Datos Categóricos

Integrantes:
Ricardo José García Medina Código: 214997
David Eduardo León Vanegas Código: 214951
David Ricardo Martínez Hernández Código: 261931

Probabilidad Y Estadística Fundamental
Sandra Vergara Cardozo

Universidad Nacional de Colombia
Facultad de Ingeniería
Bogotá

Índice

1. Objetivos

- Comprender la definición de un dato categórico y la diferencia entre datos cuantitativos.
- Conocer el método de chi-cuadrado (χ^2).
- Comprender como se debe desarrollar el análisis de datos categóricos y las herramientas que se deben utilizar para realizarlo.
- Mostrar a nuestros compañeros la importancia del análisis de datos categóricos para el desarrollo de la practica experimental.
- Saber elegir la técnica apropiada para cada situación de análisis de datos categóricos.
- Comprender e interpretar correctamente los resultados obtenidos.
- Comprender la definición y el uso de los modelos loglineal saturado y jerárquico para el desarrollo de análisis para datos categóricos.

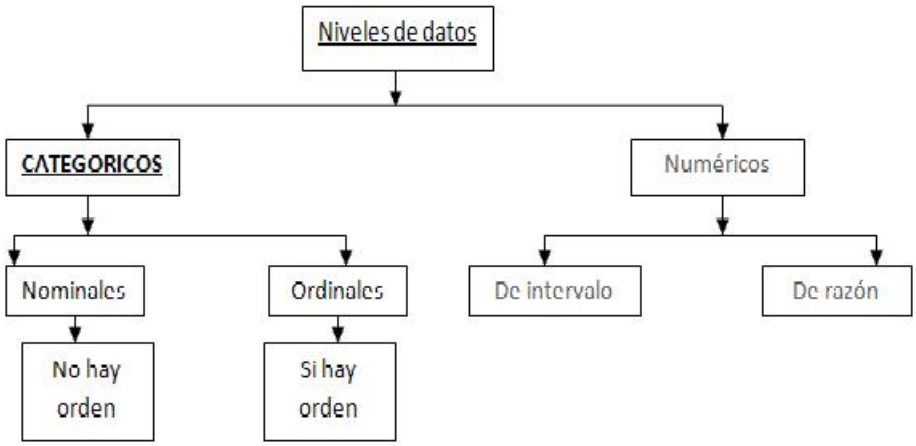
2. Introducción

Tanto una distribución de frecuencias como un histograma se pueden construir cuando el conjunto de datos es de naturaleza ***cualitativa* categórica**. en algunos casos habrá ordenamiento natural de clases, es decir que son organizados en grupos de la misma clase, en tanto que en otros casos el orden sera arbitrario, como las diferentes religiones del mundo. Con estos datos categóricos, los intervalos anteriores cuyos rectángulos se construyen, deben tener igual amplitud.[1]

3. Análisis De Datos Categóricos

Los datos categóricos son los que provienen de resultados de experimentos en los que los resultados se miden en escalas categóricas, son las mismas variables cualitativas.

- Escalas categóricas: sólo asignan una categoría o clasifican el fenómeno o propiedad a que se mide.



Cuando se tienen datos categóricos, el análisis que se hace es determinar el tipo de asociación existente, entre ciertas variables cualitativas, puede que no haya asociación alguna entonces se dice que las variables son independientes, o que haya asociación

[1] Texto tomado de [1], *Pág 22*

lo que quiere decir que algunos valores de una variable inclinan a que la otra variable tome ciertos valores más que otros, cuando hay asociación existen grados de intensidad, que hacen que la predisposición sea mayor o menor.
La independencia y dependencia es simétrica.

$$A\ ind.\ o\ depen.\ de\ B \Longrightarrow B\ ind.\ o\ depen.\ de\ A$$

3.1. Tablas de Contingencia

3.2. Esquemas

3.2.1. Muestreo

3.2.2. Multinomial

3.2.3. Poisson

4. Análisis de dos variables

Cuando se quiere saber si dos variables cualitativas están asociadas o no se utilizan las *tablas de contingencia* o de valores observados. En estas tablas, las variables se representan al lado izquierdo y en la parte superior. Supóngase que se tienen dos eventos A con i opciones y B con j opciones, entonces la tabla de contingencia para estas variables seria.

Tabla 1. Tabla de contingencia

$B \diagdown A$	1	...	$i-1$	i	$Total$
1	o_{11}				
. . .		. . .			
$j-1$			o_{i-1j-1}		
j				o_{ij}	$o_{i\bullet}$
Total				$o_{\bullet j}$	n

Donde o_{ij} es la frecuencia que tiene la característica ij, $o_{i.}$ y $o_{.j}$ son las frecuencias totales para la característica i y j, respectivamente, estas reciben el nombre de frecuencias marginales y n es el total de observaciones.

$$o_{i\bullet} = \sum_{i=1}^{i} o_{ij}$$

$$o_{\bullet j} = \sum_{j=1}^{j} o_{ij}$$

$$n = \sum_{i=1}^{i}\sum_{j=1}^{j} o_{ij} = \sum_{i=1}^{i} o_{i\bullet} = \sum_{j=1}^{j} o_{\bullet j}$$

Luego de que se tiene la tabla de valores esperados, se debe hallar la tabla de valores esperados que se obtiene con.

$$e_{ij} = \frac{o_{i\bullet}o_{\bullet j}}{n}$$

así la tabla de valores esperados es parecida a la tabla de contingencia solo que se remplaza o_{ij} por e_{ij} para todo o_{ij} de la tabla, en cuanto a las frecuencias marginales se hace la sumatoria en cada fila y columna con los valores de las frecuencias esperadas, mientras que n será el mismo valor y puede ser rectificado con la suma de las frecuencias marginales.
Con la tabla de frecuencias esperadas puede saber si existe relación entre las variables ya que si la tabla de contingencia es idéntica a la tabla de frecuencias esperadas las variables no tiene ninguna relación por consiguiente son **independientes**, en caso contrario hay **asociación**, el problema que surge ahora es saber si las tablas presentan mucha diferencia o no, para esto se utiliza un estadístico que nos elimine este problema.

4.1. Estadístico chi-cuadrado χ^2

$$\chi^2 = \sum_{i=1}^{i}\sum_{j=1}^{j}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} = \sum_{i=1}^{i}\sum_{j=1}^{j}\frac{o_{ij}^2}{e_{ij}} - n$$

El chi-cuadrado (χ^2) es una medida que refleja el grado de diferencia entre las tablas de frecuencias obtenidas y esperadas, es un estadístico de asociación. Si este estadístico es 0 entonces las tablas son idénticas, es decir que las variables son netamente **independientes** y ha medida que éste sea más grande habrá mayor **asociación** entre las variables, el problema del χ^2 radica en que no tiene un limite superior, por consiguiente no se sabrá si este valor es grande.

$$0 \leq \chi^2 < \infty$$

para saber si este estadístico es grande o pequeño se debe utilizar otro estadístico que es el grado de libertada.

$$Grado\ de\ libertad = (N^o\ de\ filas - 1)(N^o\ de\ columnas - 1)$$

Conociendo el grado de libertad y χ^2, solo basta con mirar en la tabla de χ^2, la cual por cada grado de libertad y la probabilidad para rechazar erróneamente el hecho de que las variable sean **independientes**, se obtiene un valor de χ^2. Si este valor es mayor que el dado en la tabla, entonces existe relación entre las variables.

DISTRIBUCION DE χ^2

Grados de libertad	Probabilidad										
	0,95	0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,01	0,001
1	0,004	0,02	0,06	0,15	0,46	1,07	1,64	2,71	3,84	6,64	10,83
2	0,10	0,21	0,45	0,71	1,39	2,41	3,22	4,60	5,99	9,21	13,82
3	0,35	0,58	1,01	1,42	2,37	3,66	4,64	6,25	7,82	11,34	16,27
4	0,71	1,06	1,65	2,20	3,36	4,88	5,99	7,78	9,49	13,28	18,47
5	1,14	1,61	2,34	3,00	4,35	6,06	7,29	9,24	11,07	15,09	20,52
6	1,63	2,20	3,07	3,83	5,35	7,23	8,56	10,64	12,59	16,81	22,46
7	2,17	2,83	3,82	4,67	6,35	8,38	9,80	12,02	14,07	18,48	24,32
8	2,73	3,49	4,59	5.53	7,34	9,52	11,03	13,36	15,51	20,09	26,12
9	3,32	4,17	5,38	6,39	8,34	10,66	12,24	14,68	16,92	21,67	27,88
10	3,94	4,86	6,18	7,27	9,34	11,78	13,44	15,99	18,31	23,21	29,59
	No significativo								Significativo		

usualmente se utilizan los valores con la probabilidad del $0{,}05$, ya que con éste se puede decir que las variables son independientes o no con mayor confianza. Aunque con ésto prácticamente se puede saber si las variables son independientes o no, no se puede saber cual es el grado de asociación que existe. Cuando existe una relación entre las variables el método mas apropiado es el *Coeficiente V de Cramer*.

- Coeficiente V de cramer

$$V = \sqrt{\frac{\chi^2}{n(k-1)}} \qquad ; 0 \leq V \leq 1$$

Donde K es el mínimo entre el número de filas y el numero de columnas.
Cuando V toma el valor 0 significa que no hay asociación alguna (son independientes) y entre mayor sea este valor, mayor será el grado de intensidad en la asociación.

4.2. Tabla de Contingencia Bidimensionales

5. Conclusiones y Recomendaciones

- .
- .
- .
- .
- .

Referencias

[1] Devore, Jay L. « *Probabilidad y estadística para ingeniería y ciencias* ». THOMSON, Sexta edición, 2004.

[2] Agresti, Alan. « *An Introduction to Categorical Data Analysis* ». JOHN WILEY y SONS, Inc., Segunda edición, 2002.

[3] Sito Web: http://www.jorgegalbiati.cl/enero_07/Categoricos.pdf