regA_d}^{(1)}$	0.991	0.002
	$\hat{P}_{rA_d}^{(c1)}$	0.984	0.000
	$\hat{P}_{regA_d}^{(c1)}$	0.999	0.000
Costa grande	$\hat{P}_{rA_d}^{(1)}$	0.878	0.003
	$\hat{P}_{regA_d}^{(1)}$	0.883	0.001
	$\hat{P}_{rA_d}^{(c1)}$	0.910	0.001
	$\hat{P}_{regA_d}^{(c1)}$	0.902	0.000
Costa chica	$\hat{P}_{rA_d}^{(1)}$	0.925	0.008
	$\hat{P}_{regA_d}^{(1)}$	0.923	0.004
	$\hat{P}_{rA_d}^{(c1)}$	0.916	0.004
	$\hat{P}_{regA_d}^{(c1)}$	0.918	0.002
Acapulco	$\hat{P}_{rA_d}^{(1)}$	0.494	0.002
	$\hat{P}_{regA_d}^{(1)}$	0.488	0.002
	$\hat{P}_{rA_d}^{(c1)}$	0.500	0.001
	$\hat{P}_{regA_d}^{(c1)}$	0.496	0.001

Bibliografía

- [1] Chandra, H., Chambers, R. and Salvati, N., Small area estimation of proportions in Business Survey. Center for Statistical and Survey Methodology, Working paper, 15, (2009).
- [2] Datta (2009), G.S., Sample Surveys: Inference and Analysis, Vol. 29B *in Handbook of Statistics Vol. 29B* D. Pfeffermann and C.R. Rao (editors), The Netherlands: North-Holland.
- [3] Gonzalez-Manteiga, W., Lombardia, M. J., Molina, I., Morales, D. and Santamaria, L., Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model, *Computational Statistics and Data Analysis*, 51, 2720-2733, (2007)
- [4] Malec, D. Sedransk, J., Moriarity, C. L., and LeClere, F. B., Small Area Inference for Binary Variables in National Health Interview Survey, *Journal of the American Statistical Association*, 92, 815-826, (1997)
- [5] Muñoz, J.F., Arcos, A., Álvarez-Verdejo, E., Rueda, M, Martínez-Puertas, S. (2010). Estimators and confidence intervals for the proportion using auxiliary information with applications to the estimation of prevalences. *Journal of Biopharmaceutical Statistics* 1, 1-32
- [6] Rao, J. N. K., *Small Area Estimation*, New York: Wiley, (2003)
- [7] Särndal, C. E., Design-consistent versus model-dependent estimation for small domains, *Journal of the American Statistical Association* 79, 624-631, (1984)
- [8] Särndal, C. E. and Hidiroglou, M.A., Small Domain Estimation: A Conditional Analysis, *Journal of the American Statistical Association*, 84, 266-275, (1989)
- [9] Särndal, C. E., Design-consistent versus model-dependent estimation for small domains, *Journal of the American Statistical Association* 79, 624-631, (1984)
- [10] C.E. Särndal and A. Veijanen, Model calibration and generalized regression estimation for domains and small areas, Invited paper, SAE2009, 2009

5.2. Modelos para mortalidad y tasas enfermedad en el estado de Guerrero.

Grupo de investigación: Agustín Santiago¹, Antonio Arcos², María del Mar Rueda², Silvia González³, Carlos N. Bouza⁴, José Maclovio Sautto¹, Sira Allende⁴, Dante Covarrubias¹, José Legorreta¹

1. Universidad Autónoma de Guerrero, México. 2. Universidad de Granada, España. 3. Universidad de Jaén, España, 4. Universidad de la Habana, Cuba.

5.2.1. Introducción.

El presente trabajo constituye una selección de métodos que se han diseñado de forma explícita para la estimación de proporciones en áreas pequeñas, ver por ejemplo, Rao, (2003) y Florentino García Santos: In Memoriam, 2011, págs.177-184. Éstos métodos y los que hemos propuesto en la tesis “Aportaciones a la estimación en áreas pequeñas: estimación de proporciones” (A. Santiago, 2011), constituirá el conjunto de métodos que nos permitirán desarrollar distintos proyectos de investigación en el área de la salud y proponer soluciones a problemas que cobran relevancia debido a las condiciones de marginación y pobreza en que vive un importante segmento de la población del estado de Guerrero, México. Además de problemas de salud, con esta metodología es posible abordar otros problemas, como puede percibirse de los posibles proyectos de investigación que pueden derivarse del presente trabajo, y que se proponen al final.

La mortalidad y las tasas de enfermedad de áreas pequeñas en una región o un país se usan con frecuencia para construir mapas de enfermedad, como por ejemplo, mapas del cáncer. Tales mapas se usan para visualizar la variabilidad geográfica de una enfermedad e identificar áreas de proporciones altas, garantizando así la intervención. Un modelo simple de área pequeña se obtiene asumiendo que las frecuencias observadas de área pequeña, $y_d = \sum_{U_d} A_j$, son variables independientes Poisson con media condicional $E(y_d|\lambda_d) = n_d\lambda_d$ y que $\lambda_d \stackrel{\text{iid}}{\sim} \text{gamma}(\alpha, \nu)$. Aquí λ_d y n_d son la tasa verdadera y el número de expuestos en el área d , y (α, ν) son los parámetros de escala y forma de la distribución gamma. Bajo este modelo, se obtienen estimadores suavizados de λ_d usando los métodos de BE o BJ (Clayton y Kaldor, 1987; Datta, Ghosh y Waller, 2000). También se han propuesto modelos espaciales autoregresivos condicionales (CAR) sobre el log de la razón $\theta_d = \log(\lambda_d)$ (Klayton y Kaldor, 1987). El modelo sobre λ_d puede extenderse para incorporar covariables de nivel de área \mathbf{z}_d , por ejemplo, $\theta_d = \mathbf{z}_d^T \boldsymbol{\beta} + v_d$ con $v_d \stackrel{\text{iid}}{\sim} N(0, \sigma_v^2)$. Nandram, Sedransk y Pickle (1999), estudiaron los modelos de regresión sobre el log de las

tasas de edad específica $\theta_{dj} = \log \lambda_{dj}$ incluyendo pendientes aleatorias, donde j denota edad. Pueden modelarse también tasas de mortalidad conjunta de las frecuencias observadas en dos sitios diferentes (y_{1d}, y_{2d}) , asumiendo que las (y_{1d}, y_{2d}) son condicional e independientemente ditiibuídas sobre $(\lambda_{1d}, \lambda_{2d})$ y $\theta_d = (\log \lambda_{1d}, \log \lambda_{2d})^T \overset{\text{iid}}{\sim} N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Además, y_{1d} y y_{2d} se asumen como variables de Poisson condicionalmente independientes con $E(y_{1d}|\lambda_{1d}) = n_{1d}\lambda_{1d}$ y $E(y_{2d}|\lambda_{2d}) = n_{2d}\lambda_{2d}$. Como un ejemplo de este modelo bivariado, y_{1d} y y_{2d} denotan el número de muertes debidas al cáncer en los sitios 1 y 2 y (n_{1d}, n_{2d}) la población en riesgo en los sitios 1 y 2. DeSouza (1992), mostró que el modelo bivariado conduce a estimadores mejorados de las tasas $(\lambda_{1d}, \lambda_{2d})$ comparado a estimadores basados en modelos univariantes separados.

5.2.2. Modelos de la familia exponencial

Ghosh, Natarajan, Stroud y Carlin (1998), propusieron modelos lineales generalizados con efectos aleatorios de área. Los estadísticos muestrales de las observaciones en el dominio $A_{dj} (j = 1, \dots, n_d; d = 1, \dots, D)$ condicionales sobre los θ_{dj} , se asumen independientemente distribuidos con función de densidad de probabilidad que pertenece a la familia exponencial con parámetros canónicos θ_{dj} , es decir,

$$f(A_{dj}|\theta_{dj}) = \exp \left[\frac{1}{\phi_{dj}} (\theta_{dj} A_{dj} - a(\theta_{dj})) + b(A_{dj}, \phi_{dj}) \right] \quad (5.2.1)$$

para $\phi_{dj} (> 0)$ y funciones conocidas $a(\cdot)$ y $b(\cdot)$. La familia exponencial (5.2.1) cubre distribuciones conocidas incluyendo las distribuciones normal, binomial y Poisson. Por ejemplo, $\theta_{dj} = \text{logit}(p_{dj})$ y $\phi_{dj} = 1$ si A_{dj} es Binomial (n_{dj}, p_{dj}) , y $\theta_{dj} = \log(\lambda_{dj})$ y $\phi_{dj} = 1$ si A_{dj} es Poisson (λ_{dj}) . Los θ_{dj} son modelados como

$$\theta_{dj} = \mathbf{B}_{dj}^T \boldsymbol{\beta} + v_d + u_{dj}, \quad (5.2.2)$$

donde v_d y u_{dj} son mutuamente independientes con $v_d \overset{\text{iid}}{\sim} N(0, \sigma_v^2)$ y $u_{dj} \overset{\text{iid}}{\sim} N(0, \sigma_u^2)$. El objetivo aquí es hacer inferencias sobre los parámetros θ_{dj} del área pequeña. Por ejemplo, $\theta_{dj} = \text{logit}(p_{dj})$ y p_{dj} denota la proporción asociada con la variable binaria en la j -ésima categoría de edad y sexo en la región d . Ghosh, Natarajan, Waller y Kim (1999), extendieron el modelo de enlace (5.2.2) para manejar datos espaciales, y aplicaron el modelo del mapa de la enfermedad.

5.2.3. Modelos semiparamétricos

También se han propuesto modelos semiparamétricos basados solamente en la especificación de los dos primeros momentos de las respuestas A_{dj} , condi-

cionadas sobre la media de área pequeñas μ_d , y de los μ'_d s. En la ausencia de covariables, Ghosh y Lahiri (1987), asumen el siguiente modelo:

- Para cada d , condicionada sobre los θ_d , los A_{dj} son iid con media θ_d y varianza $\mu_2(\theta_{dj})$, denotada $A_{dj}|\theta_d \stackrel{\text{iid}}{\sim}(\theta_d, \mu_2(\theta_d)), j = 1, \dots, N_d; d = 1, \dots, D$;
- $\theta_d \stackrel{\text{iid}}{\sim}(\mu, \sigma_v^2)$;
- $0 < \sigma_e^2 = E\mu_2(\theta_i) < \infty$.

Raghunathan (1993), incorporó información de covariables de nivel de área \mathbf{z}_d como sigue:

- condicionada sobre θ_d , $A_{dj} \stackrel{\text{iid}}{\sim}(\theta_d, b_1(\phi, \theta_d, a_{dj}))$ donde $b(\cdot)$ es una función positiva conocida de un parámetro de “dispersión” ϕ , medias de área pequeña θ_d y constantes conocidas a_{dj} ;
- $\theta_d \stackrel{\text{iid}}{\sim}(\tau_d = h(\mathbf{z}_d^T \boldsymbol{\beta}), b_2(\psi, \tau_d, a_d))$ donde $h(\cdot)$ es una función conocida y $b_2(\cdot)$ es una función positiva conocida de un parámetro de “dispersión” ψ , la media τ_d y una constante conocida a_d .

El modelo “longitudinal” con covariables de nivel unidad puede ser generalizado fijando

$$E(A_{dj}|\mathbf{v}_d) = \mu_{dj}, \quad V(A_{dj}|\mathbf{v}_d) = \phi b(\mu_{dj}) \quad (5.2.3)$$

y

$$h(\mu_{dj}) = \mathbf{B}_{dj1}^T \boldsymbol{\beta} + \mathbf{B}_{dj2}^T \mathbf{v}_d, \quad \mathbf{v}_d \stackrel{\text{iid}}{\sim}(\mathbf{0}, \boldsymbol{\Sigma}_v); \quad (5.2.4)$$

es decir, \mathbf{v}_d son independientes e idénticamente distribuidas con media $\mathbf{0}$ y matriz de covarianzas $\boldsymbol{\Sigma}_v$ (Breslow y Claiton, 1993).

5.2.4. Modelos para datos binarios

En esta subsección describiremos modelos de nivel unidad, que han sido pensados especialmente para respuestas binarias, A_{dj} , es decir, $A_{dj} = 1$ o 0 . En este caso, los modelos lineales mixtos no son apropiados y se han propuesto modelos alternativos. Si todas la covariables \mathbf{B}_{dj} asociadas con A_{dj} son de área específica, es decir $\mathbf{B}_{dj} = \mathbf{B}_d$, entonces podemos transformar las proporciones de la muestra de área, $\widehat{P}_{A_d} = \sum_j A_{dj}/n_d = y_d/n_d$, usando una transformación arco seno y el modelo se reduce a un modelo de nivel área. No obstante, los estimadores transformados resultantes, $\widehat{\theta}_d$, pueden no satisfacer el modelo de muestreo con errores medios de muestreo cero si n_d es pequeña.

Modelar la unidad de nivel evita las dificultades anteriores, y pueden obtenerse estimadores BE de proporciones directamente del caso general de covariables de unidad específica.

Modelos sin covariables

En este caso se asume un modelo de dos fases sobre las observaciones muestrales A_{dj} , $j = 1, \dots, n_d$; $d = 1, \dots, D$. En la primera fase, asumimos que $A_{dj}|p_d \stackrel{\text{iid}}{\sim} \text{Bernoulli}(p_d)$, $d = 1, \dots, D$. En la segunda fase se asume un modelo que liga los p_d ; en particular, $p_d \stackrel{\text{iid}}{\sim} \text{beta}(\alpha, \beta)$; $\alpha > 0$, $\beta > 0$ donde $\text{beta}(\alpha, \beta)$ denota la distribución beta con parámetros α y β :

$$f(p_d|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p_d^{\alpha-1} (1 - p_d)^{\beta-1}; \quad \alpha > 0, \beta > 0 \quad (5.2.5)$$

donde $\Gamma(\cdot)$ es la función gamma. Reducimos el vector de frecuencias observadas $\mathbf{y}_d = (y_{d1}, \dots, y_{dn_d})^T$, al total de frecuencias observadas en la muestra de dominio $y_d = \sum_j A_{dj}$, notando que y_d es un estadístico minimal suficiente para el modelo de la primera fase.

Nótese que $y_d|p_d \stackrel{\text{ind}}{\sim} \text{Binomial}(n_d, p_d)$, es decir,

$$f(y_d|p_d) = \binom{n_d}{y_d} p_d^{y_d} (1 - p_d)^{n_d - y_d}. \quad (5.2.6)$$

Se sigue de (5.2.5) y (5.2.6) que $p_d|y_d, \alpha, \beta \stackrel{\text{ind}}{\sim} \text{beta}(y_d + \alpha, n_d - y_d + \beta)$. Por lo tanto, el estimador de bayes de p_d y la varianza a posteriori de p_d están dadas por

$$\hat{p}_d^B(\alpha, \beta) = E(p_d|y_d, \alpha, \beta) = (y_d + \alpha)/(n_d + \alpha + \beta) \quad (5.2.7)$$

y

$$V(p_d|y_d, \alpha, \beta) = \frac{(y_d + \alpha)(n_d - y_d + \beta)}{(n_d + \alpha + \beta + 1)(n_d + \alpha + \beta)^2}. \quad (5.2.8)$$

Nótese que la distribución de enlace, $f(p_d|\alpha, \beta)$, es una “a priori conjugada” en el sentido que la distribución a posteriori, $f(p_d|y_d, \alpha, \beta)$, tendrá la misma forma que la distribución a priori.

Obtenemos estimadores de los parámetros del modelo de la distribución marginal: $y_d|\alpha, \beta \stackrel{\text{ind}}{\sim} \text{Beta-binomial}$ dada por

$$f(y_d|\alpha, \beta) = \binom{n_d}{y_d} \frac{\Gamma(\alpha + y_d)\Gamma(\beta + n_d - y_d)}{\Gamma(\alpha + \beta + n_d)} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}. \quad (5.2.9)$$

Los estimadores de máxima verosimilitud (MV), $\hat{\alpha}_{MV}$ y $\hat{\beta}_{MV}$ pueden obtenerse maximizando la log-verosimilitud:

$$l(\alpha, \beta) = \text{const} + \sum_{i=1}^D \left[\sum_{h=0}^{y_d-1} \log(\alpha + h) + \sum_{h=0}^{n_d-y_d-1} \log(\beta + h) - \sum_{h=0}^{n_d-1} \log(\alpha + \beta + h) \right], \quad (5.2.10)$$

donde $\sum_{h=0}^{y_d-1} \log(\alpha + h)$ se toma como cero si $y_d = 0$ y $\sum_{h=0}^{n_d-y_d-1} \log(\beta + h)$ se toma como cero si $y_d = n_d$. Una representación conveniente es en términos de la media $E(A_{dj}) = \mu = \alpha/(\alpha + \beta)$ y $\tau = 1/(\alpha + \beta)$ la cual está relacionada con la correlación intraclase $\rho = \text{Corr}(A_{dj} - A_{dk}) = 1/(\alpha + \beta + 1)$ para $j \neq k$. Usando μ y τ , (5.2.10) toma la forma

$$l(\mu, \tau) = \text{const} + \sum_{i=1}^D \left[\sum_{h=0}^{y_d-1} \log(\mu + h\tau) + \sum_{h=0}^{n_d-y_d-1} \log(1 - \mu + h\tau) - \sum_{h=0}^{n_d-1} \log(1 + h\tau) \right]. \quad (5.2.11)$$

Las expresiones explícitas para $\hat{\alpha}_{MV}$ y $\hat{\beta}_{MV}$ (ó $\hat{\mu}_{MV}$ y $\hat{\tau}_{MV}$) no existen, pero pueden obtenerse estimaciones MV por el método de Newton-Raphson o algún otro método iterativo.

Podemos usar también estimadores simples por el método de momentos de α y β . Igualamos la proporción muestral ponderada $p = \sum_d (n_d/n_T) \hat{p}_d$ y la varianza muestral ponderada $s_p^2 = \sum_d (n_d/n_T) (\hat{p}_d - \hat{p})^2$ a sus valores esperados y resolvemos las ecuaciones de momentos resultantes para α y β , donde $n_T = \sum_d n_d$. Esto lleva a estimadores de momentos, $\hat{\alpha}$ y $\hat{\beta}$, dados por

$$\hat{\alpha}/(\hat{\alpha} + \hat{\beta}) = \hat{p} \quad (5.2.12)$$

y

$$\frac{1}{\hat{\alpha} + \hat{\beta} + 1} = \frac{n_T s_p^2 - \hat{p}(1 - \hat{p})(D - 1)}{\hat{p}(1 - \hat{p})[n_T - \sum_d n_d^2/n_T - (D - 1)]}; \quad (5.2.13)$$

ver Kleinman (1973). Nótese que los estimadores de momentos no son únicos, a diferencia de los estimadores MV.

Sustituimos los estimadores de momentos $\hat{\alpha}$ y $\hat{\beta}$ en (5.2.7) para obtener un estimador BE de p_d como

$$\hat{p}_d^{BE} = \hat{p}_d^B(\hat{\alpha}, \hat{\beta}) = \hat{\gamma}_d \hat{p}_d + (1 - \hat{\gamma}_d) \hat{p}, \quad (5.2.14)$$

donde $\hat{\gamma}_d = n_d/(n_d + \hat{\alpha} + \hat{\beta})$. Observe que \hat{p}_d^{BE} es un estimador combinado del estimador directo \hat{p}_d y el estimador sintético \hat{p} y se da más peso a \hat{p}_d conforme el tamaño de muestra de área, n_d , se incrementa. Por lo tanto es similar al estimador de Fay-Herriot para el modelo de nivel básico de área, pero el peso $\hat{\gamma}_d$ evita el supuesto de varianza de muestreo conocida de \hat{p}_d . El estimador \hat{p}_d^{BE} es aproximadamente insesgado para p_d en el sentido que su sesgo, $E(\hat{p}_d^{BE} - p_d)$, es de orden D^{-1} , para D grande. Una aproximación BE sencilla usa \hat{p}_d^{BE} como el estimador de p_d , y su variabilidad se mide por la varianza a posteriori estimada $V(p_d|y_d, \hat{\alpha}, \hat{\beta}) = g_{1i}(\hat{\alpha}, \hat{\beta}, y_d)$. Sin embargo, la varianza a posteriori estimada, $g_{1i}(\hat{\alpha}, \hat{\beta}, y_d)$, puede conducir a una severa subestimación del ECM(\hat{p}_d^{BE}) porque ignora la variabilidad asociada con $\hat{\alpha}$ y $\hat{\beta}$.

Se han propuesto también modelos alternativos de dos fases. El modelo de primera fase no se cambia, pero el modelo de segunda fase se cambia a cualquier (i) $\text{logit}(p_d) = \log[p_d/(1 - p_d)] \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$ o (ii) $\Phi^{-1}(p_d) \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$, donde $\Phi(\cdot)$ es la función de distribución acumulada (FDA) de una variable $N(0, 1)$. Los modelos (i) y (ii) se llaman modelos logit-normal y probit-normal, respectivamente. La implementación de BE es más complicada para los modelos alternativos porque no existen expresiones explícitas para el estimador de bayes y la varianza a posteriori de p_d .

Para el modelo logit-normal, el estimador de bayes de p_d puede expresarse como una razón de integrales unidimensionales. Escribiendo p_d como $p_d = h_1(\mu + \sigma z_d)$, donde $h_1(a) = e^a/(1 + e^a)$ y $z_d \sim N(0, 1)$, obtenemos $\hat{p}_d^B(\mu, \sigma) = E(p_d|y_d, \mu, \sigma)$ de la distribución condicional de z_d dada y_d . Tenemos

$$\hat{p}_d^B(\mu, \sigma) = A(y_d, \mu, \sigma)/B(y_d, \mu, \sigma), \quad (5.2.15)$$

donde

$$A(y_d, \mu, \sigma) = E[h_1(\mu + \sigma z) \exp\{h_2(y_d, \mu + \sigma z)\}] \quad (5.2.16)$$

y

$$B(y_d, \mu, \sigma) = E[\exp\{h_2(y_d, \mu + \sigma z)\}], \quad (5.2.17)$$

donde $h_2(y_d, a) = ay_d - n_d \log(1 + e^a)$ y la esperanza es sobre $z \sim N(0, 1)$; ver McCulloch y Searle (2001, p. 67). Podemos evaluar (5.2.16) y (5.2.17) simulando muestras de una $N(0, 1)$. Alternativamente puede usarse, como esbozamos antes, integración numérica.

La log verosimilitud, $l(\mu, \sigma)$, para el modelo logit-normal puede escribirse como

$$l(\mu, \sigma) = \text{const} + \sum_{i=1}^D \log[B(y_d, \mu, \sigma)], \quad (5.2.18)$$

donde $B(y_d, \mu, \sigma)$ está dada por (5.2.17). Las derivadas de $l(\mu, \sigma)$, necesitarán métodos tipo Newton-Raphson para calcular estimaciones MV, y pueden ser aproximadas de una manera similar.

Usando estimadores MV de $\hat{\mu}$ y $\hat{\sigma}$, obtenemos un estimador BE de p_d como $\hat{p}_d^{BE} = \hat{p}_d^B(\hat{\mu}, \hat{\sigma})$. La varianza a posteriori, $V(p_d|y_d, \mu, \sigma)$, también puede expresarse en términos de la esperanza sobre $z \sim N(0, 1)$, notando que $V(p_d|y_d, \mu, \sigma) = E(p_d^2|y_d, \mu, \sigma) - [\hat{p}_d^B(\mu, \sigma)]^2$. Denotando $v(p_d|y_d, \mu, \sigma) = g_{1i}(\mu, \sigma, y_d)$, puede aplicarse el método jackknife para obtener un estimador aproximadamente insesgado del ECM(\hat{p}_d^{BE}).

Jiang y Lahiri (2001) usaron \hat{p}_d^{BE} basado en los estimadores de momentos del primer paso, y obtuvieron un estimador por desarrollo de Taylor del ECM(\hat{p}_d^{BE}), similar a los estimadores del ECM de segundo orden del modelo lineal mixto. Este estimador es aproximadamente insesgado, como en el caso del estimador jackknife del ECM.

Modelos con covariables

El modelo logit-normal de la subsección anterior se extiende sin dificultad al caso de covariables. En la primera fase, asumimos que $A_{dj}|p_{dj} \stackrel{\text{ind}}{\sim} \text{Bernoulli}(p_{dj})$ para $j = 1, \dots, N_d$; $d = 1, \dots, D$. Los p_{dj} están ligados en la segunda etapa asumiendo un modelo de regresión logística con efectos de área aleatorios: $\text{logit}(p_{dj}) = \mathbf{B}_{dj}^T \boldsymbol{\beta} + v_d$ donde $v_d \stackrel{\text{iid}}{\sim} N(0, \sigma_v^2)$ y \mathbf{B}_{dj} es el vector de covariables fijas. El modelo de dos variables se llama modelo logístico lineal mixto. Pertenece a la clase de los modelos lineales generalizados mixtos.

Los parámetros de área pequeña son las proporciones de área pequeña $P_d = \sum_j A_{dj}/N_d$. Como en el caso del modelo básico de nivel unidad, asumimos que los modelos se conservan para la muestra $\{(A_{dj}, \mathbf{B}_{dj}), j \in s_d; d = 1, \dots, D\}$, donde s_d es la muestra de tamaño n_d para el área d .

Expresamos P_d como $P_d = f_d p_d + (1 - f_d) p_d^*$, donde $f_d = n_d/N_d$, p_d es la proporción muestral y $p_d^* = \sum_{l \in \bar{s}_d} A_{dl}/(N_d - n_d)$ es la proporción de las unidades no muestreadas \bar{s}_d en el área d . Ahora, notando que $E(A_{dl}|p_{dl}, \mathbf{y}_d, \boldsymbol{\beta}, \boldsymbol{\sigma}_v) = p_{dl}$ para $l \in \bar{s}_d$, el estimador de bayes de p_d^* está dada por $\hat{p}_{d(c)}^B = E(p_{d(c)}|y_d, \boldsymbol{\beta}, \boldsymbol{\sigma}_v)$, donde $\hat{p}_{d(c)} = \sum_{l \in \bar{s}_d} p_{dl}/(N_d - n_d)$ y \mathbf{y}_d es el vector de valores A_j de la muestra del área d . Por lo tanto, el estimador de bayes de P_d puede expresarse como

$$\hat{P}_d^B = \hat{P}_d^B(\boldsymbol{\beta}, \boldsymbol{\sigma}_v) = f_d p_d + (1 - f_d) p_{d(c)}^B. \quad (5.2.19)$$

Si la fracción de muestreo es despreciable, podemos expresar \hat{P}_d^B como

$$\hat{P}_d^B \approx \frac{1}{N_d} E \left(\sum_{l=1}^{N_d} p_{dl} | \mathbf{y}_d, \boldsymbol{\beta}, \boldsymbol{\sigma}_v \right). \quad (5.2.20)$$

La varianza a posteriori de P_d se reduce a

$$V(P_d|\mathbf{y}_d, \boldsymbol{\beta}, \boldsymbol{\sigma}_v) = (1 - f_d)^2 E(\bar{y}_d^* - \hat{p}_{d(c)}^B)^2$$

$$= N_d^{-2} \left[E \left\{ \sum_{l \in \bar{s}_d} p_{dl}(1 - p_{dl}) | \mathbf{y}_d, \boldsymbol{\beta}, \boldsymbol{\sigma}_v \right\} + V \left\{ \sum_{l \in \bar{s}_d} p_{dl} | \mathbf{y}_d, \boldsymbol{\beta}, \boldsymbol{\sigma}_v \right\} \right]; \quad (5.2.21)$$

ver Malec et al. (1997). Nótese que (5.2.21) incluye esperanzas de la forma $E(\sum_l p_{dl}^2 | \mathbf{y}_d, \boldsymbol{\beta}, \boldsymbol{\sigma}_v)$ y $E[(\sum_l p_{dl})^2 | \mathbf{y}_d, \boldsymbol{\beta}, \boldsymbol{\sigma}_v]$ así como $E(\sum_l p_{dl} | \mathbf{y}_d, \boldsymbol{\beta}, \boldsymbol{\sigma}_v) = (N_d - n_d) \hat{p}_{dc}^B$. No existe ninguna expresión explícita para estas esperanzas. Sin embargo, podemos expresar las esperanzas como razones de integrales de una sola dimensión, similares a (5.2.15).

La estimación por máxima verosimilitud de los parámetros del modelo, $\boldsymbol{\beta}$ y $\boldsymbol{\sigma}_v$, para el modelo logístico lineal mixto y otros modelos lineales generalizados mixtos, ha recibido considerable atención en años recientes. Los métodos propuestos incluyen cuadratura numérica, algoritmo EM, Monte Carlo mediante cadenas de Markov (MCMC) y aproximación estocástica. Para detalles de los algoritmos, consultar McCulloch y Searle (2001). También se han propuesto métodos más simples, cuasi-verosimilitud penalizada (PQL), basada en maximizar la distribución conjunta de $\mathbf{y} = (y_1^T, \dots, y_D^T)^T$ y $\mathbf{v} = (v_1, \dots, v_D)^T$ con respecto a $\boldsymbol{\beta}$ y \mathbf{v} .

Farrel, MacGibbon y Tomberlin (1997a), usaron una medida bootstrap de precisión similar al bootstrap paramétrico de Laird y Louis (1987) (también llamado bootstrap tipo III). Los resultados de simulación, basados en $D = 20$ y $n_d = 50$, indicaron un buen funcionamiento del método BE propuesto. Farrel, MacGibbon y Tomberlin (1997b), relajaron el supuesto de normalidad sobre las v_d , usando el siguiente modelo de enlace: $\text{logit}(p_{dj}) = \mathbf{B}_{dj}^T \boldsymbol{\beta} + v_d$ y $v_d \stackrel{\text{iid}}{\sim}$ distribución no especificada. Usaron un método MV no paramétrico, propuesto por Laird (1978), para obtener un estimador BE de P_d . Usaron también el bootstrap tipo II de Laird y Louis (1987), para obtener una medida bootstrap de precisión.

5.2.5. Posibles proyectos de investigación.

En epidemiología y salud:

- Encuesta nacional de salud.
- Estimación de proporciones de enfermedad a nivel regional o municipal.
- Mapas de enfermedad.

En la administración pública:

- En la actualidad, prácticamente todos los gobiernos de los países del mundo realizan encuestas nacionales o estatales para obtener información sobre distintas características de la población.
- La información obtenida es utilizada en la formulación de políticas y programas, en la evaluación, en la asignación de fondos gubernamentales y en la planificación regional.
- En el sector privado se hace uso de información de encuestas para apoyar las decisiones de negocios, reorientar una campaña política, posicionar un producto en el mercado, etc.

Estadísticas oficiales:

- Índices de desarrollo humano.
- Desarrollo económico.
- Desarrollo social.
- Evaluación del desempeño gubernamental.

Estudios en educación:

- Derivados de la prueba ENLACE, PISA, etc.
- Cobertura regional.
- Nueva oferta educativa.

Bibliografía

- [1] Cadwell, B.L., Thompson, T.J. Boyd, J.P. and Barker, L.E. Bayesian Small Area Estimates of Diabetes Prevalence. *Journal of Data Science* 8, 173-188 (2010) by U.S. County, 2005
- [2] Chandra, H., Chambers, R. and Salvati, N., Small area estimation of proportions in Business Survey. Center for Statistical and Survey Methodology, Working paper, 15, (2009).
- [3] Datta (2009), G.S., Sample Surveys: Inference and Analysis, Vol. 29B in *Handbook of Statistics Vol. 29B* D. Pfeffermann and C.R. Rao (editors), The Netherlands: North-Holland.
- [4] Gonzalez-Manteiga, W., Lombardia, M. J., Molina, I., Morales, D. and Santamaria, L., Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model, *Computational Statistics and Data Analysis*, 51, 2720-2733, (2007)
- [5] Jiang, J., Lahiri, P., Mixed model prediction and small area estimation, *Test* 15, 1-96,2006
- [6] Lehtonen, R. and Veijanen, A. Logistic generalized regression estimators. *Survey Methodology* 24, 51 - 55, (1998)
- [7] MacGibbon, B. and Tomberlin, T. J., Small Area Estimation of Proportions Via Empirical bayes Techniques, *Survey Methodology*, 15, 237-252, (1989)
- [8] Malec, D. Sedransk, J., Moriarity, C. L., and LeClere, F. B., Small Area Inference for Binary Variables in National Health Interview Survey, *Journal of the American Statistical Association*, 92, 815-826, (1997)
- [9] Muñoz, J.F., Arcos, A., Álvarez-Verdejo, E., Rueda, M, Martínez-Puertas, S. (2010). Estimators and confidence intervals for the proportion using auxiliary information with applications to the estimation of prevalences. *Journal of Biopharmaceutical Statistics* 1, 1-32
- [10] Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G., Breidt, F. J., Non-parametric small area estimation using penalized spline regression, *Journal of the Royal Statistical Society Series B* 70, 265-286, (2008)
- [11] Rao, J. N. K., *Small Area Estimation*, New York: Wiley, (2003)

- [12] Saei, A. and Chambers, R., Small Area Estimation: A review of methods based on the application of mixed models. SRI Methology Working paper, (2003).
- [13] Saei, A. and R. Chambers, Small area estimation under linear and generalized linear mixed models with time and area effects, EURAREA Consortium 2004, Project Reference Volume, (2004)
- [14] Salvati, N., Chandrab, H., Ranalli, G. and Chambers, R., Small area estimation using a nonparametric model-based direct estimator, Computational Statistics and Data Analysis, 54, 2159-2171, (2010)
- [15] Särndal, C. E., Design-consistent versus model-dependent estimation for small domains, Journal of the American Statistical Association 79, 624-631, (1984)
- [16] Särndal, C. E. and Hidiroglou, M.A., Small Domain Estimation: A Conditional Analysis, Journal of the American Statistical Association, 84, 266-275, (1989)
- [17] Särndal, C. E., Design-consistent versus model-dependent estimation for small domains, Journal of the American Statistical Association 79, 624-631, (1984)
- [18] C.E. Särndal and A. Veijanen, Model calibration and generalized regression estimation for domains and small areas, Invited paper, SAE2009, 2009
- [19] M. D. Ugarte, T. Goicoa, A. F. Militino, M. Durbán, Spline smoothing in small area trend estimation and forecasting, Computational Statistics and Data Analysis, 53, 3616-3629, (2009)