Estadística II
Grado en Matemáticas, UAM, 2020-2021

Hoja 2 (Contrastes no paramétricos)

Contraste χ^2 de modelo

1. Da una demostración directa para el caso $k = 2$ de que la distribución del estadístico del contraste χ^2 de bondad de ajuste converge a una distribución χ^2_1 es decir,

$$T = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \to_d \chi^2_1 \quad \text{si } n \to \infty.$$

(Indicación: Hay que demostrar que $T = X_n^2$, donde $X_n \sim N(0,1)$. Para reducir los dos sumandos a uno, utilizar la relación existente entre O_1 y E_2 y O_2, E_2.)

2. Disponemos de una muestra $(x_1, \ldots, x_n)$, que contiene ceros y unos. Queremos contrastar la hipótesis de que se trata de una muestra aleatoria de una variable $X \sim \text{BER}(p)$, con p conocida. Escribe el valor del estadístico de Pearson (quizás te interese escribirlo en términos de la media muestral $\overline{x}$), y luego escribe la condición de rechazo para el test. Quizás te sea familiar, si observas que una χ^2 con un grado de libertad es una normal estándar al cuadrado.

3. Un cierto programa de ordenador se supone que genera cifras de cero a nueve totalmente al azar. En las 200 cifras obtenidas se observaron las siguientes frecuencias:

Números	0	1	2	3	4	5	6	7	8	9
Frecuencia	27	22	15	18	23	26	19	20	16	14

Con un nivel de significación de 0.05, ¿es "bueno" el programa?

4. Para comprobar el funcionamiento de un generador de muestras de la $N(0,1)$, obtenemos la siguiente muestra de tamaño 450:

Frecuencia	Rango
30	menores que -2;
80	entre -2 y -1;
140	entre -1 y 0;
110	entre 0 y 1;
60	entre 1 y 2;
30	mayores que 2.

¿Se puede aceptar, al nivel $\alpha = 1\,\%$, que el programa funciona correctamente?

5. Para estudiar el número de ejemplares de cierta especie en peligro de extinción que viven en un bosque, se divide el mapa del bosque en nueve zonas y se cuenta el número de ejemplares de cada zona. Se observa que 60 ejemplares viven en el bosque repartidos en las 9 zonas de la siguiente forma:

8	7	3
5	9	11
6	4	7

Mediante un contraste de hipótesis, analiza si estos datos aportan evidencia empírica de que los animales tienen tendencia a ocupar unas zonas del bosque más que otras.

6. Se clasificaron 1000 individuos de una población según el sexo y según tuvieran visión normal o fueran daltónicos. Los resultados aparecen en la tabla de la izquierda. Según un modelo genético, las probabilidades deberían ser las que aparecen a la derecha, donde $q = 1 - p =$ proporción de genes defectuosos en la población.

	Masculino	Femenino
Normal	442	514
Daltónicos	38	6

	Masculino	Femenino
Normal	$p/2$	$p^2/2 + pq$
Daltónicos	$q/2$	$q^2/2$

(a) Estima el parámetro p a partir de la muestra (solución: $\hat{p} = 91.29\,\%$).

(b) ¿Concuerdan los datos con el modelo?

7. El número de asesinatos cometidos en Nueva Jersey cada día de la semana durante el año 2003 se muestra en la tabla siguiente:

Día	Lunes	Martes	Miércoles	Jueves	Viernes	Sábado	Domingo
Frecuencia	42	51	45	36	37	65	53

(a) Contrasta al nivel $\alpha = 0.05$, mediante un test χ^2, la hipótesis nula de que la probabilidad de que se cometa un asesinato es la misma todos los días de la semana.

(b) Contrasta la hipótesis nula de que la probabilidad de que se cometa un asesinato es la misma desde el lunes hasta el viernes, y también es la misma los dos días del fin de semana (pero no es necesariamente igual en fin de semana que de lunes a viernes).

8. Según los datos de un exhaustivo estudio de mercado que se realizó en una gran ciudad, las ventas de impresoras para ordenadores personales de uso doméstico se dividen entre cuatro marcas (A, B, C y D) cuyos porcentajes del total de las ventas son 18 %, 22 %, 35 % y 25 %, respectivamente.

Un año después, se quiere analizar de nuevo la situación pero no se dispone de dinero para repetir un estudio de mercado a gran escala. Se decide, por tanto, observar la marca adquirida por 200 compradores de impresoras elegidos al azar, obteniéndose que de ellos 28 habían elegido la marca A, 48 la B, 77 la C y 47 la D.

¿Hay suficiente evidencia estadística, al nivel 5 %, para afirmar que el reparto del mercado ya no es el mismo que el año anterior?

Otros contrastes χ^2

9. Se quiere estudiar la relación entre la edad de los chicos y el tiempo T que ven semanalmente la televisión. Con una muestra de 200 chicos se obtuvieron los siguientes resultados:

Chicos	menos de 14 horas	entre 14 y 18 horas	más de 18 horas
entre 6 y 10 años	20	30	30
entre 11 y 15 años	20	40	60

¿Hay evidencia estadística significativa con nivel $\alpha = 0.05$ de que existe relación entre la edad y el tiempo que ven la televisión?

10. En un estudio sobre el uso de sistemas operativos se han seleccionado al azar 150 profesores universitarios (PU), 150 profesionales técnicos (PT) de grado medio que trabajan en la industria y 150 personas con cargos directivos (CD) en empresas. A cada uno se le preguntó por el sistema operativo (A, B o C) que utiliza habitualmente en su trabajo. Los resultados fueron los siguientes:

	A	B	C
PU	52	40	58
PT	42	45	63
CD	28	47	75

¿Hay suficiente evidencia estadística, al nivel 0.05, para concluir que existe alguna asociación entre el status profesional y la preferencia por un sistema operativo?

11. Se quiere comparar la biodiversidad de dos montes cercanos. En uno de los montes se eligen al azar 50 zonas, de 4 m^2 cada una, y se hace el recuento del número de especies vegetales diferentes que hay en cada una, con los resultados de la tabla de la izquierda. En el otro monte se hace el mismo recuento en otras 40 zonas, obteniéndose los resultados de la tabla de la derecha:

Número de zonas	Número de especies
20	menos de 6
17	entre 6 y 8
13	más de 8

Número de zonas	Número de especies
12	menos de 6
20	entre 6 y 8
8	más de 8

¿Son similares los dos montes en lo que se refiere a su biodiversidad? Haz el contraste correspondiente con un nivel de significación del 10 %.

12. Un estudio sobre tabaquismo en tres comunidades, mediante tres muestras aleatorias de tamaño 100, proporciona los siguientes resultados:

Comunidad	fumadores	no fumadores
A	13	87
B	17	83
C	18	82

¿Pueden considerarse homogéneas las tres poblaciones en cuanto a sus hábitos fumadores?

13. Se ha clasificado una muestra aleatoria de 500 hogares de acuerdo con su situación en la ciudad (Sur o Norte) y su nivel de renta (en miles de euros) con los siguientes resultados:

Renta	Sur	Norte
0 a 10	42	53
10 a 20	55	90
20 a 30	47	88
más de 30	36	89

(a) A partir de los datos anteriores, contrasta a nivel $\alpha = 0.05$ la hipótesis nula de que en el sur los hogares se distribuyen uniformemente en los cuatro intervalos de renta considerados.

(b) A partir de los datos anteriores, ¿podemos afirmar a nivel $\alpha = 0.05$ que la renta de los hogares es independiente de su situación en la ciudad?

14. Se ha realizado una encuesta para estudiar posibles relaciones entre el nivel educativo (educación superior, media o primaria) de las personas y el nivel de consumo (bajo, medio o alto) de un determinado producto. Los resultados, para 400 personas seleccionadas al azar, han sido:

	Bajo	Medio	Alto
Superior	31	41	44
Media	28	79	125
Primaria	16	17	19

Contrasta estadísticamente (nivel 0.01) la independencia entre nivel educativo y nivel de consumo.

15. Se desea evaluar la efectividad de una nueva vacuna antigripal. Para ello se decide suministrar dicha vacuna, de manera voluntaria y gratuita, a una pequeña comunidad. La vacuna se administra en dos dosis, separadas por un período de dos semanas, de forma que algunas personas han recibido una sola dosis, otras han recibido las dos y otras personas no han recibido ninguna. La siguiente tabla indica los resultados que se registraron durante la siguiente primavera en 1000 habitantes de la comunidad elegidos al azar.

	No vacunados	Una dosis	Dos dosis
Gripe	24	9	13
No gripe	289	100	565

¿Proporcionan estos datos suficiente evidencia estadística (al nivel de significación 5 %) para indicar una dependencia entre la clasificación respecto a la vacuna y la protección frente a la gripe?

16. Se ha llevado a cabo una encuesta a 100 hombres y 100 mujeres sobre su intención de voto. De las 100 mujeres, 34 quieren votar al partido A y 66 al partido B. De los 100 hombres, 50 quieren votar al partido A y 50 al partido B.

(a) Utiliza un contraste basado en la distribución χ^2 para determinar si con estos datos se puede afirmar a nivel $\alpha = 0.05$ que el sexo es independiente de la intención de voto.

(b) Determina el intervalo de valores de α para los que la hipótesis de independencia se puede rechazar con el contraste del apartado anterior.

Contrastes de Kolmogorov–Smirnov

En los ejercicios que siguen, las variables X tienen función de distribución continua $F(x)$. Dada una muestra aleatoria $(X_1, \ldots, X_n)$ de X, la función de distribución muestral $F_n(x)$ viene dada por

$$F_n(x) = \frac{1}{n} \#\{1 \leq i \leq n : X_i \leq x\}.$$

El estadístico de Kolmogorov–Smirnov se define como

$$\Delta_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|.$$

Las discrepancias laterales se definen como

$$\Delta_n^+ = \sup_{x \in \mathbb{R}} (F_n(x) - F(x))^+ \qquad \text{y} \qquad \Delta_n^- = \sup_{x \in \mathbb{R}} (F_n(x) - F(x))^-.$$

Al final de esta hoja de ejercicios aparece una tabla con percentiles de la distribución de Δ_n.

---

17. Calcula la distribución exacta bajo la hipótesis nula del estadístico de Kolmogorov–Smirnov para muestras de tamaño 1.

18. Se dispone de una muestra de tamaño dos: (x_1, x_2), donde $x_1 = -1$ y de x_2 se sabe que es un número mayor que 1. Para contrastar la hipótesis de que es una muestra aleatoria de tamaño 2 de una normal estándar, se calcula el valor del estadístico de Kolmogorov–Smirnov, obteniéndose un valor de 0.4192. ¿Cuánto vale x_2?

19. Comprueba que la variable $4n(\Delta_n^+)^2$ tiende en distribución, cuando $n \to \infty$, a una variable exponencial de parámetro 1/2.

20. Sea $x_1, \ldots, x_n$ una muestra aleatoria obtenida a partir de una variable aleatoria X con función de distribución $F(x)$ continua. La muestra ya está ordenada de menor a mayor.

Prueba que el estadístico "discrepancia lateral" Δ_n^+ se calcula, para la muestra dada, como sigue:

$$\max_{i=1,\ldots,n} \left(\frac{i}{n} - F(x_i) \right)^+.$$

¿Cómo se calcularía Δ_n^-?

21. Se desea contrastar la hipótesis nula de que *una única* observación x procede de una distribución $\mathcal{N}(0, 1)$. Si se utiliza para ello el contraste de Kolmogorov–Smirnov, determina para qué valores de x se rechaza la hipótesis nula a nivel $\alpha = 5\,\%$.

22. La hoja de cálculo `hoja2_EstadII_2020-21.xlsx` adjunta contiene algunas series de datos y propuestas de contrastes de hipótesis usando el test de Kolmogorov–Smirnov. Usa Excel para darles cumplida respuesta.

23. En un estudio de simulación se han generado 10 000 muestras aleatorias de tamaño 10 de una distribución $\mathcal{N}(0, 1)$. Para cada una de ellas se ha calculado el estadístico de Kolmogorov–Smirnov para contrastar la hipótesis nula de que los datos proceden de una distribución normal estándar, y el correspondiente p-valor.

(a) Determina un valor x tal que la proporción de estadísticos de Kolmogorov–Smirnov mayores que x, entre los 10000 obtenidos, sea aproximadamente igual a 5 %. ¿Cuál es el valor teórico al que se debe aproximar la proporción de p-valores menores que 0.1 entre los 10000 p-valores obtenidos?

(b) ¿Cómo cambian los resultados del apartado anterior si en lugar de considerar la distribución normal estándar se considera una distribución uniforme en el intervalo $(0, 1)$?

Ejercicio adicional

24. En este ejercicio se pide probar la identidad $(\star)$ siguiente:

$$\Psi(x) := 1 - 2\sum_{j=1}^{\infty}(-1)^{j+1}\, e^{-2j^2x^2} \overset{(\star)}{=} \frac{\sqrt{2\pi}}{x}\sum_{k=1}^{\infty} e^{-\frac{\pi^2}{8x^2}(2k-1)^2} \qquad \text{para cada } x > 0,$$

usando la *fórmula de sumación de Poisson*: para una función f suficientemente "buena" (por ejemplo, de la clase de Schwartz),

$$\sum_{n\in\mathbb{Z}} f(n) = \sum_{n\in\mathbb{Z}} \widehat{f}(n), \qquad \text{donde} \quad \widehat{f}(n) = \int_{-\infty}^{\infty} f(t)\, e^{-2\pi i n t}\, dt.$$

(Sugerencia: comprueba que, para $x > 0$ fijo, $\Psi(x) = \sum_{n\in\mathbb{Z}} f(n)$, donde $f(z) = e^{i\pi z}\, e^{-2x^2z^2}$. Luego, a calcular transformadas de Fourier, claro)

Percentiles de la función de distribución del estadístico de Kolmogorov–Smirnov

$n \backslash \alpha$	0.001	0.01	0.02	0.05	0.1	0.15	0.2
1		0.99500	0.99000	0.97500	0.95000	0.92500	0.90000
2	0.97764	0.92930	0.90000	0.84189	0.77639	0.72614	0.68377
3	0.92063	0.82900	0.78456	0.70760	0.63604	0.59582	0.56481
4	0.85046	0.73421	0.68887	0.62394	0.56522	0.52476	0.49265
5	0.78137	0.66855	0.62718	0.56327	0.50945	0.47439	0.44697
6	0.72479	0.61660	0.57741	0.51926	0.46799	0.43526	0.41035
7	0.67930	0.57580	0.53844	0.48343	0.43607	0.40497	0.38145
8	0.64098	0.54180	0.50654	0.45427	0.40962	0.38062	0.35828
9	0.60846	0.51330	0.47960	0.43001	0.38746	0.36006	0.33907
10	0.58042	0.48895	0.45662	0.40925	0.36866	0.34250	0.32257
11	0.55588	0.46770	0.43670	0.39122	0.35242	0.32734	0.30826
12	0.53422	0.44905	0.41918	0.37543	0.33815	0.31408	0.29573
13	0.51490	0.43246	0.40362	0.36143	0.32548	0.30233	0.28466
14	0.49753	0.41760	0.38970	0.34890	0.31417	0.29181	0.27477
15	0.48182	0.40420	0.37713	0.33760	0.30397	0.28233	0.26585
16	0.46750	0.39200	0.36571	0.32733	0.29471	0.27372	0.25774
17	0.45440	0.38085	0.35528	0.31796	0.28627	0.26587	0.25035
18	0.44234	0.37063	0.34569	0.30936	0.27851	0.25867	0.24356
19	0.43119	0.36116	0.33685	0.30142	0.27135	0.25202	0.23731
20	0.42085	0.35240	0.32866	0.29407	0.26473	0.24587	0.23152
25	0.37843	0.31656	0.30349	0.26404	0.23767	0.22074	0.20786
30	0.34672	0.28988	0.27704	0.24170	0.21756	0.20207	0.19029
35	0.32187	0.26898	0.25649	0.22424	0.20184	0.18748	0.17655
40	0.30169	0.25188	0.23993	0.21017	0.18939	0.17610	0.16601
45	0.28482	0.23780	0.22621	0.19842	0.17881	0.16626	0.15673
50	0.27051	0.22585	0.21460	0.18845	0.16982	0.15790	0.14886
OVER 50	**1.94947** / $\sqrt{n}$	1.62762 / $\sqrt{n}$	1.51743 / $\sqrt{n}$	1.35810 / $\sqrt{n}$	1.22385 / $\sqrt{n}$	1.13795 / $\sqrt{n}$	1.07275 / $\sqrt{n}$