EJERCICIOS

hoja 1: estadística descriptiva

1. Interesa estudiar la variable X=«Tiempo de vida en días» de una especie de insectos. En una muestra pequeña de 11 insectos, los resultados muestrales fueron:

   20, 25, 13, 18, 32, 25, 20, 15, 28, 40, 27

   Halla el tiempo medio de vida medio, el tiempo de vida mediano, la cuasi-varianza y la desviación típica muestral.

2. Con el fin de controlar la contaminación de un río, todas las semanas se hace una medición del nivel de ácido úrico del agua. Las mediciones durante 9 semanas fueron:

   13 10 7 5 12 7 9 5 5

   Halla el nivel medio de ácido úrico, el nivel mediano, la cuasi-varianza y la desviación típica.

3. En 1778, H. Cavendish realizó una serie de 29 experimentos, con una balanza de torsión, con objeto de medir la densidad de la tierra. Sus resultados, tomando como unidad la densidad del agua, fueron:

5,50	5,61	4,88	4,07	5,26	5,55	5,36	5,29	5,58	5,65
5,57	5,53	5,62	5,29	5,44	5,34	5,79	5,10	5,27	5,39
5,42	5,47	5,63	5,34	5,46	5,30	5,75	5,86	5,85	

   a) Representa los datos por medio de un histograma.

   b) Representa los datos por medio de un diagrama de caja (*boxplot*).

   c) Halla la media y la desviación típica.

   d) ¿Es la distribución aproximadamente normal?

4. Para evaluar la viabilidad de un proyecto de reforestación de una zona sometida a una fuerte actividad turística, se analiza la composición en mg por cm^3 de desechos orgánicos del territorio. Los datos que se obtienen son:

10,87	9,01	22,50	12,35	17,39	31,05	17,19	16,74	20,33
19,32	23,18	25,15	15,49	20,30	2,38	13,55	9,33	22,72
10,96	25,90	27,66	9,74	18,65	9,31	24,60	17,41	24,86
15,34	23,34	22,81	17,86	30,72	32,60	8,96	32,71	15,86
16,71	5,48	8,25	20,57	4,57	2,30	32,56	7,92	4,84
4,57	26,45	23,58	19,27	9,79	3,03	19,40	23,92	22,45
22,05	21,18	18,85	8,38	15,01	18,12	4,24	3,39	7,17
22,71	22,44	15,89	24,20	24,75	28,08	19,73	13,22	17,69
5,53	11,42	5,58	3,15	14,06	5,83	19,42	21,13	18,32
23,31	11,89	23,95	19,30	12,22	21,45	9,84	4,78	38,63
12,65	13,89	23,82	16,91	28,09	15,73	12,53	16,52	9,48
4,08								

   Efectuar un estudio descriptivo como en el ejercicio anterior.

5. El maíz es un elemento importante en la alimantación animal. Sin embargo, este cereal carece de algunos aminoácidos que son esenciales. Un grupo de científicos desarrolló una nueva variedad que sí contenía niveles apreciables de dichos aminoácidos. Para comprobar la utilidad de esta nueva variedad, se llevó a cabo el siguiente experimento: a un grupo de 20 pollos de 1 día se les suministró un pienso que contenía harina de maíz de la nueva variedad. A otro grupo de 20 pollos (grupo de control) se le alimentó con un pienso que sólo se diferenciaba del anterior en que no contenía harina de la variedad mejorada de maíz. Los resultados que se obtuvieron sobre las ganancias de peso de los pollos (en gramos) al cabo de 21 días de alimentación fueron los siguientes:

   - *Variedad normal*

     380 321 366 356 283 349 402 462 356 410 329 399 350 384 316 272 345 455 360 431

   - *Variedad mejorada*

     361 447 401 375 434 403 393 426 406 318 467 407 427 420 477 392 430 339 410 326

   a) Para comparar las dos distribuciones, representa los dos diagramas de caja y bigotes en un mismo gráfico. ¿Qué se puede observar en estos diagramas?

   b) ¿Cuáles son las medias y desviaciones típicas de los datos de ambos grupos? ¿Qué diferencias hay entre ambos?

6. Las autoridades sanitarias de un municipio están interesadas en evaluar la calidad del agua para consumo en términos de colonias de bacterias tróficas en un acuífero próximo a la ciudad. Se consideran dos zonas diferentes del acuífero y se obtienen los siguientes resultados (número de colonias por 1000 mm de agua):

zona 1	194	199	191	202	215	214	197
	204	199	202	230	193	194	209
zona 2	158	161	143	174	220	156	156
	156	198	161	188	139	147	116

Realiza un estudio comparativo de la calidad del agua en ambas zonas, utilizando resúmenes numéricos y diagramas de cajas. ¿Se observan diferencias entre ambas zonas?

7. Se quieren comparar dos métodos, A y B, para determinar el calor latente de fusión del hielo. La siguiente tabla da los resultados obtenidos (en calorías por gramo de masa para pasar de -0.72°C a 0°C) utilizando ambos métodos independientemente:

Método A	79,98	80,04	80,02	80,04	80,03	80,03	80,04	79,97	80,05	80,03	80,02	80,00	80,02
Método B	80,02	79,94	79,98	79,97	79,97	80,03	79,95	79,97					

a) Compara descriptivamante de los dos métodos.

b) Si por un error al transcribir los datos, se modifica el primer dato de cada método y se escribe 799,8 y 800,2, ¿cómo varían la media y la mediana?

8. El tiempo de vida de ciertos insectos parece que se ve influido por el porcentaje de humedad del hábitat en que se encuentran. Para estudiarlo, se toman muestras en 10 hábitats diferentes de las variables X=«Porcentaje medio de humedad» e Y='«Tiempo medio de vida'», y se obtienen un porcentaje medio de humedad del 59 % y un tiempo medio de vida 28,7 días. Además:

$$\sum x_i^2 = 35\,140; \qquad \sum y_i^2 = 8\,573; \qquad \sum x_i y_i = 17\,260$$

a) Expresa el tiempo medio de vida en función de la humedad mediante la recta de regresión. Evalua el ajuste.

b) ¿Cuál sería el tiempo medio de vida estimado si la humedad es del 65 %?

9. Se quiere calibrar una nueva técnica experimental indirecta para medir presiones en relación con un método estándar directo. Para esto, se han realizado 9 tomas de presión (en mm de Hg) por el método estándar directo (X) y por la nueva técnica experimental indirecta (Y). Los resultados obtenidos se resumen a continuación:

$$\sum x_i = 343 \qquad \sum y_i = 325 \qquad \sum x_i^2 = 17\,693 \qquad \sum y_i^2 = 16\,367 \qquad \sum x_i y_i = 16\,992$$

a) Calcula la recta de regresión de Y sobre X. Para una presión de 55 mm de Hg, medida con el método estándar, ¿qué presión cabría esperar con la nueva técnica?

b) ¿Qué podemos decir del ajuste de la recta de regresión a nuestros datos?

10. El manatí de Florida (*Trichechus manatus latirostris*) es un mamífero marino de gran tamaño (hasta 4,5 m y 1 500 kg) que vive a lo largo de la costa de Florida. Cada año las lanchas motoras hieren o matan muchos ejemplares. A continuación se presenta una tabla que da, anualmente, el número de licencias para motoras expedidas en Florida (expresado en miles de licencias) y el número de manatíes muertos en el periodo 1977—1990.

Año	Licencias	Manatíes	Año	Licencias	Manatíes
1977	447	13	1984	559	34
1978	460	21	1985	585	33
1979	481	24	1986	614	33
1980	498	16	1987	645	39
1981	513	24	1988	675	43
1982	512	20	1989	711	50
1983	526	15	1990	719	47

a) Se quiere analizar la relación entre el número de licencias expedidas anualmente en Florida y el número de manatíes muertos por lanchas motoras. ¿Cuál será la variable explicativa? Dibuja un diagrama de dispersión con esos datos. ¿Qué indica el diagrama sobre la relación entre esas dos variables? ¿Están las variables asociadas positiva o negativamente?

b) Halla la recta de regresión para expresar el número de manatíes muertos en función del número de licencias.

c) Describe la fuerza de la relación. ¿Se puede predecir con precisión el número de manatíes muertos cada año conociendo el número de licencias expedidas ese año? Si Florida decidiera congelar el número de licencias en 700 000, ¿cuál sería la predicción del número medio de manatíes muertos por lanchas motoras?

11. Los buenos corredores dan más pasos por segundo a medida que aumentan la velocidad. He aquí el promedio de pasos por segundo de un grupo de corredoras de élite a distintas velocidades. La velocidad se expresa en metros por segundo.

Velocidad (m/s)	4,83	5,14	5,33	5,67	6,08	6,42	6,74
Pasos por segundo	3,05	3,12	3,17	3,25	3,36	3,46	3,55

a) Se quiere predecir el número de pasos por segundo a partir de la velocidad. En primer lugar, dibuja un diagrama de dispersión.

b) Halla la recta de regresión del número de pasos por segundo con relación a la velocidad.

c) Halla el coeficiente de correlación lineal.

HOJA 2: PROBABILIDAD

1. Sea X una variable aleatoria cuantitativa cualquiera. Describe los siguientes sucesos:

a) X vale *a lo sumo* 5. b) X vale *al menos* 2.

c) $\{X < 10\}^c$ d) $\{X \leq 10\}^c$

e) $\{X > 15\}^c$ f) $\{X \geq 15\}^c$

2. Una urna contiene seis bolas rojas y cuatro negras. Se extraen dos bolas sin reemplazamiento.

a) Si se sabe que la primera es roja, ¿cuál es la probabilidad de que la segunda sea roja?

b) Sabiendo que la segunda bola ha sido negra, ¿cuál es la probabilidad de que la primera haya sido roja?

c) Halla la probabilidad de que la segunda bola sea roja.

3. Se sabe que, en cierta población, el número de personas que padecen la enfermedad E es del 1 %. Se ha investigado una prueba diagnóstica que ha resultado positiva en el 97 % de las personas que padecen la enfermedad E y en el 2 % de las personas sanas. Calcula la probabilidad de que a una persona la prueba le de resultado positivo. Calcula la probabilidad de que una persona que ha dado positivo esté enferma.

4. Supongamos que se clasifica a los individuos de cierta especie animal en tres grupos A, B y C de distintas características biológicas. La probabilidad de que un individuo tomado al azar pertenezca al grupo A, B o C es respectivamente 1/2, 1/3 y 1/6. La probabilidad de que un individuo del grupo A, B o C contraiga cierta enfermedad S es respectivamente 1/10, 1/15 y 1/12. Calcúla la probabilidad de que:

a) Un individuo contraiga la enfermedad S.

b) Un individuo enfermo sea del grupo A.

c) Un individuo sano sea del grupo A.

5. Durante la temporada de setas muchos aficionados a la micología se dedican a su recolecta. En cierta región se está realizando un estudio para identificar las *setas comestibles* y las *setas tóxicas* (algunas incluso mortales), para ello se ha dividido el terreno en 3 zonas. La proporción de setas que hay en la primera zona es p, en la segunda es q y en la tercera es r. La probabilidad de que una seta sea comestible para cada una de las zonas es P, Q o R respectivamente. Calcula la probabilidad de que una seta tóxica no pertenezca a la primera zona.

6. Para averiguar el tamaño N de una población de lagartos se utiliza el método siguiente de captura-marcaje-recaptura: Se capturan k lagartos, se les marca y se les reincorpora a su población. Un tiempo después se realizan n avistamientos independientes y se observa el número de ellos que están marcados.

a) Plantea cuál es la variable aleatoria y cuál es su modelo de distribución.

b) Si $N = 1000$, $k = 550$ y $n = 6$, halla la probabilidad de que entre los 6 reptiles:

(i) al menos uno no esté marcado; (ii) al menos uno sí esté marcado; (iii) al menos dos estén marcados y uno no lo esté; (iv) ninguno esté marcado si sabemos que al menos uno de ellos no lo está.

7. El «tiempo de vida activa en días» de un plaguicida, X, viene representado por la función de densidad:

$$f(x) = \begin{cases} \frac{1}{500}e^{-\frac{x}{500}} & \text{si } x > 0 \\ 0 & \text{en el resto} \end{cases}$$

   a) Calcula la mediana y la media del tiempo de vida activa.

   b) Determina la probabilidad de que el plaguicida tenga un periodo de duración entre 50 y 500.

8. La variable aleatoria X=«Tiempo transcurrido (en horas) hasta el fallo de una pieza» tiene función de densidad

$$f(x) = \begin{cases} \frac{1}{15000}\, e^{-\frac{x}{15000}} & \text{si } x > 0 \\ 0 & \text{en el resto.} \end{cases}$$

   Calcula el tiempo medio transcurrido hasta el fallo.

9. En el 2010 la Consejería de Sanidad registró que el 0,04 % de la población que se vacunó de la Gripe A tuvo una reacción alérgica a la vacuna. Si solamente el 10 % de la población está vacunada, halla la probabilidad de que en 5 000 individuos seleccionados al azar tengan reacción alérgica:

   a) exactamente tres; b) más de 2.

10. Una enfermedad *rara* es aquella que afecta a un pequeño número absoluto de personas o a una proporción reducida de la población. Los diversos países y regiones del mundo tienen definiciones legales diferentes. En Europa se considera *rara* a una enfermedad que afecta a menos de 1 de cada 2000 personas. ¿Cuál es la probabilidad de que haya más de tres afectados en un grupo de 10.000 europeos seleccionados aleatoriamente? ¿Cuál es el número medio de afectados en dicho grupo?

11. Sea Z una v.a. normal estándar. Determina:

   a) $P\{Z > 1\}$ b) $P\{Z < -1\}$ c) $P\{1 < Z < 3\}$ d) $P\{-2 < Z < 1\}$ e) $P\{|Z| > 1{,}5\}$.

12. En 1969 se descubrió que los faisanes de Montana (EE.UU) padecían una apreciable contaminación por mercurio que podía deberse a que habían comido semillas de plantas que fueron tratadas durante su crecimiento con metilo de mercurio. Sea X el nivel de mercurio de un ejemplar en partes por millón. Supongamos que X tiene distribución normal con media 0,25 y desviación típica 0,08. Calcular $P\{X \leq 3\}$, $P\{X \geq 0{,}17\}$, $P\{0{,}2 \leq X \leq 0{,}4\}$ y $P\{0{,}01 \leq X \leq 0{,}49\}$. ¿Hay alguna razón para suponer que la normalidad no es adecuada?

13. Se considera que la variable aleatoria «kg de algodón recogidos por parcela» sigue una distribución $N(\mu = 100; \sigma = 10)$.

   Hallar el porcentaje de parcelas en las que el número de kg recogidos será inferior a 115.

14. Se supone que el número de bacterias por mm^3 de agua en un estanque es una variable aleatoria X con distribución de Poisson de parámetro $\lambda = 0{,}5$.

   a) ¿Cuál es la probabilidad de que en un mm^3 de agua del estanque no haya ninguna bacteria?.

   b) En 40 tubos de ensayo se toman muestras de agua del estanque (1 mm^3 de agua en cada tubo). ¿Qué distribución sigue la variable Y = «Número de tubos de ensayo, entre los 40, que no contienen bacterias»? Calcular, aproximadamente, $P(Y \geq 20)$.

15. Un zoólogo estudia una cierta especie de ratones de campo. Para ello captura ejemplares de una población grande en la que la proporción de dicha especie es p.

   a) Si $p = 0,3$, hallar la probabilidad de que en 6 ejemplares capturados haya al menos 2 de los que le interesan.

   b) Si $p = 0,04$, calcular la probabilidad de que en 200 haya exactamente 3 de los que le interesan.

   c) Si $p = 0,4$, calcular la probabilidad de que en 200 haya entre 75 y 100 de los que le interesan.

16. En una población, la cantidad de plomo, X, presente en la sangre de una persona elegida al azar es una variable aleatoria con función de densidad:

$$f(x) = \begin{cases} x/300 & \text{si } 0 < x < 20 \\ (50 - x)/1350 & \text{si } 20 < x < 50 \\ 0 & \text{en el resto} \end{cases}$$

   a) Cantidad media de plomo en la sangre de los individuos de la población.

   b) Probabilidad de que en 40 personas elegidas al azar, haya entre 20 y 30 personas con una cantidad de plomo inferior a 20.

17. La capacidad de enrollar la lengua está controlada por una pareja de genes: el gen E que determina su enrollamiento y el gen e que lo impide. El gen E es dominante, de modo que una persona Ee será capaz de enrollar la lengua.

    En una ciudad grande se sabe que aproximadamente el 40 % no puede enrollar la lengua y el 60 % si puede.

    Si elegimos 200 personas al azar, ¿cuál es la probabilidad de que más de 70 no puedan enrollar su lengua?

18. Un fabricante produce varillas y recipientes para insertar las varillas. Ambos tienen secciones circulares. Los diámetros de las varillas siguen una $N(\mu = 1; \sigma = 0,2)$; los diámetros de los recipientes siguen una $N(\mu = 1,05; \sigma = 0,15)$. Un ingeniero selecciona al azar una varilla y un recipiente. ¿Cuál es la probabilidad de que la varilla pueda insertarse en el recipiente?

19. Supóngase que las estaturas, en pulgadas, de las mujeres de una cierta población siguen una distribución $N(65, 1)$ y que las estaturas de los hombres siguen una distribución $N(68, 2)$. Se selecciona al azar una mujer e, independientemente, se selecciona al azar un hombre. Determinar la probabi-lidad de que la mujer sea más alta que el hombre.

20. Una línea eléctrica se avería cuando la tensión sobrepasa la capacidad de la línea. Si la tensión es $N(100;\ 20)$ y la capacidad es $N(140;\ 10)$, calcular la probabilidad de avería.

21. Una máquina de envasado llena sacos de fertilizante de aproximadamente 30 Kg. La "cantidad de fertilizante por saco" sigue una distribución $N(\mu = 30; \sigma = 1)$.

    a) Se desea que la cantidad de fertilizante por saco esté entre 29 y 31 Kg. Calcular la probabilidad de que esté dentro de esos límites.

    b) Una empresa realiza un pedido de 80 de estos sacos de fertilizante. Calcular la probabilidad de que más de 50 estén dentro de los límites indicados.

HOJA 3: ESTIMACIÓN PUNTUAL & INTERVALOS DE CONFIANZA

1. Sea Z una v.a. normal tipificada: $Z \sim N(0,\ 1)$. Si z_α denota el valor tal que $P\{Z > z_\alpha\} = \alpha$, calcula:

    a) $z_{0,025}$ b) $z_{0,05}$ c) $z_{0,95}$ d) $z_{0,1}$.

2. Sea $X \sim N(1, 2)$, calcular el valor C tal que $P\{X > C\} = 0{,}96$ (Observación: $C \neq z_{0,96}$).

3. Para estudiar la proporción p de caballos afectados por la peste equina se les va a someter a una prueba. Se sabe que la prueba resulta positiva si el animal está enfermo. Además, si el animal está sano, hay una probabilidad 0,04 de que la prueba resulte positiva.

    a) Estudia la relación entre la probabilidad p de que un caballo esté enfermo y la probabilidad q de que la prueba resulte positiva.

    b) Si se realizó la prueba a 500 caballos y resultó positiva en 95 casos, calcula una estimación de p utilizando el resultado del apartado (a)).

4. Un test para detectar si el agua presenta cierto tipo de contaminación resulta positivo con probabilidad 0,99 si el agua está realmente contaminada (*sensibilidad del test*). Si el agua no está contaminada, resulta negativo con probabilidad 0,97 (*especificidad del test*). La sensibilidad y la especificidad se conocen debido a que se tiene mucha experiencia en el uso de la prueba.

    a) ¿Qué relación existe entre la probabilidad de que el test dé positivo y la de que el agua esté contaminada?

    b) Se aplica el test a muestras de agua de 15 lagos y resulta positivo en 2 de las muestras. Utiliza la relación del apartado (a)) para estimar el porcentaje de lagos contaminados.

5. Unos laboratorios desarrollan una prueba sencilla para detectar la *gripe aviar*. La prueba tiene una fiabilidad muy aceptable: proporciona un 4 % de falsos positivos (prueba positiva cuando el ave está sana) y un 0 % de falsos negativos (prueba negativa cuando el ave está enferma).

    En una granja avícola, se detecta un brote de *gripe aviar*. Mediante la utilización de la prueba sencilla que se ha descrito anteriormente, se quiere estimar la incidencia de la enfermedad en esa granja. Para esto, se seleccionan al azar 100 pollos, se les efectúa la prueba y se obtienen 20 casos positivos. Estima la proporción de pollos enfermos en la granja, explicando todo el proceso seguido.

6. Se sabe que el nivel de tensión sanguínea diastólica (mm Hg) en una población sigue una distribución normal de media $\mu = 87$ y desviación típica $\sigma = 7{,}5$. Un individuo se clasifica como pre-hipertenso cuando su tensión está entre 80 y 89 mm Hg. Se seleccionan aleatoriamente cuatro individuos de la población y se promedian sus presiones sanguíneas diastólicas. Calcula la probabilidad que el promedio esté entre los límites que indican pre-hipertensión.

7. Con los datos del experimento de Cavendish de 1778 (ejercicio 1.3) y suponiendo normalidad,

   a) Obtener estimaciones insesgadas para la media y la varianza de la densidad.

   b) Obtener un intervalo de confianza de nivel 0.95 para la media. ¿Cuál es el error máximo?

   c) Determinar cuántas mediciones debería realizar Cavendish para reducir el error del apartado anterior al menos a la mitad.

8. En una explotación minera, las rocas excavadas se someten a un análisis químico para determinar su contenido porcentual de cadmio. Se puede suponer que este contenido es una variable con distribución normal de media μ y varianza σ^2. Después de analizar 25 rocas se obtiene un contenido porcentual medio de 9,77 con una cuasidesviación típica de 3,164.

   (a) Construye un intervalo de confianza de nivel 95 % para el contenido porcentual medio de cadmio en la mina.

   (b) Construye un intervalo de confianza de nivel 95 % para σ.

9. Se estudia la proporción p de individuos alérgicos al polen de las acacias en una determinada población. En 200 individuos tomados al azar se observaron 8 alérgicos. Calcula un intervalo de confianza del 95 % para p.

10. Un equipo de investigadores quiere estimar la proporción p de vacas que sufren el mal de las vacas locas en una gran explotación ganadera, mediante un intervalo con un error máximo de 0,015 y nivel de confianza 0,95. ¿A cuántas vacas deben analizar para alcanzar aproximadamente este objetivo, sabiendo que en un pequeño sondeo orientativo (muestra piloto) resultó que el 15 % de las vacas estaban afectadas por la enfermedad?

11. Con los datos del ejercicio 1.5 y suponiendo normalidad e igualdad de varianzas, obtener un intervalo de confianza (al 95 %) para estimar la diferencia entre las ganancias medias de peso de los pollos con las dos variedades de pienso.

12. Se desea estimar la proporción p de ánades en la población de un parque natural que presenta altos niveles de contaminación por metales pesados. Para ello se realiza un sondeo preliminar con 50 ejemplares, de los cuales 9 resultaron tener altos niveles de contaminación.

    a) Construir un intervalo de confianza, de nivel 0,95, para p a partir de los resultados.

    b) ¿Qué tamaño muestral debería utilizarse en un nuevo sondeo para estimar p con un error máximo de 2,5 puntos porcentuales y un nivel de confianza de 0,92?

13. En una población, el consumo de agua anual por individuo es una variable aleatoria con distribución normal de media μ y desviación típica $\sigma = 7{,}5$. ¿Qué tamaño muestral hace falta para obtener un intervalo de confianza para μ con un margen de error ± 2 y un nivel de confianza de 0,90?

14. Un estudio sobre cicatrización en tritones dió los siguientes resultados (velocidad de cicatrización en $\mu m/h$)

    25 13 44 45 57 42 50 36 35 38 43 31 26 48

    Información resumida: $\sum x_i = 533$ $\quad \sum x_i^2 = 22023$.

    a) Asumiendo Normalidad, calcula un intervalo de confianza del 95 % para la media de la velocidad de cicatrización.

    b) ¿Cuántos tritones habría que muestrear para estimar la media con una confianza del 95 % y un error inferior a 2 unidades.

    c) Representa los datos por medio de un diagrama de barras y de un *boxplot*. ¿Qué puedes decir sobre la hipótesis de normalidad?

15. La concentración media, X, de un determinado contaminante en la atmósfera es de 8 ppm (partes por millón). Un día se hacen mediciones en 10 puntos de una ciudad (alejados unos de otros), obteniéndose los resultados que se resumen a continuación:

$$\sum_{i=1}^{10} x_i = 85{,}2 \qquad \sum_{i=1}^{10} x_i^2 = 773{,}82$$

Suponiendo normalidad, calcula intervalos de confianza al nivel 0,95 para la concentración media y para la varianza del contaminante.

Continúa en el ejercicio 4.1.

16. Se admite que el número de microorganismos en una muestra de 1 mm cúbico de agua de un río sigue una distribución de Poisson de parámetro λ. En 40 muestras se han detectado, en total, 833 microorganismos. Calcula un estimador puntual y un intervalo de confianza al 90 % para λ.

17. Se quiere estudiar la influencia de la hipertensión de los padres sobre la presión sanguínea de los hijos. Para ello se seleccionan dos grupos de niños, unos con padres de presión sanguínea normal (grupo 1) y otros con uno de sus padres hipertenso (grupo 2), obteniéndose las siguientes presiones sistólicas:

Grupo 1	104	88	100	98	102	92	96	100	96	96
Grupo 2	100	102	96	106	110	110	120	112	112	90

Indicación: $\sum x_i = 972$, $\sum y_i = 1058$, $\sum(x_i - \bar{x})^2 = 201{,}6$, $\sum(y_i - \bar{y})^2 = 707{,}4$

Suponiendo normalidad,

a) ¿podemos afirmar que la presión media es mayor en los niños con padres hipertensos? (se supone que las varianzas en las dos poblaciones de niños son iguales).

b) Al nivel 90 % ¿es correcto suponer varianzas iguales?

18. En un estudio sobre el estado de la salud dental en una ciudad, se tomó una muestra elegida al azar de 280 varones entre 35 y 44 años, y se contó el número de piezas dentarias de cada individuo. Tras la revisión pertinente, los dentistas informaron que había 70 individuos con 28 o más dientes.

a) Realiza una estimación por intervalos de confianza de la proporción de individuos de esta ciudad con 28 dientes o más, con un nivel de confianza 0.95.

b) Utilizando los datos anteriores como muestra piloto, calcula cuál sería el tamaño muestral necesario para efectuar la anterior estimación con un error inferior a 0,01 (al mismo nivel de confianza).

19. La producción de trigo (en Tm/Ha) por parcela en cierta región sigue una distribución Normal.

Se escogen 8 parcelas al azar y se obtienen las siguientes producciones:

11,04 11,13 9,04 10,60 11,26 8,78 9,51 10,78

Indicación: $\sum x_i = 82{,}14$, $\sum x_i^2 = 850{,}3646$.

a) Halla un intervalo de confianza del 99 % para la media de la producción por parcela.

b) ¿Cuál debe ser el número de parcelas observadas para estimar la media con un error menor que 0,3 y un nivel de confianza del 99 %?

HOJA 4: CONTRASTES DE HIPÓTESIS PARAMÉTRICOS

1. Con los datos del ejercicio 3.15 y suponiendo normalidad, ¿resulta aceptable, al nivel de significación $\alpha = 0{,}05$, que el grado de contaminación medio es el habitual?

2. Se analiza un envío de botellas sobre las que se afirma que contienen 100 cl. de agua. Examinada una muestra de 5 botellas se obtiene que la media es de 95 cl. y la cuasivarianza es $s^2 = 1{,}1$. Al nivel de significación 5 %, ¿existe evidencia empírica para afirmar que la cantidad media de agua no es de 100 cl.?

3. Un sociólogo afirma que el 40 % de los universitarios han viajado al extranjero al menos una vez. En una muestra de 100 universitarios, se observa que 36 han salido del país en alguna ocasión. Contrastar la hipótesis del sociólogo para un nivel de significación del 10 %.

4. La concentración media de dióxido de carbono en el aire a una determinada altura es habitualmente de unas 355 ppm (partes por millón). Se sospecha que esta concentración es mayor en la capa de aire más próxima a la superficie. Para contrastar esta hipótesis se analiza el aire en 20 puntos próximos al suelo elegidos aleatoriamente. Se obtiene una media muestral de 580 ppm y una desviación típica muestral muestral de 180 ppm. ¿Proporcionan estos datos suficiente evidencia estadística, al nivel 0,01, a favor de la hipótesis de que la concentración es mayor cerca del suelo? Indica razonadamente si el p-valor es mayor o menor que 0,01.

5. Un fabricante de materiales para insonorización produce dos tipos A y B. De los 1000 primeros lotes vendidos, 560 fueron del tipo A. ¿Proporcionan estos datos suficiente evidencia estadística (al nivel de significación 0.01) para concluir que los consumidores prefieren mayoritariamente el tipo A?

6. Con los datos del ejercicio 1.7, ¿se puede afirmar, a un nivel de significación del 10 %, que existen diferencias significativas entre los resultados medios proporcionados por los dos métodos?

7. Con los datos del ejercicio 1.5, continuado en el ejercicio 3.11, y suponiendo normalidad e igualdad de varianzas, ¿se puede considerar que hay suficiente evidencia estadística para afirmar que la ganancia media de peso es mayor con la variedad mejorada? Da una respuesta con un nivel de significación 0,10.

   Al nivel de significación 0,10 ¿era razonable aceptar la hipótesis de igualdad de varianzas?

8. Se desea comparar la proporción de viviendas con calefacción en Extremadura y en Galicia. Se hace un muestreo en las dos comunidades con los siguientes resultados:

   Extremadura: De 500 viviendas elegidas al azar, 300 disponen de calefacción.

   Galicia: De 1000 viviendas elegidas al azar, 680 disponen de calefacción.

   ¿Hay suficiente evidencia estadística para concluir, con un nivel de significación del 5 %, que es menor la proporción de viviendas con calefacción en Extremadura que en Galicia?

9. Se van a probar dos medicamentos, A y B, contra una enfermedad. Para esto se tratan 100 ratones enfermos con el medicamento A y otros 100 con el medicamento B. El número medio de horas que sobreviven con A es $\bar{x} = 1\,200$ y el número medio con B es $\bar{y} = 1\,400$. Suponiendo normalidad en ambos casos y sabiendo que:

$$\sum (x_i - \bar{x})^2 = 900\,000 \qquad \sum (y_i - \bar{y})^2 = 950\,000$$

   a) ¿Se puede aceptar igualdad de varianzas con $\alpha = 0{,}10$?

   b) ¿Es más efectivo el medicamento B? Plantea el contraste adecuado al nivel de significación 0,05.

10. Con objeto de estudiar si el número de pulsaciones por minuto en hombres (X) puede considerarse menor que en mujeres (Y), se toman muestras de 16 hombres y 16 mujeres, obteniéndose los siguientes resultados:

$$\sum x_i = 1\,248 \qquad \sum x_i^2 = 97\,570 \qquad \sum y_i = 1\,288 \qquad \sum y_i^2 = 103\,846$$

    ¿Qué se puede decir al respecto?

11. Una compañía americana de distribución de gasolina quiere probar el rendimiento de un nuevo combustible. Para esto hace pruebas con 8 modelos de coches diferentes en una autopista. El número de millas recorridas por galón de gasolina con cada coche con el antiguo combustible y con el nuevo combustible se da a continuación:

Coche	1	2	3	4	5	6	7	8
Antiguo combustible	58	52	50	50	50	49	44	42
Nuevo combustible	60	55	52	51	48	50	42	46

    Suponiendo normalidad, ¿se puede concluir que el nuevo combustible proporciona un mejor rendimiento? Dar una respuesta con un nivel de significación de 0,05.

    Cambia la variable de estudio de «millas por galón» a «litros de consumo por cada 100 km». Suponiendo de nuevo normalidad realiza el mismo contraste.

    ¿Qué se puede decir de la hiótesis de normalidad en ambos casos?

12. Con el objeto de estudiar la efectividad de un agente diurético, se eligieron al azar 11 pacientes, aplicando a 6 de ellos dicho fármaco y un placebo a los restantes. La variable observada en esta experiencia fue la concentración de sodio en la orina a las 24 horas, la cual dio los resultados siguientes:

Diurético	20.4	62.5	61.3	44.2	11.1	23.7
Placebo	1.2	6.9	38.7	20.4	17.2	

    Se supone que las concentraciones de sodio, en ambos casos, tienen una distribución $\mathcal{N}(\mu_1, \sigma)$ y $\mathcal{N}(\mu_2, \sigma)$ respectivamente. Contrasta, a un nivel de significación del 5 %, si existe diferencia en el efecto medio al usar el agente diurético.

13. Se desea estudiar si la utilización de tratamientos para reducir el nivel de colesterol reduce también el riesgo de sufrir infartos. Para ello, 2 051 hombres de mediana edad recibieron un tratamiento para reducir el nivel de colesterol a base de *gemfibrozil* mientras que un grupo de control de 2 030 hombres recibió un placebo. Durante los cinco años siguientes, 56 hombres del grupo de *gemfibrozil* y 84 del grupo del placebo sufrieron infartos.

    (a) ¿Existe evidencia empírica, al nivel $\alpha = 0{,}05$ de que el *gemfibrozil* es eficaz para reducir el riesgo de sufrir infartos?

    (b)Determina razonadamente si el p-valor del contraste es mayor o menor que 0,05.

14. Algunos estudios parecen sugerir que tomar cada día una aspirina puede tener efectos beneficiosos para la salud. En un amplio estudio sobre 22 071 personas sanas de más de 40 años, se administró una aspirina diaria durante cierto periodo de tiempo a 11 037 personas mientras que el resto recibió un placebo. En la siguiente tabla aparece el número de personas que sufrieron infartos y embolias para cada grupo:

	Grupo aspirina	Grupo placebo
Infartos	139	239
Embolias	98	119

    (a) ¿Aportan estos datos suficiente evidencia empírica al nivel $\alpha = 0{,}05$ de que tomar una aspirina es beneficioso para evitar sufrir embolias?

    (b) Calcula un intervalo de confianza de nivel 0,95 para la la diferencia de proporciones de individuos que sufren un infarto entre los que han tomado y no han tomado aspirinas.

15. Una compañía petrolífera está considerando la posibilidad de introducir un nuevo aditivo en su gasolina, esperando incrementar el kilometraje medio por litro. Los ingenieros del grupo de investigación prueban 10 coches con la gasolina habitual y otros 10 coches con la gasolina con el nuevo aditivo. El resumen de los resultados es:

    «Kilometraje medio sin aditivo»= 14,2 km/l $\quad s_1^2 = 3{,}24$

    «Kilometraje medio con aditivo»= 15,4 km/l $\quad s_2^2 = 5{,}76$

    a) ¿Se puede considerar probado que el nuevo aditivo aumenta el kilometraje medio por litro? Plantea el modelo correspondiente (suponiendo normalidad e igualdad de varianzas) al nivel de significación 0,05.

    b) Con los datos disponibles, ¿era razonable trabajar con la hipótesis de igualdad de varianzas en el apartado anterior? Da una respuesta razonada con un nivel de significación de 0,10.

16. Se quieren comparar dos métodos rápidos para estimar la concentración de una hormona en una solución. Se preparan en el laboratorio 10 dosis de la hormona y se mide la concentración de cada una con los dos métodos. Se obtienen los siguientes resultados:

Dosis	1	2	3	4	5	6	7	8	9	10
Método A	10,7	11,2	15,3	14,9	13,9	15,0	15,6	15,7	14,3	10,8
Método B	11,1	11,4	15,0	15,1	14,3	15,4	15,4	16,0	14,3	11,2

    Al nivel de significación 0,05, contrasta si los dos métodos proporcionan, en media, las mismas estimaciones.

17. El contenido medio habitual de arsénico en un parque nacional es de 9 ppm. Se cree que últimamente este contenido medio ha podido aumentar. Para estudiar este posible aumento de la contaminación, se toma una muestra del contenido de arsénico en 20 puntos diferentes del parque. Se obtiene una media muestral de 10 ppm, y una cuasi-varianza de 2,1 ppm.

    ¿Son los datos lo suficientemente concluyentes como para poder afirmar que, efectivamente, ha habido una contaminación, es decir, que el contenido medio en la actualidad es superior al habitual de 9 p.p.m.? Da una respuesta, al nivel de significación 0,05, suponiendo normalidad en los datos.

    ¿Es el p-valor de estos datos superior o inferior a 0,05? Da una respuesta razonada sin hacer cálculos adicionales.

18. Se desea estudiar la efectividad de un insecticida ecológico contra los áfidos en los cultivos de patata. Para ello, se tratan por un lado 40 plantas con el insecticida, obteniéndose una media de 15 áfidos por planta con una cuasi-varianza de 9,1. Por otro lado, sobre un grupo de control de 30 plantas que no reciben tratamiento, se obtiene una media de 18,3 áfidos por planta con una cuasi-varianza de 10,5.

    a) Suponiendo normalidad e igualdad de varianzas, ¿se puede concluir que el insecticida es eficaz? Es decir, ¿se puede concluir que el número medio de áfidos por planta es menor cuando se utiliza el insecticida ecológico? Da una respuesta razonada con un nivel de significación de 0,10.

    b) ¿Era aceptable asumir la igualdad de varianzas? Da una respuesta razonada al nivel de significación 0,10.

19. En otro lugar, se lleva a cabo otro estudio diferente con el insecticida ecológico del ejercicio anterior. En este nuevo estudio, queremos comparar proporciones de plantas afectadas.

    Por un lado, se tratan 100 plantas con el insecticida y se comprueba que 10 de ellas tienen áfidos.

    Por otro lado, se tiene un grupo de control de otras 100 plantas diferentes que no reciben tratamiento y se comprueba que 15 de ellas tienen áfidos.

    ¿Podemos concluir, al nivel de significación 0,05, que la proporción de plantas con áfidos es menor cuando son tratadas con el insecticida?

20. La producción de trigo (en T/ha) que se obtiene un año en 5 parcelas es la siguiente:

    11,04 15,13 9,04 20,60 31,26

    Se quiere estudiar la efectividad de un nuevo fertilizante. Para ello, se observa la producción de trigo que se obtiene al año siguiente, usando el nuevo fertilizante. Los resultados en las mismas 5 parcelas son los siguientes:

    12,08 17,28 10,82 18,90 32,03

    Se puede afirmar que el nuevo fertilizante aumenta la producción media? Responde razonadamente, al nivel de significación 0,01.

    Haz una crítica razonada del diseño.

21. El nivel ceráunico de un lugar es el número de días al cabo del año en los que hay tormenta (se considera día con tormenta a aquel en el que al menos se oye un trueno). En cierta comarca, se conoce por datos históricos que el nivel ceráunico sigue una distribución de Poisson con una media (λ) de 20 días.

    Sin embargo, se piensa que últimamente su nivel ceráunico ha aumentado, como consecuencia del cambio climático. En un seguimiento de los últimos 50 años, se ha obtenido que el número medio de días al año con tormenta ha sido de 22. ¿Se puede afirmar que efectivamente se ha producido un aumento del nivel ceráunico? Da una respuesta razonada, al nivel de significación del 5 %.

HOJA 5: CONTRASTES DE HIPÓTESIS NO PARAMÉTRICOS

1. Estamos interesados en comprobar la perfección de un dado cúbico: un dado normal de 6 caras. Para esto realizamos 300 lanzamientos del dado anotando los puntos obtenidos en cada lanzamiento. A la vista de los resultados obtenidos, y con un nivel de significación del 5 % ¿hay razones para rechazar que el dado sea perfecto?

Puntuación del dado (X)	1	2	3	4	5	6
Frecuencias	43	49	56	45	66	41

2. Se quiere comprobar que un programa de ordenador genera observaciones de una distribución $N(0;1)$. Para ello se obtiene una muestra aleatoria de 450 observaciones mediante dicho programa, que proporciona los siguientes resultados:

Valor x de la observación	$x<-2$	$-2<x<-1$	$-1<x<0$	$0<x<1$	$1<x<2$	$2<x$
Número de observaciones	30	80	140	110	60	30

    ¿Se puede aceptar, al nivel $\alpha = 0{,}01$, que el programa funciona correctamente?

3. Con los datos del experimento de Cavendish (ejercicios 1.3 y 3.7), ¿se puede aceptar, al nivel de significación 0,05, que la densidad de la tierra se ajusta a una distribución normal?

4. Para estudiar si el grupo sanguíneo de los individuos se relaciona con la predisposición a padecer diabetes, se seleccionan al azar 400 sujetos de los que se determina el grupo sanguíneo y el nivel de glucosa en sangre en idénticas condiciones experimentales. Clasificada la segunda medida en niveles bajo, medio y alto, se obtiene:

	Bajo	Medio	Alto	Total
0	137	86	35	258
A	42	23	11	76
B	19	17	7	43
AB	14	7	2	23
Total	212	133	55	400

    Comprueba si existe independencia entre el grupo sanguíneo y el nivel de glucosa al nivel de significación 0,05.

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 desea estudiar el número de accidentes por día que se producen en cierto regimiento. Para ello se toman al azar los partes de 200 días dentro de los últimos 5 años, encontrando los siguientes resultados:

Número de accidentes/día	0	1	2	3	4	5	6
Número de días	58	75	44	18	3	1	1

a) ¿Se puede aceptar, con nivel de confianza del 90 %, que el número de accidentes por día sigue una distribución de Poisson?

b) Suponiendo que el número de accidentes por día sigue una Poisson(λ), ¿hay suficiente evidencia estadística (tomar nivel de significación $\alpha = 0{,}05$) de que el verdadero valor medio λ del número de accidentes por día es menor que 1,35? El p-valor, ¿es mayor o es menor que 0,05?

7. 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 0,05) para indicar una dependencia entre la clasificación respecto a la vacuna y la protección frente a la gripe?

8. Se quiere comparar la biodiversidad de dos montes cercanos. Para ello se sigue el procedimiento siguiente.

En uno de los montes se eligen al azar 50 zonas, de 4 m^2 cada una, se recuenta el número de especies vegetales diferentes que hay en cada una y se observa que: en 20 zonas había menos de 6 especies; en 17 zonas había entre 6 y 8 especies; y en 13 zonas había más de 8 especies.

En el otro monte se hace el mismo recuento en 40 zonas elegidas con las mismas características y se observa que: en 12 zonas había menos de 6 especies diferentes; en 20 zonas había entre 6 y 8 especies diferentes; y en 8 zonas había más de 8 especies diferentes.

¿Son similares los dos montes en lo que se refiere a su biodiversidad? Haz el contraste correspondiente con un nivel de significación del 0,10.

9. Se han clasificado 1000 individuos de una población según su sexo y según fueran normales o daltónicos.

	Masculino	Femenino
Normal	442	514
Daltónicos	38	6

Según un modelo genético, las probabilidades deberían ser:

$$\begin{array}{ll} \frac{1}{2}\,p & \frac{1}{2}\,p^2 + pq \\[1em] \frac{1}{2}\,q & \frac{1}{2}\,q^2 \end{array}$$

donde $q = 1 - p =$ proporción de genes defectuosos en la población.

A partir de la muestra se ha estimado que $q = 0{,}087$. ¿Concuerdan los datos con el modelo?

10. Se estudia la distribución de los grupos sanguíneos O, A, B, AB en dos comunidades. Se obtienen los resultados siguientes.

	O	A	B	AB
Comunidad 1	121	120	79	33
Comunidad 2	118	95	121	30

a) ¿Se puede considerar que son homogéneas ambas comunidades?

b) Consideremos ahora sólo los datos de la comunidad 1. El modelo teórico asigna las siguientes probabilidades a cada uno de los grupos:

$$\begin{array}{cccc} O & A & B & AB \\ r^2 & p^2+2pr & q^2+2qr & 2pq \end{array} \qquad (p+q+r=1)$$

A partir de los datos de la muestra se han obtenido las siguientes estimaciones de los parámetros: $\hat{p} = 0,2465$ y $\hat{q} = 0,1732$. Obtener las frecuencias esperadas según el modelo teórico y contrastar la hipótesis de que los datos se ajustan a él.

11. Se ha desarrollado un modelo teórico para las diferentes clases de una variedad de moscas. Este modelo indica que la mosca puede ser de tipo L con probabilidad p^2, de tipo M con probabilidad q^2 y de tipo N con probabilidad $2pq$ $(p+q=1)$. Para confirmar el modelo experimentalmente se toma una muestra de 100 moscas, y se obtienen 10, 50 y 40 moscas de los tipos L, M y N, respectivamente.

    ¿Se ajustan los datos al modelo teórico, al nivel de significación 0,05?

12. Un Ayuntamiento decide poner 4 contenedores para reciclar papel en una zona de la ciudad con la idea de que sean utilizados por la misma cantidad de personas (aproximadamente). Para ver si esto es cierto, hace una encuesta en la zona a 300 personas, preguntándoles que contenedor utilizan. Los resultados obtenidos son los siguientes:

El contenedor 1 es utilizado por 80 personas	El contenedor 2 es utilizado por 70 personas
El contenedor 3 es utilizado por 85 personas	El contenedor 4 es utilizado por 65 personas

a) Como consecuencia de estos resultados, ¿resulta aceptable que los 4 contenedores tienen el mismo nivel de utilización? Da una respuesta razonada, con un nivel de significación de 0,10.

b) El p-valor del contraste anterior, ¿es inferior o superior a 0,10? Dar una respuesta razonada.

13. Se han clasificado los alumnos de un curso según su grupo sanguíneo y su Rh con los resultados siguientes.

Rh/ Grupo	A	B	AB	0
+	48	32	12	40
-	16	6	5	12

a) ¿Son las variables «Rh» y «grupo sanguíneo» independientes? Da una respuesta razonada al nivel de significación 0,05.

b) El p-valor de los datos ¿es inferior o superior a 0,05? Da una respuesta razonada.

14. En una ciudad existen 4 puntos limpios para recogida de residuos: A, B, C y D. Por la forma en que están repartidos en la población, se piensa que las proporciones o probabilidades de uso de estos 4 puntos limpios siguen el modelo:

$$P(A) = p \qquad P(B) = \frac{1}{2} - p \qquad P(C) = \frac{1}{2} - p \qquad P(D) = p.$$

Se extrae una muestra aleatoria de 120 hogares y se obtienen las siguientes frecuencias:

Punto limpio	A	B	C	D
O_i	18	32	41	29

a) ¿Se ajustan bien los datos al modelo propuesto? Da una respuesta al nivel de significación $\alpha = 0{,}05$, (se ha estimado $\hat{p} = 0,20$).

b) Decide razonadamente si el p-valor es inferior o superior a 0,05.