Problemas de Estadística

Alfredo Sánchez Alberca
asalber@ceu.es
https://aprendeconalf.es

Tabla de contenidos

Prefacio

Colección de problemas de Estadística aplicada a la Economía para el Master en Análisis y Comunicación de Datos.

1 Estimación de parámetros

Ejercicio 1.1. El número medio de trabajadores en las PYMES españolas es 5 y su varianza 4. Realizado un muestreo aleatorio de 16 PYMES, calcular:

a. La esperanza y varianza de la media muestral.

b. La esperanza de la varianza y de la cuasivarianza muestral.

c. Mínimo tamaño que ha de tener la muestra para que exista una probabilidad mayor o igual al 95% de que la media muestral se desvíe de la media poblacional a lo sumo 0.5 unidades.

d. Si realizamos un muestreo aleatorio de tamaño 320 obtener $P(4.9 \leq \bar{x} \leq 5.2)$.

Solución

Sea X la variable aleatoria que mide el número de trabajadores en una muestra de 16 PYMES españolas.

a. $E(\bar{x}) = 5$ y $Var(\bar{x}) = \frac{4}{16} = 0.25$.
b. $E(s^2) = 3.75$ y $E(\hat{s}^2) = 4$.
c. $P(|\bar{x} - 5| \leq 0.5) = 0.95$ si $n \geq 320$.
d. Sea Y la variable aleatoria que mide el número de trabajadores en una muestra de 320 PYMES españolas. Entonces, $Y \sim N\left(5, \sqrt{\frac{4}{320}}\right)$ y $P(4.9 \leq \bar{x} \leq 5.2) = 0.777$.

Ejercicio 1.2. El precio de las acciones de Iberpapel se distribuyen según un modelo normal $N(\mu, 2)$. Si se analizan 16 sesiones de la Bolsa de Madrid elegidas aleatoriamente, ¿cuál es la probabilidad de que la cuasivarianza muestral del precio de las acciones sea mayor o igual que 2.136?

Solución

Sabemos que $\frac{ns^2}{\sigma^2} \sim \chi^2(n-1)$. Por tanto, $P(\hat{s}^2 \geq 2.136) = P\left(\frac{16 \cdot s^2}{2} \geq 4 \cdot 2.136\right) = P\left(\chi^2(15) \geq 8.544\right) = 0.9$.

Ejercicio 1.3. El porcentaje de votantes con preferencia de un determinado partido es del 5% en una región A, y el 10% en otra B. Consultados 100 electores de la región A y 150 de la B, determinar la probabilidad de que el porcentaje de electores consultados favorables a dicho partido en la segunda región supere en más de 2% al porcentaje de electores favorables a dicho partido en la primera.

Solución

Sea X_1 proporción de votantes favorables al partido en la región A en una muestra de 100 electores y X_2 proporción de votantes favorables al partido en la región B en una muestra de 150 electores. Entonces, $X_1 \sim N\left(0.05, \sqrt{\frac{0.05\cdot(1-0.05)}{100}}\right)$ y $X_2 \sim N\left(0.1, \sqrt{\frac{0.1\cdot(1-0.1)}{150}}\right)$, y $P(X_2 - X_1 > 0.02) = 0.8212$.

Ejercicio 1.4. Se sabe que el gasto mensual en carne de porcino en las familias españolas se distribuye de manera normal. Se ha realizado un muestreo aleatorio simple en el que se ha preguntado a 20 familias sobre el gasto mensual en carne de porcino, y se ha obtenido una media de 170.31 € y una cuasidesviación típica de 36 €.

a. Obtener el intervalo de confianza para el gasto medio mensual en carne de porcino con un 95% de confianza.

b. ¿Cómo podría obtenerse un intervalo de confianza más preciso para el gasto medio suponiendo que no varían la media muestral, la cuasivarianza muestral y el tamaño de la muestra? ¿Y si no varían la media muestral, la cuasivarianza muestral y el nivel de confianza?

c. Obtener el intervalo de confianza para la desviación típica del gasto mensual en carne de porcino con un 95% de confianza.

Solución

Sea $X \sim N(\mu, \sigma)$ la variable aleatoria que mide el gasto mensual en carne de porcino en una muestra de 20 familias españolas.

a. Usando el intervalo de confianza de la t de Student se tiene $\mu \in (153.46, 187.16)$ con una confianza del 95%.

b. Para obtener un intervalo de confianza más preciso para el gasto medio manteniendo la misma media muestral, cuasivarianza muestral y tamaño de la muestra, se puede reducir el nivel de confianza.

   Y para obtener un intervalo de confianza más preciso para el gasto medio manteniendo la misma media muestral, cuasivarianza muestral y nivel de confianza, se puede aumentar el tamaño de la muestra.

c. Usando el intervalo de confianza de la χ^2 se tiene $\sigma \in (27.37, 52.58)$ con una confianza del 95%.

Ejercicio 1.5. La OMS ha obtenido una muestra de los pesos de 50 niños de 12 años, que proporciona una media muestral de 47 kg y una cuasidesviación típica muestral de 11 kg. Suponiendo que la población sigue una distribución normal:

a. Obtener un intervalo de confianza para la media poblacional con un 95% de nivel de confianza.

b. El director de la OMS considera que el intervalo es poco preciso, pero quiere mantener el nivel de confianza. Por ello decide reducir a la mitad la amplitud del intervalo. En estas condiciones, ¿cuál debería ser el tamaño de la muestra para cumplir los objetivos del director?

c. Los resultados obtenidos en los análisis anteriores siguen sin convencer al director de la OMS y le pide a su equipo que establezca un intervalo de confianza para la media poblacional con un 99% de nivel de confianza, manteniendo la misma muestra del primer apartado.

d. El director decide reducir en un tercio la amplitud del intervalo anterior, pero quiere mantener el nivel de confianza ¿cuál debería ser el tamaño de la muestra para cumplir dicho objetivo?

Solución

Sea $X \sim N(\mu, \sigma)$ la variable aleatoria que mide el peso de los niños de 12 años.

a. Usando el intervalo de confianza de la t de Student se tiene $\mu \in (43.88, 50.12)$ con una confianza del 95%.

b. Para reducir a la mitad la amplitud del intervalo manteniendo el nivel de confianza, se debe cuadruplicar el tamaño de la muestra, es decir, se necesita $n = 200$ niños.

c. Usando el intervalo de confianza de la t de Student se tiene $\mu \in (42.83, 51.16)$ con una confianza del 99%.

d. Para reducir en un tercio la amplitud del intervalo manteniendo el nivel de confianza, se debe multiplicar por 9 el tamaño de la muestra, es decir, se necesita $n = 450$ niños.

Ejercicio 1.6. Un fabricante de bicicletas quiere estimar la media de ventas de bicicletas en un año. Para ello, ha tomado una muestra aleatoria simple de 17 establecimientos, y ha obtenido una media muestral 3650 bicicletas con una cuasidesviación típica muestral de 55 bicicletas. Suponiendo que las ventas de bicicletas siguen una distribución normal:

a. Calcular el intervalo de confianza para la media con un nivel de confianza del 95%.

b. Calcular el intervalo de confianza para la varianza con un grado de confianza del 95%.

> **Solución**
>
> Sea $X \sim N(\mu, \sigma)$ la variable aleatoria que mide las ventas de bicicletas en un año.
>
> a. Usando el intervalo de confianza de la t de Student se tiene $\mu \in (3621.70, 3678.30)$ con una confianza del 95%.
>
> b. Usando el intervalo de confianza de la χ^2 se tiene $\sigma^2 \in (1512.5, 7006.38)$ con una confianza del 95%.

Ejercicio 1.7. Un banco quiere saber el nivel de implantación de la criptomoneda Bitcoin y para ello se ha realizado un muestreo aleatorio simple de 100 españoles, resultando que 15 tienen bitcoins.

a. Obtener un intervalo de confianza del 95% para la proporción poblacional de españoles que poseen bitcoins.

b. ¿A cuántos españoles se debería encuestar para lograr una semiamplitud del intervalo de 0.02, utilizando un nivel de confianza del 90%?

> **Solución**
>
> Sea $X \sim B(100, p)$ la variable aleatoria que mide el número de españoles que poseen bitcoins.
>
> a. Usando el intervalo de confianza de la normal se tiene $p \in (0.0801, 0.2199)$ con una confianza del 95%.
>
> b. Para lograr una semiamplitud del intervalo de 0.02 con un nivel de confianza del 90%, se necesita encuestar a $n = 861$ españoles.

Ejercicio 1.8. Para conocer la prevalencia de la COVID en una población se ha tomado una muestra aleatoria de 500 personas y se ha observado que 36 tenían COVID. ¿Qué precisión tiene el intervalo de confianza del 95% para la proporción de personas infectadas en la población? ¿Qué tamaño muestral habría que tomar para doblar la precisión del intervalo?

Solución

Sea $X \sim B(500, p)$ la variable aleatoria que mide el número de personas infectadas por COVID.
Usando el intervalo de confianza de la normal se tiene el error en la estimación es $E = 0.0227$, es decir, un 2.27%.
Para doblar la precisión del intervalo se necesita un tamaño muestral de $n = 1993$ personas.

Ejercicio 1.9. Tras la liberalización del transporte ferroviario de pasajeros en las líneas de alta velocidad en España, la compañía francesa SNCF estudia la proporción de clientes que utiliza al menos una vez al mes el servicio de alta velocidad. A tal efecto la empresa realiza un muestreo aleatorio en el que se seleccionan 50 usuarios y en el que resulta que 35 de ellos afirma utilizar este servicio una vez al mes como mínimo. Calcular el intervalo de confianza del 98% para la proporción poblacional de usuarios que utilizan la alta velocidad al menos una vez al mes.

Solución

Sea $X \sim B(50, p)$ la variable aleatoria que mide el número de usuarios que utilizan la alta velocidad una vez al mes como mínimo.
Usando el intervalo de confianza de la normal se tiene $p \in (0.549, 0.851)$ con una confianza del 98%.

Ejercicio 1.10. Leemos en un periódico que la intención de voto a un partido político está entre el 25% y el 31% con un 95% de confianza. ¿Cuál es el tamaño muestral que se ha utilizado para dar esta estimación?

Solución

Usando el intervalo de confianza de la normal se tiene que el tamaño muestral necesario para obtener este intervalo para la proporción de personas que votaría al partido es $n = 861$.

Ejercicio 1.11. En una encuesta realizada a 1000 personas sobre la intención de voto en unas elecciones, 350 comentaron que votarían al partido A y 390 al partido B. Calcular los intervalos de confianza del 95% para el porcentaje de voto a cada partido. ¿Se puede afirmar con un 95% de confianza que el partido B ganaría las elecciones?

> **Solución**
>
> Sea $A \sim B(500, p_A)$ la variable aleatoria que mide el número de personas que votarían al partido A y $B \sim B(1000, p_B)$ la variable aleatoria que mide el número de personas que votarían al partido B en una muestra de 1000 personas.
> Usando el intervalo de confianza de la normal se tiene que el porcentaje de voto al partido A está entre el 32.04% y el 37.96% y el porcentaje de voto al partido B está entre el 35.98% y el 42.02 con una confianza del 95%.

Ejercicio 1.12. Para ver si una campaña de publicidad sobre un fármaco ha influido en sus ventas, se tomó una muestra de 8 farmacias y se midió el número de fármacos vendidos durante un mes, antes y después de la campaña, obteniéndose los siguientes resultados:

Antes	147	163	121	205	132	190	176	147
Después	150	171	132	208	141	184	182	145

Obtener la variable diferencia y construir un intervalo de confianza para la media de la diferencia con un nivel de significación 0.05. ¿Existen pruebas suficientes para afirmar con un 95% de confianza que la campaña de publicidad ha aumentado las ventas?

> **Solución**
>
> Sea X la variable aleatoria que mide la diferencia entre el número de fármacos vendidos en una farmacia en un mes antes y después de la campaña de publicidad.
> Usando el intervalo de confianza de la t de Student se tiene que $\mu \in (-1.75, 21.75)$ con una confianza del 95%.

Ejercicio 1.13. Se ha realizado un estudio para comparar los ingresos medios de las personas de dos ciudades. Para ello, se ha tomado una muestra de 100 personas en una ciudad y 120 en la otra. En la primera ciudad se ha observado una media de 1630 € mensuales y una cuasidesviación típica de 150 €, mientras que en la segunda ciudad se ha observado una media de 1780€ y una cuasidesviación típica de 160 €. Calcular el intervalo de confianza del 95% para la diferencia de medias de ingresos mensuales entre las dos ciudades.

> **Solución**
>
> Sea $X_1 \sim N(\mu_1, \sigma_1)$ la variable aleatoria que mide los ingresos mensuales de las personas en la primera ciudad y $X_2 \sim N(\mu_2, \sigma_2)$ la variable aleatoria que mide los ingresos mensuales de las personas en la segunda ciudad.

El intervalo de confianza del %95% para el cociente de varianzas es $\frac{\sigma_1{}^2}{\sigma_2^2} \in (0.6273, 1.2492)$ por lo que podemos asumir varianzas iguales, y el intervalo de confianza para para la diferencia de medias es $\mu_1 - \mu_2 \in (-191.20, -108.80)$ con una confianza del 95%.

Ejercicio 1.14. La siguiente tabla muestra el porcentaje de industrias españolas y europeas según el consumo de energía primaria de las mismas durante el año 2002 (se estudiaron 1000 industrias españolas y 1000 del resto de europa).

Fuente energética	España	Resto de Europa
Petróleo	52.2%	40.4%
Carbón	15.2%	14.8%
Nuclear	13.0%	15.2%
Gas	12.8%	23.5%
Energías Renovables	6.5%	6.1%

Estudiar, según los intervalos de confianza para diferencia de proporciones, para qué energías el porcentaje de industrias de España es significativamente diferente del resto de Europa.

Solución

Los intervalos de confianza del 95% para la diferencia de proporciones de industrias que usan las diferentes fuentes de energía en España y en el resto de Europa son:

- Petróleo: $(0.0747, 0.1613)$. Hay diferencias significativas.
- Carbón: $(-0.0273, 0.0353)$. No hay diferencias significativas.
- Nuclear: $(-0.0525, 0.0085)$. No hay diferencias significativas.
- Gas: $(-0.1405, -0.0735)$. Hay diferencias significativas.
- Energías Renovables: $(-0.0173, 0.0253)$. No hay diferencias significativas.

2 Contrastes de hipótesis paramétricos

Ejercicio 2.1. Sabiendo que el año pasado el consumo per cápita de azúcar en España fue de 4.8 kg y que este consumo sigue una distribución normal, hemos seleccionado aleatoriamente a 20 españoles obteniendo una media muestral de 5 kg y una cuasidesviación típica muestral de 0.4 kg. Contrastar la hipótesis de que el consumo de azúcar per cápita de este año en España no ha variado utilizando un nivel de significación del 10% en cada uno de los casos siguientes.

a. Suponiendo que la alternativa es que el consumo de azúcar per cápita sea distinto.

b. Suponiendo que la alternativa es que el consumo de azúcar per cápita sea mayor.

Solución

Sea $X \sim N(\mu, \sigma)$ la variable aleatoria que mide el consumo per cápita de azúcar en España.

a. La hipótesis nula es $H_0 : \mu = 4.8$ y la hipótesis alternativa es $H_1 : \mu \neq 4.8$. Utilizando el contraste de la t de student, el estadístico de contraste es $t = 2.24$ y el p-valor es $p = 2P(t(19) > 2.24) = 0.037$. Como el p-valor es menor que el nivel de significación $\alpha = 0.1$, se rechaza la hipótesis nula y se concluye que el consumo de azúcar per cápita de este año en España ha variado.

b. En este caso, la hipótesis nula es $H_0 : \mu = 4.8$ y la hipótesis alternativa es $H_1 : \mu > 4.8$. El estadístico del contraste es el mismo pero ahora el p-valor es $p = P(t(19) > 2.24) = 0.0185$. Como el p-valor es menor que el nivel de significación $\alpha = 0.1$, se rechaza la hipótesis nula y se concluye que el consumo de azúcar per cápita de este año en España ha aumentado.

Ejercicio 2.2. En una clase de Estadística se ha comprobado que el 20% del alumnado falta a clase. Para disminuir esta preocupante cifra, los profesores han incorporado un sistema de evaluación continua que tendrá en cuenta las notas de clase de los alumnos en la nota final. Contraste con un nivel de significación del 5% que la incorporación de este método no es efectiva, es decir, el absentismo antes y después de la evaluación continua es el mismo, sabiendo que el porcentaje medio de no asistencia en 50 días tomados al azar ha sido del 17%.

💡 Solución

Sea $X \sim B(50, p)$ la variable aleatoria que mide el número de días que falta un alumno a clase. La hipótesis nula es $H_0 : p = 0.2$ y la hipótesis alternativa es $H_1 : p \neq 0.2$. Utilizando el contraste para la proporción de una población, el estadístico de contraste es $z = -1.5$ y el p-valor es $p = 2P(Z < -0.53) = 0.298$. Como el p-valor es mayor que el nivel de significación $\alpha = 0.05$, no se rechaza la hipótesis nula y se concluye que la incorporación de este método no es efectiva.

Ejercicio 2.3. Una empresa fabricante de detergente afirma que el contenido de cada paquete de detergente sigue una distribución normal de media 2150 g, pero una Asociación de Consumidores no está conforme con esta afirmación, por lo que realiza un estudio consistente en obtener una muestra aleatoria simple de 121 paquetes de detergente, obteniendo un contenido medio muestral de 2070 g y una cuasidesviación típica muestral de 130 grs. Contraste esta hipótesis con un nivel de significación del 5%.

💡 Solución

Sea $X \sim N(\mu, \sigma)$ la variable aleatoria que mide el contenido de cada paquete de detergente. La hipótesis nula es $H_0 : \mu = 2150$ y la hipótesis alternativa es $H_1 : \mu \neq 2150$. Utilizando el contraste de la t de student, el estadístico de contraste es $t = -6.77$ y el p-valor es $p = 2P(t(120) < -6.77) = 0.00005$. Como el p-valor es menor que el nivel de significación $\alpha = 0.05$, se rechaza la hipótesis nula y se concluye contenido promedio de los paquetes de detergente es significativamente diferente de los 2150 g declarados por la empresa.

Ejercicio 2.4. El número de aprobados en una asignatura de un determinado curso ha sido del 64%. En uno de los grupos de ese curso se han presentado al examen 40 alumnos de los que 31 aprobaron. ¿Puede afirmarse con un nivel de significación del 5% que los alumnos de dicho grupo han obtenido mejores calificaciones que el resto de los alumnos del curso?

💡 Solución

Sea $X \sim B(40, p)$ la variable aleatoria que mide el número de aprobados en el grupo. La hipótesis nula es $H_0 : p = 0.64$ y la hipótesis alternativa es $H_1 : p > 0.64$. Utilizando el contraste para la proporción de una población, el estadístico de contraste es $z = 1.78$ y el p-valor es $p = P(Z > 1.78) = 0.0376$. Como el p-valor es menor que el nivel de significación $\alpha = 0.05$, se rechaza la hipótesis nula y se puede afirmar que los alumnos de dicho grupo han obtenido mejores calificaciones que el resto de los alumnos del curso.

Ejercicio 2.5. Se sabe que el consumo anual de helado correspondiente a cada español sigue una distribución normal y que el año pasado el consumo medio fue de 20 kg. Queremos contrastar si este año se va a mantener el consumo medio de helado que el año pasado, y para comprobarlo se efectúa una muestra aleatoria de 22 españoles, obteniéndose los siguientes resultados:

15, 18, 24, 31, 22, 12, 6, 35, 42, 2, 16, 25, 20, 10, 17, 19, 14, 30, 14, 23, 15, 19.

Realizar el contraste con un nivel de significación de 0.10.

Solución

Sea $X \sim N(\mu, \sigma)$ la variable aleatoria que mide el consumo anual de helado correspondiente a cada español. La hipótesis nula es $H_0 : \mu = 20$ y la hipótesis alternativa es $H_1 : \mu \neq 20$. Utilizando el contraste de la t de student, el estadístico de contraste es $t = -0.25$ y el p-valor es $p = 2P(t(21) < -0.25) = 0.8033$. Como el p-valor es mayor que el nivel de significación $\alpha = 0.1$, no se rechaza la hipótesis nula y se concluye que no hay pruebas evidentes en la muestra de que el consumo de helados este año vaya a ser diferente del año pasado.

Ejercicio 2.6. Telefónica ha constatado que el consumo de datos de los clientes que han contratado el paquete Fusión (fibra simétrica, TV, teléfono fijo y móvil. sigue una distribución normal cuya dispersión viene determinada por σ=3 Gb. Sin embargo, tras la incorporación de Netflix a la oferta de Movistar TV se ha observado que la dispersión habría podido cambiar, por lo que se ha llevado a cabo un muestreo con 15 clientes en el que la cuasidesviación típica es igual a 3.5 Gb.

Determine si efectivamente la varianza en el consumo de datos de los clientes Fusión ha cambiado tras la incorporación de Netflix y especifique la región crítica óptima utilizando un nivel de significación del 2%.

Solución

Sea $X \sim N(\mu, \sigma)$ la variable aleatoria que mide el consumo de datos de los clientes que han contratado el paquete Fusión. La hipótesis nula es $H_0 : \sigma^2 = 9$ y la hipótesis alternativa es $H_1 : \sigma^2 \neq 9$. Utilizando el contraste de la χ^2, el estadístico de contraste es $\chi^2 = 19.06$ y para un nivel de significación $\alpha = 0.02$, se obtienen las regiones críticas delimitadas por $\chi^2_{14}(0.01) = 5.629$ y $\chi^2_{14}(0.99) = 26.119$. Como el estadístico del contraste cae entre estos dos valores, dentro de la región de aceptación, no se rechaza la hipótesis nula y se concluye que no hay pruebas evidentes en la muestra de que la varianza en el consumo de datos de los clientes Fusión haya cambiado tras la incorporación de Netflix.

Ejercicio 2.7. La OMS está preocupada por el incremento de la obesidad entre los niños de 12 años. La variable peso de los niños se distribuye según una normal. El equipo formula la hipótesis de que el peso medio de los niños de 12 años es de 46.5 kg. Seleccionada una muestra aleatoria de 40 niños, se obtiene que la media muestral es 49 kg y la cuasidesviación típica muestral vale 7 kg.

a. Realizar el contraste de hipótesis adecuado para ver si el peso medio de los niños de 12 años es mayor de 46.5 kg.

b. Si se pretende detectar al menos un incremento de 2 kg en el peso medio de los niños de 12 años, ¿qué potencia tiene el contraste?

c. ¿Qué tamaño muestral sería necesario para conseguir una potencia del al menos el 80%?

Solución

Sea $X \sim N(\mu, \sigma)$ la variable aleatoria que mide el peso de los niños de 12 años.

a. La hipótesis nula es $H_0 : \mu = 46.5$ y la hipótesis alternativa es $H_1 : \mu > 46.5$. Utilizando el contraste de la t de student, el estadístico de contraste es $t = 2.26$ y el p-valor es $p = P(t(39) > 2.26) = 0.0147$. Como el p-valor es menor que el nivel de significación $\alpha = 0.05$, se rechaza la hipótesis nula y se concluye que el peso medio de los niños de 12 años es mayor de 46.5 kg.

b. Cuando hay un incremento de al menos 2 kg en el peso medio de los niños el estadístico del contraste es $t = 1.81$, y la potencia del contraste es $1 - P(t(39) < 1.81) = 0.039$. Esta potencia es muy baja, lo cual indica que es poco probable que esta prueba detecte una diferencia de ese tamaño si realmente existe.

c. Para conseguir una potencia del al menos el 80% se necesita un tamaño muestral de $n = 76$ niños.

Ejercicio 2.8. Al lanzar un nuevo producto al mercado, el fabricante duda entre que sea adquirido por el 20% de la población o que sea adquirido por el 30%. Seleccionada al azar una muestra de 400 posibles compradores del producto se obtiene que la demandaría el 22%. La regla de decisión que utilizaremos es aceptar la hipótesis nula si adquieren el producto menos del 25% de los consultados.

a. ¿Se puede aceptar la hipótesis nula?

b. Calcular nivel de significación del contraste.

c. Obtener la potencia del contraste.

Solución

Sea $X \sim B(400, p)$ la variable aleatoria que mide el número de compradores del producto.

a. La hipótesis nula es $H_0 : p = 0.2$ y la hipótesis alternativa es $H_1 : p > 0.2$. Si se decide fijar aceptar la hipótesis nula con una proporción muestral menor de 0.25 entonces se aceptaría la hipótesis ya que la proporción muestral observada es 0.22.

b. El nivel de significación del contraste es la máxima probabilidad de cometer un error de tipo I, es decir, de rechazar la hipótesis nula cuando es cierta, que en este caso es $P(\hat{p} > 0.25|p = 0.2) = 0.0062$.

c. La potencia del contraste es $1 - P(\hat{p} < 0.25|p = 0.3) = 0.985$, por lo que se trata de un buen contraste para detectar al menos un incremento del 10% en la proporción de compradores.

Ejercicio 2.9. Según los datos publicados por REE (Red Eléctrica Española. el consumo medio diario de electricidad en los hogares españoles es de 9.5 watios/hora. La compañía comercializadora de electricidad Holaluz sabe que el citado consumo sigue una distribución normal, pero debe determinar cuál será el consumo medio en el que basará su estrategia de cara a 2021, ya que el aumento del teletrabajo podría provocar que el referido consumo medio varíe en el nuevo ejercicio y se incremente a 11 watios/hora, por lo que ha llevado a cabo un muestra aleatoria con 50 clientes en el que la media muestral ha resultado ser igual a un consumo de 10.57 watios/hora mientras que la cuasivarianza es igual a 9.

a. ¿Con que hipótesis debería trabajar Holaluz en 2021 si se considera un nivel de significación del 1%?

b. ¿Cuál es la potencia del contraste?

Solución

a. Sea $X \sim N(\mu, \sigma)$ la variable que mide el consumo diario de electricidad de los hogares españoles.

La hipótesis nula es $H_0 : \mu = 9.5$ y la hipótesis alternativa es $H_1 : \mu > 9.5$.
Utilizando el contraste de la t de student, el estadístico de contraste es $t = 2.522$ y el p-valor es $p = P(t(49) > 2.522) = 0.0075$. Como el p-valor es menor que el nivel de significación $\alpha = 0.01$, se rechaza la hipótesis nula y se concluye que el consumo medio de electricidad en los hogares españoles en 2021 será mayor de 9.5 watios/hora.

a. Suponiendo que la media del consumo es 11, el estadístico del contraste es $t = 3.536$ y la potencia del contraste suponiendo que el consumo medio es 11 watios/hora, es 0.965.

Ejercicio 2.10. Se ha desarrollado un aditivo para gasolina que reduce la emisión de CO en la combustión. Para comprobar la efectividad del aditivo, se realiza un estudio en el que se mide en una muestra de 10 coches la cantidad de CO emitida (en Kg/l), tanto con gasolina con aditivo, como con gasolina sin aditivo, obteniendo los siguientes resultados:

Sin aditivo:	0.38	0.42	0.41	0.39	0.45	0.47	0.44	0.38	0.40	0.50
Con aditivo:	0.38	0.36	0.38	0.32	0.39	0.45	0.39	0.39	0.35	0.48

Realizar un contraste de hipótesis para comprobar la efectividad del aditivo.

Solución

Sea $X \sim N(\mu, \sigma)$ la variable aleatoria que mide la diferencia entre la cantidad de CO emitida en la combustión de la gasolina sin aditivo y con aditivo. La hipótesis nula es $H_0 : \mu = 0$ y la hipótesis alternativa es $H_1 : \mu > 0$.
Utilizando el contraste de la t de student para la comparación de medias en poblaciones pareadas, el valor crítico que define las regiones de aceptación y rechazo para $\alpha = 0.05$ es $t^9_{0.95} = 1.833$. El estadístico del contraste es $t = 4.072$ que claramente cae en la región de rechazo al ser mayor que el valor crítico, y el p-valor es $p = P(t(9) > 4.072) = 0.0014$. Como el p-valor es menor que el nivel de significación $\alpha = 0.05$, se rechaza la hipótesis nula y se concluye que el aditivo es efectivo.

Ejercicio 2.11. En un estudio sobre el consumo de alcohol entre los jóvenes durante los fines de semana, se preguntó a 100 chicos y a 125 chicas, de los que 63 chicos y 59 chicas contestaron que consumían. En vista de estos datos, ¿existe alguna diferencia significativa entre las proporciones de chicos y chicas que consumen alcohol los fines de semana?

Solución

Sea $X_H \sim B(63, p_H)$ la variable aleatoria que mide el número de chicos que consumen alcohol durante los fines de semana en una muestra de tamaño 63 y $X_M \sim B(59, p_M)$ la variable aleatoria que mide el número de chicas que consumen alcohol durante los fines de semana en una muestra de tamaño 59. La hipótesis nula es $H_0 : p_H = p_M$ y la hipótesis alternativa es $H_1 : p_H \neq p_M$.
Utilizando el contraste de comparación de proporciones, el estadístico de contraste

es $z = 2.4026$ y el p-valor es $p = 2P(Z > 2.4026) = 0.0163$. Como el p-valor es menor que el nivel de significación $\alpha = 0.1$, se rechaza la hipótesis nula y se concluye que hay diferencias significativas entre el consumo de alcohol de chicos y chicas.

Ejercicio 2.12. Se ha realizado un estudio para comparar los ingresos medios de las personas de dos ciudades. Para ello, se ha tomado una muestra de 100 personas en una ciudad y 120 en la otra. En la primera ciudad se ha observado una media de 1630 € mensuales y una cuasidesviación típica de 150 €, mientras que en la segunda ciudad se ha observado una media de 1780€ y una cuasidesviación típica de 160 €. ¿Existen diferencias significativas entre los ingresos medios de las personas de las dos ciudades? Calcular el intervalo de confianza del 95% para la diferencia entre los ingresos medios.

Solución

Sea $X_1 \sim N(\mu_1, \sigma_1)$ la variable aleatoria que mide los ingresos mensuales de las personas en la primera ciudad y $X_2 \sim N(\mu_2, \sigma_2)$ la variable aleatoria que mide los ingresos mensuales de las personas en la segunda ciudad.
Planteamos las hipótesis $H_0 : \mu_1 = \mu_2$ y $H_1 : \mu_1 \neq \mu_2$. Para poder utilizar el contraste de la t de student para la comparación de medias en dos poblaciones independientes, primero necesitamos contrastar la igualdad de varianzas mediante el contraste F. El estadístico del contraste es $F = 0.879$ y el p-valor es $p = 2P(F(99, 119) < 0.8789) = 0.508$. Como el p-valor es mayor que el nivel de significación $\alpha = 0.05$, no se rechaza la hipótesis nula y se puede asumir que las varianzas son iguales.
Aplicando el contraste de la t de student para la comparación de medias en dos poblaciones independientes con varianzas iguales, el estadístico de contraste, la cuasidesviación típica ponderada es $\hat{s}_p = 155.5384$, el estadístico del contraste es $t = -7.123$ y el p-valor es $p = 2P(t(218) < -7.123) = 7.5 \cdot 10^{-12}$. Como el p-valor es menor que el nivel de significación $\alpha = 0.05$, se rechaza la hipótesis nula y se concluye que hay diferencias significativas entre los ingresos medios de las personas de las dos ciudades.
El intervalo de confianza del 95% para la diferencia entre los ingresos medios es $(-191.51, -108.49)$.

Ejercicio 2.13. Para ver si la ley antitabaco está influyendo en el número de cigarros que se fuman mientras se está en los bares se seleccionó una muestra en la que se midió el número de cigarros fumados por hora mientras se estaba en un bar antes de la entrada en vigor de la ley y otra muestra distinta en la que también se midió el número de cigarros fumados por hora después de la entrada en vigor de la ley (se entiende que con la ley en vigor los cigarros se fuman en el exterior de los bares). Los resultados aparecen en la siguientes tablas:

Antes

Cigarros	Personas
$0-1$	12
$1-2$	21
$2-3$	20
$3-4$	8

Después

Cigarros	Personas
$0-1$	22
$1-2$	18
$2-3$	8
$3-4$	4

a. Calcular el intervalo de confianza del 99% para el número medio de cigarros fumados por hora en los bares antes de la entrada en vigor de la ley. ¿Cuántos individuos serían necesarios para poder estimar dicha media con un margen de error no mayor de ± 0.1 cigarros por hora?

b. Contrastar si la nueva ley ha reducido significativamente el consumo medio de tabaco en los bares. ¿Cuánto vale el p-valor del contraste?

Solución

Sean X e Y las variables que miden el número medio de cigarros fumados por hora antes y después de la entrada en vigor de la ley respectivamente.

a. Puesto que se trata de una muestra grande de $n_x = 61$, la fórmula del intervalo de confianza para la media es

$$\bar{x} \pm z_{\alpha/2}\frac{\hat{s}}{\sqrt{n}}.$$

A partir de la tabla de frecuencias se calculan los estadísticos necesarios:

$$\begin{aligned}
\bar{x} &= \frac{\sum x_i n_i}{n_x} = \frac{0.5 \cdot 12 + \cdots + 3.5 \cdot 8}{61} = \frac{115.5}{61} = 1.8934,\\
s_x^2 &= \frac{\sum x_i^2 n_i}{n_x} - \bar{x}^2 = \frac{0.5^2 \cdot 12 + \cdots + 3.5^2 \cdot 8}{61} - 1.8934^2 = \frac{273.25}{61} - 3.585 = 0.8944,\\
\hat{s}_x^2 &= \frac{n_x}{n_x - 1}s_x^2 = \frac{61}{60}0.8944 = 0.9093,\\
\hat{s}_x &= \sqrt{0.9093} = 0.9536.
\end{aligned}$$

Como se pide un nivel de confianza del 99% se tiene que $\alpha = 0.01$ y $\alpha/2 = 0.005$, de modo que buscando en la tabla de la función de distribución de la normal estándar se tiene que $z_{\alpha/2} = 2.5758$. Así pues, sustituyendo en la fórmula del intervalo se obtiene

$$\bar{x} \pm z_{\alpha/2}\frac{\hat{s}}{\sqrt{n}} = 1.8934 \pm 2.5758\frac{0.9536}{\sqrt{61}} = 1.8934 \pm 0.3145 = (1.5789, 2.2079).$$

Por otro lado, el número de individuos necesario para estimar la media con un margen de error no mayor de ± 0.1 cigarros por hora y una confianza del 99% es

$$n = \frac{4z_{\alpha/2}^2\hat{s}^2}{A^2} = \frac{4 \cdot 2.5758^2 \cdot 0.9093}{(2 \cdot 0.1)^2} = 603.2974,$$

es decir, se necesitarían como mínimo 603 individuos.

b. Para contrastar si la nueva ley ha reducido significativamente el consumo medio de tabaco en los bares hay que realizar un contraste unilateral de comparación de medias

$$\begin{aligned} H_0 &: \mu_x = \mu_y \\ H_1 &: \mu_x > \mu_y \end{aligned}$$

Como los tamaños muestrales son grandes, $n_x = 61$ y $n_y = 52$, y no se conocen las varianzas poblacionales, el estadístico de contraste es

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\hat{s}_x^2}{n_x} + \frac{\hat{s}_y^2}{n_y}}},$$

que sigue una distribución normal estándar.

Para calcularlo se necesitan, además de los estadísticos de X calculados en el apartado anterior, los estadísticos de Y, que son:

$$\begin{aligned} \bar{y} &= \frac{\sum y_j n_j}{n_y} = \frac{0.5 \cdot 22 + \cdots + 3.5 \cdot 4}{52} = \frac{72}{52} = 1.3846, \\ s_y^2 &= \frac{\sum y_j^2 n_j}{n_y} - \bar{y}^2 = \frac{0.5^2 \cdot 22 + \cdots + 3.5^2 \cdot 4}{52} - 1.3846^2 = \frac{145}{52} - 1.9171 = 0.8713, \\ \hat{s}_y^2 &= \frac{n_y}{n_y - 1}s_y^2 = \frac{52}{51}0.8713 = 0.8884, \end{aligned}$$

de manera que sustituyendo en la fórmula del estadístico de contraste se obtiene

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\hat{s}_x^2}{n_x} + \frac{\hat{s}_y^2}{n_y}}} = \frac{1.8934 - 1.3846}{\sqrt{\frac{0.9093}{61} + \frac{0.8884}{52}}} = 2.8447.$$

Como el estadístico sigue una distribución normal estándar, la región de aceptación para un nivel de significación $\alpha = 0.05$ es $Z \leq z_\alpha = 1.6449$, y como el estadístico cae fuera de esta región se rechaza la hipótesis nula y se concluye que hay pruebas significativas de que la nueva ley ha reducido el consumo medio de tabaco en los bares.

Finalmente, el p-valor del contraste es

$$P(Z > 2.8447) = 1 - P(Z \leq 2.8447) = 1 - F(2.8447) = 0.0022.$$

Ejercicio 2.14. Para ver si una campaña de publicidad sobre un fármaco ha influido en sus ventas, se tomó una muestra de 8 farmacias y se midió el número de fármacos vendidos durante un mes, antes y después de la campaña, obteniéndose los siguientes resultados:

Antes	147	163	121	205	132	190	176	147
Después	150	171	132	208	141	184	182	145

Realizar un contraste de hipótesis para comprobar si la campaña de publicidad ha aumentado las ventas medias.

Solución

Sea X la variable aleatoria que mide la diferencia entre el número de fármacos vendidos en una farmacia en un mes después y antes de la campaña de publicidad. Planteamos las hipótesis $H_0 : \mu = 0$ y $H_1 : \mu > 0$. Utilizando el contraste de la t de student para la una media poblacional, el estadístico del contraste es $t = 1.965$ y el p-valor es $p = P(t(7) > 1.965) = 0.045$. Como el p-valor es menor que el nivel de significación $\alpha = 0.05$, se rechaza la hipótesis nula y se concluye que hay pruebas significativas de que la campaña de publicidad ha aumentado las ventas.

Ejercicio 2.15. En una encuesta sobre la intención de voto un partido político, se han tomado dos muestra de 1000 personas en dos ciudades A y B. En la ciudad A se han obtenido 350 votos a favor del partido y en la ciudad B se han obtenido 390 votos a favor del partido. ¿Existen diferencias significativas entre las proporciones de votos a favor del partido en las dos ciudades?

Solución

Sea $X_A \sim B(1000, p_A)$ la variable aleatoria que mide el número de votos a favor del partido en la ciudad A y $X_B \sim B(1000, p_B)$ la variable aleatoria que mide el número de votos a favor del partido en la ciudad B.
La hipótesis nula es $H_0 : p_A = p_B$ y la hipótesis alternativa es $H_1 : p_A \neq p_B$. Utilizando el contraste para la comparación de dos poblaciones, se tiene que el estadístico de contraste es $z = -1.853$ y el p-valor es $p = 2P(Z < -1.853) = 0.0639$. Como el p-valor es mayor que el nivel de significación $\alpha = 0.05$, no se rechaza la hipótesis nula y se concluye que no hay diferencias significativas entre las proporciones de votos a favor del partido en las dos ciudades.