Probabilidad e inferencia estadística

Indice

Probabilidad e inferencia estadística

I. Introducción

El análisis y la inferencia estadística son parte imprescindible de la comprensión de los datos. Aportan herramientas consistentes y un marco teórico que conforma el esqueleto del análisis de información.

Para hacer un análisis más riguroso, se realizan asunciones sobre la naturaleza de los datos, verificando que son una aproximación razonable y aplicándose técnicas específicas del tipo de dato asumido.

II. Objetivos específicos

- Entender lo que es la probabilidad e intuir sus aplicaciones.
- Comprender las distribuciones de probabilidad básicas y sus entornos de uso.
- Formarse una idea sobre la inferencia estadística y cómo se utiliza en análisis de datos para dar estimaciones y tomar decisiones.
- Aplicar en R estos conceptos.

III. Probabilidad

La probabilidad es una abstracción matemática que se adjudica a un suceso. Se representa como un número entre 0 y 1 e indica la propensión a la ocurrencia de dicho suceso.

Así, por ejemplo, si tenemos 3 bolas negras en una bolsa con 7 bolas blancas más, la probabilidad de obtener una bola negra al extraer una al azar de la bolsa sería 3 / 10 = 0.3. Si el suceso definido es N, el valor de probabilidad será:

- P(N) = 0.3

Este valor de propensión al suceso o probabilidad señala con qué frecuencia se dará el suceso si repetimos el experimento indefinidas veces.

Cuando se habla de probabilidad, previamente ha de definirse un *espacio muestral*, que es el conjunto de todos los posibles sucesos que pueden darse en un experimento. En el caso anterior de la bolsa de bolas, el espacio muestral es sacar una bola blanca o sacar una bola negra.

Existe toda una axiomática de las reglas de la probabilidad y sus propiedades que no se expondrán aquí, ya que previamente es necesaria la comprensión del concepto.

La probabilidad y la estadística tratan el estudio de procesos aleatorios de maneras diferentes:

Características de la probabilidad

- Lógica autocontenida.
- Reglas para calcular probabilidades.
- Una sola respuesta correcta.

Características de la estadística

- Imprecisa, más parecida al arte.
- Trabaja con datos experimentales e intenta obtener conclusiones probabilísticas.
- No existe la respuesta correcta.

IV. Distribuciones de probabilidad

Una distribución de probabilidad (o variable aleatoria) es una ley según la cual se pueden generar sucesos numéricos.

Por ejemplo, cada vez que tiramos un dado equilibrado se produce como resultado un número, ese número se obtiene en cada tirada de acuerdo con una misma regla que tiene que ver con las propiedades físicas del dado y de la tirada —suponiendo que puedan darse las mismas circunstancias al reproducir el experimento—. En este caso, los sucesos posibles o espacio muestral son los números de 1 a 6 y sus probabilidades —si el dado está equilibrado— valen 1/6 para cada valor numérico posible.

Hay dos tipos de variables aleatorias básicas, según el tipo de números que toma:

Discreta

Los valores son aislados. Ejemplos: tirada de un dado, número de conexiones de un router, número de veces en que una moneda cae de cara al lanzarla 10 veces, un paciente tiene una enfermedad o no, etc.

Continua

Los valores pertenecen a un intervalo numérico, esto es, conforman un continuo numérico en un rango de valores posibles. Ejemplos: altura de una persona, tiempo que transcurre hasta que una máquina falla, dinero que gana un agente automático de *trading*, precio de la electricidad, etc.

Anotación: Leyes de distribución que incluye R

Algunas de las leyes de distribución que incluye R son las siguientes:

Función	Distribución
Beta	Beta
Binom	Binomial
Exp	Exponencial
Chisq	Chi-cuadrado
Gamma	Gamma
Fisher	F
Lnorm	Lognormal
Pois	Poisson
T	T-student
Unif	Uniforme
Norm	Normal

Tabla 3.1. Leyes de distribución que incluye R. *Fuente:* elaboración propia.

Cada distribución de probabilidad tiene distintas formulaciones y sirve para modelar un tipo distinto de fenómeno.

Distribuciones de probabilidad discretas

Se estudiarán las siguientes funciones de probabilidad discretas:

- **Bernouilli**: realizar un experimento que tiene como conclusión las opciones éxito o fracaso, de modo que la probabilidad o propensión de éxito siempre será la misma. Ejemplo: tirar una moneda y considerar éxito si cae de cara y fracaso si sale cruz. Si la moneda es equilibrada, la probabilidad de éxito es 0.5.
- **Binomial:** realizar una cantidad fija de experimentos de Bernouilli y contar el número de éxitos que se obtienen. Ejemplo: se lanza un dado 20 veces y se cuenta el número de ocasiones en las que el resultado es superior a 4.
- **Uniforme discreta:** el experimento puede tomar valores en un conjunto finito o limitado, siendo la probabilidad de cada suceso la misma. Ejemplo: al lanzar un dado equilibrado, se pueden obtener números de 1 a 6 siendo cada cara igual de probable.
- **Poisson:** modelan eventos que tienen tasas de frecuencia regulares, indican el número de sucesos en un intervalo temporal. Ejemplo: número de visitas a una página web en un día, número de coches que pasan por un tramo de carretera en una hora.

Distribuciones de probabilidad continua

Se expondrán las siguientes distribuciones de probabilidad continua:

- **Exponencial:** modelan el tiempo entre eventos que se generan en un proceso de Poisson (como los casos vistos antes). También se aplica cuando se trata de un experimento que toma cualquier valor positivo. Ejemplos: tiempo hasta que se realiza una visita en una página web, tiempo hasta que pasa un coche por la carretera, vida útil de una bombilla.
- **Uniforme continua:** el experimento puede tomar valores en un rango acotado: tiene un extremo superior e inferior. Es igual de probable cada uno de los posibles valores. Ejemplo: la distancia de un dardo al centro de una diana si suponemos que el dardo cae con la misma probabilidad en todos los puntos —esto no es del todo preciso matemáticamente, ¿por qué?—.
- **Normal:** es la distribución probabilística más importante, ya que puede agrupar a todas las demás bajo los supuestos pertinentes. Se aplica cuando es más probable que se obtengan valores numéricos en torno a un centro y valores exponencialmente menos probables al alejarse del mismo.

Cada distribución de probabilidad posee fórmulas relacionadas y funciones atribuidas. Las fundamentales son:

Función de distribución de probabilidad acumulada (cdf)

Tiene como entrada un valor numérico y como respuesta la probabilidad de que la distribución de probabilidad tomará valores iguales o menores que el valor dado. La notación es la siguiente:

F(x) = “probabilidad de que la distribución tome un valor menor o igual que x”

Ejemplo: la cdf de una tirada de dados, que es la distribución de probabilidad de “número obtenido al tirar un dado” y corresponde al tipo uniforme discreta, sería la siguiente:

- *F(1) = 1 / 6*
- *F(1.2) = 1 /6*
- *F(2) = 2 / 6*
- *F(0.3) = 0*
- *F(5.6) = 5 / 6*
- *F(7) = 1*

Las gráficas de las cdf de variables discretas dan saltos, mientras que las de las variables continuas son suaves y no tienen cortes:

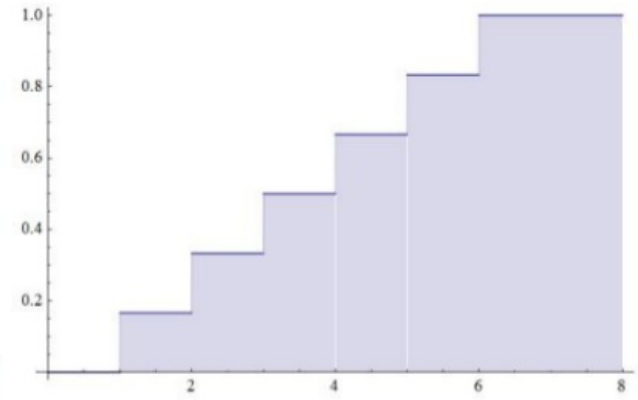

Imagen 3.1. Función de distribución. *Fuente:* elaboración propia.

Ejemplo: la cdf de la "vida útil de una bombilla" es un ejemplo de distribución exponencial, que es uno de los tipos de variable continua. Solo toma valores mayores que cero. Observa cómo para valor 0 o inferiores es 0, y conforme se mueve a la derecha se aproxima a 1:

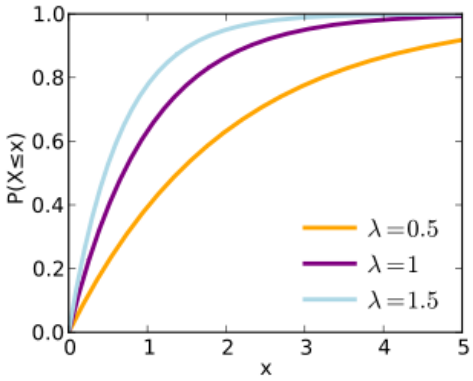

Imagen 3.2. Función de distribución.
Fuente: www.wikipedia.com

Función de masa de probabilidad o de densidad de probabilidad (pmf y pdf, respectivamente)

Las distribuciones de probabilidad poseen, además de la función de distribución de probabilidad acumulada, una de estas dos funciones descriptivas de la concentración de probabilidad puntual. Pueden ser de dos tipos:

Variable aleatoria discreta

Como concentran su probabilidad en determinados valores numéricos, tienen **función de masa de probabilidad**, la cual indica la probabilidad de tomar un valor concreto.

Ejemplo: en el caso de "resultado al tirar una moneda", la función de masa de probabilidad o pmf es:

- p(cara) = 0.5
- p(cruz) = 0.5

Variable aleatoria continua

Las variables continuas toman valores en un intervalo numérico. Esto hace que no acumulen probabilidad en ninguno de los puntos. Sin embargo, cada valor posible tiene una densidad probabilística asociada, que viene representada por la función de masa de probabilidad.

Ejemplo: en el caso de "tiempo de vida útil de una bombilla", la **función de densidad** probabilística es:

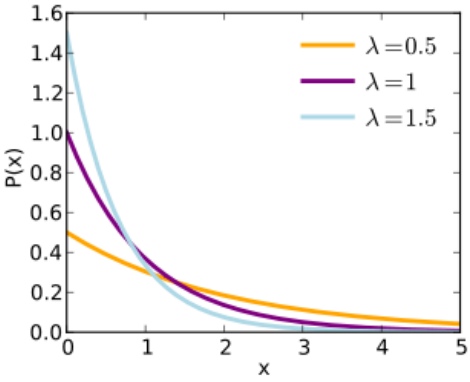

Imagen 3.3. Función de densidad de probabilidad. *Fuente:* www.wikipedia.com

Observa cómo la densidad se acumula más fuertemente cerca de 0 y decrece rápidamente. Esto refleja que es difícil tener bombillas muy longevas y que conforme se aumente el tiempo de duración, la densidad será mucho menor.

Función de cuantiles (qf)

Dado un valor numérico, indica qué porcentaje de valores deja por debajo en el ordenamiento probabilístico que genera la distribución. Es la función inversa de la cdf.

Ejemplo: en el caso de la "vida útil de la bombilla", si queremos saber cuánto tiempo de vida tiene la bombilla que dura más que el 70 % de las bombillas que menos duran, este valor nos lo da la función de cuantiles: qf(0.7) = 23, lo que significa que el 70% de bombillas duran 23 horas o menos.

A continuación, se muestra un ejemplo de la pmf y cdf de la tirada de dados (variable discreta) en una imagen:

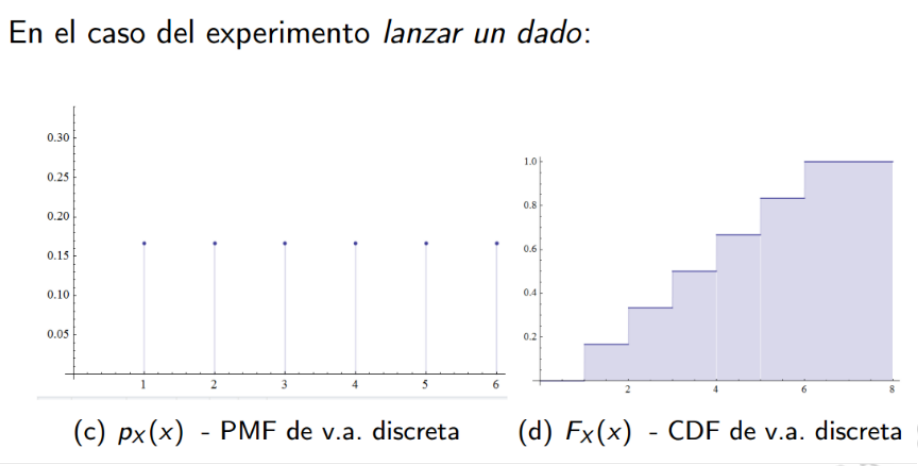

Imagen 3.4. Función de probabilidad y de distribución. *Fuente:* elaboración propia.

En el ejemplo de la pdf y cdf de una distribución de probabilidad normal, se puede apreciar que se concentran en torno a un punto céntrico y su densidad de probabilidad decae rápidamente al alejarse de ese centro:

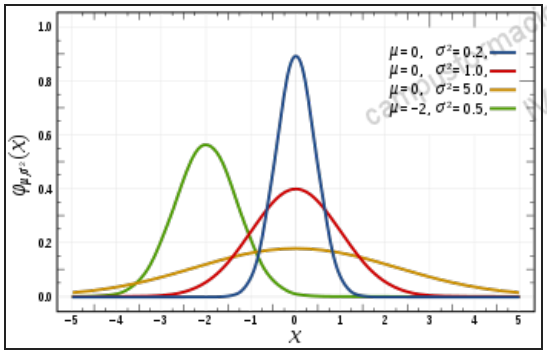

Imagen 3.5. Distribución normal. *Fuente:* www.wikipedia.com

Las distribuciones descritas dependen de una serie de **parámetros** que definen y concretan cada caso de uso:

Bernouilli

Parámetro *p*: indica la probabilidad de éxito.

Binomial

- Parámetro *p*: indica la probabilidad de éxito.
- Parámetro *n*: indica el número de experimentos.

Uniforme discreta

Parámetro *n*: indica el valor máximo que puede tomar la distribución, que tendrá como sucesos posibles 1, 2, ..., n.

Poisson

Parámetro λ: tasa de incidencia del suceso, cuyo conteo se realiza.

Exponencial

Parámetro λ: misma tasa de incidencia del suceso de conteo que genera los intervalos de tiempo que mide la exponencial.

Uniforme continua

- Parámetro *a*: extremo inferior del intervalo donde toma valores.
- Parámetro *b*: extremo superior del intervalo donde toma valores.

Normal

- Parámetro *μ*: valor central o medio que toma la distribución.
- Parámetro *σ*: dispersión respecto al valor central.

Además de las funciones descritas, las distribuciones de probabilidad tienen otras señas de identidad importantes:

Esperanza matemática (μ)

Es el concepto que formaliza la idea de valor medio de un fenómeno aleatorio.

Ejemplo: si tenemos “número de conexiones de un router en 10 minutos” como una distribución de Poisson, la esperanza matemática es el valor esperado de conexiones o valor medio, que dependerá del parámetro λ.

Varianza (σ)

Representa la dispersión cuadrática respecto al valor medio de un fenómeno aleatorio.

A continuación, un notebook de estadísticos descriptivos sobre datos y otro con el diseño de las variables aleatorias consultadas:

Descarga: Consulta el notebook UD3 N00

Descárgate el archivo UD3_N00 y ejecútalo en R. También puedes verlo en.html

Descarga: Consulta el notebook UD3 N01

Descárgate el archivo UD3_N01 y ejecútalo en R. También puedes verlo en.html

V. Inferencia estadística

La inferencia estadística se encarga del estudio de los métodos para la obtención del modelo de probabilidad que sigue una distribución de una determinada población, a través de una muestra obtenida de la misma.

El esquema seguido sería:

Población --> Muestra --> Inferencia estadística --> Población

Dos problemas fundamentales de este estudio son:

- **Intervalos de confianza:** estimar los parámetros que rigen la distribución de probabilidad asociada a la población.
- **Contraste de hipótesis:** enfrentar dos hipótesis sobre una población y obtener una conclusión basada en la información estadística que da una muestra.

Existen dos tipos de inferencia estadística según la situación:

Inferencia paramétrica

Se conoce la forma de la distribución de probabilidad de la población y solo hemos de estimar los parámetros mediante métodos ajustados a esa distribución.

Inferencia no paramétrica

No se conoce la forma de la distribución y se usan métodos genéricos.

Nuestro único método para estudiar la distribución de probabilidad de una población es a través de muestras (extracciones empíricas medidas) de la misma. Se presentan los siguientes conceptos:

Población

Es la variable o familia de variables a analizar.

Ejemplos: longitud de la pierna de una rana, tiempo de frenada de un vehículo de 100 km/h a 0 km/h, probabilidad de que un fumador enferme con EPOC, etc.

Distribución probabilística que aproxima a la población

Es el tipo de distribución de probabilidad elegido basándonos en criterios de adecuación al caso, para aproximar las variables a estudiar.

Ejemplo: respectivamente, atendiendo a los casos anteriores, distribución normal, distribución exponencial y distribución Bernouilli.

Muestra aleatoria simple

Consiste en una extracción de la población tomando las medidas de las variables a analizar, de manera que cada extracción no dependa de las anteriores y se produzca en las mismas condiciones.

Ejemplo: tomamos aleatoriamente 100 ranas de una charca.

Estadístico

Es una función que se aplica sobre la muestra y da un indicador numérico que supone información para el análisis.

Ejemplo: la media o el máximo de las longitudes de pata de esas 100 ranas son estadísticos muestrales.

5.1. Inferencia paramétrica

Cuando la población a estudiar tiene una distribución probabilística asociada conocida y depende de un parámetro, usamos los métodos de la inferencia paramétrica. Esta trata de extraer conclusiones teóricas sobre la distribución probabilística a partir de la muestra empírica.

Hay que entender que a la hora de realizar estimaciones de los valores de parámetros o decidir sobre los mismos para tomar decisiones, hay que valorar la incertidumbre que las muestras tienen, ya que no son representaciones perfectas de la población, si no que pueden tener imperfecciones y ruido.

Como bajo ciertas condiciones, las distribuciones se pueden considerar aproximadas por la Normal, se dará prioridad a los métodos basados en normalidad, que además son los más usados.

Intervalo de confianza

Un intervalo de confianza es un rango de valores en los que se estima que un parámetro de la distribución probabilística asumida se encuentre con una seguridad o confianza determinada, denotada por 1 - α.

El intervalo se calcula mediante fórmulas que están basadas en estadísticos muestrales como la media muestral o la varianza muestral.

Ejemplo: los ingresos de una web siguen una Normal(μ,σ) con media μ desconocida. Una muestra de 100 días tiene una media muestral de 30 €/día. Esto es un estimador pobre del valor de la media teórica si no se completa con un rango de valores en los que se espera que esté la media, así como una confianza. Al crear un intervalo de confianza al 95%, obtenemos los valores (28.5, 31.5), esta información es mucho más rica para evaluar los rendimientos de la página web, ya que indica que con una probabilidad de 0.95 al menos, la media teórica de la distribución probabilística se encuentra en ese intervalo.

Nos limitaremos a describir intervalos de confianza para una población normal de distintos tipos. Términos relevantes en esto son:

- **Nivel de confianza y significación:** 1 - α - probabilidad de que el parámetro esté en el intervalo de confianza. α - probabilidad de que el parámetro no esté en el intervalo de confianza.
- **Parámetro sobre el que creamos el intervalo.** Generalmente μ o σ, aunque también lo haremos sobre proporciones de Bernouilli p.

No solo se pueden crear intervalos de confianza para una sola población, también se pueden hacer intervalos de confianza para la diferencia de medias de dos poblaciones o para el cociente de la varianza de las mismas. Esto sirve para comparar dos grupos.

Hay que estimar, con una cierta seguridad probabilística, la diferencia de las medias de altura entre la población A y la población B. Para ello, se puede crear un intervalo de confianza de esta diferencia con el que se pueden tener cotas superior e inferior de esta diferencia con una confianza determinada que puede servir para especular sobre resultados que dependan de esta diferencia.

Contraste de hipótesis

Una hipótesis estadística es una afirmación sobre la población. En el contexto paramétrico, se referirá a una proposición sobre el valor o rango de valores de un parámetro que rige la distribución.

Un contraste de hipótesis enfrenta dos hipótesis estadísticas complementarias **H0 (hipótesis nula)** y **H1 (hipótesis alternativa)** sobre una o más poblaciones. El objetivo es saber si los resultados de una muestra, que avala siempre a H1, son suficientemente significativos para concluir que H0, su hipótesis complementaria, no es cierta.

Tomando como información una muestra y usando los estadísticos adecuados, dictamina si los efectos de la muestra son relevantes para adoptar H1.

$$\begin{cases} H_0: & \text{hipótesis nula} \\ H_1: & \text{hipótesis alternativa} \end{cases}, \quad \text{significación } \alpha$$

Para cada tipo de contraste, hay un estadístico apropiado que se obtiene de la muestra y dos regiones complementarias: región de aceptación y región de rechazo o crítica, que dependen a su vez del nivel de significación α establecido.

El proceso del contraste de hipótesis se podría resumir en:

1. Obtención de muestra.
2. Cálculo del estadístico del contraste sobre la muestra.
3. Observar en qué zona cae y concluir:
    a. Aceptamos H0 si cae en la zona de aceptación.
    b. Rechazamos H0 si cae en la región crítica. Esto es equivalente a decidir que los efectos de la muestra que avalan H1 son relevantes o significativos.

Términos importantes en el contraste de hipótesis:

- **P-valor:** representa la probabilidad de tener los resultados de la muestra o más extremos (en la dirección de H1), si suponemos como cierta la hipótesis nula H0. Es intuitivo pensar que, si este valor es muy bajo, deberíamos rechazar la hipótesis nula.

En general, rechazaremos la hipótesis nula H0 siempre que el p-valor sea menor que 0.05.

- **Tipos de errores:** son los distintos errores de la decisión del contraste en función de la verdad subyacente y la conclusión tomada. Estos son:
    - Error tipo I: rechazar H0 cuando es cierta.
    - Error tipo II: no rechazar H0 cuando es falsa.

Una empresa afirma que en sus cajas de tornillos la proporción de piezas defectuosas es menor o igual que 0.03 (3 %). La organización del consumidor toma una serie de muestras y contrasta si la empresa está mintiendo sobre su producto, planteando el contraste de hipótesis siguiente:

H0: p <= 0.03

H1: p>0.03

Esta calcula el estadístico adecuado para establecer si la muestra avala H1 con suficiente relevancia y, observando que cae en la región crítica, rechaza H0. Por tanto, se puede concluir que la empresa miente sobre sus productos.

Descarga: Consulta el notebook UD3 N02

Descárgate el archivo UD3_N02 y ejecútalo en R. También puedes verlo en.html

Descarga: Consulta el notebook UD3 N03

Descárgate el archivo UD3_N03 y ejecútalo en R. También puedes verlo en.html

5.2. Inferencia no paramétrica

La inferencia no paramétrica se realiza cuando no conocemos la distribución probabilística.

Se estudiarán tres casos de contrastes no paramétricos que representan la idea de este tipo de inferencia:

Test chi cuadrado de bondad de ajuste

Si tenemos una muestra de frecuencias de incidencia en categorías, una distribución teórica o ideal de las mismas, este test indica si se puede considerar que la muestra se ajusta a la distribución de frecuencias ideal.

Test chi cuadrado de homogeneidad

Contrasta si varias muestras pueden provenir de una misma distribución.

Test chi cuadrado de independencia

Contrasta si varias muestras son independientes. Esto es, si se puede considerar que los valores que toma cada una no influyen en los que toma la otra.

Descarga: Consulta el notebook UD3 N04

Descárgate el archivo UD3_N04 y ejecútalo en R. También puedes verlo en.html

5.3. ANOVA

El análisis de varianza (ANOVA) es un procedimiento estadístico para comparar las medias de dos o más poblaciones.

Los ANOVAS evalúan la influencia y diferenciación que aportan distintos factores sobre una variable numérica, dictaminando si son influyentes en el valor de la misma.

Hay diversos tipos de ANOVA, aquí se verá el *one-way* ANOVA.

Este plantea el contraste de hipótesis que tiene por hipótesis nula la igualdad de medias de varios grupos determinados por factores respecto a una única variable continua:

$$H_0 : \mu_1 = \cdots = \mu_a$$

¿Influye el color de ojos sobre la estatura en un ser humano?

Descarga: Consulta el notebook UD3 N05

Descárgate el archivo UD3_N05 en R y este csv. También puedes verlo en .html

VI. Resumen final

Se han mostrado el esqueleto probabilístico y los conceptos más inmediatos dentro de la inferencia estadística, aplicándolos en casos. Si bien es un campo muy conceptual y sórdido dentro de las matemáticas y la estadística, su aplicación lo convierte en intuitivo.

Además, los resultados estadísticos en forma de intervalos de confianza e intervalos de hipótesis refuerzan mucho las presentaciones en análisis de datos y, dependiendo del objetivo, pueden resultar imprescindibles.

Ejercicios

Caso práctico

Como repaso del tema y preparación para el Caso práctico final, se presenta el siguiente caso práctico.

Descárgate el archivo ACTIVIDAD3_UD3 en R. También puedes verlo en .html

Cuando lo hayas realizado, puedes descargar su solución y comprobar tus resultados.

Solución

En los siguientes archivos dispones de la solución de la actividad propuesta:

- Solución en .html
- Solución en R

Recursos

Glosario.

- **Contraste de hipótesis**: Juicio estadístico que se realiza entre dos hipótesis, que basa su fallo en los valores muestrales y su significación.
- **Distribución probabilística**: Ley con respecto a la cual se generan números aleatorios.
- **Intervalo de confianza**: Rango de valores en los que se espera que un parámetro de una distribución probabilística se halle con una probabilidad dada.
- **Nivel de confianza**: Valor de probabilidad con el que se dictamina una ley estadística.
- **P-valor**: Valor asociado a una muestra y un contraste de hipótesis. Si es menor que 0.05 rechazamos la hipótesis nula H0.
- **Probabilidad de un suceso**: Propensión en forma de ratio a la ocurrencia de un suceso.