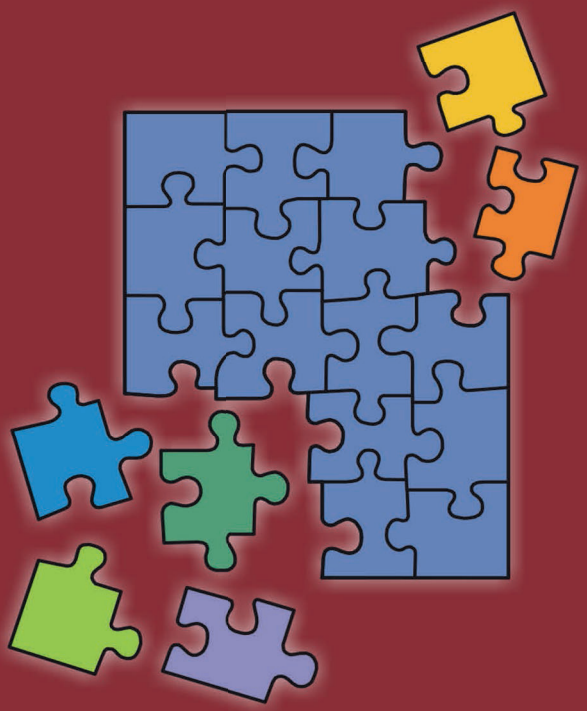
colección textos

ANÁLISIS ESTADÍSTICO DE DATOS CATEGÓRICOS

textos textos textos textos textos textos

Luis Guillermo Díaz Monroy
Mario Alfonso Morales Rivera



UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ
FACULTAD DE CIENCIAS

Facultad de Ciencias
Saber más y formar mejor

ANÁLISIS ESTADÍSTICO DE DATOS CATEGÓRICOS

ANÁLISIS ESTADÍSTICO DE DATOS CATEGÓRICOS

LUIS GUILLERMO DÍAZ MONROY

MARIO ALFONSO MORALES RIVERA

Departamento de Estadística
Facultad de Ciencias

Universidad Nacional de Colombia
sede Bogotá

Análisis estadístico de datos categóricos

Primera edición, 2009
Bogotá, Colombia
ISBN 978-958-719-186-8

Diseño de carátula: Andrea Kratzer M.

Catalogación en la publicación Universidad Nacional de Colombia

Díaz Monroy, Luis Guillermo, 1958-
Análisis estadístico de datos categóricos / Luis Guillermo Díaz Monroy, Mario Alfonso Morales Rivera. – Bogotá : Universidad Nacional de Colombia. Facultad de Ciencias, 2009
xvii, 376 p.

ISBN : 978-958-719-186-8

1.Análisis de regresión logística 2.Tablas de contingencia 3.Modelos loglineales 4. Análisis de correspondencias (Estadística) 5. Modelos lineales (Estadística)
I. Morales Rivera, Mario Alfonso, 1965- II. Tít

CDD-21 519.536 / 2009

A Daniel, Camila, Diego y Pilar, mi única categoría.
Luis G. Díaz

A Nevis, mi ángel guardián.
Mario A. Morales

Contenido

Tablas

Figuras

Introducción

La distinción entre los llamados datos *cualitativos* y los denominados *cuantitativos* no siempre es clara, pues en algunos casos variables de tipo cuantitativo pueden considerarse como variables categóricas al dividir su rango de valores en intervalos o categorías, esto corresponde a una categorización de una variable cuantitativa. Un tratamiento recíproco puede considerarse para las variables cualitativas, es decir que pueden transformarse a variables cuantitativas, este procedimiento se muestra con el análisis de correspondencias. En estas notas se hace una revisión, bastante panorámica, sobre algunas metodologías estadísticas que coadyuvan al esclarecimiento e interpretación de la información contenida en datos categóricos.

El texto ha sido elaborado pensando en un lector que demande el uso de algunas herramientas estadísticas, útiles para el análisis de la información, principalmente de tipo categórico, producto de algún trabajo de investigación. No obstante que los primeros destinatarios son las personas que trabajen en torno a problemas de la salud y la biología, el material estadístico que se ofrece puede ser empleado por investigadores de otras disciplinas, pues basta cambiar el escenario de los ejemplos e ilustraciones, para hacer de este texto un instrumento de apoyo a varias disciplinas.

La primera parte contiene algunos conceptos generales junto con el tratamiento clásico de datos categóricos a través del análisis de tablas de contingencia, los cuales se desarrollan en los capítulos 1 y 2. Tratamientos alternativos a las tablas de contingencia se desarrollan en los capítulos 3 y 4 mediante el análisis de correspondencias y los modelos log-lineales. En el capítulo 5 se presenta el modelamiento con variable respuesta de tipo categórico el cual se hace a través de la regresión logística. El capítulo 6 contiene algunas de las técnicas de discriminación de uso más frecuente. En el capítulo 7 se esquematizan algunos contrastes de tipo no paramétrico sobre estadísticas de localización. Se tratan, en el capítulo 8, algunas técnicas estadísticas para datos de conteo. Finalmente, en el capítulo 9, se desarrolla la técnica de emparejamiento de datos.

Para el desarrollo de los cálculos que las estimaciones y estadísticas requie-

ren, se hace uso, principalmente, de los paquetes SAS y R. En cada capítulo se presenta la sintaxis pertinente para la ejecución de tales cómputos. Se debe advertir que existen otras herramientas computacionales igualmente útiles, tales como SPSS, BMDP, MINITAB, STATA, entre otras.

Agradecemos al Grupo de Investigación en "Estadística aplicada en la investigación experimental, la industria y la biotecnología", al Departamento de Estadística de la Universidad Nacional y al Departamento de Matemáticas y Estadística de la Universidad de Córdoba por posibilitar y permitir el ofrecimiento de estas notas.

Las notas se deben principalmente a la bibliografía que se anexa al final y a las preguntas, comentarios y sugerencias de nuestros colegas y de nuestros estudiantes. Este es un material que se puede mejorar en la medida que sea leído y cuestionado, por tanto agradecemos los comentarios y sugerencias que surjan de su estudio.

Luis Guillermo Díaz Monroy

Mario Alfonso Morales Rivera

Capítulo 1

Conceptos preliminares

1.1 Introducción

En este capítulo se presentan los aspectos fundamentales, a manera de elementos estadísticos básicos, para el desarrollo de los demás temas. Se revisa la naturaleza de los datos categóricos, el modelo probabilístico *binomial*, el de *Poisson* y la inferencia sobre una *proporción*.

1.2 Escala de medida

La escala de medida de una variable categórica es un elemento importante para la selección del análisis estadístico apropiado. Una selección inadecuada de la escala de medida puede conducir a una estrategia estadística inapropiada que arrojaría conclusiones erróneas acerca de la realidad contenida en los datos.

1.2.1 Dicotómicas

Son variables que tienen dos posibles respuestas, frecuentemente corresponden a la presencia o no de cierto atributo. Por ejemplo: ¿Desarrolló el sujeto la enfermedad? ¿En los últimos tres meses ha fumado alguna vez, o no? ¿Está afiliado actualmente al régimen contributivo de salud, o no?, etc.

Por ejemplo, el objetivo de un ensayo clínico para un nuevo medicamento contra la gripe es saber si los pacientes alivian sus dolencias. La tabla 1.1

contiene información sobre 124 pacientes, quienes recibieron tratamiento (medicamento) o un placebo (sin medicamento).

Tabla 1.1: Resultados respiratorios.

Tratamiento	Favorable	Desfavorable	Total
Placebo	16	48	64
Prueba	40	20	60

El grupo placebo consta de 64 pacientes; mientras que el grupo de prueba del medicamento contiene 60 pacientes.

1.2.2 Ordinal

En muchas ocasiones las variables categóricas representan más de dos posibles resultados, y a veces estos resultados poseen un orden propio. Tales variables tienen una escala de medida *ordinal*.

El estado de mejoría o progreso de un paciente se puede calificar como marcado (3), regular (2), ninguno (1). Este es el caso de un ensayo clínico en el que se investiga un tratamiento para la artritis reumatoidea. A hombres y mujeres les fue asignada una actividad (tratamiento) o un placebo (no actividad). Se midió el nivel de progreso o mejoría conseguido al final del ensayo; los datos están dispuestos en la tabla 1.2.

Tabla 1.2: Datos de artritis.

		Progreso			
Sexo	Tratamiento	Marcado	Regular	Ninguno	Total
Femenino	Actividad	16	5	6	27
Femenino	Placebo	6	7	19	32
Masculino	Actividad	5	2	7	14
Masculino	Placebo	1	0	10	11

Fuente: Stokes, Davis y Koch (1997: 218)

Note que las variables categóricas pueden manejarse de diferentes formas. Por ejemplo, en la tabla 1.2 se pueden fusionar las columnas Marcado y Regular para producir una variable dicotómica: "Progreso" frente a "No progreso". Este tipo de agrupamiento se hace generalmente durante el análisis cuando hay interés por esta clase de respuestas o cuando se quiere obtener información adicional sobre los datos.

1.2.3 Conteos discretos

Corresponde a los casos en los que en lugar de registrar categorías, los resultados son números enteros. Por ejemplo, una investigación sobre enfermedades respiratorias en niños de diferentes zonas visitados dos veces determina si ellos mostraron síntomas de la enfermedad. La respuesta medida fue si los niños exhibieron síntomas en 0, 1 o 2 periodos. La tabla 1.3 contiene estos resultados.

Tabla 1.3: Niños con problemas respiratorios.

		Periodos			
Sexo	Tratamiento	0	1	2	Total
Femenino	Rural	45	64	71	180
Femenino	Urbana	80	104	116	300
Masculino	Rural	84	124	82	290
Masculino	Urbana	106	117	87	310

1.2.4 Nominal

Si se dispone de variables con más de dos categorías, a las cuales no se les atribuye un orden, se tiene una variable de tipo *nominal*. Por ejemplo, el tipo de sangre de una persona y la región geográfica donde nació; la tabla 1.4 muestra esta información. En este tipo de variables la relación

Tabla 1.4: Tipo de sangre por región de procedencia.

	Tipo de sangre				
Región	O	A	B	AB	Total
Norte	40	160	150	50	400
Centro	50	130	100	40	320
Sur	70	180	90	60	400
Total	160	470	340	150	**1.120**

que se puede establecer entre sus categorías es estrictamente de igualdad (o desigualdad).

1.3 Esquema de muestreo

Cuando el interés en el estudio es de tipo inferencial, los datos categóricos pueden proceder de diferentes esquemas de muestreo, sea este probabilístico o no. La naturaleza del muestreo determina los supuestos que pueden hacerse para desarrollar y aplicar un análisis estadístico determinado. Generalmente, los datos se ubican en uno de tres esquemas muestrales: datos históricos, datos experimentales y datos de encuestas.

Los datos históricos se refieren a estudios en los cuales los datos tienen una definición geográfica o circunstancial. Por ejemplo, la ocurrencia de una enfermedad infecciosa en una área determinada, los niños atendidos en un centro de salud, o el número de accidentes durante un periodo específico.

Los datos experimentales son extraídos de estudios que involucran la asignación aleatoria de tratamientos a un grupo de sujetos. Es el caso en el que a los sujetos se les administra una dosis entre varias dosis de un medicamento. En ciencias de la salud, los datos experimentales pueden incluir pacientes a quienes se les administra un placebo o un medicamento de acuerdo con sus condiciones médicas.

En estudios por encuestas, los individuos son seleccionados aleatoriamente desde una población objetivo. Por ejemplo, se seleccion una muestra de los usuarios de determinado medicamento para investigar algunos rasgos físicos de estos. El investigador puede seleccionar aleatoriamente una población de estudio y luego asignar aleatoriamente tratamientos a los individuos que resulten para el estudio.

La principal diferencia entre los tres esquemas de muestreo asociados a los ejemplos anteriores es el empleo de la aleatorización para obtenerlos. Los datos históricos no involucran aleatorización; en consecuencia, es difícil asumir que ellos representan determinada población. Los datos experimentales tienen una buena cobertura de la población, la cual está restringida por los tratamientos considerados en el protocolo del estudio; en el muestreo por encuestas, los datos tienen una muy buena cobertura de alguna población grande.

La unidad de aleatorización (en conexión con la unidad de observación) puede ser un sujeto o un conglomerado de sujetos. Además, la aleatorización puede aplicarse a sujetos, llamados estratos o bloques, con igual o desigual probabilidad. En muestreo por encuestas, esto puede conducir a diseños complejos, como el muestreo aleatorio estratificado, o un diseño de conglomerados por múltiples etapas. En estudios de diseño experimental, tales consideraciones llevan a estudios de medidas repetidas, datos longitudinales, entre otros.

1.4 Modelos de muestreo

El análisis de datos categóricos, o casi cualquier tipo de análisis estadístico, requiere supuestos acerca del mecanismo de aleatorización que genera los datos; esto es, el modelo probabilístico desde el cual se asume que son generados los datos. Se presentan a continuación las distribuciones de probabilidad de uso más frecuente en el análisis de datos categóricos.

1.4.1 Distribución de Poisson

La distribución de Poisson es un modelo probabilístico adecuado para evaluar la probabilidad de ocurrencia de un evento en intervalos de tiempo, longitud, superficie o volumen. Por ejemplo, el número de accidentes por semana en un tramo de carretera, el número de aves muertas por kilómetro cuadrado en una región. Si X es la variable aleatoria que cuenta el número de veces que un evento ocurre por intervalo (tiempo, longitud, superficie, etc), y si se tiene que el número promedio de eventos por intervalo es μ, las probabilidades de los posibles resultados $k = 0, 1, 2, \ldots$ se calculan mediante la expresión

$$P(X = k) = \frac{e^{-\mu}\mu^k}{k!}, \quad \text{para } k = 0, 1, 2, \ldots \tag{1.1}$$

El término $k!$ es llamado el *factorial* de k y denota el producto de los k–primeros enteros, es decir, $k! = 1 \times 2 \times 3 \times \cdots \times k$, con $0! = 1$. El término $e^{-\mu}$ denota la *función exponencial*, algunas veces expresada como $\exp(-\mu)$; siendo $e \approx 2.7182$ el cual es la base de los *logaritmos naturales*. Esto último significa que $e^a = b$ si y solo si $\ln(b) = a$. Por ejemplo $e^{0.7} = \exp(0.7) = 2.0$ corresponde a que $\ln(2.0) = 0.7$.

Suponga que el número de personas con infarto que acuden a una clínica tiene una tasa promedio de 2 por semana. Mediante el modelo de Poisson con $\mu = 2$, (i) la probabilidad de 0 infartos $(k = 0)$, y, (ii) de a lo más un infarto, en una semana cualquiera, por (1.1), es igual, respectivamente, a:

$$(i) \quad P(X = 0) = \frac{e^{-2}2^0}{0!} = e^{-2} = 0.1353$$
$$(ii) \quad P(X \leq 1) = \frac{e^{-2}2^0}{0!} + \frac{e^{-2}2^1}{1!} = 0.1353 + 0.2707 = 0.4060.$$

donde $P(X = 1) = e^{-2}(2^1)/1! = 0.2707$ es la probabilidad que se presente un infarto durante una semana. De (i) la probabilidad de que hayan infartos es: $P(X \geq 1) = 1 - P(X = 0) = 1 - 0.1353 = 0.8647$

El valor esperado de la variable aleatoria X con distribución de Poisson es igual a su varianza; es decir,

$$\mathrm{E}(X) = \mathrm{var}(X) = \mu, \quad \sigma(X) = \sqrt{\mu} \tag{1.2}$$

Para el caso de los infartos por semana, si la tasa de ocurrencia de estos permanece constante de una semana a otra, entonces en un periodo largo el conteo de estos tendría una media de alrededor de 2 y una desviación estándar cercana a $\sqrt{2} = 1.41$.

De acuerdo con los parámetros dados en (1.2) se observa que la varianza se incrementa a medida que la media lo hace; los conteos tienden a variar más cuando el nivel de sus promedios es alto. Así, cuando el número de infartos por semana es 10, se observa que la variabilidad es más grande que cuando el número es igual a 2 por semana.

1.4.2 Distribución binomial

En el ejemplo anterior, el número de infartos fatales semanal es aleatorio. El número de infartos semanal, fatales o no, es también aleatorio. En muchas aplicaciones se tiene como fijo el número de veces que se presenta un fenómeno. En cada caso, el resultado es un evento A o no es el evento A (es A^c); entonces, se quiere registrar las veces que este fenómeno ocurre con la característica determinada (A) en un número fijo de observaciones (n).

Un ejemplo es el caso de infarto fatal (A) o no fatal (A^c) y, en general, la ocurrencia o no de un evento; también es común hablar de *"éxito"* o *"fracaso"* para referirse al evento A o al A^c, respectivamente.

Una variable aleatoria X tiene distribución *binomial* si su función de probabilidad está dada por

$$P(X = k) = \binom{n}{k} \pi^k (1-\pi)^{n-k}, \text{para } k = 0, 1, 2, \ldots, n. \tag{1.3}$$

donde los dos parámetros n y $\pi = P(A)$ son tales que n es un entero no negativo y $0 \leq \pi \leq 1$. Se escribe $X \sim B(n, \pi)$. Para $n = 1$ la variable aleatoria se denomina de *Bernoulli*; es decir, que una variable aleatoria binomial es una suma de variables independientes tipo Bernoulli.

La cantidad $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ es el número de posibles arreglos de k-elementos (subconjuntos) que se pueden formar a partir de un conjunto que tiene n-elementos. Así, por ejemplo, el número de formas como se puede conformar un comité de 3 personas, escogidas de un grupo de 5 personas, es

$$\binom{5}{3} = \frac{5!}{3!2!} = 10$$