

Análisis de frecuencia: las tablas de contingencia

Chi square statistic

Cross Tabulation Data (Table -2)			
	Smartphones	Laptops	Total Sales (in \$M)
Region 2	132	14	146
Region 3	92	16	108
Total Sales (in \$M)	224	30	254

$$\chi^2 = \sum \left[\frac{(fo - fe)^2}{fe} \right]$$

This formula says that chi-square is the sum we will get if we,

1. Subtract fe from fo for each of the values in the table.
2. Square each of the differences
3. Divide each squared differences by fe
4. Sum all eight of the answers.

Oscar Centeno Mora

Preámbulo

- ¿Podemos probar si hay diferencias entre la región de residencia y el tipo de beca que se otorga?
- ¿Hay diferencia entre la inclinación política y el lugar de nacimiento?
- ¿Existen diferencia entre el sexo y el tipo de enfermedad viral (asma, tos crónica, etc.)?
- ¿Podemos probar las diferencias significativas entre el tipo de obesidad y la etnia ?
- Sin duda alguna, un análisis de frecuencias mediante tablas de contingencia es, el más común en el área de salud...



Preámbulo

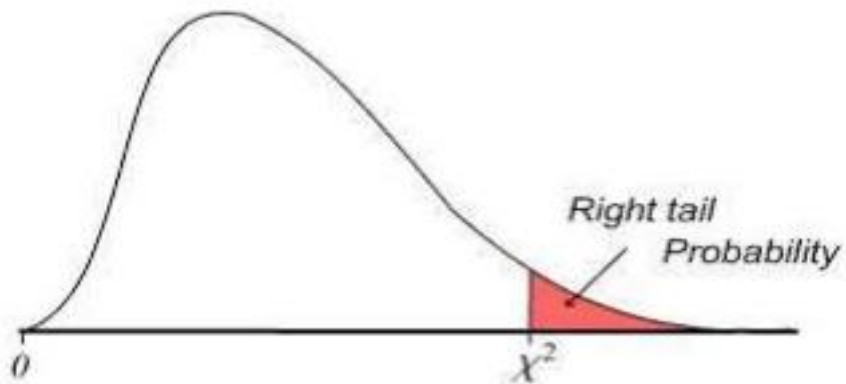
- Hasta ahora hemos realizado análisis de más de dos variables con la presencia de por lo menos una variable cuantitativa (¿cuáles?).
- ¿Podemos someter a una prueba de hipótesis a dos o más variables cualitativas?
- El análisis de frecuencia por tablas de contingencia nos permite lo anterior.
- En este capítulo utilizaremos una nueva función de probabilidad: la Chi-cuadrado (Ji-cuadrado), con tal de procesar el análisis de una tabla de frecuencias.



Preámbulo

$$\chi^2 = \sum \frac{(o - e)^2}{e}$$

Chi-Square Test



Índice

1

La Chi Cuadrada

4

Otros análisis de
frecuencias: prueba
exacta de Fisher

2

Prueba de
independencia

3

Prueba de
homogeneidad

Índice

1

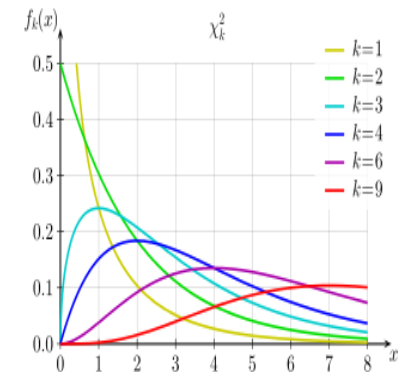
La Chi Cuadrada

Propiedades de la Chi-Cuadrado

- El estadístico de Chi-cuadrado es la función de distribución de mayor uso para el análisis de conteo en datos provenientes de tablas de frecuencias.
- La distribución Chi-cuadrada puede deducirse a partir de la distribución normal. Suponga que a partir de una variable aleatoria Y que sigue una distribución normal, con media μ y desviación estándar σ , eligen muestras aleatorias e independientes de tamaño $n = 1$.
- Cada valor seleccionado puede transformarse en la variable normal estándar a través de la formula:

$$Z = \frac{y_i - \mu}{\sigma}$$

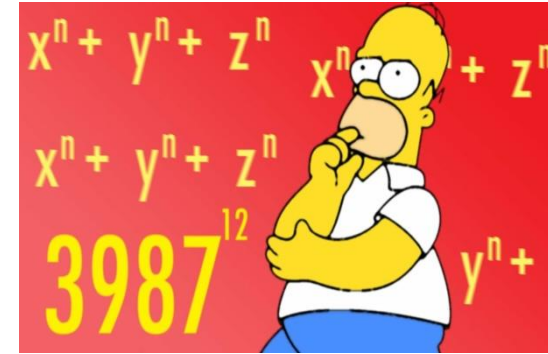
$$\chi^2 = \sum_{j=1}^k \frac{(f_{b_j} - f_{e_j})^2}{f_{e_j}}$$



Propiedades de la Chi-Cuadrado

- Cada valor de z puede elevarse al cuadrado para obtener una z^2 . Cuando se estudia la distribución muestral de z^2 , se obtiene que sigue una distribución Ji o Chi-Cuadrada con 1 grado de libertad. Eso es:

$$\chi^2_{(1)} = \left(\frac{y - \mu}{\sigma} \right)^2 = z^2$$



- Una Chi cuadrada es un caso especial de una Z elevada al cuadrado. Existen muchas propiedades y teoremas sobre esta distribución, pero esto no será abordado en este curso, lastimosamente...
- Ahora, ¿qué tipos de pruebas se llevan a cabo mediante la presente distribución?



Propiedades de la Chi-Cuadrado

- Hacemos uso de la presente distribución para las siguientes pruebas:

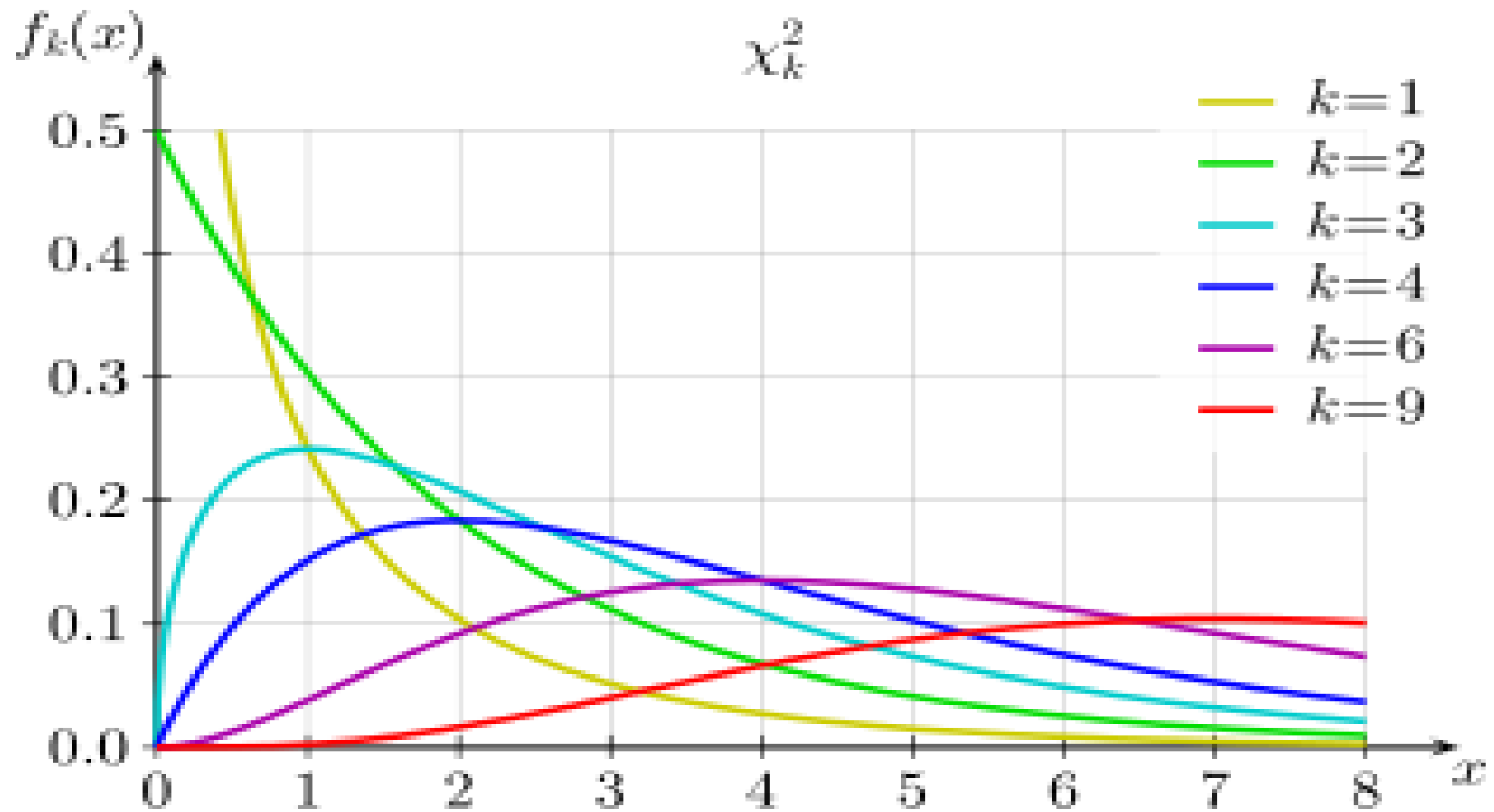
1. Prueba de bondad de ajuste.
2. Prueba de independencia
3. Prueba de homogeneidad
4. Prueba exacta de Fischer

- Se pone de manifiesto que, en cierto sentido, todas las pruebas de Ji-cuadrada que se utilizan pueden ser consideradas como pruebas de bondad de ajuste con las que se prueba precisamente la bondad de ajuste en las frecuencias observadas con respecto a las frecuencias que se esperarían si los datos se obtuvieran bajo alguna hipótesis o teoría en particular.

$$t = \frac{(X_1 - X_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

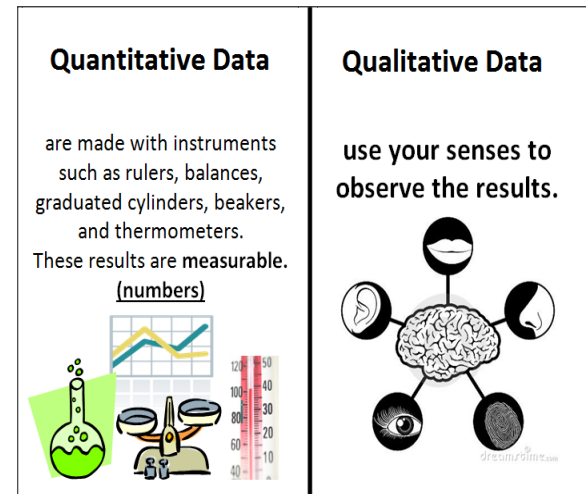
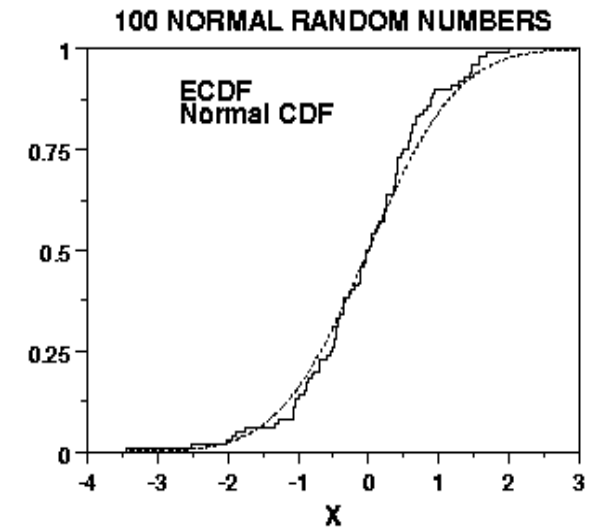


Chi- cuadrada



Propiedades de la Chi-Cuadrado

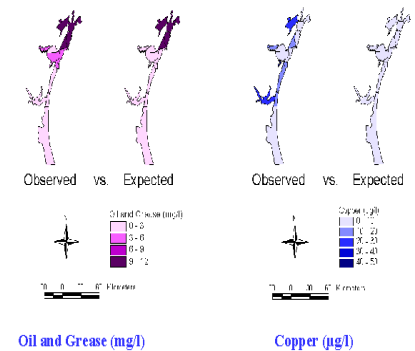
- Sin embargo, se reserva la expresión "bondad de ajuste" para utilizarla en un sentido más estricto, es decir para referirse a la comparación de la distribución de una muestra con alguna distribución teórica que se supone describe a la población de la cual se extrajo.
- El fundamento de la Ji o Chi-Cuadrado es para ser utilizada con datos cualitativos.
- Existen dos tipos de frecuencias en las que se centra el interés de esta parte de: frecuencias observadas y frecuencias esperadas.



Propiedades de la Chi-Cuadrado

- **Las frecuencias observadas** son el número de objetos o individuos en la muestra que caen dentro de las diversas categorías de la variable de interés. Por ejemplo, si se tiene una muestra de 100 pacientes hospitalizados se puede observar que 50 son casados, 30 son solteros, 15 son viudos y cinco son divorciados.
- **Las frecuencias esperadas** son el número de individuos u objetos en la muestra que se esperaría observar en alguna hipótesis nula respecto a la variable es verdadera. Por ejemplo, la hipótesis nula puede ser que las cuatro categorías de estado civil tienen igual representación dentro de la población de la que se extrajo la muestra. En este caso se esperaría que en este ejemplo hubiera 25 casados, 25 solteros, 25 viudos y 25 divorciados.

Observed vs. Expected



**JUST DOX
AS EXPECTED.**

Propiedades de la Chi-Cuadrado

¿Cómo obtenemos las frecuencias esperadas? A partir de las marginales, pronto lo veremos.



Propiedades de la Chi-Cuadrado

- El estadístico para cualquier tipo de las pruebas de antes, es el siguiente:

$$\chi^2_{(1)} = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

- Cuando la hipótesis nula es verdadera, X^2 sigue una distribución casi como X^2 con $k - r$ grados de libertad. En la determinación de los grados de libertad, k es igual al número de grupos para los que las frecuencias observadas y esperadas están disponibles, y r es el número de restricciones impuestas sobre las comparaciones dadas.
- Una restricción es impuesta cuando se fuerza la suma de las frecuencias esperadas para que sea igual a la suma de frecuencias observadas, y la restricción adicional es impuesta para cada parámetro que sea estimado a partir de la muestra.

Propiedades de la Chi-Cuadrado

- Bajo la cantidad o el estadístico: $\chi^2_{(1)} = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right]$
- Si dicho estadístico es pequeño si las frecuencias observadas y esperadas están muy cerca y será muy grande si las diferencias son muy grandes.
- El valor calculado de X^2 se compara contra el valor tabulado de X^2 con $k - r$ grados de libertad. La regla de decisión, entonces, es: rechazar H_0 si X^2 es mayor o igual que el valor tabulado de X^2 para el valor seleccionado de α . Bajo la prueba de independencia es un tanto distinto.



Propiedades de la Chi-Cuadrado

- Estamos en presencia de una prueba de hipótesis como las anteriores. ¿Qué entendemos entonces por el estadístico de $\chi^2_{(1)}$?



Chi- cuadrada

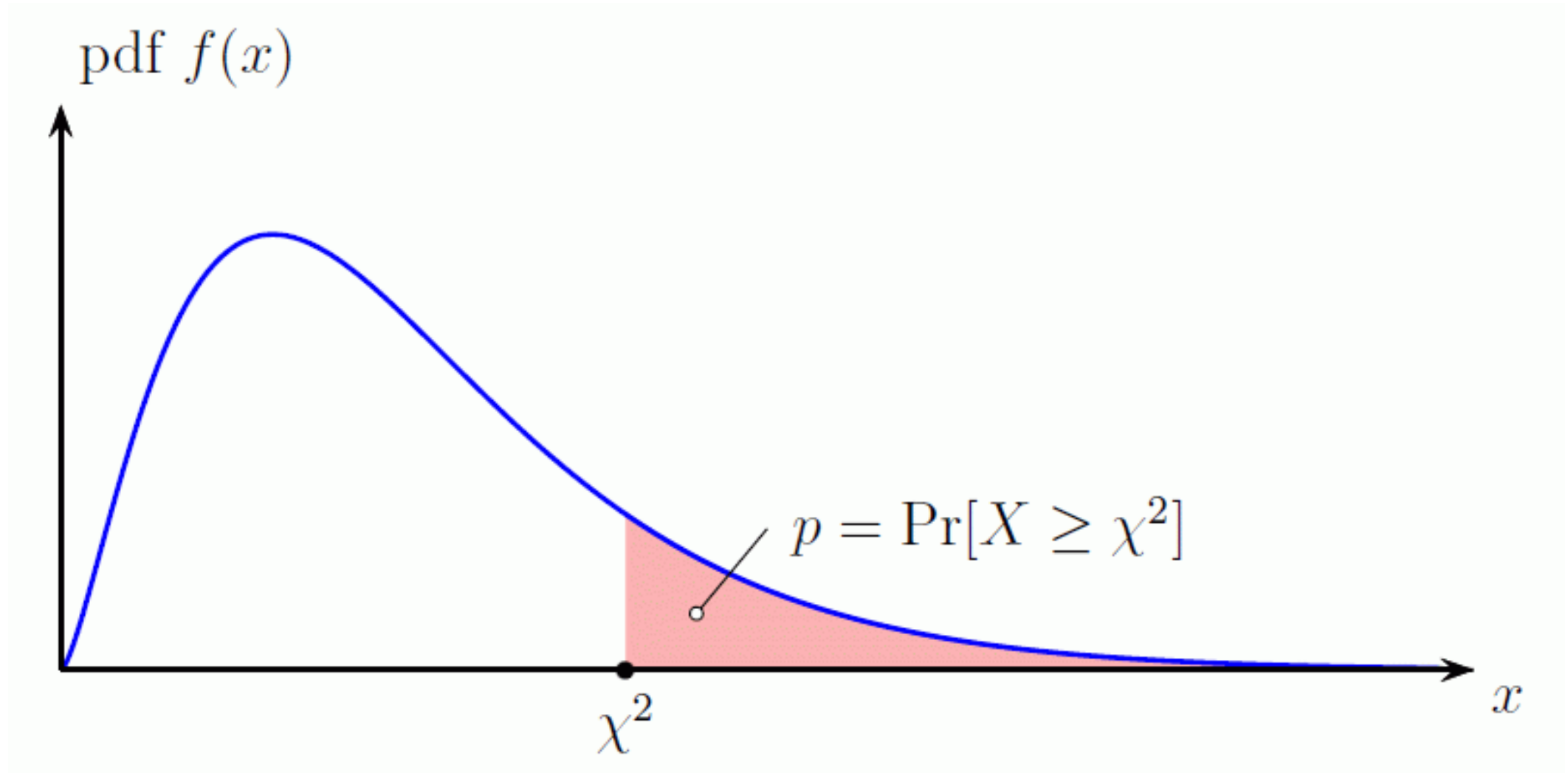


TABLE IV								
Degrees of Freedom	Chi-Square (χ^2) Distribution Area to the Right of Critical Value							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892

Índice

1

La Chi Cuadrada

2

Prueba de
independencia

Prueba de independencia

- Otro uso, quizá el mas frecuente, de la distribución ji-cuadrada es el de probar la hipótesis nula que indica que dos criterios de clasificación son independientes cuando se aplican al mismo conjunto de entidades.
- Se dice que dos criterios de clasificación son independientes si la distribución de un criterio es la misma, sin importar cómo sea la distribución del otro.
- Por ejemplo, si el estado socioeconómico y el área de residencia de los habitantes de cierta ciudad son independientes, se esperaría encontrar la misma proporción de familias en los grupos socioeconómicos bajo, medio y alto en todas las áreas de la ciudad.



Prueba de independencia

- **Tabla de contingencia:** La clasificación de un conjunto de entidades, de acuerdo con dos criterios, por ejemplo personas, se representa mediante una tabla en la que los r renglones representan los diversos niveles de uno de los criterios de clasificación, y las c columnas representan los diversos niveles del segundo criterio. Dicha tabla se conoce generalmente como *tabla de contingencia*.

Segundo criterio del nivel de clasificación	Primer criterio del nivel de clasificación					Total
	1	2	3	...	c	
1	N_{11}	N_{12}	N_{13}	...	N_{1c}	$N_{1.}$
2	N_{21}	N_{22}	N_{23}	...	N_{2c}	$N_{2.}$
3	N_{31}	N_{32}	N_{33}	...	N_{3c}	$N_{3.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮
r	N_{r1}	N_{r2}	N_{r3}		N_{rc}	$N_{r.}$
Total	$N_{.1}$	$N_{.2}$	$N_{.3}$...	$N_{.c}$	N

Prueba de independencia

- Se tiene interés en probar la hipótesis nula según la cual, en la población, los dos criterios de dosificación son independientes. Si la hipótesis es rechazada, se podrá concluir que los dos criterios de clasificación no son independientes. Se extrae una muestra de tamaño n de la población de entidades, y la frecuencia de ocurrencia de las entidades en la muestra, que corresponden a las casillas formadas por la intersección de los renglones y columnas de la tabla anterior.

Segundo criterio del nivel de clasificación	Primer criterio del nivel de clasificación					Total
	1	2	3	...	c	
1	n_{11}	n_{12}	n_{13}	...	n_{1c}	$n_{1.}$
2	n_{21}	n_{22}	n_{23}	...	n_{2c}	$n_{2.}$
3	n_{31}	n_{32}	n_{33}	...	n_{3c}	$n_{3.}$
...
r	n_{r1}	n_{r2}	n_{r3}	...	n_{rc}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$...	$n_{.c}$	n

Prueba de independencia

**¡Lo importante es
trabajar con los
marginales!**

Prueba de independencia

- ***Cálculo de las frecuencias esperadas*** Para cada celda se calculan las frecuencias esperadas, bajo la hipótesis nula de que los dos criterios de clasificación son independientes.
- Bajo la suposición de independencia, por ejemplo, se calcula la probabilidad de que uno de los n individuos representados en el renglón 1 y columna 1 de la tabla (es decir, casilla **11**) mediante la multiplicación de la probabilidad de que el individuo sea contabilizado en el renglón 1 por la probabilidad de que el individuo sea contabilizado en la columna 1.
- Esto es la obtención del número esperado por la marginal, y se denota como:

$$\frac{(n_{1.}n_{.1})}{n}$$

Prueba de independencia

- Se puede ver que para obtener la frecuencia esperada para una casilla dada, se multiplica el total del renglón en el que la casilla está localizada por el total de la columna en donde está la casilla, y se divide el producto entre el gran total.
- La comparación entre las frecuencias observadas y esperadas se llevan a cabo mediante el estadístico de la χ^2

$$\chi^2_{(1)} = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

- Si la discrepancia es suficientemente "pequeña", puede sostenerse la hipótesis nula. Si la discrepancia es suficientemente "grande", se rechaza la hipótesis nula y se concluye que los dos criterios de clasificación no son independientes.

Prueba de independencia

- Es posible demostrar que la X^2 definida de esta forma esta distribuida aproximadamente como una X^2 con $(r - 1)(c - 1)$ grados de libertad cuando la hipótesis nula es verdadera. Si el valor calculado X^2 es mayor que el valor tabulado de X^2 para alguna α , se rechaza la hipótesis nula en el nivel de significación. El procedimiento se ilustra con el ejemplo siguiente

	$P(X \leq x)$							
	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990
r	$\chi^2_{0.99}(r)$	$\chi^2_{0.975}(r)$	$\chi^2_{0.95}(r)$	$\chi^2_{0.90}(r)$	$\chi^2_{0.10}(r)$	$\chi^2_{0.05}(r)$	$\chi^2_{0.025}(r)$	$\chi^2_{0.01}(r)$
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34
4	0.297	0.484	0.711	1.064	7.779	9.488	11.14	13.28
5	0.554	0.831	1.145	1.610	9.236	11.07	12.83	15.09
6	0.872	1.237	1.635	2.204	10.594	12.59	14.45	16.81
7	1.239	1.690	2.167	2.833	12.017	14.07	16.01	18.48
8	1.646	2.180	2.733	3.490	13.362	15.51	17.54	20.09
9	2.088	2.700	3.325	4.168	14.684	16.92	19.02	21.67
10	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21

En resumen...

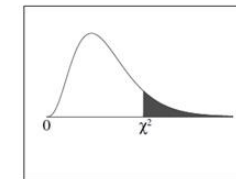
	Univer.	Secund.	
Uno	93 77,3 3,18	32 47,6 5,11	125
Dos	197 187,5 0,48	106 115,5 0,78	303
Ninguno	72 97,1 6,48	85 59,8 10,61	157
	362	223	585



$$\chi^2_{(1)} = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right]$$



Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_{\alpha}$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955

Observado
Esperado
(Obs-Esp)² / Esp

Prueba de independencia: ejemplo

- El propósito de un estudio realizado por Vermund era investigar la hipótesis de que las mujeres infectadas con VIH que también están infectadas con el papilomavirus humano (PVR) detectado mediante hibridación molecular, tienen mas probabilidad de tener anomalías citológicas cervicales que las mujeres con uno de los dos virus mencionados. Los datos que se muestran en la tabla siguiente tabla. Se pretende saber si es posible concluir que existe relación entre el estadio de PVR y la etapa de infección por VIH.



Prueba de independencia

- 1. Datos: la siguiente tabla

PVH	Seropositivo, sintomático	Seropositivo, asintomático	Seronegativo	Total
Positivo	23	4	10	37
Negativo	10	14	35	59
Total	33	18	45	96

Prueba de independencia

- 2. Supuestos: Se considera que la muestra disponible para el análisis es equivalente a una muestra aleatoria extraída de la población de interés.
- 3. H_0 : el estadio del PVH y la etapa de infección por VIH son independientes
 H_a : las dos variables son independientes.
- 4. La prueba de hipótesis:

$$\chi^2_{(1)} = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

Prueba de independencia

- Distribución de la prueba estadística: cuando H_0 es verdadera, la X^2 con $(r - 1)(c - 1)$, para el presente caso X^2 con $(2-1)(3-1) = (1)(2) = 2$ grados de libertad.
- **Regla de decisión:** se rechaza H_0 si el valor calculado de X^2 es mayor que o igual que 5.991.
- **Cálculo de las estadísticas de prueba:** La frecuencia esperada para la primera casilla es $(33 \cdot 37) / 96 = 12,72$. Las otras frecuencias esperadas se calculan de manera similar. Véase la siguiente tabla.

Prueba de independencia

PVH	VIH		Seronegativo	Total
	Seropositivo, sintomático	Seropositivo, asintomático		
Positivo	23 (12.72)	4 (6.94)	10 (17.34)	37
Negativo	10 (20.28)	14 (11.06)	35 (27.66)	59
Total	33	18	45	96

Prueba de independencia

- La estadística de prueba es

y esta se obtiene a partir de la ecuación

$$= \frac{(23 - 12.72)^2}{12.72} + \frac{(4 - 6.94)^2}{6.94} + \dots + \frac{(35 - 27.66)^2}{27.66}$$

$$= 8.30805 + 1.24548 + \dots + 1.94778 = 20.60081$$

Prueba de independencia

- 8. **Decisión estadística:** Se rechaza H_0 porque $20.60081 > 5.991$.
- 9. **Conclusión:** Se concluye que H_0 es falsa y que sí hay relación entre el estadio de PVH y la etapa de infección por VIH.
- 10. **Valor de p :** no se muestra por tablas.

Prueba de independencia

- Ejemplo 1: Como parte de su estudio, los investigadores reunieron información respecto al uso de agujas intercambiables por parte de adictos a drogas inyectables. Obtuvieron información para localizarlos a través de los archivos de instituciones de tratamiento para drogadictos y a través de investigaciones diseñadas para hacer participar a individuos que no reciben asesorías. ¿Es posible concluir, a partir de estos datos, que hay relación entre el uso de agujas intercambiables y el ser asesorados por la institución?

	Uso de agujas intercambiables			
	Regular	Ocasional	Nunca	No se sabe
Agencia	56	15	20	24
No agencia	19	6	16	53

Prueba de independencia

- Ejemplo 2: una muestra de 500 estudiantes universitarios participaron en un estudio para evaluar el nivel de conocimientos respecto a determinado grupo de enfermedades comunes. La tabla siguiente presenta la clasificación de los estudiantes de acuerdo con su principal campo de estudio y el nivel de conocimientos sobre el grupo de enfermedades. ¿Sugieren estos datos que existe una relación entre el conocimiento del grupo de enfermedades y el principal campo de estudio de los estudiantes de nivel superior de los cuales se extrajo esta muestra?

Campo de estudio	Conocimientos de enfermedades		Total
	Buena	Deficiente	
Premédico	31	91	122
Otro	19	359	378
Total	50	450	500

Índice

1

La Chi Cuadrada

2

Prueba de
independencia

3

Prueba de
homogeneidad

No lo veremos en este curso: si les interesa consultar el libro de “Bioestadística” de W. Daniel.

12.5 PRUEBA DE HOMOGENEIDAD

Una de las características de los ejemplos y ejercicios de la sección anterior es que, en cada caso, se supuso que la muestra total había sido extraída antes de que las entidades se agruparan de acuerdo con los criterios de clasificación. Es decir, se determinó el número observado de entidades que caen en cada casilla después que se extrajo la muestra. Como resultado, los totales de renglones y columnas son cantidades aleatorias que no están bajo el control del investigador. Se considera que la muestra extraída en estas condiciones es una sola muestra tomada de una sola población. Sin embargo, en ocasiones, los totales de renglones o columnas pueden estar bajo el control del investigador, es decir, el investigador puede especificar que las muestras independientes sean extraídas de varias poblaciones. En este caso, se dice que un conjunto de totales marginales es *fijo*, mientras que el otro conjunto, que corresponde al criterio de clasificación aplicado a las muestras, es *aleatorio*. Como se ha visto, el primer procedimiento conduce a una prueba de independencia de ji-cuadrada. La segunda situación conduce a una *prueba de homogeneidad* de ji-cuadrada. Las dos situaciones no sólo comprenden procedimientos de muestreo distintos, sino que también conducen a preguntas e hipótesis nulas distintas. La prueba de independencia se ocupa de la pregunta: ¿son independientes los criterios de clasificación? La prueba de homogeneidad se ocupa de la pregunta: ¿las muestras ex-

Índice

1

La Chi Cuadrada

4

Otros análisis de
frecuencias: prueba
exacta de Fischer

2

Prueba de
independencia

3

Prueba de
homogeneidad

No lo veremos en este curso: si les interesa consultar el libro de “Bioestadística” de W. Daniel.

12.6 PRUEBA EXACTA DE FISHER

En ocasiones se tienen datos que pueden resumirse en una tabla de contingencia de 2×2 , pero los datos provienen de muestras muy pequeñas. La prueba de ji-cuadrada no es un método adecuado de análisis si los requerimientos mínimos de frecuencias esperadas no se cumplen. Si, por ejemplo, n es menor que 20 o si n está entre 20 y 40 y una de las frecuencias esperadas es menor que 5, se debe evitar el uso de la prueba de ji-cuadrada.

Fisher (7, 8) propuso a mediados de 1930, casi al mismo tiempo que Irwin (9) y Yates (10), una prueba que sustituye a la prueba ji-cuadrada cuando los requeri-

Conclusión

- El presente capítulo presenta una técnica analítica para analizar datos cualitativos.
- Se presenta la distribución de Chi-cuadrado, y su fundamento en el análisis de la información.
- La prueba de bondad de ajuste, prueba de independencia y prueba de homogeneidad son algunas herramientas para datos cualitativos que se quieren analizar mediante técnicas estadísticas.
- La prueba de independencia fue la estudiada en el presente curso.



*The
End*