Estadística II & Inferencia Estadística

Unidad de Formación en Matemática y Estadística (UFME)

Coordinación de Estadística

Actualizado al 13-03-2023

Índice general

Presentación **5**

Modalidad **7**

1. Pruebas de hipótesis **9**
1.1. Concepto 9
1.1.1. Elaboración 9
1.1.2. Errores tipo I y II 11
1.1.3. Procedimiento de prueba 12
1.1.4. Intervalos de confianza 14
1.2. Pruebas de hipótesis para la media 15
1.3. Pruebas de hipótesis para la diferencia de medias 23
1.4. Prueba de hipótesis para comparación de varianzas 23
1.5. Prueba de hipótesis para la diferencia de proporciones 23

2. Regresión Lineal **25**
2.1. Análisis descriptivo de datos 26
2.1.1. Gráficos 26
2.1.2. Tablas 26
2.2. Covarianza 26
2.3. Correlación 26
2.4. Regresión lineal simple 26
2.4.1. Suma cuadrática de errores 26
2.4.2. Estimadores de mínimos cuadrados 26
2.4.3. Pruebas de hipótesis 26
2.4.4. Supuestos 26
2.4.5. Error estándar residual 26
2.4.6. Coeficiente de determinación 26
2.5. Regresión lineal múltiple 26
2.5.1. Suma cuadrática de errores 26
2.5.2. Suma cuadrática total 26
2.5.3. Estimadores de mínimos cuadrados 26
2.5.4. Pruebas de hipótesis 26

2.5.5. Supuestos 26
2.5.6. Error estándar residual 26
2.5.7. Coeficiente de determinación ajustado 26
2.6. Selección de variables 26
2.6.1. Forward 26
2.6.2. Backward 26

3. ANOVA y pruebas no paramétricas 27
3.1. ANOVA de un factor 27
3.1.1. Varianzas 27
3.1.2. Condiciones 27
3.1.3. Tamaño del efecto 27
3.2. ANOVA de dos factores 27
3.2.1. Varianzas 27
3.2.2. Condiciones 27
3.2.3. Tamaño del efecto 27
3.3. Análisis de homogeneidad de varianzas 27
3.3.1. Prueba de Levene y Barlett 27
3.4. Pruebas no paramétricas 27
3.4.1. Pruebas de los rangos signados 27
3.4.2. Prueba de W-M-W 27
3.4.3. Prueba de Kruskall - Wallis 27

Presentación

La asignatura Estadística II & Inferencia Estadística, es el segundo curso estadístico de la carrera de Ingeniería Comercial e Ingeniería en Control de Gestión respectivamente. Estos cursos tienen un enfoque práctico con un fuerte énfasis en inferencia a partir de datos. Los cursos, se concentran en pruebas de hipótesis, modelos de regresión lineal, Análisis de varianzas y Pruebas no paramétricas. La metodología de aprendizaje se basa en clases interactivas - participativas, las que están basadas en el uso de R como programa estadístico para el análisis de datos.

Modalidad

La modalidad de trabajo consta de los siguientes elementos:

1. El **documento** web cuenta con el desarrollo de todos los tópicos de curso, además de ejemplificaciones y ejercicios.

2. En su mayoría, los ejemplos y ejercicios presentes en el documento hacen **uso de R** como programa de análisis estadístico. El desarrollo de los ejercicios por parte del estudiante serán en Google Colab R. Esta plataforma no cuenta con una opción de configuración interna para R, sin embargo, en el siguiente enlace se puede acceder a un documento con una configuración preestablecida para este lenguaje. El archivo que se genera se guardará automáticamente en la cuenta de Gmail predeterminada, por lo que se recomienda que dicha cuenta corresponda a la asociada a la UDP; otra opción en caso de no querer modificar su cuenta predeterminada, es descargar el archivo y cargarlo manualmente en la carpeta de Drive que estime conveniente. Los aspectos relacionados con el uso de Google Colab R serán abordados en el **Taller Introductorio**.

3. Se cuenta con **talleres de práctica**, lo cuales se desarrollarán en de Google Colab R. Estos talleres cuentan con tres elementos: ejercicios desarrollados, ejercicios propuestos para desarrollar en clases y ejercicios para el trabajo independiente del estudiante. Para estos últimos, **NO** habrá pauta, ya que se espera que el estudiante sea capaz de evaluar críticamente la solución obtenida.

4. El curso cuenta con **bibliografía** obligatoria y suplementaria:

    - (Obligatoria) *"Estadística para Administración y Economía"* (Anderson et al., 2008)
    - (Obligatoria,) *"Probabilidad y Estadística para Ingeniería y Ciencias"* (Devore, 2008)
    - (Complementaria) *"R Programming for Data Science"* (Peng, 2016)
    - (Complementaria) *"The R Software: Fundamentals of Programming and Statistical Analysis"* (de Micheaux et al., 2013)

- (Complementaria) *"ggplot2: Elegant Graphics for Data Analysis"* (Wickham, 2009)

Además, a lo largo del documento se añaden citas que refuerzan el contenido presentando. Al final de cada sección se encuentra el detalle de cada una de ellas.

5. Las **bases de datos** a utilizar en el curso se encuentran disponible en un repositorio web de libre acceso.

Unidad 1

Pruebas de hipótesis

1.1. Concepto

Una **hipótesis estadística** o simplemente *hipótesis* es una pretensión o aseveración sobre el valor de un solo parámetro (característica de la población o característica de una distribución de la población) o sobre los valores de varios parámetros (Devore, 2008, página 285) (Anderson et al., 2008, página 340).

En cualquier cualquier problema de prueba de hipótesis, existen dos hipótesis contradictorias consideradas, la hipótesis nula y la alternativa.

La **hipótesis nula** denotada por H_0, es la pretensión de que inicialmente se supone cierta (la pretensión de "creencia previa"). La **hipótesis alternativa** denotada por H_1 (o H_a), es la aseveración contradictoria a H_0.

La hipótesis nula será rechazada en favor de la hipótesis alternativa solo si la evidencia muestral sugiere que H_0 es falsa. Si la muestra no contradice fuertemente a H_0, se continuará creyendo en la verdad de la hipótesis nula. Las dos posibles conclusiones derivadas de un análisis de prueba de hipótesis son entonces *rechazar* H_0 o *no rechazar* H_0.

1.1.1. Elaboración

En algunas aplicaciones no parece obvio cómo formular la hipótesis nula y alternativa. Se debe tener cuidado en estructurar la hipótesis apropiadamente de manera que la conclusión de la prueba de hipótesis proporcione la información que el investigador o la persona encargada de tomar las decisiones desea. A partir de la situación, las pruebas de hipótesis pueden tomar tres formas (tabla 1.1), las cuales se diferencian en el desigualdad o igualdad empleada en la hipótesis alternativa.

Tabla 1.1: Planteamiento de las pruebas de hipótesis

Caso 1	Caso 2	Caso 3
$H_0 : \theta = \theta_0$	$H_0 : \theta = \theta_0$	$H_0 : \theta = \theta_0$
$H_1 : \theta \neq \theta_0$	$H_1 : \theta > \theta_0$	$H_1 : \theta < \theta_0$

En diversas ocasiones, H_1 se conoce como la "hipótesis del investigador", puesto que es la pretensión que al investigador en realidad le gustaría validar. La palabra *nulo* "significa sin valor", la que sugieres que H_0 no deberá ser identificada con la hipótesis de ningún cambio.

Ejemplo 1.1. Considérese, que el 10 % de todas las tarjetas de circuito producidas por un cierto fabricante durante un periodo de tiempo reciente estaban defectuosas. Un ingeniero ha sugerido un cambio en el proceso de producción en la creencia de que dará por resultado una proporción reducida del proceso cambiado.

La hipótesis alternativa (posición del investigador) es $H_1 : p < 0.10$, la pretensión de que la modificación del procesos redujo la proporción de las tarjetas defectuosas. Una opción natural para H_0 en esta situación es la pretensión contraria a la establecida en H_1, es decir, $p \geq 0.1$. En su lugar se considera $H_0 : p = 0.1$ contra $H_1 : p < 0.1$, tal como se expuso en la tabla anterior.

Ejercicio 1.1. El gerente de Danvers-Hilton Resort afirma que la cantidad media que gastan los huéspedes en un fin de semana es menos de $600 dólares. Un miembro del equipo de contadores observó que en los últimos meses habían aumentado tales cantidades. El contador emplea una muestra de cuentas de fin de semana para probar la afirmación del gerente.

a. ¿Qué forma de hipótesis deberá usar para probar la afirmación del gerente? Explique.

Caso 1	Caso 2	Caso 3
$H_0 : \mu = 600$	$H_0 : \mu = 600$	$H_0 : \mu = 600$
$H_1 : \mu \neq 600$	$H_1 : \mu > 600$	$H_1 : \mu < 600$

b. ¿Cuál es la conclusión apropiada cuando no se puede rechazar la hipótesis nula H_0?
c. ¿Cuál es la conclusión apropiada cuando se puede rechazar la hipótesis nula H_0?

Ejercicio 1.2. El gerente de un negocio de venta de automóviles está pensando en un nuevo plan de bonificaciones, con objeto de incrementar el volumen de ventas. Al presente, el volumen medio de ventas es 14 automóviles por mes. El gerente desea realizar un estudio para ver si el plan de bonificaciones incrementa

el volumen de ventas. Para recolectar los datos, una muestra de vendedores venderá durante un mes bajo el nuevo plan de bonificaciones.

a. Dé las hipótesis nula y alternativa más adecuadas para este estudio.
b. Comente la conclusión resultante en el caso en que H_0 no pueda rechazarse.
c. Comente la conclusión que se obtendrá si H_0 puede rechazarse.

Ejercicio 1.3. Debido a los costos y al tiempo de adaptación de la producción, un director de fabricación antes de implantar un nuevo método de fabricación, debe convencer al gerente de que ese nuevo método de fabricación reducirá los costos. El costo medio del actual método de producción es $220 por hora. En un estudio se medirá el costo del nuevo método durante un periodo muestral de producción,

a. Dé las hipótesis nula y alternativa más adecuadas para este estudio.
b. Haga un comentario sobre la conclusión cuando H_0 no pueda rechazarse.
c. Dé un comentario sobre la conclusión cuando H_0 pueda rechazarse.

1.1.2. Errores tipo I y II

Las hipótesis nula y alternativa son afirmaciones opuestas acerca de la población. Una de las dos, ya sea la hipótesis nula o la alternativa es verdadera, pero no ambas. Lo ideal es que la prueba de hipótesis lleve a la aceptación de H_0 cuando H_0 sea verdadera y al rechazo de H_0 cuando H_1 sea verdadera. Por desgracia, las conclusiones correctas no siempre son posibles. Como la prueba de hipótesis se basa en una información muestral debe tenerse en cuenta que existe la posibilidad de error.

Los dos tipos de errores que se pueden cometer son:

- **Error tipo I:** Rechazar H_0 cuando H_0 es verdadera.
- **Error tipo II:** No rechazar H_0 cuando H_0 es falsa.

Es posible el error que se desea cometer, es decir, es posible establecer la probabilidad de cometer un error tipo I o II, pero no ambos. El **nivel de significancia** es la probabilidad de cometer un error tipo I cuando la hipótesis nula es verdadera. Para denotar el nivel de significancia se usa la letra griega α, y los valores que se suelen usar para α con 0.0 y 0.01.

Ejemplo 1.2. Walter Williams, columnista y profesor de economía en la universidad George Mason indica que siempre existe la posibilidad de cometer un error tipo I o un error tipo II al tomar una decisión (*The Cincinnati Enquirer*, 14 de agosto de 2005). Hace notar que la Food and Drug Administration corre el riesgo de cometer estos errores en sus procedimientos para la aprobación de medicamentos.

Cuando comete un error tipo I, la FDA no aprueba un medicamento que es seguro y efectivo. Al cometer un error tipo II, la FDA aprueba un medicamento que presenta efectos secundarios imprevistos. Sin importar la decisión que se tome, la probabilidad de cometer un error costoso no se puede eliminar.

Ejercicio 1.4. Nielsen informó que los hombres jóvenes estadounidenses ven diariamente 56.2 minutos de televisión en las horas de mayor audiencia (*The Wall Street Journal Europe*, 18 de noviembre de 2003). Un investigador cree que en Alemania, los hombres jóvenes ven más tiempo la televisión en las horas de mayor audiencia. Este investigador toma una muestra de hombres jóvenes alemanes y registra el tiempo que ven televisión en un día. Los resultados muestrales se usan para probar las siguientes hipótesis nula y alternativa.

$$H_0 : \mu = 56.2$$
$$H_1 : \mu > 56.2$$

a. En esta situación, ¿cuál es el error tipo I? ¿Qué consecuencia tiene cometer este error?
b. En esta situación, ¿cuál es el error tipo II? ¿Qué consecuencia tiene cometer este error?

Ejercicio 1.5. Suponga que se va a implantar un nuevo método de producción si mediante una prueba de hipótesis se confirma la conclusión de que el nuevo método de producción reduce el costo medio de operación por hora.

a. Dé las hipótesis nula y alternativa adecuadas si el costo medio de producción actual por hora es $220.
b. En esta situación, ¿cuál es el error tipo I? ¿Qué consecuencia tiene cometer este error?
c. En esta situación, ¿cuál es el error tipo II? ¿Qué consecuencia tiene cometer este error?

1.1.3. Procedimiento de prueba

Un procedimiento de prueba es un regla, basada en datos muestrales, para decidir si rechazar H_0. Este proceso consta de dos elementos:

- **Estadístico de prueba:** Función de los datos muestrales en los cuales ha de basarse la decisión.
- **Región de rechazo:** Conjunto de todos los valores estadísticos de prueba por los cuales H_0 será rechazada.

Para decidir si H_0 es finalmente rechazada es posible ocupar dos métodos.

1. **Método del valor p**

Un valor-p es una probabilidad que porta a una medida de evidencia suministrada por la muestra contra la hipótesis nula. Valores pequeños indican una evidencia mayor contra la hipótesis nula.

Además de representar un probabilidad, el valor-p puede ser vista como una porción de área bajo la curva. La figura 1.1 muestra la relación entre los distintos elementos ya mencionados.

La curva corresponde a la función de probabilidad de los datos. Los valores centrales son aquellos que son más probables de observar (parte más alta de la curva), mientras que los valores extremos (derecha e izquierda) son los menos probables de observar. El punto de color rojo corresponde al estadístico de prueba, función que nos dará un valor con el que seremos capaces de rechazar o no H_0. Finalmente el área de color verde corresponde al área bajo la curva desde el estadístico observado hacia la izquierda (en este caso).

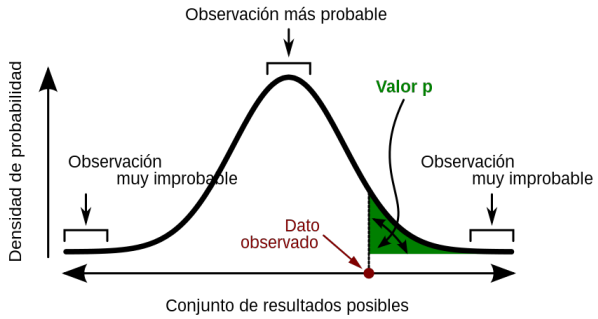

Figura 1.1: Estadístico de prueba para un prueba altenativa con signo $>$

La tabla 1.2, da cuenta de la relación que existe entre las pruebas de hipótesis y la ubicación del valor-p en el gráfico presentado.

Tabla 1.2: Hipótesis alternativa, valor-p y estadístico de prueba

Signo de comparación en H_1	Referencia	Ubicación del estadístico de prueba y valor-p
$>$	Unilateral derecha	A la derecha del gráfico
$<$	Unilateral izquierda	A la izquierda del gráfico
$\neq$	Bilateral	A ambos lados del gráfico

La regla de rechazo usando el valor-p es

$$\text{Rechazar } H_0 \text{ si el valor-p } \leq \alpha$$

En la figura 1.2, se puede observar los tres casos posibles para las distintas hipótesis alternativas, en las cuales se ejemplifica un valor-p en cada uno de los casos. De izquierda a derecha, las hipótesis alternativas correspondientes son unilateral izquierda, unilateral derecha, y bilateral. La decisión de si en cada uno de los casos se rechaza o no la hipótesis nula, depende del valor elegido para la significancia.

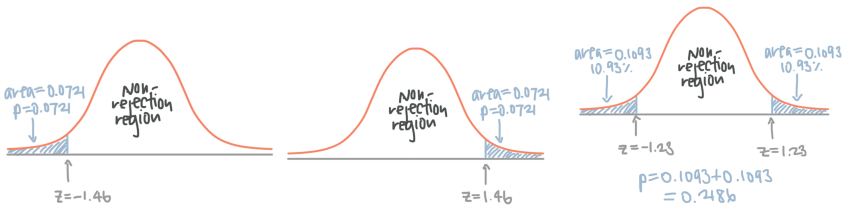

Figura 1.2: Estadísticos de prueba

2. **Método del valor crítico**

Este método consiste en comparar el estadístico de prueba con un número fijo llamado **valor crítico**. El valor crítico es un punto de referencia para determinar si el valor del estadístico de prueba es lo suficientemente pequeño para rechazar la hipótesis nula. El valor crítico corresponde a una porción del área bajo la curva, llamada α (el cual es definido por el investigador), y está ubicada en el mismo sector que el valor-p.

El intervalo de números generado a partir del valor crítico es lo denominado **región de rechazo**. En la figura 1.3, se observa que un hipótesis es rechazada cuando el valor-p es menor a α, lo cual, es equivalente a decir, que estadístico de prueba es mayor (o menor) al valor crítico, a esto se le denomina "caer en la región de rechazo".

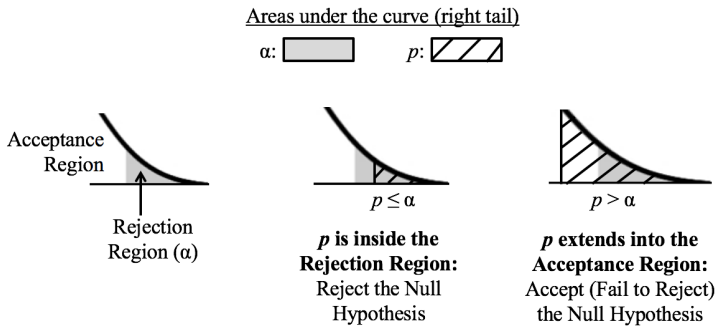

Figura 1.3: Valor-p, el área bajo la curva

Los lineamiento de como construir un estadístico de prueba y un valor-p se darán a conocer a partir de la sección 1.2.

1.1.4. Intervalos de confianza

Existe un relación directa entre las pruebas de hipótesis y los intervalos de confianza, ya que estos pueden ser utilizados para rechazar o no H_0. La tabla 1.3, da cuenta de del tipo de intervalo de confianza que se debe elaborar para cada tipo de prueba de hipótesis.

Tabla 1.3: Hipótesis alternativa e Intervalo de confianza

Signo de comparación en H_1	Tipo de intervalo de confianza
$>$	(a, ∞)
$<$	$(-\infty, b)$
$\neq$	(a, b)

A lo largo de las distintas pruebas, se abordarán los distintos métodos de prueba, incluyendo el uso de intervalos de confianza.

1.2. Pruebas de hipótesis para la media

Esta sección se centra en el planteamiento y prueba de hipótesis relacionadas a la parámetro de media. Para cada uno de estos casos, se detalla el procedimiento en R y los distintos métodos de prueba para la decisión de rechazo de H_0. En particular, las pruebas para este parámetro requieren asumir la suposición de distribución normal de los datos.

Pruebas de hipótesis para la media de una distribución normal con varianza poblacional conocida

Aun cuando la suposición de que el valor de σ^2 es conocido, rara vez se cumple en la práctica. Este caso proporciona un buen punto de partida debido a la facilidad con que los procedimientos generales y sus propiedades pueden ser desarrollados. La hipótesis nula en los tres casos propondrá que μ tiene un valor numérico particular, el valor nulo, el cual será denotado por μ_0.

El estadístico de prueba y los valores críticos de comparación están dados en la tabla 1.4.

Tabla 1.4: Criterios de rechazo para la prueba de una media con varianza poblacional conocida

Hipótesis nula	***Estadístico***
$H_0 : \mu = \mu_0$	$Z_0 = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
Hipótesis alternativa	***Criterio de rechazo***
$H_1 : \mu \neq \mu_0$	$\lvert Z_0 \rvert \geq Z_{1-\alpha/2}$
$H_1 : \mu > \mu_0$	$Z_0 \geq Z_{1-\alpha}$
$H_1 : \mu < \mu_0$	$Z_0 \leq Z_{\alpha}$

Ejemplo 1.3. El índice Rockwell de dureza para acero se determina al presionar

una punta de diamante en el acero y medir la profundidad de la penetración, el cual tiene un varianza de medición de 6. Para 50 especímenes de una aleación de acero, el índice Rockwell de dureza promedió 62. El fabricante dice que esta aleación tiene un índice de dureza promedio menor a 64. Asumiendo que el índice de dureza sigue una distribución normal, ¿hay suficiente evidencia para refutar lo dicho por el fabricante con un nivel de significancia de 1 %?

Al plantear la prueba de hipótesis se debe tener en cuenta que la hipótesis del investigador ha de estar reflejada en H_1, tal como se muestra a continuación.

$$H_0 : \mu = 64 H_1 : \mu < 64$$

Luego, se desarrolla la expresión del estadístico de prueba, para conocer su valor numérico.

$$Z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{62 - 64}{\sqrt{6}/\sqrt{50}} = -5.774$$

Ocupando el método del valor crítico, escribimos el criterios de rechazo correspondiente. Sin embargo, aún está la tarea de determinar el valor crítico.

$$Z_0 \leq Z_\alpha$$
$$-5.774 \leq Z_{0.01}$$

Para determinar el valor de $Z_{0.01}$ (tal como menciona en los cursos de Estadística I y Estadística descriptiva), el comando a ocupar

```
qnorm(p = 0.01)
```

```
## [1] -2.326348
```

Es claro que -5.774 es menor a -2.32, es decir, que al cumplirse la condición de rechazo, esto implica que se rechaza H_0. Por lo tanto, existe suficiente evidencia estadística para rechazar la hipótesis nula, es decir, se apoya la postura del fabricante respecto a un indice de dureza promedio menor a 64, con una significancia del 1 % (o confianza del 99 %).

En caso de que deseemos utilizar el método del valor-p, es necesario apoyarnos en R para realizar el calculo de este. El comando necesario para calcular el valor depende la prueba que estemos llevando a cabo, por lo que en el siguiente documento podrán encontrar un resumen para las distintas pruebas.

```
pnorm(-5.774)
```

```
## [1] 3.870572e-09
```

El valor-p obtenido es evidentemente menor a la significancia (0.01), obteniéndose la misma conclusión antes expuesta.

Respecto al intervalo de confianza, es posible determinarlo dada la siguiente expresión.

$$\left(-\infty, \bar{x} + Z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right) \tag{1.1}$$

Reemplazando los valores dados en el enunciado se tiene

$$\left(-\infty, 62 + Z_{0.99}\frac{6}{\sqrt{50}}\right) = (-\infty, 62.71)$$

Al observar el intervalo de confianza, se debe observar si el valor de μ_0 está dentro o fuera de este. En este caso, $\mu_0 = 64$ no se encuentra dentro del intervalo. Por lo tanto, se rechaza la hipótesis nula, obteniéndose la misma conclusión antes obtenida.

Al igual que el valor-p, la forma en la que se debe usar el intervalo de confianza varía dependiendo del tipo de prueba de hipótesis que se realiza, por lo que en el siguiente documento podrán encontrar un resumen para las distintas pruebas, dicho documento incluye los distintos comando en R para obtener los resultados de una prueba de hipótesis de manera automática.

Para este tipo de pruebas, no hay comandos en R que permitan hacer el trabajo de manera automática. Esto es debido a lo expuesto en un principio: **difícilmente se conoce la varianza poblacional en la práctica**.

Ejercicio 1.6. Sea el estadístico de prueba Z con una distribución normal estándar cuando H_0 es verdadera. Dé el nivel de significación en cada una de las siguientes situaciones:

a. $H_1 : \mu > \mu_0$, región de rechazo $z \geq 1.88$.
b. $H_1 : \mu < \mu_0$, región de rechazo $z \leq -2.75$.
c. $H_1 : \mu \neq \mu_0$, región de rechazo $z \geq 2.88$ o $z \leq -2.88$.

Ejercicio 1.7. Un fabricante de cajas de cartón afirma que sus cajas tienen un peso promedio de 5 kg. Para verificar esta afirmación, un cliente selecciona al azar 25 cajas y encuentra que el peso promedio es de 4.8 kg con una desviación estándar conocida de 0.5 kg. ¿Hay suficiente evidencia para rechazar la afirmación del fabricante al nivel de significancia del 5 %?

Ejercicio 1.8. Un entrenador de baloncesto cree que sus jugadores pueden hacer tiros libres con una precisión de al menos el 80 %. Para probar su hipótesis, el entrenador toma una muestra aleatoria de 30 tiros libres y encuentra que la proporción de tiros exitosos es del 75 %. Si la desviación estándar conocida de la proporción de tiros exitosos es del 5 %, ¿hay suficiente evidencia para rechazar la hipótesis del entrenador al nivel de significancia del 1 %?

Ejercicio 1.9. Un cirujano afirma que sus pacientes se recuperan en un promedio de 5 días después de una cirugía. Para probar su afirmación, el cirujano toma una muestra aleatoria de 20 pacientes y encuentra que la duración promedio de recuperación es de 6 días, con una desviación estándar conocida de 1.5 días. ¿Hay suficiente evidencia para rechazar la afirmación del cirujano al nivel de significancia del 10 %?

Ejercicio 1.10. Se cree que la cantidad promedio de cafeína en una taza de café es de 100 mg. Para probar esta hipótesis, se toma una muestra aleatoria de 50 tazas de café y se encuentra que la cantidad promedio de cafeína es de 105 mg, con una desviación estándar conocida de 15 mg. ¿Hay suficiente evidencia para rechazar la hipótesis nula al nivel de significancia del 5 %?

Ejercicio 1.11. Se desea evaluar si la altura promedio de una población de girasoles es de 150 cm. Para ello, se selecciona una muestra aleatoria de 30 girasoles y se encuentra que la altura promedio es de 155 cm, con una desviación estándar conocida de 5 cm. ¿Hay suficiente evidencia para rechazar la hipótesis nula al nivel de significancia del 1 %?

Pruebas de hipótesis para la media de una distribución normal con varianza poblacional desconocida

De igual manera los expuesto en el primer caso, los pasos a seguir para probar una hipótesis son los mismo, y se mantendra así para cualquier caso.

1. Plantear las hipótesis nula y alternativa
2. Identificar o establecer el nivel de significancia.
3. Identificar los datos muestrales y poblacionales con los que se cuenta.
4. Utilizar alguna de las reglas de decisión (Estadístico de prueba, valor-p o intervalo de confianza).
5. Concluir

En la situación de una prueba de hipótesis de la media, en la cual lo datos distribuyen normal y la varianza poblacional es desconocida, los criterios de rechazo son similares a los vistos anteriormente, sin embargo, cambia la distribución del estadístico de prueba, tal como se muestra en la tabla xxx.

Tabla 1.5: Criterios de rechazo para la prueba de una media con varianza poblacional conocida

Hipótesis nula	***Estadístico***
$H_0 : \mu = \mu_0$	$t_0 = \dfrac{\bar{x} - \mu_0}{S/\sqrt{n}}$
Hipótesis alternativa	***Criterio de rechazo***
$H_1 : \mu \neq \mu_0$	$\lvert t_0 \rvert \geq t_{1-\alpha/2,n-1}$
$H_1 : \mu > \mu_0$	$t_0 \geq t_{1-\alpha,n-1}$
$H_1 : \mu < \mu_0$	$t_0 \leq t_{\alpha,n-1}$

donde n corresponde al tamaño de la muestra.

Ejemplo 1.4. (Portafolio 2022, segundo semestre, pregunta 2) La base de datos *imacec.csv*, contiene los datos del Imacec del área de producción de bienes en los sectores de Minería e Industria, desde enero de 2018 hasta junio de 2022. Establezca si hay suficiente evidencia estadística, para afirmar que, el valor promedio del Imacec de cada sector por separado, es mayor al promedio general de todos los sectores (denote el promedio general por μ_0). Establezca las hipótesis respectivas, estadísticos y criterios de rechazo, utilizando una significancia del 10 %. Asume que las variables distribuyen normal y tienen varianza poblacional desconocida.

En este caso al contar con una base de datos (y para este tipo de prueba), podemos hacer uso directamente de R para obtener el estadístico de prueba, valor-p e intervalo de confianza asociado.

```
# Carga de la base de datos
df =
 ↪  read.csv("https://raw.githubusercontent.com/Dfranzani/Bases-de-datos-para-cursos/main/2022-2/
```

En primer lugar, obtenemos el valor de mu_0 para poder plantear la prueba de hipótesis

```
# Promedio de todos los sectores
promedio_general = mean(c(df$Mineria, # Valores del Imacec de
 ↪  Minería
                          df$Industria)) # Valores del Imacec de
 ↪  Industria
promedio_general # Mu_0
```

```
## [1] 98.54167
```

Iniciamos con la prueba de hipótesis para el sector de minería.

$$H_0 : \mu_{\text{Imacec-Mineria}} = 98.54167$$
$$H_1 : \mu_{\text{Imacec-Mineria}} > 98.54167$$

Luego, haciendo uso de R obtenemos los elementos necesario para rechazar o no H_0.

```
# Minería
t.test( # Prueba de hipótesis para el estadístico con
 ↪  distribución t-student
  x = df$Mineria, # Valores del Imacec de Minería
  alternative = "greater", # Signo de desigualdad de la hipótesis
   ↪  alternativa
  mu = 98.54167, # Valor del Mu_0
  conf.level = 0.9 # Confianza = 1 - alfa
  )
```

```
##
##  One Sample t-test
##
## data:  df$Mineria
## t = -1.2773, df = 53, p-value = 0.8965
## alternative hypothesis: true mean is greater than 98.54167
## 90 percent confidence interval:
##  96.21024      Inf
## sample estimates:
## mean of x
##  97.38519
```

El estadístico reportado es $t = -1.27373$ con un valor-p de 0.8965, el cual, al ser mayor a la significancia de 0.05, implica que no existe suficiente evidencia estadística para rechazar H_0, por lo que se asume que, el valor promedio del Imacec del sector de Minería no es mayor al promedio de general de ambos sectores.

Utilizando el intervalo de confianza $(96.21, \infty)$, se observa que el valor de $\mu_0 = 98.54$ se encuentra dentro del intervalo, por ende, no existe suficiente evidencia estadística para rechazar H_0, obteniéndose la misma conclusión que al usar el valor-p.

La prueba de hipótesis para el sector de industria es la siguiente.

$$H_0 : \mu_{\text{Imacec-Industria}} = 98.54167$$
$$H_1 : \mu_{\text{Imacec-Industria}} > 98.54167$$

```r
# Industria
t.test( # Prueba de hipótesis para el estadístico con
 ↪  distribución t-student
  x = df$Industria, # Valores del Imacec de Industria
  alternative = "greater", # Signo de desigualdad de la hipótesis
   ↪  alternativa
  mu = 98.54167, # Valor del Mu_0
  conf.level = 0.9 # Confianza = 1 - alfa
  )
```

```
##
##  One Sample t-test
##
## data:  df$Industria
## t = 1.3678, df = 53, p-value = 0.08857
## alternative hypothesis: true mean is greater than 98.54167
## 90 percent confidence interval:
##  98.60095      Inf
## sample estimates:
## mean of x
##  99.69815
```

El estadístico reportado es $t = 1.3678$ con un valor-p de 0.08857, el cual, al ser mayor a la significancia de 0.1, implica que existe suficiente evidencia estadística para rechazar H_0, por lo que se asume que, el valor promedio del Imacec del sector de Industria es mayor al promedio de general de ambos sectores.

Utilizando el intervalo de confianza $(98.6, \infty)$, se observa que el valor de $\mu_0 = 98.54$ no se encuentra dentro del intervalo, por ende, no existe suficiente evidencia estadística para rechazar H_0, obteniéndose la misma conclusión que al usar el valor-p.

Ejercicio 1.12. Utilizando la base del ejemplo 1.4. Establezca si hay suficiente evidencia estadística para afirmar que, el valor promedio del Imacec de cada sector durante el año 2022, es mayor al promedio general de todos los sectores en el mismo año (denote el promedio general por μ_0). Establezca las hipótesis respectivas, estadísticos y criterios de rechazo, utilice una significancia del 7 %. Asuma que las variables distribuyen normal y tienen varianza poblacional desconocida. Concluya.

Ejercicio 1.13. El control de emisión de residuos ha sido un tema que ha cobrado gran importancia en los últimos 20 años debido a los efectos del calentamiento global. Uno de los tantos residuos que contamina el aire es el Metano (CH4). Para estudiar este fenómeno haremos uso de la base *metano.csv*, la cual contiene los siguientes datos:

- Año: año en el que se realiza la medición de emisión de CH4.
- Mes: mes del año en el que se realiza la medición de emisión de CH4.

- CH4: concentración de CH4 (partes por miles de millones) en un muestra de aire.

Establezca si hay suficiente evidencia estadística para afirmar los siguiente:

1. La concentración promedio de metano es distinta a 1700 partes por millón.
2. La concentración promedio de metano del año 2021 es superior a 1780 partes por millón.
3. La concentración promedio de metano del periodo en el periodo de años 2019 - 2022 (inclusive) es inferior a 1750 partes por millón.

Establezca las hipótesis respectivas, estadísticos y criterios de rechazo, utilice una significancia del 7 %. Asuma que las variables distribuyen normal y tienen varianza poblacional desconocida. Concluya.

Ejercicio 1.14. La base de datos *consumidor.csv* contiene registros del Índice de Confianza del Consumidor (ICC). Este indicador de confianza del consumidor proporciona una indicación de la evolución futura del consumo y el ahorro de los hogares. Un indicador por encima de 100 señala un aumento en la confianza de los consumidores hacia la situación económica futura, como consecuencia de la cual son menos propensos a ahorrar y más inclinados a gastar dinero en compras importantes en los próximos 12 meses. Los valores por debajo de 100 indican una actitud pesimista hacia la evolución futura de la economía, lo que posiblemente resulte en una tendencia a ahorrar más y consumir menos.

Las variables que contiene la base de datos son las siguientes:

- Locacion: lugar en donde se mide el ICC (FRA = Francia, POL = Polonia, OECD = OCDE, ESP = España, BEL = Bélgica, ITA = Italia, DEU = Alemania).
- Mes: corresponde al mes en el que se realiza la medición del índice.
- Ano: corresponde al año en el que se realiza la medición del índice.
- ICC: valor del índice de confianza del consumidor.

Establezca si hay suficiente evidencia estadística para afirmar los siguiente:

1. El promedio del ICC es distinto a 100 puntos.
2. El promedio del ICC en Francia es menor a 105 puntos.
3. El promedio del ICC en Alemania es mayor a 107 puntos.

Establezca las hipótesis respectivas, estadísticos y criterios de rechazo, utilice una significancia del 12 %. Asuma que las variables distribuyen normal y tienen varianza poblacional desconocida. Concluya.

1.3. Pruebas de hipótesis para la diferencia de medias

Prueba de hipótesis para la diferencia de medias de dos distribuciones normales con varianzas poblacionales conocidas

Prueba de hipótesis para la diferencia de medias de dos distribuciones normales con varianzas poblacionales desconocidas e iguales

Prueba de hipótesis para la diferencia de medias de dos distribuciones normales con varianzas poblacionales desconocidas y distintas

1.4. Prueba de hipótesis para comparación de varianzas

1.5. Prueba de hipótesis para la diferencia de proporciones

Unidad 2

Regresión Lineal

2.1. Análisis descriptivo de datos

2.1.1. Gráficos

2.1.2. Tablas

2.2. Covarianza

2.3. Correlación

2.4. Regresión lineal simple

2.4.1. Suma cuadrática de errores

2.4.2. Estimadores de mínimos cuadrados

2.4.3. Pruebas de hipótesis

2.4.4. Supuestos

Linealidad

Normalidad

Homocedasticidad

Independencia

2.4.5. Error estándar residual

2.4.6. Coeficiente de determinación

2.5. Regresión lineal múltiple

2.5.1. Suma cuadrática de errores

2.5.2. Suma cuadrática total

2.5.3. Estimadores de mínimos cuadrados

2.5.4. Pruebas de hipótesis

Unidad 3

ANOVA y pruebas no paramétricas

3.1. ANOVA de un factor

3.1.1. Varianzas

3.1.2. Condiciones

3.1.3. Tamaño del efecto

3.2. ANOVA de dos factores

3.2.1. Varianzas

3.2.2. Condiciones

3.2.3. Tamaño del efecto

3.3. Análisis de homogeneidad de varianzas

3.3.1. Prueba de Levene y Barlett

3.4. Pruebas no paramétricas

3.4.1. Pruebas de los rangos signados

3.4.2. Prueba de W-M-W

3.4.3. Prueba de Kruskall - Wallis

Bibliografía

Anderson, D. R., Sweeney, D. J., and Williams, T. A. (2008). *Estadística para administración y economía*. Cengage Learning, México, 10a ed edition.

de Micheaux, P. L., Drouilhet, R., and Liquet, B. (2013). *R and Its Documentation*, pages 141–150. Springer New York, New York, NY.

Devore, J. L. (2008). *Probability and statistics for engineering and the sciences*. Thomson/Brooks/Cole, Belmont, CA, 7th ed edition.

Peng, R. D. (2016). *R programming for data science*. Leanpub, Victoria, BC, Canada.

Wickham, H. (2009). *Ggplot2: elegant graphics for data analysis*. Use R! Springer, New York.