eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

TAREA 4

EJERCICIOS

1. Un cierto tipo de planta puede presentarse en cuatro variedades, cuyas probabilidades respectivas de acuerdo a una teoría genética son: $\frac{1}{2} + \theta/4$, $(1-\theta)/4$, $(1-\theta)/4$, y $\theta/4$, en que θ es un parámetro, no especificado por la teoría, con valor entre 0 y 1. Una muestra de n plantas proporcionó $n_0, \ldots, n_3$ ejemplares respectivamente de las cuatro variedades.

   a) Obtén estadísticos suficientes para θ.

   b) Obtén el estimador MV de θ y un estimador de su varianza.

   c) En este caso es posible la obtención analítica del estimador MV (que se reduce a resolver una ecuación de sgundo grado). Emplea no obstante el método de Newton-Raphson para resolver la ecuación de verosimilitud y llegar al mismo resultado numéricamente, cuando $n_0 = 5$, $n_1 = 4$, $n_2 = 3$ y $n_3 = 1$.

   d) No hay aquí datos perdidos, y en consecuencia parecería que el uso del algortimo EM está fuera de lugar. Observa, sin embargo, que si imaginamos la primera celda (la de probabilidad $\frac{1}{2} + \theta/4$) como compuesta por otras dos de las que hemos perdido los valores n_{01} y n_{02} y sólo conocemos $n_0 = n_{01} + n_{02}$, el problema se simplifica. Escribe la verosimilitud para comprobarlo.

2. La Tabla 1 (procede de [5], pág. 22) proporciona datos relativos a una muestra de 7477 mujeres de entre 30 y 39 años, clasificadas de acuerdo con la visión en cada uno de sus dos ojos. Hay varios modos

Ojo	Ojo izquierdo				
derecho	Muy buena	Buena	Regular	Mala	Total
Muy buena	1520	266	124	66	1976
Buena	234	1512	432	78	2256
Regular	117	362	1772	205	2456
Mala	36	82	179	492	789
Total	1907	2222	2507	841	7477

Cuadro 1: Visión no corregida de 7477 mujeres entre 30 y 39 años

en que una tal tabla puede haberse generado. Podríamos haber decidido que queríamos examinar a, precisamente, 7477 mujeres, y haberlas tomado al azar de la población objeto de análisis. Podríamos también haber examinado mujeres tomadas al azar *sin fijar de antemano el tamaño muestral*, que entonces también sería aleatorio. Por ejemplo, podríamos haber examinado ojos de mujeres hasta agotar nuestro presupuesto, o hasta finalizar un número de días que teníamos asignados para el examen.

a) Sea θ_{ij} la probabilidad de que una mujer tomada al azar sea clasificada en la casilla ij. Muestra que con los dos tipos de muestreo mencionados (multinomial y Poisson) la verosimilitud se maximiza para los mismos valores estimados de los θ_{ij}.

b) Supón que el muestreo ha sido multinomial (tamaño muestral fijo). ¿Cuántos parámetros hay? ¿Cuántos son libres?

c) ¿Cuáles son estadísticos suficientes para dichos parámetros?

d) ¿Cuáles son los estimadores insesgados de varianza mínima?

e) Podríamos pensar en distintos modelos como generadores de nuestras observaciones. Uno de ellos es el modelo de independencia. ¿Cuántos parámetros precisaríamos si suponemos que la probabilidad de estar en una determinada fila no depende de la columna en que se esté (es decir, que la visión en ambos ojos es completamente independiente)? ¿Cuáles serían estimadores insesgados para dichos parámetros?

f) Otro modelo que sería natural considerar es el de simetría: $\theta_{ij} = \theta_{ji}$ para todo i,j. ¿Cuántos parámetros hay? ¿Cuántos son libres? ¿Cuáles serían estimadores insesgados de mínima varianza para dichos parámetros? ¿Maximo verosímiles?

g) ¿Implicaría el supuesto de simetría la homogeneidad de probabilidades marginales? ¿Y viceversa? Da una demostración o un contraejemplo.

3. Considera la siguiente tabla tridimensional en que los pasajeros del único viaje del *Titanic* se clasifican por su condición de supervivientes o no, su sexo y su *status* (clase en que viajaban o si eran tripulantes). Los datos los tienes desglosados en la dataframe `titanic.frame` (lee con la función

Tripulante	Muertos		Supervivientes		
o clase	Mujer	Varón	Mujer	Varon	Total
Tripulación	3	670	20	192	885
Primera	4	118	141	62	325
Segunda	13	154	93	25	285
Tercera	106	422	90	88	706

Cuadro 2: Pasajeros del *Titanic*

`dget()`), en el lugar habitual.

a) ¿Cuál sería una estimación insesgada del número esperado de supervicientes en cada combinación status/sexo si la probabilidad de superviviencia hubiera sido la misma para todos? ¿Qué observas?

b) Sea θ_{ijk} la probabilidad de que un sujeto pertenezca a la clase de supervivencia i, status j y sexo k. Considera los siguientes modelos:

$$\theta_{ijk} = \theta_{i++}\theta_{+j+}\theta_{++k} \quad (1)$$

$$\theta_{ijk} = \theta_{i++}\theta_{+jk}; \quad (2)$$

Expresa en palabras los supuestos implícitos en cada uno. Determina cuantos parámetros libres hay en el modelo (2), estímalos y estima el número esperado de supervivientes bajo dicho modelo.

4. Considera un experimento en que se administra a un colectivo de personas un tratamiento o un placebo, examinando a continuación quienes contraen una enfermedad y quienes no. Los resultados se resumen en la Tabla 3. Parece claro que el tratamiento es efectivo ¿no? (¿por qué?).

	Enferman	No enferman	Total
Tratamiento	5950	9005	14955
Placebo	5050	1095	6145
Total	11000	10100	21100

Cuadro 3: Datos agregados

	Enferman	No enferman	Total
Tratamiento	950	9000	9950
Placebo	50	1000	1050
Total	1000	10000	11000

Cuadro 4: Datos para mujeres

Sin embargo, llevado de tu afán de averiguar más desglosas los resultados en hombres y mujeres, porque piensas que quizá el tratamiento pueda ser más efectivo en unos que en otros. Y obtienes las Tablas 4 y 5.

	Enferman	No enferman	Total
Tratamiento	5000	5	5005
Placebo	5000	95	5095
Total	10000	100	10100

Cuadro 5: Datos para hombres

¿Qué está pasando?

AYUDAS, SUGERENCIAS Y COMPLEMENTOS

1. Sobre análisis de datos categóricos y tablas de contingencia hay mucha bibliografía. Un pequeño manual muy legible es [4]. Mucho más avanzados son [2] o [5], [1], [3]. En la sección 519.235 de biblioteca podrás encontrar bastantes más libros al respecto.

   Si quieres, puedes buscar en algún manual bajo “paradoja de Simpson”; pero no dejes de pensar quince minutos por tu cuenta sobre la aparente anomalía que ilustra la Cuestión 4.

2. Para emplear el método de Newton-Raphson, puedes escribir un bucle `while() { ... }`, comprobando al comienzo si la diferencia entre dos valores sucesivos de $\hat{\theta}_n$ y $\hat{\theta}_{n+1}$ es o no menor que una tolerancia prefijada (en cuyo caso acabarás la iteración).

3. Puesto que previamente has calculado las derivadas precisas, las puedes emplear sin más. En la practica, frecuentemente te encontrarás con expresiones que son mucho más largas y tediosas de derivar, y una de estas dos aproximaciones te será de utilidad:

   - Obtener las derivadas precisas mediante un programa especializado en cálculo simbólico. En la Facultad dispones de Mathematica: invócalo como `math`. Hay alternativas de dominio público como Maxima.
   - Entregar a R (o S-PLUS) la expresión original y dejar que calcule las derivadas precisas: mira las funciones `deriv` y `deriv3` en cualquiera de los dos paquetes.

4. A partir de la *dataframe* `titanic.frame` puedes obtener fácilmente una tabla de contingencia mediante aplicación de la función `table`. Observa que obtendrás una tabla con cuatro dimensiones —hay información también acerca de la edad—. Puedes eliminar directamente la dimensión que sobra sumando,

   ```
   titanic.orig <- table(titanic.frame)
   titanic.reducida <- titanic.orig[,1,,] + tabla.orig[,2,,]
   ```

   (si quisieras colapsar los dos niveles de la segunda dimensión, por ejemplo), o mediante la función `apply`,

   ```
   titanic.reducida <- apply(titanic.orig,c(1,3,4),sum)
   ```

   para lograr lo mismo de otro modo. En este último caso has de indicar los márgenes que conservas.

5. La notación θ_{i++} es similar a la empleada en ANOVA (allí utilizábamos puntos en lugar de $+$ para indicar los índices sobre los que se suma). Por ejemplo, $\theta_{ij+} = \sum_k \theta_{ijk}$.

Referencias

[1] A. Agresti. *Categorical Data Analysis*. Wiley, 1990. Signatura: 519.235 AGR.

[2] Y.M.M. Bishop, S.E. Fienberg, and P.W. Holland. *Discrete Multivariate Analysis. Theory and Practice*. MIT Press, Cambridge, Mass., 1975.

[3] R. Christensen. *Log-Linear Models*. Springer-Verlag, 1990. Signatura: 519.235 CHR.

[4] S.E. Fienberg. *The Analysis of Cross-Classified Categorial Data*. MIT Press, Cambridge, Mass., 1980.

[5] R.L. Plackett. *The Analysis of Categorical Data*. Griffin, London, 1974.