CAPÍTULO 7

INFERENCIA ESTADÍSTICA CON DATOS DE FRECUENCIAS

7.1 INTRODUCCIÓN

En el presente capítulo se aborda el tema de la realización de inferencias respecto a una población en situaciones en las que los datos disponibles son las **frecuencias** (o sea, el número de veces) con las que se han presentado determinados sucesos.

En general el objetivo perseguido en estos estudios es el de obtener conclusiones respecto a los valores poblacionales de una o más probabilidades. Estos valores se estiman en las muestras por las proporciones relativas con las que se presentan los sucesos correspondientes, por lo que en general los temas abordados en esta unidad tienen que ver con el contraste de hipótesis respecto a proporciones.

Se aborda, en primer lugar, el problema del **estudio de la hipótesis de que unas determinadas probabilidades tienen en la población ciertos valores predeterminados (contraste de proporciones)**. Se introducen al respecto, en primer lugar, ciertos conceptos fundamentales en Inferencia Estadística, como los de Hipótesis Nula y Riesgo de 1ª y de 2ª especie Para el contraste de de proporciones se presenta un test muy importante y conocido, desarrollado hace más de 100 años por Karl Pearson. Este test se basa en la distribución Gi-dos (χ^2), que se presenta de forma sucinta.

En el siguiente apartado se desarrolla cómo el test Gi-dos puede aplicarse para estudiar la **independencia entre dos variables cualitativas**, a partir de las frecuencias observadas de las distintas combinaciones recogidas en una **tabla de contingencia**. También se estudia la aplicación de dicho test para investigar la **homogeneidad** en una tabla de contingencia cuyas filas no constituyen valores aleatorios de una característica cualitativa, sino tratamientos alternativos fijados voluntariamente.

Finalmente se comentan muy sumariamente ciertas generalizaciones de las técnicas vistas en la sección anterior, aplicables al estudio de tablas de frecuencias de grandes dimensiones o al de hipertablas de contingencia.

7.2 EJEMPLOS DE TESTS DE CONTRASTE DE PROPORCIONES

Ejemplo 1: al lanzar 50 veces un dado se ha obtenido 10 veces el uno, 5 veces el dos, 8 veces el tres, 8 veces el cuatro, 7 veces el cinco y las 12 veces restantes el 6. ¿Es admisible la hipótesis de que el dado es simétrico?

Ejemplo 2: De los 80 nacimientos de padres chinos que se produjeron un año en Valencia 48 resultaron ser varones y los 32 restantes mujeres. ¿Demuestra este resultado que en los nacimientos chinos los varones son más frecuentes que las mujeres o, por el contrario, la muestra es compatible con la hipótesis de que las proporciones de ambos sexos son iguales?

Ejemplo 3: en cruces entre híbridos de ratones blancos y negros se obtuvieron 15 ratones blancos y 25 negros. ¿Son compatibles estos resultados con la hipótesis de que la herencia

del color en los ratones sigue un modelo mendeliano simple que implica unas proporciones teóricas de 1/4 para el carácter recesivo (el blanco) y 3/4 para el dominante (el negro)?

En los tres ejemplos expuestos existen J sucesos mutuamente excluyentes que constituyen una partición de las respectivas poblaciones.

Autoevaluación: ¿Cuáles son estos J sucesos en cada caso?

Se trata de estudiar a partir de los resultados de una muestra si es admisible la hipótesis de que las probabilidades p_j de los diferentes sucesos tienen unos valores p_{j0} previamente postulados:

$$p_1=p_{10} \quad p_2=p_{20} \ldots . \ p_J=p_{J0}$$

frente a la hipótesis alternativa de que al menos un p_j es diferente de p_{j0}

7.3 HIPÓTESIS NULA

La hipótesis cuya admisibilidad se quiere analizar a partir de unos datos se denomina en Estadística **Hipótesis Nula** (simbolizándose en general como $\mathbf{H_0}$).

En general H_0 suele reflejar el estado actual de conocimiento (quizás sería más preciso decir de desconocimiento) sobre la cuestión en estudio.

Así, en los dos primeros ejemplos parece lógico asumir (**¡mientras no "demuestre" lo contrario!)**, que el dado es simétrico, o que los varones y las mujeres son igual de frecuentes en los nacimientos

En contextos científicos, la Hipótesis Nula refleja en general una posición "conservadora" frente a nuevas hipótesis o teorías:

Por ejemplo, si se está investigando la posible eficacia de un nuevo fármaco frente a la hipertensión, la posición intelectual de partida en la investigación (recogida en la $\mathbf{H_0}$) es que el fármaco no es eficaz, a no ser que haya en los datos recogidos una "evidencia fuerte" en contra de dicha hipótesis.

En contextos industriales la $\mathbf{H_0}$ refleja frecuentemente una actitud de prudente escepticismo. Por ejemplo: no nos creemos que un nuevo proceso B sea mejor que el proceso tradicional A, a no ser que en los datos experimentales haya una "fuerte evidencia" contra la $\mathbf{H_0}$ que supone que los dos procesos son iguales (¡mientras no "se demuestre" lo contario!).

¿Y cómo "se demuestra" a partir de datos que la Hipótesis Nula" es falsa?

Idea intuitiva: Viendo que $\mathbf{H_0}$ es difícilmente compatible con los datos observados, es decir, viendo que unos datos como los observados serían muy poco probables si la $\mathbf{H_0}$ fuera cierta

7.4 RIESGOS DE 1ª Y DE 2ª ESPECIE

Al realizar un test estadístico sobre una determinada Hipótesis Nula H_0, es posible llegar a dos tipos de conclusiones erróneas:

Error de 1ª especie: el que se comete cuando se rechaza como falsa una H_0 que en realidad es cierta

Error de 2ª especie: el que se comete cuando se acepta como cierta una H_0 que en realidad es falsa

- En un estudio para ver si un dado es simétrico con el fin de detectar si un jugador que lo está utilizando es un tramposo:

  ¿Qué sería un error de 1ª especie? ¿Cuáles serían las consecuencias de cometerlo?

  ¿Qué sería un error de 2ª especie? ¿Cuáles serían las consecuencias de cometerlo?

- En un estudio para ver si un nuevo fármaco es efectivo contra la hipertensión:

  ¿Qué sería un error de 1ª especie? ¿Cuáles serían las consecuencias de cometerlo?

  ¿Qué sería un error de 2ª especie? ¿Cuáles serían las consecuencias de cometerlo?

A la probabilidad que un determinado procedimiento estadístico tiene de cometer un error de 1ª especie se le denomina **Riesgo de 1ª especie**, representándose en general mediante la letra griega α. En general, para rechazar una H_0 se exige que el α asociado a dicha decisión sea "bajo" (es muy frecuente que se utilice 0.05 como valor límite para α)

A la probabilidad que un determinado procedimiento estadístico tiene de cometer un error de 2ª especie se le denomina **Riesgo de 2ª especie**, representándose en general mediante como β (y se dice que **1 -** β es la **potencia** del procedimiento estadístico).

7.5 CONTRASTE DE PROPORCIONES

Como ya se ha indicado, se trata de estudiar a partir de las frecuencias $x_1, \ldots, x_I$ con las que se han presentado en los N individuos de una muestra un conjunto de I sucesos $A_1, \ldots, A_I$ mutuamente excluyentes y cuya reunión es el suceso seguro (los A_i implican, por lo tanto, una partición de la población global), si es admisible la hipótesis nula de que las probabilidades p_j de los diferentes sucesos tienen unos valores p_{j0} previamente postulados:

$$\mathbf{H_0:}\ p_1=p_{10}\quad p_2=p_{20}\ \ldots.\ p_J=p_{J0}$$

(Lógicamente se habrá de cumplir que $\sum_{i=1}^{I} p_{io} = 1$)

La hipótesis alternativa es que algunos de los p_j son diferentes de p_{j0}

Si la probabilidad de un suceso A_i es realmente p_{io}, en un total de N observaciones el suceso A_i debe aparecer en promedio Np_{io} veces.

A Np_{i0} se le denomina: valor “teórico” en el caso de ser cierta la H_o.

Intuitivamente parece razonable que si todas las x_i muestrales son “parecidas” a las Np_{j0} teóricas la hipótesis nula H_0 será aceptable, mientras que si algunas de las x_i “difieren mucho” de sus respectivos valores teóricos Np_{j0} la hipótesis nula deberá rechazarse.

Sin embargo...

¿qué debemos entender como “parecidas” o como “diferir mucho”?

Vamos a ver a continuación un test que permite contrastar la hipótesis nula que estamos estudiando.

Este test define una particular medida “d” de la discrepancia entre las frecuencias observadas x_i y las teóricas Np_{j0}.

Se conoce, al menos aproximadamente, la distribución que sigue “d” cuando es cierta la hipótesis nula, y se sabe que dicho estadístico tiende (como es lógico) a tomar valores más elevados si la hipótesis nula es falsa. Dicho estadístico “d” puede por tanto utilizarse, como vamos a ver, para estudiar si H_0 es o no admisible.

Test Gi-dos

Estudiando este problema Karl Pearson propuso un estadístico "d" para medir la discrepancia entre los valores observados en la muestra x_j y los valores teóricos bajo la hipótesis nula Np_{j0}. Dicho estadístico tiene por expresión:

$$d = \sum_{j=1}^{j=J} \frac{(observados - teoricos)^2}{teoricos}$$

Obviamente "d" sólo será igual a cero si coinciden totalmente los valores observados con los teóricos, tomando valores tanto más elevados cuanto mayor sea la diferencia entre éstos y aquéllos.

Pearson demostró que cuando H_0 es cierta el estadístico "d" sigue, aproximadamente, una determinada distribución estadística denominada distribución Gi-dos con I-1 grados de libertad (χ^2_{I-1}). (Esta distribución se estudia en el apartado siguiente.)
Por el contrario, si H_0 es falsa, los valores que toma "d" tienden a ser más elevados que los que cabría esperar para una variable Gi-dos con dichos grados de libertad.

Por lo tanto H_0 se rechazará si si d resulta "demasiado grande" para ser el valor de una χ^2_{I-1}, o sea, si la probabilidad de que una χ^2_{I-1} sea tan grande como "d" es pequeña (menor que el riesgo de 1ª especie α que se esté dispuesto a asumir.

Nota

¿Qué quiere decir que el estadístico "d" sigue una determinada distribución estadística?

Como veremos en el Apartado 8.6 del capítulo siguiente, a cualquier cantidad que pueda obtenerse operando con los datos de una muestra se le denomina un "estadístico". Asumiendo que la muestra en cuestión ha sido obtenida de forma aleatoria, todo estadístico es, en el fondo, una variable aleatoria definida sobre la población de todas las posibles muestras que podrían haberse obtenido

En nuestro caso, en concreto, el estadístico "d" sigue, aproximadamente, una distribución especial, la Gi-dos de Pearson que se estudia en el apartado siguiente

Calidad de la aproximación

Tanto en el test de la razón de verosimilitud generalizada como en el de Pearson, la distribución aproximada Gi-dos de "d" cuando H_0 es cierta, se basa en aproximar las x_j que son binomiales por variables normales. Para que esta aproximación no sea muy grosera es preciso que los valores teóricos Np_{j0} no sean muy bajos. Tradicionalmente se acostumbra a exigir que, a ser posible, estos valores sean como mínimo iguales a 5.

7.6 LA DISTRIBUCIÓN GI-DOS

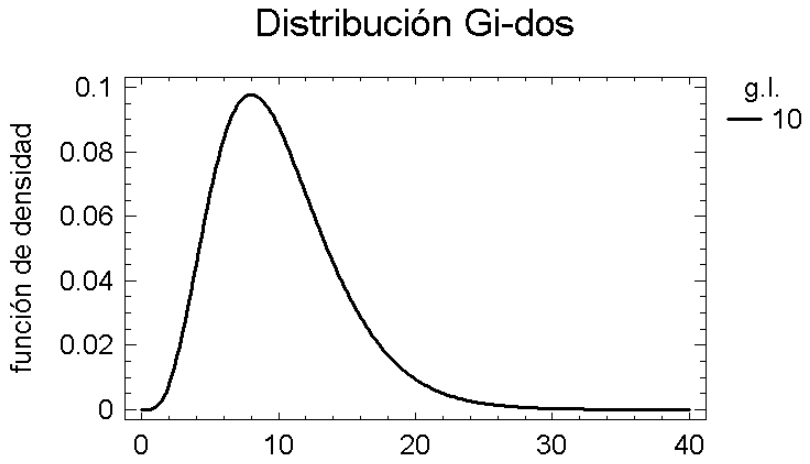
(Esta distribución se denomina también Gi-cuadrado, Chi-dos o Chi-cuadrado, representándose mediante la letra griega Gi (o Chi) como χ^2_ν, siendo ν los "grados de libertad de la distribución)

Por definición una variable aleatoria **Y** sigue una distribución Gi-dos con ν grados de libertad si es la suma de los cuadrados de ν variables N[0,1] independientes.

Así, si $X_1,...,X_\nu$ son variables N[0,1] independientes:

$$\mathbf{Y} = X_1^2 + ... + X_\nu^2 \sim \text{Gi-dos con } \nu \text{ grados de libertad } \chi^2_\nu$$

La siguiente figura refleja la forma de la función de densidad de una variable Gi-dos con 10 grados de libertad χ^2_{10}



Como puede apreciarse la distribución sólo toma valores positivos (lo que era obvio por tratarse de una suma de cuadrados) y es asimétrica positiva. Dicha asimetría, sin embargo, decrece a medida que aumentan los grados de libertad de la variable.

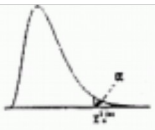
Se demuestra que la media de una variable Gi-dos con ν grados de libertad es precisamente igual a ν, y la varianza es igual a 2ν.

La tabla de la hoja siguiente da, para diferentes valores de α y de ν, el valor x_α tal que la probabilidad de que una Gi-dos con ν grados de libertad sea mayor que x_α es igual a α.

Evaluación: Calcular la probabilidad de que una variable Gi-dos con 10 grados de libertad esté comprendida entre 3.94 y 23.21

DISTRIBUCIÓN χ^2

D. E. I. O.



Grados de libertad - n	Probabilidad de una cola - α												
	0,005	0,010	0,025	0,050	0,100	0,250	0,500	0,750	0,900	0,950	0,975	0,990	0,995
1	7,8794	6,6349	5,0239	3,8415	2,7055	1,3233	0,4549	0,1015	0,0158	0,0039	0,0010	0,0002	0,0000
2	10,597	9,210	7,378	5,991	4,605	2,773	1,386	0,575	0,211	0,103	0,051	0,020	0,010
3	12,838	11,345	9,348	7,815	6,251	4,108	2,366	1,213	0,584	0,352	0,216	0,115	0,072
4	14,860	13,277	11,143	9,488	7,779	5,385	3,357	1,923	1,064	0,711	0,484	0,297	0,207
5	16,750	15,086	12,832	11,070	9,236	6,626	4,351	2,675	1,610	1,145	0,831	0,554	0,412
6	18,548	16,812	14,449	12,592	10,645	7,841	5,348	3,455	2,204	1,635	1,237	0,872	0,676
7	20,278	18,475	16,013	14,067	12,017	9,037	6,346	4,255	2,833	2,167	1,690	1,239	0,989
8	21,955	20,090	17,535	15,507	13,362	10,219	7,344	5,071	3,490	2,733	2,180	1,647	1,344
9	23,589	21,666	19,023	16,919	14,684	11,389	8,343	5,899	4,168	3,325	2,700	2,088	1,735
10	25,188	23,209	20,483	18,307	15,987	12,549	9,342	6,737	4,865	3,940	3,247	2,558	2,156
11	26,757	24,725	21,920	19,675	17,275	13,701	10,341	7,584	5,578	4,575	3,816	3,054	2,603
12	28,300	26,217	23,337	21,026	18,549	14,845	11,340	8,438	6,304	5,226	4,404	3,571	3,074
13	29,819	27,688	24,736	22,362	19,812	15,984	12,340	9,299	7,042	5,892	5,009	4,107	3,565
14	31,319	29,141	26,119	23,685	21,064	17,117	13,339	10,165	7,790	6,571	5,629	4,660	4,075
15	32,801	30,578	27,488	24,996	22,307	18,245	14,339	11,037	8,547	7,261	6,262	5,229	4,601
16	34,267	32,000	28,845	26,296	23,542	19,369	15,339	11,912	9,312	7,962	6,908	5,812	5,142
17	35,718	33,409	30,191	27,587	24,769	20,489	16,338	12,792	10,085	8,672	7,564	6,408	5,697
18	37,156	34,805	31,526	28,869	25,989	21,605	17,338	13,675	10,865	9,390	8,231	7,015	6,265
19	38,582	36,191	32,852	30,144	27,204	22,718	18,338	14,562	11,651	10,117	8,907	7,633	6,844
20	39,997	37,566	34,170	31,410	28,412	23,828	19,337	15,452	12,443	10,851	9,591	8,260	7,434
21	41,401	38,932	35,479	32,671	29,615	24,935	20,337	16,344	13,240	11,591	10,283	8,897	8,034
22	42,796	40,289	36,781	33,924	30,813	26,039	21,337	17,240	14,041	12,338	10,982	9,542	8,643
23	44,181	41,638	38,076	35,172	32,007	27,141	22,337	18,137	14,848	13,091	11,689	10,196	9,260
24	45,558	42,980	39,364	36,415	33,196	28,241	23,337	19,037	15,659	13,848	12,401	10,856	9,886
25	46,928	44,314	40,647	37,652	34,382	29,339	24,337	19,939	16,473	14,611	13,120	11,524	10,520
26	48,290	45,642	41,923	38,885	35,563	30,435	25,336	20,843	17,292	15,379	13,844	12,198	11,160
27	49,645	46,963	43,195	40,113	36,741	31,528	26,336	21,749	18,114	16,151	14,573	12,878	11,808
28	50,994	48,278	44,461	41,337	37,916	32,620	27,336	22,657	18,939	16,928	15,308	13,565	12,461
29	52,336	49,588	45,722	42,557	39,087	33,711	28,336	23,567	19,768	17,708	16,047	14,256	13,121
30	53,672	50,892	46,979	43,773	40,256	34,800	29,336	24,478	20,599	18,493	16,791	14,953	13,787
31	55,002	52,191	48,232	44,985	41,422	35,887	30,336	25,390	21,434	19,281	17,539	15,655	14,458
32	56,328	53,486	49,480	46,194	42,585	36,973	31,336	26,304	22,271	20,072	18,291	16,362	15,134
33	57,648	54,775	50,725	47,400	43,745	38,058	32,336	27,219	23,110	20,867	19,047	17,073	15,815
34	58,964	56,061	51,966	48,602	44,903	39,141	33,336	28,136	23,952	21,664	19,806	17,789	16,501
35	60,275	57,342	53,203	49,802	46,059	40,223	34,336	29,054	24,797	22,465	20,569	18,509	17,192
36	61,581	58,619	54,437	50,998	47,212	41,304	35,336	29,973	25,643	23,269	21,336	19,233	17,887
37	62,883	59,893	55,668	52,192	48,363	42,383	36,336	30,893	26,492	24,075	22,106	19,960	18,586
38	64,181	61,162	56,895	53,384	49,513	43,462	37,335	31,815	27,343	24,884	22,878	20,691	19,289
39	65,475	62,428	58,120	54,572	50,660	44,539	38,335	32,737	28,196	25,695	23,654	21,426	19,996
40	66,766	63,691	59,342	55,758	51,805	45,616	39,335	33,660	29,051	26,509	24,433	22,164	20,707
41	68,053	64,950	60,561	56,942	52,949	46,692	40,335	34,585	29,907	27,326	25,215	22,906	21,421
42	69,336	66,206	61,777	58,124	54,090	47,766	41,335	35,510	30,765	28,144	25,999	23,650	22,138
43	70,616	67,459	62,990	59,304	55,230	48,840	42,335	36,436	31,625	28,965	26,785	24,398	22,860
44	71,892	68,710	64,201	60,481	56,369	49,913	43,335	37,363	32,487	29,788	27,575	25,148	23,584
45	73,166	69,957	65,410	61,656	57,505	50,985	44,335	38,291	33,350	30,612	28,366	25,901	24,311
46	74,440	71,202	66,616	62,830	58,641	52,056	45,335	39,220	34,215	31,439	29,160	26,657	25,041
47	75,704	72,443	67,821	64,001	59,774	53,127	46,335	40,149	35,081	32,268	29,956	27,416	25,775
48	76,969	73,683	69,023	65,171	60,907	54,196	47,335	41,079	35,949	33,098	30,755	28,177	26,511
49	78,231	74,919	70,222	66,339	62,038	55,265	48,335	42,010	36,818	33,930	31,555	28,941	27,249
50	79,490	76,154	71,420	67,505	63,167	56,334	49,335	42,942	37,689	34,764	32,357	29,707	27,991
60	91,952	88,379	83,298	79,082	74,397	66,981	59,335	52,294	46,459	43,188	40,482	37,485	35,534
70	104,215	100,425	95,023	90,531	85,527	77,577	69,334	61,698	55,329	51,739	48,758	45,442	43,275
80	116,321	112,329	106,629	101,879	96,578	88,130	79,334	71,145	64,278	60,391	57,153	53,540	51,172
90	128,299	124,116	118,136	113,145	107,565	98,650	89,334	80,625	73,291	69,126	65,647	61,754	59,196
100	140,170	135,810	129,561	124,342	118,498	109,141	99,334	90,133	82,358	77,929	74,222	70,065	67,328

7.7 ANÁLISIS DESCRIPTIVO DE TABLAS DE FRECUENCIAS

En el análisis de **variables aleatorias bidimensionales de naturaleza cualitativa** los datos recogidos sobre una muestra de individuos pueden presentarse como una tabla de frecuencias, cuyo elemento genérico x_{ij} indica el número de veces que ha aparecido asociada la variante i de la primera variable con la correspondiente a la variante j de la segunda..

A este tipo de tablas se les denomina en Estadística *tablas de contingencia*.

A este tipo de tablas de frecuencias, que son especialmente útiles cuando las dos variables estudiadas son de naturaleza cualitativa, se les denomina en Estadística **tablas de contingencia.**

La tabla siguiente está obtenida a partir de las respuestas dadas por los alumnos a la encuesta, y cruza las variables SEXO y POLITICA.

Como hemos señalado cada casilla recoge el número de individuos que tienen los valores correspondientes para las dos variables (SEXO y POLITICA). A la derecha de la tabla se recogen las frecuencias totales, tanto absolutas como relativas (estas últimas expresadas en porcentaje), para las dos alternativas de la variable SEXO. A estas frecuencias se les denomina **frecuencias marginales**. En la parte inferior de la tabla se recogen las frecuencias marginales para las diferentes alternativas de la variable POLITICA.

Con el fin de estudiar si la orientación política es similar en los dos sexos conviene calcular la frecuencia relativa de cada casilla respecto al total de la fila correspondiente. Estas frecuencias relativas, que se recogen en la tabla en porcentaje en la parte inferior de cada casilla, se denominan **frecuencias relativas condicionales** de POLITICA en función de los valores de SEXO.

Es importante que las frecuencias relativas se calculen adecuadamente respecto al total de la fila o de la columna correspondiente, según sea relevante para los objetivos perseguidos en un determinado estudio

Autoevaluación: obtener la tabla de frecuencias anterior, pero calculando las frecuencias relativas respecto al total de cada columna en vez de respecto al total de cada fila. ¿Se deduce de esta nueva tabla que los chicos son más de izquierdas que las chicas, porque el 60% de los de izquierdas son chicos y sólo el 40% son chicas?

Autoevaluación: Cuando se dice que el 80% de los matrimonios que se divorcian en España son católicos, ¿de qué tipo de frecuencia relativa se está hablando? ¿Qué frecuencias relativas consideras que sería interesante comparar con el fin de estudiar la relación entre religión y divorcio?

Autoevaluación: Se ha constatado que el 30% de los conductores implicados en accidentes de tráfico tenían una tasa de alcohol en sangre superior a la permitida. ¿Se deduce de ello que el alcohol es responsable del 30% de los accidentes de tráfico? ¿Se deduce, por el contrario, que puesto que el 70% de los conductores accidentados tenían una tasa de alcohol inferior a la máxima permitida y sólo el 30% la tenían superior, el no beber alcohol incrementa la probabilidad de accidentes? (Ver respuesta en el Anejo al final del Tema)

Autoevaluación: Estudiar, a partir de la tabla cruzada correspondiente, la relación existente en la muestra entre POLITICA y PROBLEMA. ¿Crees que las conclusiones obtenidas respecto a la relación entre "posición política" y "problema considerado más importante" se repetirían en línea generales en otras muestras extraídas de la población constituida por la juventud universitaria española?

7.8 INFERENCIA EN TABLAS DE FRECUENCIAS

En el análisis a nivel inferencial de estas tablas se presentan dos problemas, aparentemente muy parecidos pero formalmente claramente diferentes, según que la pertenencia de los individuos observados a una u otra fila de la tabla tenga o no carácter aleatorio:

- ***Tests de Independencia***
- ***Tests de Homogeneidad***

7.8.1 Test de independencia

En una población existen definidas dos variables aleatorias de tipo cualitativo, la primera de ellas con I variantes y la segunda con J variantes. La tabla de contingencia recoge las frecuencias observadas de cada una de las IxJ combinaciones posibles de ambas variables en una muestra de N individuos seleccionados al azar de dicha población.

Autoevaluación: En la población de los estudiantes universitarios valencianos se consideran las variables aleatorias cualitativas POLITICA y TRANSPORTE, que fueron definidas en la encuesta realizada en la primera clase del curso. Los 131 alumnos de una determinada clase en el curso 89/90 pueden considerarse una muestra representativa de esta población, estando sus respuestas guardadas en el fichero CURS8990. Construir la tabla de contingencia correspondiente a estas dos variables.

En el primer ejemplo de la autoevaluación anterior se desea estudiar la hipótesis de si en la población el medio de transporte utilizado es independiente de la opción política del alumno. Es decir, ¿puede admitirse que poblacionalmente la proporción de alumnos que utilizan los distintos medios de transporte es la misma en los alumnos de las diferentes opciones políticas? O, dicho de otra forma equivalente, es admisible que poblacionalmente la proporción de alumnos de las diferentes opciones políticas es la misma entre los que vienen en coche propio, que entre los que vienen en autobús, o entre los que utilizan cualquiera de los otros métodos de transporte.

El problema que se trata de estudiar en estos casos es el de la posible independencia entre las dos variables cualitativas consideradas en la población.

Tal como se ve a continuación, este problema puede considerarse como un caso particular de ajuste a una distribución teórica, en el que hay que estimar los parámetros de la misma.

Obtención del estadístico “d”

Si denominamos $p_{i.}$ a la probabilidad de la variante i-ésima de la primera variable, y $p_{.j}$ a la de la variante j-ésima de la segunda variable, la probabilidad p_{ij} asociada a la casilla (i,j) de la tabla será, si es cierta la H_0 de independencia entre ambas variables, igual a $p_{i.}xp_{.j}$

$$p_{ij} = p_{i.}xp_{.j} \qquad \text{si es cierta la } H_0 \text{ de independencia}$$

Por tanto el estadístico “d” para medir la discrepancia entre los valores observados x_{ij} y los teóricos bajo la hipótesis de independencia será

$$d = \sum_{i=1}^{i=I}\sum_{j=1}^{j=J} \frac{(x_{ij} - Np_{i.}\,p_{.j})^2}{Np_{i.}\,p_{.j}}$$

donde los parámetros $p_{i.}$ y $p_{.j}$ deben sustituirse por sus estimadores obtenidos a partir de los datos

Se demuestra que estos estimadores son:

$p^*_{i.}$ = (total de la fila i)/N = $x_{i.}$/N

$p^*_{.j}$ = (total de la columna j)/N = $x_{.j}$/N

con lo que la expresión de los valores teóricos esperados para cada casilla bajo la hipótesis de independencia será :

$$t_{ij} = Np^*_{i.}p^*_{.j} = N\frac{x_{i.}}{N}\frac{x_{.j}}{N} = \frac{x_{i.}x_{.j}}{N} = \frac{(\text{total fila i})(\text{total columna j})}{\text{total general}}$$

Se demuestra que el estadístico “d” sigue, si es cierta la H_0 de independencia, una distribución Gi-dos con (I-1)(J-1) grados de libertad

Autoevaluación: Estudiar a partir de los datos del fichero CURS8990 si es admisible la hipótesis de que el lugar de residencia es independiente del sexo.

Autoevaluación: Estudiar a partir de los datos del mismo fichero si es admisible la hipótesis de independencia entre POLITICA y TRANSPORTE. Repetir este estudio pero considerando sólo los estudiantes que no se declaran de derechas.

7.8.2 Test de homogeneidad

Un problema diferente se presenta cuando las filas no representan diferentes alternativas seleccionadas al azar de una variable aleatoria cualitativa, sino distintos tratamientos dados a los individuos de la población o, por ejemplo, muestras recogidas en diferentes condiciones.

Se trata en estos caso de contrastar la hipótesis de los distintos tratamientos (o las diferentes muestras en el segundo ejemplo) son homogéneos respecto a las características representadas por las columnas.

Ejemplo 1: Se compararon 3 tratamientos para curar cierta enfermedad, obteniéndose los resultados que se recogen en la tabla adjunta.

	Curados	No Curados
Tratam. 1	25	12
Tratam. 2	15	18
Tratam. 3	11	10

¿Pueden considerarse los tratamientos igual de efectivos?

Ejemplo 2: dos encuestas de consumo realizadas seleccionando al azar hogares en dos poblaciones diferentes obtuvieron las siguientes clasificación de hogares según el nivel de renta :

	Renta anual (millones de pts)			
	$< 1_1$	$1_1 - 3_1$	$3_1 - 5_1$	$> 5_1$
Encuesta 1	300	180	65	8
Encuesta 2	58	31	22	5

¿Pueden considerarse homogéneas las poblaciones muestreadas? Es decir, ¿puede considerarse la proporción poblacional de hogares en las diferentes categorías de renta similar en las dos poblaciones?

En general si I es el número de las filas y J el de las columnas, la distribución de la variable asociada a las columnas en cada uno de las I filas o tratamientos, vendrá caracterizada por las probabilidades $p_{j/i}$ de cada una de las columnas dentro de esa fila.

La hipótesis nula de homogeneidad implica que las $p_{j/i}$ son las mismas en las I filas, e implica la existencia de un conjunto de sólo J-1 parámetros independientes (los $p_{.j}$) asociados a la tabla.

Si $x_{i.}$ es el número total de observaciones de la fila i, el valor teórico esperado en la casilla (i,,j) bajo la hipótesis nula es $x_{i.}p^*_{.j}$, donde $p^*_{.j}$ es el estimador de $p_{.j}$, bajo la hipótesis nula de

homogeneidad, que se demuestra fácilmente que es $p^*_{.j} = x_{.j}/N$, siendo N el total de observaciones.

El valor teórico correspondiente bajo H_0 a la casilla (i,j) será por tanto

$$t_{ij} = x_{i.}\frac{x_{.j}}{N} = \frac{x_{i.}x_{.j}}{N} = \frac{(\text{total fila i})(\text{total columna j})}{\text{total general}}$$

que es exactamente el mismo que se obtuvo en el caso del test de independencia.

El estadístico "d" para contrastar la hipótesis de homogeneidad será, por tanto,

$$d = \sum_{i=1}^{i=I}\sum_{j=1}^{j=J}\frac{(x_{ij} - t_{ij})^2}{t_{ij}}$$

con las t_{ij} dadas por la expresión anterior, y es idéntico al utilizado en el test de independencia.

Se demuestra que el estadístico "d" sigue, si es cierta la H_0 de homogeneidad, una distribución Gi-dos con (I-1)(J-1) grados de libertad, que son idénticos a los obtenidos en el test de independencia.

Autoevaluación : estudiar si es admisible la hipótesis de homogeneidad en los dos ejemplos planteados al principio de esta sección.

7.8.3. Generalizaciones

Las técnicas vistas en la sección anterior pueden generalizarse al análisis de tablas de contingencia con más de dos dimensiones.

Estas generalizaciones tienen una gran importancia práctica porque, por ejemplo, los resultados de una encuesta con preguntas cerradas siempre pueden sintetizarse como una tabla de contingencia multidimensional (tantas dimensiones como preguntas) cuyo análisis puede llevarse a cabo mediante estas técnicas.

Los Modelos log-linear constituyen la técnica estadística más poderosa y general para el análisis de tablas de frecuencias múltiples. Mediante la misma es posible investigar la dependencia conjunta entre diversas variables cualitativas, haciendo estos modelos respecto a datos de frecuencias un papel similar al que hacen las técnicas de Análisis de la Varianza (que se ven en un tema posterior) para el estudio de una variable continua que depende varios factores.

Para el análisis descriptivo de tablas de contingencia de grandes dimensiones (por ejemplo la tabla 50x10 delos votos obtenidos en unas elecciones en las 50 provincias españolas por los 10 partidos principales) tiene gran interés la técnica de Análisis Factorial de Correspondencias, que permite profundizar y representar gráficamente la naturaleza de las relaciones existentes entre las filas y entre las columnas.

El desarrollo de estas técnicas se sale por su carácter avanzado, de los límites de este curso.

7.A EJERCICIOS

7.A.1 Ejercicio resuelto

Se sabe que la incidencia de determinados tipos de cancer en adultos entre 20 y 30 años es de 1.5 por 10.000 al año. En 28.000 jóvenes que participaron recientemente en un conflicto bélico se han constatado al año siguiente 8 casos de estos cánceres.
¿Puede afirmarse que la incidencia de la enfermedad en estos jóvenes es significativamente más elevada (para un riesgo de 1ª especie α = 0.01) que en la población normal?
¿Qué tipo de test ha sido el realizado para responder a la pregunta anterior?
En este ejemplo ¿qué sería cometer un error de 2ª especie? (contestar utilizando el lenguaje del problema y no tecnicismos estadísticos)

Solución

Población: Jóvenes que participan en un conflicto bélico

Sucesos: A. padecer cierto cáncer B: no padecer ese cáncer

Hipótesis Nula: H_0: P(A) = 0.00015 (o sea, el haber estado en esa guerra no afecta a la probabilidad de sufrir el cáncer) y por tanto P(B) = 0.99985

Frecuencias observadas de los sucesos: $O_A = 8$ $O_B = 27992$ Total =28000

Frecuencias teóricas si es cierta H_0 : $T_A = 28000x0.00015 = 4.2$ $T_B = 28000x0.99985 = 27995.8$

Estadístico de "discrepancia" respecto a H_0

$$d = \frac{(8-4.2)^2}{4.2} + \frac{(27992-27995.8)^2}{27995.8} = 3.43$$

Como d = 3.43 es menor que $\chi^2_1(0.01) = 6.63$ la H_0 es admisible. No puede afirmarse (para un riesgo de 1ª especie α = 0.01) que la incidencia de la enfermedad en los jóvenes que fueron a la guerra es significativamente más elevada que en la población normal

El tipo de test que se ha realizado es un test de contraste de proporciones

un error de 2ª especie es el que cometería afirmando que ir a la guerra no influye sobre la probabilidad de tener un cáncer cuando realmente sí que influye

7.A.2 Ejercicios adicionales

Se ha seleccionado al azar una muestra de 300 alumnos de una universidad. El 60% de los mismos fueron chicos y el 40% restantes chicas. A la pregunta sobre su tendencia política el 30% de los chicos se manifestaron de derechas el 35% de centro y el 35% restante de izquierdas. De las chicas el 35% se manifestaron de derechas el 25% de centro y el 40% de izquierdas.

a) *¿Puede afirmarse que existe una relación estadísticamente significativa (para un riesgo de 1ª especie α = 0.05) entre sexo y tendencia política en la universidad investigada?*
b) *En el problema anterior explicar (utilizando el lenguaje del problema) lo que sería cometer un error de 1ª especie*

Se sabe que la incidencia de un determinado problema de audición en niños en edad escolar es del 1%. En los 500 niños de un colegio situado cerca de una estación ferroviaria se han registrado 15 casos de niños con dicho problema de audición. ¿Puede afirmarse a partir de este dato que la situación del colegio ha afectado significativamente a la incidencia de la enfermedad?
En este problema anterior, ¿qué hubiera sido cometer un error de 1ª especie? (contestar utilizando el lenguaje del problema y no tecnicismos estadísticos)

La incidencia de una cierta enfermedad en la población joven de un país es de dos casos al año por cada 1000 personas. En 3000 jóvenes que estuvieron expuestos a una emisión de gases contaminados se registraron en un año 15 casos de dicha enfermedad.
¿Puede afirmarse que la exposición a dicho gas aumenta significativamente (para un riesgo de 1ª especie α = 0.05) la probabilidad de contraer la mencionada enfermedad?
¿Qué tipo de test es el que permite contestar a la pregunta anterior?

De los 1280 nacimientos registrados en un hospital en cierto número de años, 660 correspondieron a chicos y 620 a chicas.
a)¿ Es admisible la hipótesis de que la frecuencia de nacimientos de chicos y chicas en la población es la misma?
b) El 20% de las chicas que nacieron necesitaron una cesárea, mientras que este tipo de intervención sólo fue precisa en el 15% de los nacimientos de chicos. ¿Puede afirmarse (para un riesgo de 1ª especie α=0.05) que existe una relación entre el sexo y el hecho de que el parto sea por cesárea?
c) En el apartado b) ¿qué sería cometer un error de 2ª especie? (explicarlo en el lenguaje del problema, sin utilizar tecnicismos estadísticos)

En un hospital se trataron un determinado año 400 casos de infarto, de los que un 60% correspondieron a varones y el 40% restante a mujeres. Se constató que el 85% de las mujeres sobrevivieron al infarto, mientras que sólo lo hicieron el 70% de los varones.
a) ¿Puede afirmarse a partir de los datos anteriores que la incidencia del infarto de miocardio es más elevada en varones que en mujeres? (operar con un riesgo de 1ª especie α = 0.01)
b) ¿Es significativa la diferencia de supervivencia entre sexos? (operar con un riesgo de 1ª especie α = 0.01)
c) En este segundo problema, ¿que sería un error de segunda especie?