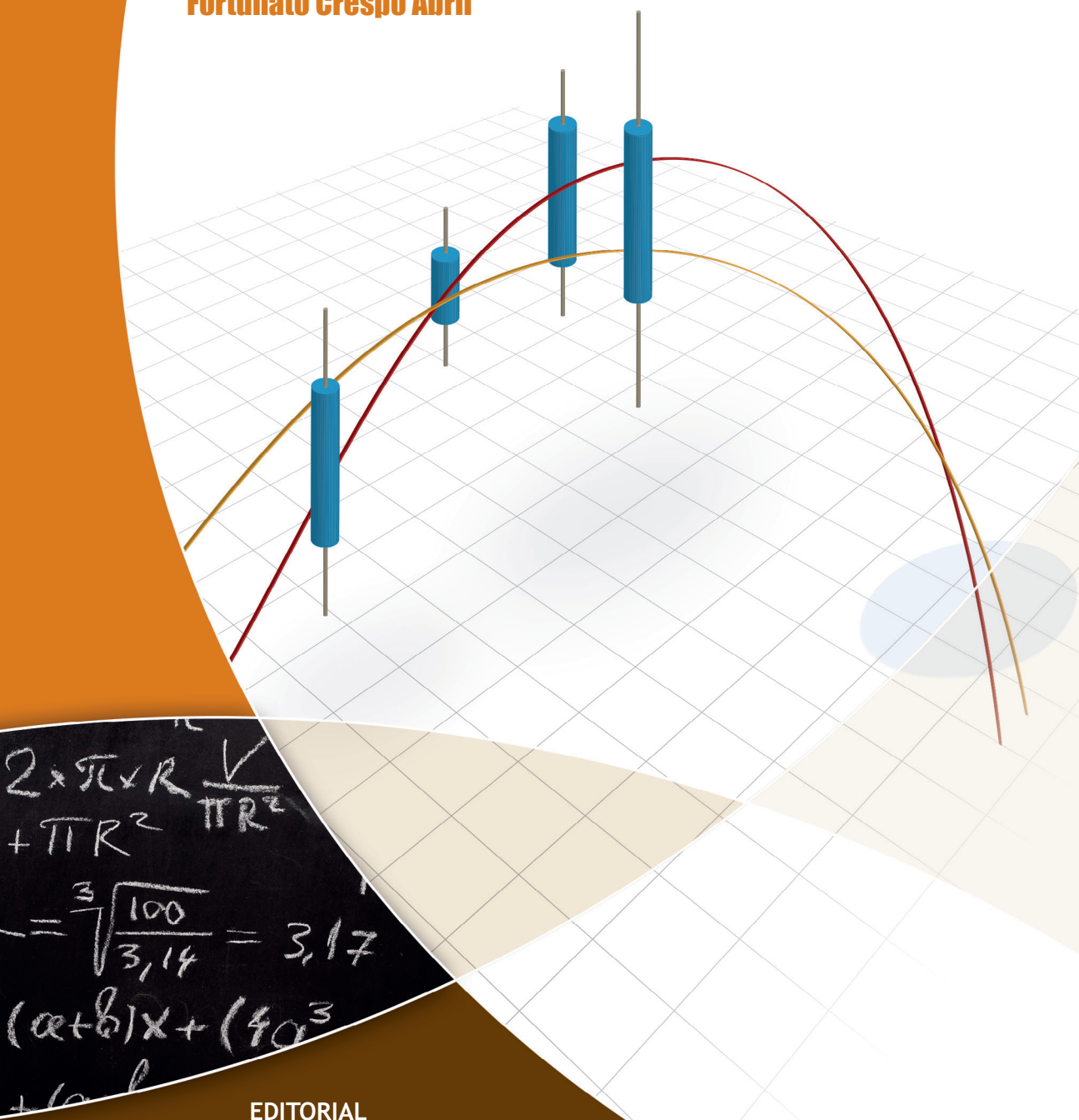
Métodos estadísticos: ejercicios resueltos y teoría

Fortunato Crespo Abril



EDITORIAL
UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Fortunato Crespo Abril

Métodos estadísticos:

ejercicios resueltos y teoría

EDITORIAL
UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Colección *Académica*

Los contenidos de esta publicación han sido revisados por el Departamento de Estadística e Investigación Operativa Aplicadas y Calidad de la Universitat Politècnica de València

Para referenciar esta publicación utilice la siguiente cita:
Crespo Abril, Fortunato (2017*). Métodos Estadísticos: ejercicios resueltos y teoría.* Valencia: Editorial Universitat Politècnica de València

distribución: www.lalibreria.upv.es / Ref.: 0291_05_01_01

Imprime: Byprint Percom, sl

ISBN: 978-84-9048-667-2
Impreso bajo demanda

Impreso en España

Resumen

Mi experiencia como docente, impartiendo las asignaturas de Introducción a la Estadística, Métodos Estadísticos en Economía, y Econometría, en el grado de Administración y Dirección de Empresas de la Universitat Politècnica de València, me ha permitido detectar los puntos y conceptos que mayor dificultad entrañan para los alumnos a la hora de resolver problemas de Estadística.

Los métodos estadísticos descritos en este libro son algunos de los métodos de inferencia clásicos más utilizados en la práctica: técnicas de inferencia para medias y para proporciones (en problemas de una y de dos muestras independientes), ANOVA de un factor, prueba de Kruskal-Wallis, y algunas pruebas de hipótesis basadas en la distribución χ^2 (problemas de tablas de contingencia y test de bondad de ajuste).

Existen muchos libros que explican estos métodos y que incluyen una buena colección de problemas. Sin embargo, la mayoría de mis alumnos siguen demandando más problemas resueltos y, sobre todo, problemas en los que se desarrolle paso a paso los procedimientos utilizados.

Este libro presenta una colección de problemas de inferencia básica resueltos con un gran nivel de detalle, pero, al mismo tiempo es algo más. La primera sección de cada capítulo introduce los procedimientos que se utilizarán en la resolución de los problemas del mismo, haciendo un especial hincapié en las condiciones que la teoría estadística exige para poder aplicarlos, y también, y esto es quizás aún más interesante, los requisitos que deben cumplirse en la práctica para que dichos procedimientos sean seguros y fiables.

He hecho un esfuerzo en crear un conjunto de esquemas que permitan visualizar los conceptos fundamentales en los que se sustentan los métodos estadísticos de inferencia comentados a lo largo de los diferentes capítulos. Cada esquema presenta la distribución muestral de un estadístico, resultado base en el que se apoya el método propuesto, junto con las condiciones que deben cumplirse para que dicha distribución sea, precisamente, la que es. Creo que estos esquemas son un buen resumen de todo aquello que debe tenerse en cuenta a la hora de aplicar cada uno de los métodos estadísticos descritos.

Otro aspecto importante en el que se centran los problemas resueltos es mostrar cómo deben redactarse las conclusiones de un estudio de inferencia. He utilizado un lenguaje sencillo para explicar los resultados a alguien que no sepa estadística y, también, en la mayoría de los ejercicios se muestra la forma en que estas conclusiones deben aparecer en un informe técnico o en una publicación de carácter científico.

Soy consciente de que los alumnos pueden caer en la trampa de creer que los ejercicios son más sencillos de lo que en realidad son: cuando se realizan estudios de inferencia, no sólo hay que saber cómo aplicar un determinado método, sino que el primer paso es conocer qué método se debe aplicar y, esta dificultad adicional se pierde si uno sabe de antemano que los ejercicios de un capítulo aplican unos métodos concretos, y no otros. Por ello, he añadido un capítulo que presenta los enunciados de todos los ejercicios desordenados (al azar), de este modo el lector puede poner realmente a prueba sus conocimientos a la hora de resolver los ejercicios.

Tanto en las clases de teoría, como en las prácticas que imparto, utilizo el programa R para aplicar los métodos descritos en este libro. R es también el programa que utilizo en mi tarea de investigador. Es el software estadístico que prefiero utilizar: por su potencia, por su versatilidad y, porque además, es gratuito.

En muchos de los ejercicios, he añadido los comandos de R que permiten obtener unos resultados más precisos que los obtenidos al realizar los cálculos utilizando las tablas de probabilidades y percentiles que aparecen al final del libro, pero éste no pretende ser ningún manual sobre cómo utilizar este programa.

Cómo utilizar este libro

La primera sección de cada capítulo presenta un poco de teoría, con los aspectos más importantes a tener en cuenta antes de aplicar los métodos estadísticos que se describen. Esta sección debe estudiarse con detalle para poder resolver los problemas sin dificultad. Una vez aprendida, el esquema que aparece en la misma puede utilizarse, a modo de resumen, para repasar de forma rápida todos estos aspectos.

Una vez estudiada esta sección, pueden abordarse los ejercicios. Los primeros ejercicios de cada capítulo sirven de guía para intentar resolver el resto de forma autónoma, antes de ver la solución de los mismos.

Al comienzo de cada ejercicio se indica la página en la que aparece el enunciado del mismo, sin incluir su solución. De este modo, se pueden resolver los problemas sin tener ningún tipo de pista sobre su resolución.

> En muchos de los ejercicios aparecen recuadros en gris como este. Estos recuadros muestran cómo pueden redactarse las conclusiones de un análisis estadístico utilizando una terminología estadística precisa. En estos recuadros encontraréis la forma habitual en la que aparecen los resultados de un análisis estadístico en publicaciones científicas y en la prensa.

El capítulo 8 presenta, a modo de resumen, y para facilitar su consulta, las fórmulas requeridas por los diferentes métodos estadísticos utilizados a lo largo

del libro. Las tablas utilizadas en los ejercicios para obtener probabilidades y percentiles, son las que aparecen en este capítulo. Al final del mismo se presenta una colección de problemas que os ayudará a saber cómo utilizar de forma correcta estas tablas.

En algunos ejercicios aparecen recuadros como el siguiente:

comando en R

resultados

Este tipo de recuadros muestran los comandos que pueden utilizarse, desde el programa R, para obtener resultados más precisos que los que obtenemos al realizar los cálculos a mano.

Cuando en los enunciados de los ejercicios se proporcionan los datos de la muestra utilizada para hacer el análisis estadístico pertinente, estos recuadros presentan, además, los comandos que permiten obtener los intervalos de confianza y el *valor p* de las pruebas de hipótesis que corresponda.

Una forma rápida de obtener ayuda adicional sobre cualquier comando de R, es utilizar el comando `help()`, indicando, dentro del paréntesis, el nombre del comando sobre el que pedimos ayuda.

help(t.test)

Utilizando la orden del ejemplo anterior, obtendremos ayuda sobre el comando `t.test`, que permite aplicar los procedimientos t de inferencia para medias en problemas de una y de dos muestras independientes.

Si tenéis dificultades con algún método estadístico en concreto, podéis resolver primero los ejercicios del capítulo que lo describe, pero, recordad que ponerse a prueba exige no saber de antemano el método estadístico que se debe aplicar en cada caso.

Índice general

Capítulo 1

Inferencia para medias Problemas de una muestra

Los métodos que se describen en este capítulo centran su atención en obtener conclusiones válidas respecto a la media μ (valor desconocido) de una variable aleatoria numérica (generalmente continua) de una población de referencia.

Los problemas a los que se aplican estos métodos son problemas de una sola muestra.

1.1 Procedimientos Z (asumimos que σ es conocida)

Los procedimientos basados en la distribución Normal estándar, denominados por ello *procedimientos* Z, en los problemas de inferencia sobre la media μ de una población, encajan en el esquema presentado en la Figura 1.1, y son los siguientes:

- Estadístico utilizado para realizar un contraste de hipótesis:

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

- Intervalo con un nivel de confianza del $(1-\alpha)\cdot 100\,\%$ para la media μ de la población:

$$\overline{x} \pm z^{\alpha/2} \cdot \frac{\sigma}{n}$$

donde:

$\overline{x}$ es la media de la muestra

n es el tamaño de la muestra (el número de unidades observadas)

σ es el valor que asumimos conocer para la desviación típica de la población estudiada

$z^{\alpha/2}$ es el valor crítico que buscaremos en la tabla de la distribución Normal estándar (pag. 302)

μ_0 es el valor considerado para la media de la población en la hipótesis nula H_0

El esquema de la Figura 1.1 refleja un conjunto de aspectos clave para entender los fundamentos de los métodos estadísticos utilizados:

1. Partimos de una población de elementos, o individuos, de los cuales se estudia una característica aleatoria numérica X, cuyo valor medio μ es desconocido. Éste es el parámetro de interés, y sobre el cual queremos realizar inferencia.

2. Con el fin de obtener conclusiones válidas respecto del parámetro μ, se obtendrá información de una muestra aleatoria simple (m.a.s.) de tamaño n extraída de esta población.

3. El resultado clave en el que se basan los métodos estadísticos es saber qué resultados esperaríamos obtener si seleccionásemos muchas muestras del mismo tamaño de la población objetivo.

   La teoría estadística indica que bajo ciertas condiciones, (las que aparecen dentro de las elipses azules en el esquema) la distribución de las medias $\overline{x}$ de las posibles muestras de tamaño n sigue un modelo de distribución Normal: $\overline{X} \sim N\left(\mu\ ,\ \frac{\sigma}{\sqrt{n}}\right)$

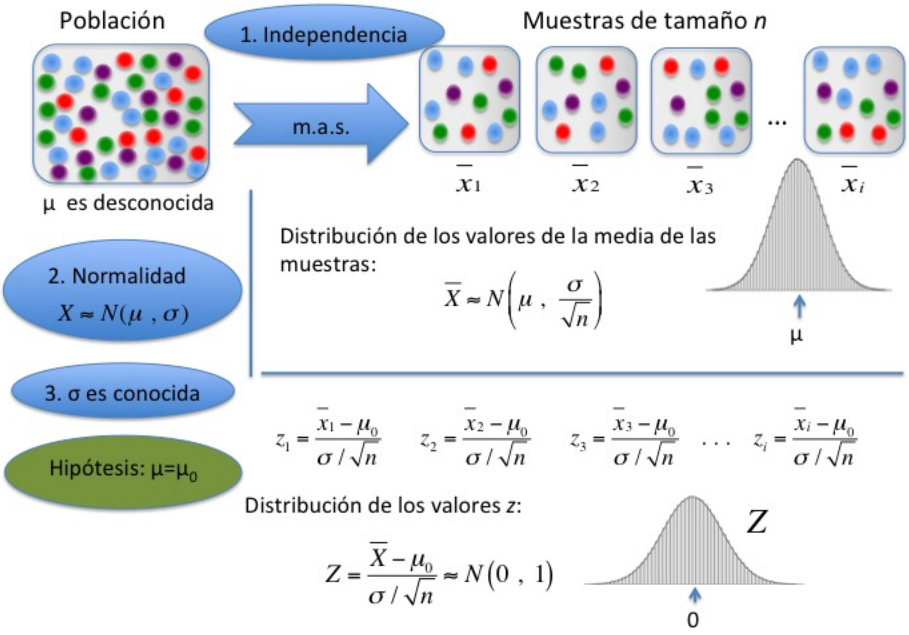


Figura 1.1: Procedimientos Z para medias en problemas de una muestra.

4. A partir de este resultado, y si el valor de la media de la población μ fuese el valor μ_0, tipificando obtendremos la distribución muestral del estadístico z:

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0\ ,\ 1)$$

   Si $\mu = \mu_0$, ¿es razonable obtener, simplemente por la variación que ocasiona el azar del muestreo, unos resultados como los obtenidos en nuestra muestra?

   Conocer la distribución muestral del estadístico Z nos permite responder esta pregunta en términos de probabilidades (calcular el *valor p*), y obtener un intervalo para estimar el valor de μ con un nivel de confianza determinado.

Como hemos indicado, para poder aplicar los procedimientos Z, la teoría estadística exige que se cumplan las siguientes condiciones (elipses azules en la Figura 1.1):

1. Independencia. La muestra debe ser elegida utilizando un muestreo aleatorio simple (m.a.s.). Esta forma de selección garantiza que los elementos o individuos de la muestra han sido seleccionados de manera independiente.

   Asumimos, además, que el tamaño de la población estudiada es infinito.

2. Normalidad. Asumimos que la variable X, estudiada en la población, sigue un modelo de distribución Normal, con una media μ desconocida. Es decir, se asume que $X \sim N(\mu\ ,\ \sigma)$.

3. Asumimos que la desviación típica σ es un valor conocido.

¿Son seguros estos procedimientos en la práctica?

Las condiciones que exige la teoría estadística para aplicar los procedimientos anteriores difícilmente se cumplen en la práctica. Tiene más sentido preguntarse si los métodos descritos siguen siendo fiables aunque las condiciones exigidas no se cumplan de manera estricta.

Estas son las condiciones que debes verificar para utilizar de forma segura estos procedimientos en la práctica:

1. **Independencia**. La muestra debe haber sido seleccionada utilizando un m.a.s. La utilización de otros tipos de muestreo probabilístico (estratificado, sistemático, por conglomerados, etc.) exige cambios en la fórmula utilizada para obtener el estadístico utilizado en el contraste, aunque las fórmulas vistas pueden ser una buena aproximación.

   La selección aleatoria de las unidades que forman la muestra garantiza que éstas no están relacionadas y que las observaciones son independientes.

   Aplicar estos métodos a muestras seleccionadas de cualquier forma, utilizando, por ejemplo, un muestreo de conveniencia o muestras de voluntarios, tiene, en el mejor de los casos, un carácter meramente especulativo, y los resultados obtenidos pueden ser totalmente erróneos.

2. **Normalidad**. La comprobación de este supuesto exige la realización de un análisis descriptivo de los datos de la muestra. Esto nos permite observar las regularidades de los mismos y detectar la presencia de datos anómalos.

   Dependiendo del tamaño de la muestra:

   - Muestras pequeñas ($n < 25$): utiliza los procedimientos Z sólo si no existen datos anómalos que tengan un peso excesivo en el cálculo de la media $\overline{x}$, y cuando la distribución tenga un sólo pico y no presente una marcada asimetría.

     Obtener un gráfico de la función de densidad empírica, un gráfico de cuantiles y un diagrama de caja es útil para comprobar estos aspectos.

   - Muestras de más de 25 elementos: puedes utilizar los procedimientos Z con seguridad, incluso cuando la distribución de los datos sea asimétrica y existan datos anómalos.

     El *Teorema Central del Límite* garantiza que, cuando el tamaño de las muestras seleccionadas n es elevado, la distribución de la media $\overline{X}$ de las muestras es Normal, incluso cuando la distribución de la variable estudiada X presente asimetrías importantes.

   De todos modos, aplica el sentido común: el valor 25 es simplemente un valor de referencia, que puede servirte de guía, pero resulta obvio que

una muestra con 27 datos que presente muchos datos anómalos y con una marcada asimetría nos hará dudar del cumplimiento de la condición de Normalidad y será arriesgado aplicar los procedimientos Z a esta muestra de datos.

Siempre debes hacer un análisis descriptivo de tus datos porque, además de comprobar los aspectos mencionados anteriormente, puede ayudarte a detectar errores ocasionados por la captura incorrecta de los mismos.

La presencia de datos anómalos (especialmente si éstos son valores muy extremos) puede sesgar los resultados obtenidos, ya que la media muestral $\overline{x}$, es un estadístico que se ve muy afectado por la presencia de estos valores.

Si se detectan datos anómalos debe estudiarse por qué se han producido y, salvo que se trate de errores realizados al tomar los datos, éstos no pueden eliminarse de la muestra sin más. Lo más honesto es realizar un segundo análisis, sin incluir los valores anómalos, para ver cómo cambian las conclusiones, y presentar las conclusiones de ambos estudios (con y sin datos anómalos) para contrastar los resultados.

3. σ **conocida**. Asumir que se conoce la desviación típica de la variable estudiada en la población suele ser muy poco realista.

   En algunas ocasiones el valor de σ se estima a partir de datos de estudios previos; en estos casos, debería garantizarse que las condiciones en que se realizaron estos estudios son las mismas que las actuales, y que el valor de σ no ha cambiado.

   En los problemas que permiten utilizar estos procedimientos debe expresarse, de forma clara, que se asume que el valor de σ es conocido.

   En general, será preferible utilizar los procedimientos t para medias como alternativa a estos procedimientos, ya que utilizan la desviación típica de la muestra s, para calcular el valor del estadístico de referencia, en lugar de recurrir a la desviación típica de la población.

El *valor p* de una prueba de hipótesis

El *valor p* obtenido al realizar un contraste de hipótesis, es una medida de la fuerza que proporcionan los datos de la muestra seleccionada en contra de la hipótesis nula y a favor de la hipótesis alternativa que plantea el test.

Esta fuerza se expresa en términos de probabilidad (de ahí el nombre que recibe: *p* por probabilidad), y su valor es la probabilidad, asumiendo que la hipótesis nula es cierta, de obtener, simplemente por azar, unos resultados como los observados en la muestra o más extremos, en el sentido que indica la hipótesis alternativa.

El *valor p* se calcula atendiendo al efecto que buscamos en los datos, y que queda definido por el signo que aparece en la hipótesis alternativa H_1. La Tabla 1.1 muestra cómo se calcula el *valor p* según el tipo de test planteado.

Pruebas de una cola		Test de 2 colas
$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$
$H_1 : \mu > \mu_0$	$H_1 : \mu < \mu_0$	$H_1 : \mu \neq \mu_0$
$valor\ p = P(Z > z)$	$valor\ p = P(Z < z)$	$valor\ p = 2 \cdot P(Z > \lvert z \rvert)$

Tabla 1.1: Cálculo del *valor p* utilizando los procedimientos *Z*

La Figura 1.2 muestra cómo interpretar correctamente el *valor p* asociado a una prueba de hipótesis. Cuanto menor es el *valor p*, mayor es la evidencia que muestran los datos de la muestra en contra de la hipótesis nula que se plantea en el test. El valor obtenido suele compararse con un valor de referencia α, que recibe el nombre de *nivel de significación*:

- Si el *valor* $p < \alpha$, nuestra decisión será rechazar la hipótesis nula, y aceptar la hipótesis alternativa. Diremos, en este caso, que los resultados del estudio son significativos para un nivel de significación α.
- En cambio, cuando el *valor* $p > \alpha$, no tendremos suficiente evidencia para poder rechazar H_0. Los resultados del estudio no serán significativos para ese nivel de significación.

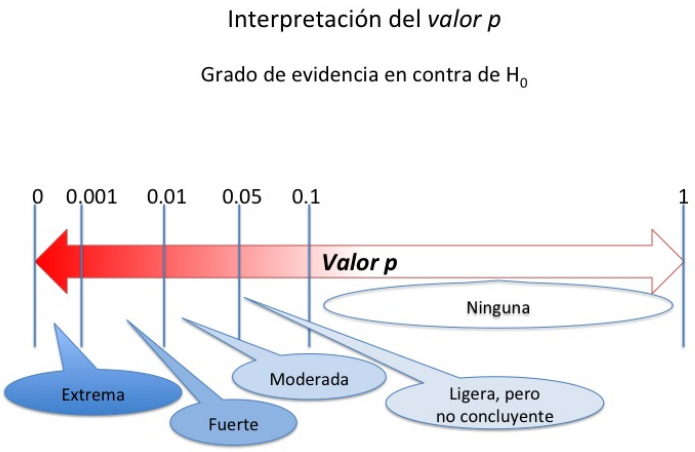


Figura 1.2: Cómo interpretar el *valor p* en una prueba de hipótesis

El nivel de significación del test lo fija el investigador que realiza la prueba. Generalmente suele utilizarse un valor $\alpha = 5\,\%$, pero lo que realmente importa al realizar una prueba de hipótesis es el *valor p* obtenido. Cuanto más pequeño es el *valor p* mayor es el grado de evidencia que proporcionan los datos de la muestra en contra de la hipótesis nula. Es decir, obtener un *valor* $p = 0{,}045$ o un *valor* $p = 0{,}055$ es, en términos prácticos, lo mismo, y esto es lo que realmente importa.

1.1.1 Consumo de agua en Valencia

Enunciado pág. 272

Según datos oficiales del Instituto Nacional de Estadística (INE), el consumo medio de agua en los hogares españoles durante el año 2014 fue de 40 litros por persona y día.

¿Es el consumo de agua similar para los habitantes de la Comunitat Valenciana?

Con el fin de responder a esta pregunta se realizó un estudio, en el que se midió durante el año 2014 el consumo de agua en una muestra aleatoria de 100 hogares

de la Comunitat Valenciana. Tras finalizar la recogida de datos, se calculó el consumo medio de agua por habitante y día en los hogares seleccionados.

a) Identificar y definir la variable aleatoria implicada en este estudio, la población objetivo y la muestra del mismo.

Muestra: los 100 hogares de la Comunitat Valenciana seleccionados de forma aleatoria para hacer el estudio.

Población: todos los hogares de la Comunitat Valenciana.

Variable aleatoria: X = Consumo de agua en cada hogar (litros por persona y día)

b) Indicar cuál sería la hipótesis nula y la alternativa de un contraste que permita responder a la pregunta planteada.

$$\begin{aligned} H_0 &: \mu = 40 \; litros \\ H_1 &: \mu \neq 40 \; litros \end{aligned}$$

La H_0 afirma que el consumo medio de agua por persona y día en la Comunitat Valenciana en el año 2014, es similar al consumo medio de agua en el Estado Español, mientras que la alternativa contempla valores superiores e inferiores a 40 litros (consumo medio de agua por persona y día en los hogares españoles en 2014). El test planteado es de dos colas.

c) Los datos recopilados muestran que el consumo medio por habitante y día en la muestra seleccionada fue de $\overline{x} = 42{,}5$ litros. Obtener el *valor p* del contraste anterior y escribir un breve párrafo con las conclusiones que se desprenden del estudio realizado, asumiendo para ello que la desviación típica del consumo de agua en los hogares de la Comunitat Valenciana es de $\sigma = 10$ litros por persona y día.

Realizar un dibujo de la distribución que corresponda, que muestre el valor crítico y las zonas de aceptación y de rechazo para un nivel de significación $\alpha = 5\,\%$, el valor del estadístico y el *valor p* del test.

Valor del estadístico, obtenido a partir de los datos de la muestra seleccionada:

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{42{,}5 - 40}{10/\sqrt{100}} = 2{,}5$$

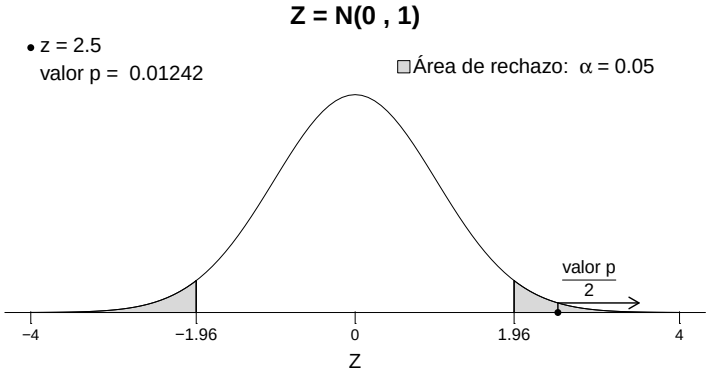


Figura 1.3: Zonas de aceptación y de rechazo, valor del estadístico z y *valor p*.

$$Valor\ p = 2 \cdot P(Z > 2{,}5) = 2 \cdot 0{,}0062 = 0{,}0124$$

2*pnorm(-2.5, lower.tail=T)

0.01241933

Como el valor p es inferior al nivel de significación ($\alpha = 0{,}05$), rechazamos la hipótesis nula planteada en el test.

Los resultados del estudio son significativos (*valor* $p = 0{,}0124$) para un nivel de significación del 5 %. El estudio realizado permite concluir que el consumo medio de agua por persona y día en los hogares de la Comunitat Valenciana no es similar al consumo medio en el resto del estado, que fue de 40 litros por persona y día en el año 2014.

d) Obtener un intervalo con una confianza del 95 % para el consumo medio de agua por persona y día en los hogares de la Comunitat Valenciana en el año 2014.

$$\overline{x} \pm z^{0{,}025}\frac{\sigma}{\sqrt{n}}$$

$$42{,}5 \pm 1{,}96\frac{10}{\sqrt{100}} = 42{,}5 \pm 1{,}96 = [40{,}54\ ,\ 44{,}46]$$

Tenemos una confianza del 95 % de que el consumo medio de agua en los hogares de la Comunitat Valenciana en el año 2014 estuvo comprendido entre los 40,54 *y los* 44,46 *litros por persona y día.*

El consumo medio de agua en los hogares de la Comunitat Valenciana durante el año 2014 fue de 42,5 litros por persona y día (margen de error de ±1,96 litros por persona y día, para una confianza del 95 %).

e) Los datos del estudio realizado, ¿permitirían admitir la hipótesis de que el consumo medio de agua por persona y día en los hogares de la Comunitat Valenciana en el año 2014 fue de 41 litros? Justifica tu respuesta.

El valor 41 está dentro del intervalo de confianza obtenido en el apartado anterior, por ello, este valor sí sería admitido al realizar una prueba de hipótesis de dos colas. Esta hipótesis sí es consistente con los datos obtenidos, al igual que cualquier otro valor que pertenezca al intervalo de confianza obtenido.

Cualquier hipótesis nula del tipo $H_0 : \mu = \mu_0$ *que considere un valor* μ_0 *que pertenece al intervalo* $[a\ ,\ b]$*, obtenido con un nivel de confianza* $(1-\alpha)100\,\%$*, será admitida en un contraste de dos colas que utilice un nivel de significación* α*.*

f) Tras realizar un análisis descriptivo de los datos de la muestra, se obtuvieron los gráficos que se presentan en la Figura 1.4. Indica, a la vista de estos gráficos, si se incumple alguna de las condiciones asumidas para poder aplicar con seguridad los procedimientos estadísticos que has utilizado.

- *La condición de* independencia *se cumple, ya que se indica que la muestra de hogares seleccionada para realizar el estudio es aleatoria. Cabe esperar, por tanto, que esta muestra sea representativa de la totalidad de los hogares de la Comunitat Valenciana.*

- *Observando el diagrama de caja vemos que la distribución del consumo de agua presenta una ligera asimetría positiva (hacia la derecha) y que además en la muestra seleccionada existe un valor anómalo. Al tratarse de un único punto aislado y no estar éste excesivamente alejado del resto de los datos, su peso en el análisis no será excesivo, por lo que no será necesario tomar ninguna medida especial.*

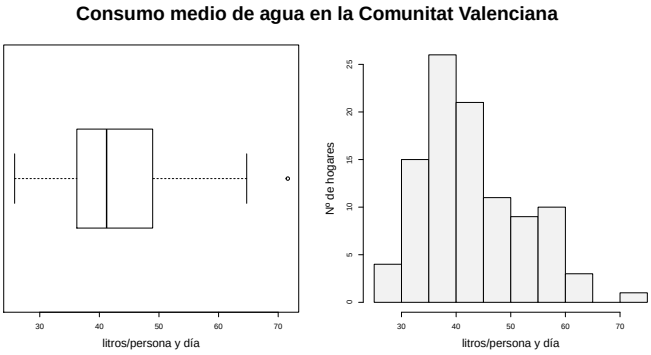


Figura 1.4: Diagrama de caja e histograma del consumo de agua por persona y día en una muestra de 100 hogares de la CV.

Observando el histograma vemos, además, que se trata de una distribución de un solo pico (alrededor de los 40 litros).

Por lo tanto, aunque la condición de Normalidad no se cumple de forma estricta, al tratarse de una muestra grande ($n = 100$), el TCL garantiza que la distribución muestral de la media sí seguirá un modelo de distribución Normal, por lo que, en la práctica, es seguro utilizar los procedimientos Z en este estudio.

- *El enunciado indica que la desviación típica del consumo de agua en los hogares de la CV es de* 10 *litros por persona y día, pero no se justifica cómo se ha obtenido este dato. ¿Se trata de un valor obtenido a partir de estudios fiables?*

  En caso de dudas, no deberíamos asumir que conocemos este dato, y utilizar los procedimientos t, que utilizan la desviación típica de la muestra, en lugar de la desviación típica de la población.

1.1.2 Holandeses, entre los más altos del mundo

Enunciado pág. 283

Los holandeses figuran entre las poblaciones más altas del mundo, con una media de 1,84 metros para los varones, y de 1,71 metros para las mujeres. Se nota en las tiendas, en la calle, y en los colegios, con adolescentes de apenas 13 años que calzan ya un 44. Todo el género a la venta llega a tallas muy grandes.

En un país pequeño, con 16,7 millones de habitantes, los altos destacan y son la norma casi a partes iguales.

¿Siguen siendo los jóvenes españoles más bajos que los varones holandeses?

Con el fin de responder a esta pregunta se seleccionó una muestra aleatoria de 55 universitarios españoles, con edades comprendidas entre los 18 y los 23 años.

a) Identificar y definir la variable aleatoria implicada en este estudio, la población objetivo y la muestra del mismo.

Muestra: los 55 universitarios españoles seleccionados para hacer el estudio.

Población: todos los varones españoles con edades comprendidas entre los 18 y 23 años.

Variable aleatoria: X = Estatura de cada individuo (en cm)

b) Indicar cuál sería la hipótesis nula y la alternativa de un contraste que permita responder a la pregunta planteada.

$$\begin{aligned} H_0 &: \mu = 184\ cm \\ H_1 &: \mu < 184\ cm \end{aligned}$$

*La H_0 afirma que la estatura media de todos los varones españoles con edades comprendidas entre los 18 y 23 años es la misma que la de los varones holandeses (*184 *cm), mientras que la alternativa contempla únicamente valores por debajo de este valor, es decir, afirma que la estatura media de la población de varones españoles en este grupo de edad es inferior a la estatura media de los varones holandeses.*

c) Sabiendo que la estatura media de los 55 estudiantes seleccionados es $\overline{x} = 178$ cm, y asumiendo que la desviación típica de la estatura en la población de jóvenes españoles con edades comprendidas entre los 18 y los 23 años es $\sigma = 8{,}5$ cm, obtener el *valor p* del contraste anterior y escribir un breve párrafo con las conclusiones que se desprenden del estudio realizado.

Valor del estadístico, obtenido a partir de los datos de la muestra seleccionada:

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{178 - 184}{8{,}5/\sqrt{55}} = -5{,}23$$

$$Valor\ p = P(Z < -5{,}23) \approx 0$$

pnorm(-5.23, lower.tail=T)
8.475502e-08

Un valor p tan pequeño nos proporciona mucha evidencia en contra de la hipótesis nula planteada. Claramente debemos rechazar H_0.

Los resultados del estudio son extremadamente significativos (*valor* $p \approx 0$). El estudio realizado permite concluir que la estatura media de los jóvenes españoles, con edades comprendidas entre los 18 y los 23 años, es inferior a la estatura media de los varones holandeses (184 cm).

d) Obtener un intervalo con una confianza del 95 % para la estatura media de la población de varones españoles con edades comprendidas entre los 18 y los 23 años.

$$\overline{x} \pm z^{0{,}025}\frac{\sigma}{\sqrt{n}}$$

$$178 \pm 1{,}96\frac{8{,}5}{\sqrt{55}} = 178 \pm 2{,}25 = [175{,}75\ ,\ 180{,}25]\ \text{cm}$$

Tenemos una confianza del 95 % de que la estatura media de los varones españoles con edades comprendidas entre los 18 y los 23 años se encuentra entre los 175,75 *y los* 180,25 *cm.*

La estatura media de los varones españoles con edades comprendidas entre los 18 y 23 años es de 178 cm (margen de error de ±2,25 cm, para una confianza del 95 %).

e) Realiza una crítica del estudio, indicando además si se incumple alguna de las condiciones asumidas para poder aplicar con seguridad los procedimientos estadísticos que has utilizado.

- *El estudio compara la estatura media de los jóvenes españoles con la de los varones holandeses. En el primer caso se especifica un rango de edad (varones de 18 a 23 años), mientras que en el segundo no se indica ningún rango de edad. Sería más razonable comparar dos grupos de individuos con edades similares.*

- *La muestra utilizada para hacer el estudio no es representativa de la población de todos los varones españoles con edades comprendidas entre los 18 y 23 años, ya que únicamente se han seleccionado universitarios. Esto puede ser un problema importante, sobre todo si la estatura de los jóvenes universitarios fuese diferente a la de la población de todos los jóvenes en general.*

- *La estatura es una variable continua y cabe esperar que su distribución sea Normal. Además, como la muestra está formada por 55 individuos, el TCL garantiza que la distribución muestral de la media será aproximadamente Normal, sin embargo, sería interesante realizar un análisis descriptivo de los datos de la muestra para detectar posibles datos anómalos y observar las regularidades que estos datos presentan.*

- *El enunciado indica que la desviación típica de la estatura para la población de jóvenes españoles con edades comprendidas entre los 18 y los 23 años es de* 8,5 *cm, pero este valor no se justifica. ¿Cómo se ha obtenido este dato?*

1.1.3 Gasto medio en las rebajas

Enunciado pág. 289

En las noticias de una cadena de TV se informa de que el gasto medio de los consumidores españoles en rebajas durante este año será inferior a 70 euros.

Para corroborar esta información se utiliza una muestra aleatoria de 400 consumidores, y se observa que la cantidad media que éstos han gastado en rebajas es de 67 euros. Se asume que la desviación típica del gasto en rebajas es de $\sigma = 20$ euros (similar a la de años anteriores).

a) Plantear una prueba de hipótesis para decidir si los datos de la muestra son coherentes con la información proporcionada por la cadena de TV.

Dibujar la distribución adecuada y marcar el área de aceptación y de rechazo, el valor del estadístico y su *valor p* asociado. Interpretar el resultado en el contexto del problema.

$X =$ *Gasto en rebajas de cada consumidor*

Información de la muestra: $n = 400$, $\overline{x} = 67$ *euros*

$$\begin{aligned} H_0 &: \mu = 70 \text{ } euros \\ H_1 &: \mu < 70 \text{ } euros \end{aligned}$$

Si H_0 *fuese cierta:*

$$\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

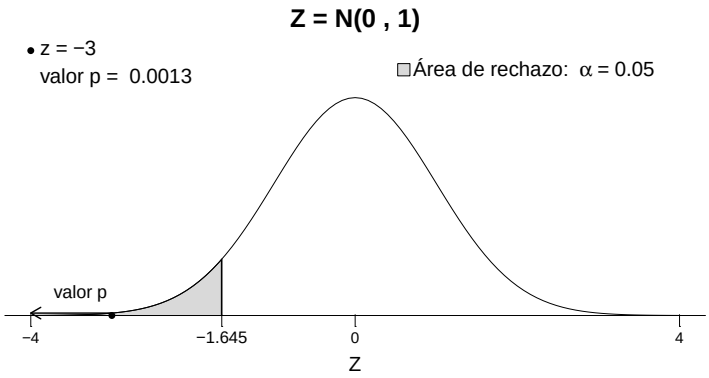


Figura 1.5: Zonas de aceptación y de rechazo, valor del estadístico z y *valor p*.

Valor del estadístico:
$$z = \frac{67 - 70}{20/\sqrt{400}} = -3$$

Valor crítico: $z^{0,05} = -1{,}645$.

$Valor\ p = P(Z < -3) = 0{,}0014$.

Trabajando con un nivel de significación del 5 %, debemos concluir que la información proporcionada por la cadena de TV sí es coherente con la información que se desprende de la muestra de 400 consumidores seleccionada: el gasto medio de los consumidores españoles en rebajas durante este año será inferior a 70 euros.

pnorm(-3, lower.tail=T)
0.001349898

Interpretación del valor p: si el consumo medio de los consumidores españoles, en el periodo de rebajas, fuese igual o superior a 70 euros, la probabilidad de obtener, simplemente por azar, una muestra de 400 consumidores como la seleccionada, sería de 0,0013. *Esta probabilidad tan baja, indica que tenemos mucha evidencia en contra de la hipótesis nula planteada. Por tanto, debemos concluir que el gasto medio no es de 70 euros, sino inferior.*

b) Obtener la potencia del test anterior si el gasto medio por consumidor en rebajas fuese de $\mu = 65$ euros. Define qué es la potencia de un test e interpreta el resultado obtenido en el contexto del problema.

La potencia de un test se define como: $1 - \beta$, *y será la probabilidad que tiene el test de rechazar, de forma correcta, una hipótesis nula que es falsa.*

Para obtenerla, calculamos en primer lugar el valor de β*:*

- *1. Zona de aceptación del test anterior antes de tipificar:*

$$\frac{\overline{x} - 70}{20/\sqrt{400}} = -1{,}645 \rightarrow \overline{x} = 68{,}355$$

Zona de aceptación: valores de $\overline{X} > 68{,}355$

- *2. Cálculo de* β*:*

$$\beta = P(\overline{X} > 68{,}355/\mu = 65) =$$

$$= P(N(0,1) > \frac{68{,}355 - 65}{20/\sqrt{400}}) = P(Z > 3{,}36) = 0{,}0004$$

```
pnorm(3.36, lower.tail=T)
```
0.9996103

La potencia del test anterior, cuando $\mu = 65$ *euros, será de* $1 - 0{,}0004 = 0{,}9996 = 99{,}96\,\%$.

Interpretación: la probabilidad que tiene el test de hipótesis planteado en el apartado a) de rechazar, de forma correcta, que el gasto medio en rebajas ha sido igual o superior a 70 euros, cuando en realidad este gasto era tan solo de 65 euros, es del 99.96 %.

c) Justificar si con la muestra de 400 consumidores seleccionada puede estimarse, con un nivel de confianza del 95 %, el gasto medio de los consumidores españoles en rebajas con un margen de error de ± 1 euros.

Como nos proporcionan el valor de σ*, un intervalo de confianza para el gasto medio de los españoles en rebajas, se obtendrá mediante la siguiente fórmula:*

$$\overline{x} \pm z^{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Trabajando con un α *del 5 %, el margen de error será:*

$$\pm 1{,}96 \frac{20}{\sqrt{400}} = \pm 1{,}96$$

Por lo tanto, no es posible estimar con un nivel de confianza del 95 %, el gasto medio en rebajas con un margen de error de tan solo ± 1 *euros utilizando una muestra de 400 consumidores. Para conseguir este objetivo deberíamos incrementar el tamaño de la muestra.*

1.2 Procedimientos t

Los procedimientos t, en los problemas de inferencia sobre la media μ de una población, encajan en el esquema presentado en la Figura 1.6, y son los siguientes:

- Estadístico utilizado para realizar un contraste de hipótesis:

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

- Intervalo con un nivel de confianza del $(1-\alpha)\cdot 100\,\%$ para la media μ de la población:

$$\overline{x} \pm t_{n-1}^{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

donde:

$\overline{x}$ es la media de la muestra

s es la desviación típica de la muestra

n es el tamaño de la muestra

$t_{n-1}^{\alpha/2}$ es el valor crítico que buscaremos en las tablas de las distribuciones t (pag. 303 – 304)

μ_0 es el valor considerado para la media de la población en la hipótesis nula H_0

El esquema de la Figura 1.6 es similar al esquema que describía los procedimientos Z, pero estos procedimientos no asumen un valor conocido para la desviación típica de la población, σ.

El estadístico utilizado es:

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

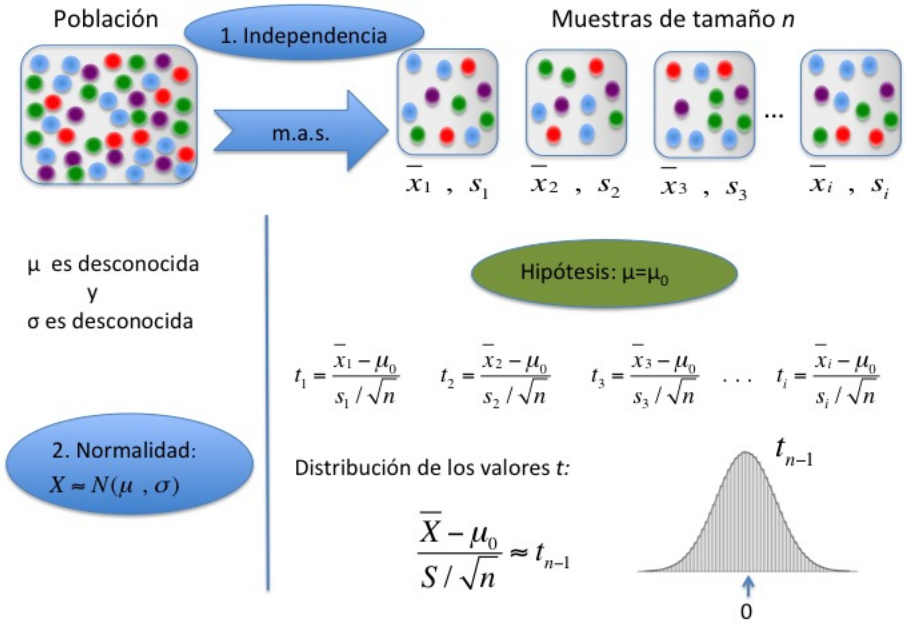


Figura 1.6: Procedimientos t para medias en problemas de una muestra.

Este estadístico, calculado a partir de la información que se desprende de las muestras de tamaño n que pueden extraerse de la población de referencia, y bajo ciertas condiciones, sigue un modelo de distribución t con $n-1$ grados de libertad.

Para poder aplicar los procedimientos t en problemas de una muestra, la teoría estadística exige que se cumplan las siguientes condiciones (elipses azules en la Figura 1.6).

1. Independencia. La muestra debe ser elegida utilizando un muestreo aleatorio simple (m.a.s.). Esta forma de selección garantiza que los elementos o individuos de la muestra han sido seleccionados de forma independiente.

   Asumimos, además, que el tamaño de la población estudiada es infinito.

2. Normalidad. Asumimos que la variable estudiada en la población X sigue un modelo de distribución Normal, con una media μ y una desviación típica σ que son desconocidas. Es decir, se asume que $X \sim N(\mu\ ,\ \sigma)$.

¿Son fiables estos procedimientos en la práctica?

Las condiciones anteriores no se cumplen nunca en la práctica, y por ello, como sucedía con los procedimientos Z, es más interesante preguntarse si los métodos descritos siguen siendo fiables aunque las condiciones exigidas no se cumplan de manera estricta.

Estas son las condiciones que debes verificar para utilizar de forma segura estos procedimientos en la práctica:

1. **Independencia**. La muestra debe haber sido seleccionada utilizando un m.a.s. La utilización de otros tipos de muestreo probabilístico (estratificado, sistemático, por *clusters*) exige cambios en la fórmula utilizada para obtener el estadístico utilizado en el contraste, aunque las fórmulas vistas pueden ser una buena aproximación.

   La selección aleatoria de las unidades que forman la muestra garantiza que éstas no están relacionadas y que las observaciones son independientes.

2. **Normalidad**. La comprobación de este supuesto exige la realización de un análisis descriptivo de los datos de la muestra, para observar las regularidades de los mismos y detectar la presencia de datos anómalos.

Dependiendo del tamaño de la muestra:

- Muestras pequeñas ($n < 40$): utiliza los procedimientos t sólo si no existen datos anómalos que tengan un peso excesivo en el cálculo de la media $\overline{x}$, y cuando la distribución tenga un sólo pico y no presente una marcada asimetría.

  Obtener un gráfico de la función de densidad empírica, un gráfico de cuantiles y un diagrama de caja es útil para comprobar estos aspectos.

- Muestras de más de 40 elementos: puedes utilizar los procedimientos t con seguridad, incluso cuando la distribución de los datos sea asimétrica y existan datos anómalos.

  El *Teorema Central del Límite* (TCL) garantiza que, cuando el tamaño de las muestras seleccionadas n es elevado, la distribución de la media $\overline{x}$ de las muestras es Normal, incluso cuando la distribución de la variable estudiada X presente asimetrías importantes.

De nuevo, es el *valor p*, obtenido al realizar un contraste de hipótesis, el resultado clave que nos permite tomar una decisión respecto al grado de evidencia que proporcionan los datos de la muestra en contra de la hipótesis nula planteada y en favor de la alternativa.

Lo único que cambia es la forma de calcularlo (ver Tabla 1.2), pero su interpretación sigue siendo la misma.

Pruebas de una cola		Test de 2 colas
$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$
$H_1 : \mu > \mu_0$	$H_1 : \mu < \mu_0$	$H_1 : \mu \neq \mu_0$
$valor\ p = P(t_{n-1} > t)$	$valor\ p = P(t_{n-1} < t)$	$valor\ p = 2 \cdot P(t_{n-1} > \lvert t \rvert)$

Tabla 1.2: Cálculo del *valor p* utilizando los procedimientos t en problemas de una muestra

1.2.1 Un nuevo diseño de página web

Enunciado pág. 242

Una empresa vende sus productos por Internet y cree que un cambio en el diseño de su página web, aumentando el número y la calidad de las fotos de los productos que exhibe, conseguirá incrementar sus ventas.

El volumen medio de ventas diarias, utilizando el diseño de página actual, es de 9300€.

Antes de tomar una decisión, decide hacer una prueba durante un mes (30 días) presentando sus productos con el nuevo diseño de página web.

El volumen medio de ventas de esos 30 días, en los que se utiliza el nuevo diseño de página, ha sido de 9700€, y la desviación típica de las cifras de ventas ha sido de 1570€.

a) Identifica y define en este contexto, la muestra, la población y la variable implicada.

La muestra: los 30 días en que se observan las ventas con el nuevo diseño.

La población: todos los días en que se puedan realizar ventas con el nuevo diseño.

La variable aleatoria: $X =$ Volumen diario de ventas (en €).

b) ¿Existe evidencia significativa a un nivel del 5 % de que el nuevo diseño de página incrementa las ventas?

Plantea una prueba de hipótesis para responder a la pregunta y realiza un dibujo marcando las zonas de aceptación y de rechazo, el valor crítico, el valor del estadístico y el *valor p* del test.

Información de la muestra: $n = 30, \quad \overline{x} = 9700€, \quad s = 1570€$

$\mu =$ *Volumen medio de las ventas diarias con el nuevo diseño.*

$$\begin{aligned} H_0 &: \mu = 9300 \text{ euros} \\ H_1 &: \mu > 9300 \text{ euros} \end{aligned}$$

Si H_0 fuese cierta:

$$\frac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim t_{29}$$

Valor del estadístico:

$$t = \frac{9700 - 9300}{1570/\sqrt{30}} = 1{,}40$$

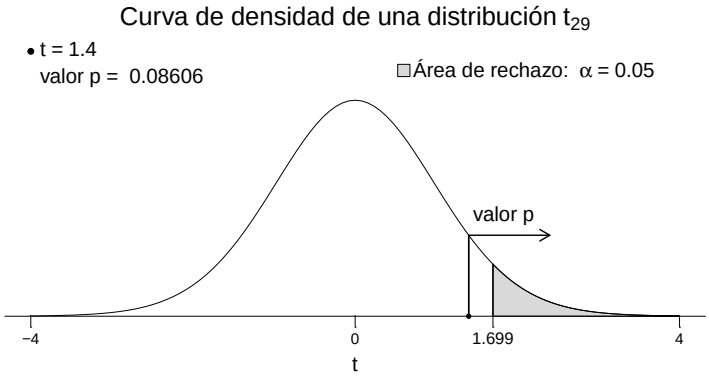


Figura 1.7: Zonas de aceptación y de rechazo, valor del estadístico t y *valor p*.

Valor crítico (en tablas): $t_{29}^{0,05} = 1{,}699$

$Valor\ p = P(t_{29} > 1{,}4) \quad \rightarrow \quad 0{,}05 < valor\ p < 0{,}1$

pt(1.4, df=29, lower.tail=F)
0.08606059

c) Explica, utilizando un lenguaje sencillo que entienda alguien que no sabe estadística, que es el *valor p* de este test. Da una interpretación en el contexto del problema.

Si el nuevo diseño de página web no incrementase las ventas medias diarias (éstas fuesen de 9300€ o inferiores), la probabilidad de obtener, simplemente por azar, unos resultados como los observados en nuestra mues-

tra o más extremos, es decir, con un volumen medio de ventas igual o superior a 9700€, sería igual a 0,086*.*

Los datos del estudio no muestran suficiente evidencia en contra de H_0*. Los resultados no son significativos para un nivel de significación del 5 %.*

c) Calcula la potencia del test utilizado en el apartado *a)* para detectar un incremento medio de las ventas por día de 500€ e interpreta el resultado en el contexto del problema.

- *Zona de aceptación en función de los valores de* $\overline{X}$*:*

  Aceptamos H_0 *si* $t_{29} < 1{,}699$*.*

  Si "destipificamos"*:*

$$\frac{\overline{x} - 9300}{1570/\sqrt{30}} = 1{,}699$$

$$\overline{x} = 1{,}699\frac{1570}{\sqrt{30}} + 9300 = 9787$$

  Aceptamos H_0 *si* $\overline{X} < 9787$€

- *Calculamos el valor de* β *si la media de la población fuese de 9300 + 500 = 9800* €

$$\beta = P(\overline{X} < 9787/\mu = 9800) = P\left(t_{29} < \frac{9787 - 9800}{1570/\sqrt{30}}\right)$$

$$\beta = P(t_{29} < -0{,}05) \approx 0{,}5$$

pt(-0.05, df=29, lower.tail=T)

0.4802326

La potencia del test, para detectar un incremento medio de las ventas de 500€, es aproximadamente igual a 0,5*. Esto significa que, si el nuevo diseño de página incrementase las ventas medias diarias en*

una cantidad igual a 500€, la probabilidad de que el test propuesto detecte esta situación (rechazando correctamente la $H_0 : \mu = 9300€$) sería igual a 0.5. El test es, por lo tanto, poco potente para detectar este incremento en las ventas.

d) Un responsable de ventas que acaba de seguir un curso de Métodos Estadísticos afirma que el diseño utilizado para obtener los datos no es adecuado. Explica cómo podría realizarse un estudio mejor que el que se ha realizado.

El estudio se ha realizado comparando datos históricos de ventas con las ventas obtenidas, durante 30 días, utilizando el nuevo diseño de página web.

Además del cambio en el diseño de página web, existen múltiples factores que pueden tener un efecto sobre las ventas: ofertas de productos, ingresos puntuales de los compradores, etc., que no están controlados.

Sería más interesante realizar un diseño de experimentos para poder determinar la existencia de relaciones del tipo causa-efecto*.*

Un estudio más eficiente sería aquel que divida (al azar) a los compradores en dos grupos: unos accederán a la página web con el diseño clásico, y otros accederán a la página web con el nuevo diseño. Tras una serie de días, se podrán comparar las ventas medias obtenidas en los dos grupos.

1.2.2 ¿Mejora la calidad del agua?

Enunciado pág. 287

El personal responsable de medio ambiente de la Generalitat Valenciana ha implantado una serie de medidas para mejorar la calidad del agua del río Xúquer.

Para realizar el estudio, se tomaron muestras de agua de 41 puntos del cauce del río que fueron seleccionados de forma aleatoria. En cada uno de estos puntos se midió la cantidad de oxígeno disuelto en el agua (partes por millón).

Un mes después de implantar las medidas propuestas, se volvieron a realizar mediciones en los mimos puntos del río.

El incremento medio de la cantidad de oxígeno disuelto en el agua en las muestras tomadas fue de +4,2 %, y la desviación típica de estos datos fue de un 13 %.

A la vista de estos resultados, ¿podemos afirmar que las medidas tomadas por el personal responsable de medio ambiente han tenido éxito? ¿Han conseguido aumentar la cantidad de oxígeno disuelto en el cauce del río Xúquer?

a) Identifica y define en este contexto, la muestra, la población y la variable implicada.

La muestra: los 41 puntos del cauce del río donde se han tomado las muestras de agua.

La población: todos los puntos del cauce del río donde puedan tomarse muestras de agua, es decir, la totalidad del río.

La variable aleatoria: X = Incremento de la cantidad de oxígeno disuelto en el agua en cada muestra (en %).

b) Plantea una prueba de hipótesis para determinar si las medidas tomadas han conseguido incrementar el contenido de oxígeno disuelto en el río. Realiza un dibujo marcando las zonas de aceptación y de rechazo, el valor crítico, el valor del estadístico y el *valor p* del test.

Información de la muestra: $n = 41, \quad \overline{x} = 4{,}2\,\%, \quad s = 13\,\%$

μ = Incremento medio de la cantidad de oxígeno disuelto en el agua del cauce del río Xúquer, un mes después de implementar las medidas propuestas.

$$\begin{aligned} H_0 &: \mu = 0\,\% \\ H_1 &: \mu > 0\,\% \end{aligned}$$

Si H_0 fuese cierta:

$$\frac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim t_{40}$$

Valor del estadístico:

$$t = \frac{4{,}2 - 0}{13/\sqrt{41}} = 2{,}07$$

Valor crítico (en tablas): $t_{40}^{0{,}05} = 1{,}684$

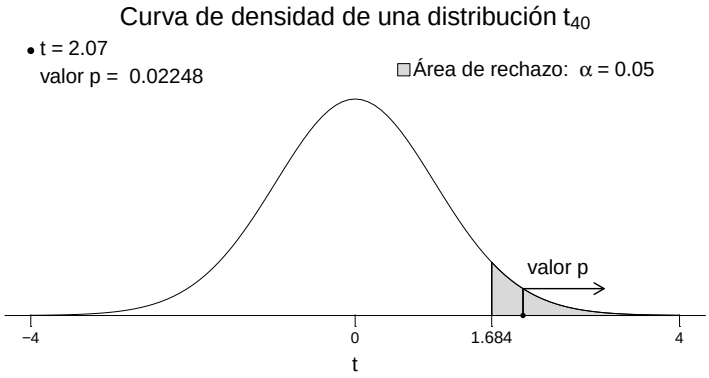


Figura 1.8: Zonas de aceptación y de rechazo, valor del estadístico t y *valor p*.

$Valor\ p = P(t_{40} > 2{,}07) \approx 0{,}025$

pt(2.07, df=40, lower.tail=F)
0.02247595

c) Explica, utilizando un lenguaje sencillo que entienda alguien que no sabe estadística, que es el *valor p* de este test. Da una interpretación en el contexto del problema.

Si las medidas implantadas no hubiesen incrementado el contenido medio de oxígeno disuelto en el cauce del río, la probabilidad de obtener, simplemente por azar, unos resultados como los observados en las muestras tomadas o más extremos, sería igual a 0,0225.

Los resultados del estudio son significativos (*valor p* = 0,02250, obtenido mediante un test t de una cola). El estudio realizado permite concluir que las medidas implantadas para incrementar el contenido de oxígeno disuelto en el cauce del río Xúquer sí han cumplido su objetivo.

c) Obtener un intervalo con una confianza del 95 % para el incremento medio de la cantidad de oxígeno disuelto en el agua del cauce del río Xúquer, un mes después de implantar las medidas propuestas.

$$\overline{x} \pm t_{40}^{0,025} \frac{s}{\sqrt{n}}$$

$$4{,}2 \pm 2{,}021 \frac{13}{\sqrt{41}} = 4{,}2 \pm 4{,}1 = [0{,}1\ ,\ 8{,}3]\,\%$$

Tenemos una confianza del 95 % de que el incremento medio de la cantidad de oxígeno disuelto en el agua del cauce del río Xúquer, un mes después de implantar las medidas propuestas, es un valor comprendido entre 0,1 % *y* 8,3 %.

El incremento medio de la cantidad de oxígeno disuelto en el agua del cauce del río Xúquer, un mes después de implantar las medidas propuestas, es del 4.2 % (margen de error de ±4,1 %, para una confianza del 95 %).

d) Comenta los resultados obtenidos en los apartados anteriores y redacta un breve párrafo con las conclusiones que consideres pertinentes.

Este estudio pone de manifiesto la importancia de diferenciar entre significación estadística *y* significación práctica.

Hemos visto que los resultados del estudio son significativos para un nivel de significación del 5 % (valor $p = 0{,}0225$*). Esto significa que es poco probable obtener unos resultados como los observados, simplemente por azar, si no se hubiese producido un incremento en el contenido de oxígeno del río tras adoptar las medidas propuestas.*

Es importante aportar, además, un intervalo de confianza para el parámetro de interés, ya que este intervalo nos permite medir la magnitud del efecto buscado. En este estudio, el margen de error del intervalo obtenido es muy elevado (±4,1 %)*. Existe mucha incertidumbre a la hora de estimar el incremento medio de la cantidad de oxígeno disuelto en el agua que se ha conseguido.*

Sería interesante reducir esta incertidumbre y ello puede conseguirse incrementando la cantidad de muestras tomadas, n.

Otro dato que llama la atención es el valor tan elevado observado para la desviación típica de las muestras ($s = 13\,\%$)*. ¿Es razonable este dato?*

Un valor tan alto puede indicar que hay puntos del río en los que no se produce un incremento de la cantidad de oxígeno disuelta, sino todo lo contrario, y es por esto por lo que la desviación típica toma un valor tan elevado. ¿Tiene sentido observar valores negativos? ¿Es posible que las medidas adoptadas no mejoren la calidad del agua en algunos puntos del cauce, e incluso que la empeoren?

1.2.3 Emisión de arsénico

Enunciado pág. 258

El arsénico es una de las 10 sustancias químicas que la OMS considera más preocupantes para la salud pública, y establece que el límite recomendado para la concentración de arsénico en el agua potable es de 10 $\mu g/l$.

Una empresa de procesamiento de vidrio necesita aprovechar el agua de un río cercano en su proceso de fabricación. Esta empresa realiza vertidos periódicos al río tras la limpieza de su maquinaria.

¿Suponen estos vertidos un problema para la salud por superar la concentración de arsénico recomendada por la OMS?

Se tomaron muestras de 121 vertidos realizados por la empresa a lo largo de tres meses.

La información que se desprende de los datos obtenidos es:

$$\overline{x} = 12\ \mu g/l \qquad s = 8\ \mu g/l$$

a) Realizar una prueba de hipótesis para determinar si estaría justificado emprender acciones legales contra la empresa por no cumplir las recomendaciones de la OMS. Justificar la respuesta.

Variable aleatoria implicada: X = concentración de arsénico en cada vertido de la empresa ($\mu g/l$).

Parámetro: μ = concentración media de arsénico en los vertidos que realiza la empresa.

$\overline{x}$ = concentración media de arsénico en las muestras tomadas en 121 vertidos de la empresa.

Prueba de hipótesis:

$$\begin{aligned} H_0 &: \mu = 10\ \mu g/l \\ H_1 &: \mu > 10\ \mu g/l \end{aligned}$$

Si H_0 fuese cierta:

$$\frac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim t_{120}$$

Valor del estadístico:

$$t = \frac{12 - 10}{8/\sqrt{121}} = 2{,}75$$

Valor crítico (en tablas): $t_{120}^{0,05} = 1{,}658$

$Valor\ p = P(t_{120} > 2{,}75) \quad \rightarrow \quad 0{,}001 < valor\ p < 0{,}005$

pt(2.75, df=120, lower.tail=F)
0.003441163

Un valor p tan pequeño proporciona mucha evidencia en contra de H_0 y a favor de H_1. Rechazamos H_0. El estudio realizado indica que la empresa está superando el límite que marca la OMS: en los vertidos que realiza, la concentración media de arsénico es superior a 10 $\mu g/l$. Por lo tanto, sí que estaría justificado emprender acciones legales contra la misma.

b) Si las emisiones de este contaminante superan la cantidad de 15 $\mu g/l$, se producen daños irreversibles a la fauna y flora marinas que las reciben. Obtener la *potencia* del test anterior para detectar esta situación. Explica las implicaciones prácticas del resultado obtenido.

- *Zona de aceptación en función de los valores de $\overline{X}$:*

  Aceptamos H_0 si $t_{120} < 1{,}658$.

$$\frac{\overline{x} - 10}{8/\sqrt{121}} = 1{,}658$$

$$\overline{x} = 1{,}658\frac{8}{\sqrt{121}} + 10 = 11{,}225$$

Aceptamos H_0 *si* $\overline{X} < 11{,}225$ $\mu g/l$

- *Calculamos el valor de* β *si la media de la población fuese de 15* $\mu g/l$

$$\beta = P(\overline{X} < 11{,}225/\mu = 15) = P\left(t_{120} < \frac{11{,}225 - 15}{8/\sqrt{121}}\right)$$

$$\beta = P(t_{120} < -5{,}19) \approx 0$$

La potencia del test es prácticamente igual a 1.

pt(-5.19, df=120, lower.tail=T)

4.338474e-07

power.t.test(n=121, delta=15-10, sd=8,
sig.level=0.05, type="one.sample",
alternative="one.sided")

One-sample t test power calculation: power = 0.9999999

La potencia que tiene el test planteado para detectar emisiones donde el contenido medio de arsénico sea superior a 15 $\mu g/l$ *es prácticamente igual a 1.*

Si la empresa realizase este tipo de emisiones tan perjudiciales para el medio ambiente, el test propuesto lo detectaría prácticamente en el 100 % de los casos.

1.2.4 Leche de cabra

Enunciado pág. 288

El gerente de un supermercado desea estimar la demanda diaria de leche de cabra (paquetes/día) que se venden en su establecimiento.

Los datos siguientes son el número de paquetes de un litro vendidos a lo largo de 15 días:

68, 59, 58, 65, 65, 51, 71, 62, 59, 75, 60, 61, 51, 51, 64

a) ¿Podemos aplicar con seguridad los procedimientos t de una muestra a este conjunto de datos? Justifica tu respuesta.

Las condiciones que deben cumplirse para poder aplicar estos procedimientos con seguridad son:

1. ***Independencia.*** *La muestra debe haberse seleccionado siguiendo alguna técnica de muestreo probabilístico para garantizar la independencia de los datos y la representatividad de la misma.*

   El enunciado no indica cómo se han seleccionado los días en que se han observado las ventas. Asumimos, para que el estudio tenga validez, que dicha muestra puede considerarse como una m.a.s. de los días en los que el supermercado está abierto al público.

2. ***Normalidad.*** *La variable estudiada es X = número de paquetes de leche de un litro vendidos diariamente en el supermercado. Esta variable es discreta y por ello, no puede seguir un modelo de distribución Normal. A pesar de esto, los procedimientos t de una muestra podrán aplicarse a estos datos si la distribución de los mismos presenta un sólo pico, no existen datos anómalos y su distribución no presenta fuertes asimetrías.*

   Podemos comprobar estas condiciones realizando un análisis descriptivo de los datos de la muestra, obteniendo un diagrama de caja, un gráfico con la función de densidad empírica y un gráfico de cuantiles (Figura 1.9).

   El diagrama de caja muestra una distribución simétrica y sin puntos anómalos.

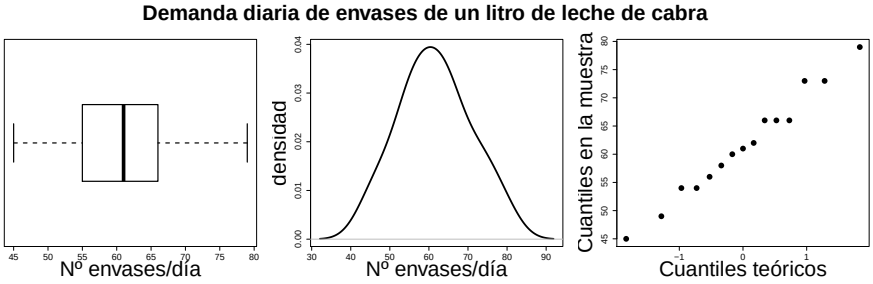


Figura 1.9: Diagrama de caja, función de densidad empírica y gráfico de cuantiles.

La función de densidad empírica muestra una distribución con un solo pico y con forma de campana, muy similar a la curva de densidad de un modelo de distribución Normal.

Al representar los datos en un gráfico de cuantiles, éstos se posicionan formando (más o menos) una línea recta en la diagonal del gráfico, y esto confirma que el modelo de la distribución Normal es adecuado para estudiar este conjunto de datos. (Al utilizar un gráfico de cuantiles, resulta difícil comprobar si éstos siguen un modelo de distribución Normal cuando se trabaja con un conjunto tan reducido de datos. Si trabajamos con muestras pequeñas debemos ser flexibles a la hora de valorar si los puntos se posicionan formando una línea recta en la diagonal del gráfico).

Por lo tanto, si asumimos que la muestra seleccionada puede considerarse una m.a.s. de la población de referencia, sí es seguro aplicar los procedimientos t de una muestra a este conjunto de datos.

b) Obtener un intervalo de confianza del 90 % para el número medio de envases de un litro de leche de cabra que se venden diariamente e interpretar su significado.

$X =$ *número de envases de un litro de leche de cabra vendidos diariamente*

Información de la muestra:

$n = 15 \quad \overline{x} = 61{,}47$ *envases* $\quad s = 9{,}33$ *envases*

Intervalo de confianza para μ del 90 %:

$$\overline{x} \pm t_{14}^{0,05} \frac{s}{\sqrt{n}}$$

$$61{,}47 \pm 1{,}761 \frac{9{,}33}{\sqrt{15}} = 61{,}47 \pm 4{,}24 = [57{,}23\ ,\ 65{,}71]\ \textit{envases}$$

```
ventas=c(60, 62, 45, 54, 79, ..., 54, 66, 66, 58, 73, 73)
t.test(ventas, conf.level=0.9)$conf.int
57.22533  65.70800
```

Interpretación: con una confianza del 90 % afirmamos que el valor real de μ (número medio de envases de leche de cabra de un litro vendidos diariamente) se encuentra entre las 57.23 y las 65.71 unidades.

El número medio de envases de leche de cabra de un litro vendidos diariamente en el supermercado es de 61,47 envases (margen de error de $\pm$4,24 envases, para una confianza del 90 %).

c) Si se desea estimar el número de unidades diarias que se venden, de forma que, en un 95 % de los casos, el error cometido no supere las 2 unidades, cual es el tamaño mínimo de la muestra que es necesario seleccionar? (Nota: Asumir un valor de $\sigma = 10$ unidades).

Deseamos obtener un intervalo de confianza del 95 % para μ de forma que el margen de error sea inferior a $\pm$2 unidades.

Como asumimos que σ es conocida, utilizamos un procedimiento Z:

$$\overline{x} \pm z^{0,025} \frac{\sigma}{\sqrt{n}}$$

$$\textit{margen de error} = z^{0,025} \frac{\sigma}{\sqrt{n}}$$

$$1{,}96 \frac{10}{\sqrt{n}} < 2$$

Para seguir leyendo haga click aquí