20 - Pruebas de Hipótesis - Ejercicios

Table of Contents

Los siguientes ejercicios fueron tomados de:

* Batista William (2020). Introducción a la inferencia estadística aplicada. 2da Ed. Editorial Facultad de Agronomía. 272 pgs. <https://www.agro.uba.ar/catalog/introducci-n-la-inferencia-estad-stica-aplicada-3ra-edici-n>

# Ejercicios

## 1. Nicotina en cigarrillos. (Resuelto)

Un fabricante de cigarrillos sostiene que el contenido promedio de nicotina de su producto es menor que 0.7 miligramos por cigarrillo y solicita autorización para incluir esta afirmación en su publicidad. Para decidir si autorizarlo o no, es necesario establecer si es razonable dar por cierta dicha afirmación mediante una prueba de hipótesis. Para ello, se decide obtener una muestra aleatoria de 30 cigarrillos de la marca en cuestión y determinar el contenido de nicotina de cada uno.

Una vez obtenida la muestra, los contenidos de nicotina observados son los siguientes (mg/cigarrillo): 0.71 0.75 0.67 0.68 0.72 0.58 0.69 0.71 0.61 0.67 0.75 0.72 0.61 0.63 0.75 0.73 0.59 0.60 0.64 0.60 0.68 0.69 0.78 0.80 0.65 0.62 0.64 0.79 0.76 0.76

1. Formular las hipótesis estadísticas apropiadas para la prueba requerida.

$$ H\_{0}: \mu\_0 \geq 0.7 \\ H\_{1}: \mu\_0 < 0.7 $$

> El fabricante quiere decir que sus cigarrilos tienen menos de 0.7 mg de nicotina por cigarrillo, por eso planteamos eso como H1, ya que al rechazar H0, estarímos aceptando esa hipótesis.

1. Elegir un nivel de significación para esta prueba.

Se trata de un asunto de salud pública, por lo tanto deberíamos ser exigentes respecto de la probabilidad de error, sin embargo también puede argumentarse que un fumador no ha decidido cuidar su salud, fuma sabiendo que es perjudicial, y esto podría justificar una mayor probabilidad de error en la prueba. En el primer caso exigiríamos como máximo un alfa = 0.01, pero en el segundo parece aceptable un alfa = 0.05. Lo dejo esto en vuestro criterio.

1. Dando por supuesto que los datos registrados son realizaciones de variables aleatorias con distribución Normal, calcular el valor-p a partir de los datos disponibles.

Primero hago el cálculo a mano. Estos datos además van a usarse para graficar. Suponemos, como indica el enunciado, que se trata de una distribución Normal. pnorm() calcula la probabilidad acumulada hasta el valor (estPrueba), consideramos la cola izquierda, por eso en este caso la usamos directamente. Si la hipótesis se planteara al revés deberíamos usar 1 - pnorm(), que calcula la cola derecha.  
  
valor\_p

## [1] 0.1182848

También es buena idea calcular la probabilidad de que un cigarrillo tenga más de 0.7 mg de Nicotina, o igualmente podemos decir que es la proporción de los cigarrillos cuya concentración de Nicotina es superior a 0.7 mg.

## [1] 0.4144566

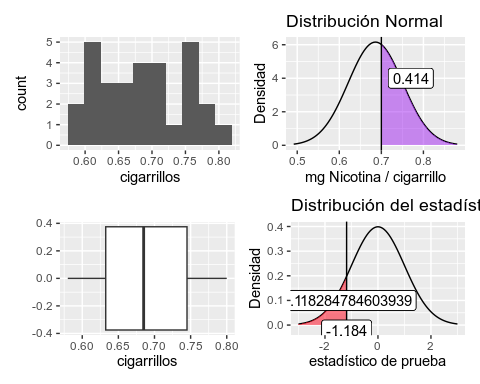
El problema pide utilizar una distribución Normal, por eso usamos z.test(), aunque al tratarse de una muestra lo correcto sería utilizar una prueba t.test() (como se hace más abajo). Nótese que las diferencias son leves y no cambian la conclusión en este caso.

La opción alternative=‘less’ indica que la hipótesis alternativa H1 dice que media0 es menor (less) a 0.7 mg/cigarrillo.

## One-sample z-Test  
##   
## data: cigarrillos  
## z = -1.1836, p-value = 0.1183  
## alternative hypothesis: true mean is less than 0.7  
## 95 percent confidence interval:  
## NA 0.7054558  
## sample estimates:  
## mean of x   
## 0.686

## One Sample t-test  
##   
## data: cigarrillos  
## t = -1.1836, df = 29, p-value = 0.1231  
## alternative hypothesis: true mean is less than 0.7  
## 95 percent confidence interval:  
## -Inf 0.7060977  
## sample estimates:  
## mean of x   
## 0.686

Siempre conviene graficar para ver mejor la información y mejorar de ese modo la interpretación. Aquí se grafica usando distribuciones Normales.



1. Concluir e interpretar aclarando el alcance de la inferencia realizada.

Con estos valores no puede rechazarse H0 con ninguno de los criterios posible, por lo tanto No se puede permitir a los fabricantes de cigarrillos que su publicidad diga que los cigarrillos tienen menos de 0.7 mg de nicotina por cigarrillo.

Además la probabilidad de que un cigarrillo tenga más de 0.7 mg de Nicotina es 0.414 (pNicotinaAlta), que es el área violeta mostrada en el gráfico superior izquierdo (g3), lo que indica que de cada 10 cigarrillos que alguien fuma, 4 están excedidos en Nicotina.

## 2. Emisiones de CO2 por automóviles

Una compañía productora de automotores asegura que los autos del modelo que ofrece para ser equipado como taxi cumplen en promedio con la norma de la Agencia de Protección Ambiental de los Estados Unidos que establece que estos vehículos deben emitir menos de 2,5 gramos de monóxido de carbono por kilómetro. Para decidir si dar por cierta esta afirmación se decide obtener una muestra aleatoria de 25 de estos vehículos matriculados como taxis en la ciudad de San Francisco y medir el nivel de emisión de monóxido de carbono de cada uno. Con los datos obtenidos se calcula la media muestral (2,44 g/km) y el estimador insesgado de la varianza (0,021 g2/k㎡).

1. Formular hipótesis estadísticas apropiadas para tomar la decisión en cuestión. Enunciar el significado de cada hipótesis en términos del fenómeno de interés.
2. Seleccionar un nivel de significación apropiado. ¿Qué implica el valor elegido?
3. ¿En qué consiste el procedimiento para generar los datos necesarios para poner a prueba las hipótesis propuestas?
4. Dando por cierto que los niveles de emisión registrados son realizaciones de variables aleatorias con distribución Normal, calcular el valor-p. ¿Qué probabilidad representa?
5. Concluir y explicar la conclusión en términos del fenómeno de interés.
6. ¿A qué población se extiende la inferencia realizada?

## 3. Margen bruto de maíz

Una empresa productora de semillas ofrece un nuevo híbrido de maíz que ha sido puesto a prueba en 12 lotes tomados al azar dentro del partido de Pergamino. En Pergamino el costo de producción de una hectárea de maiz hibrido es de US$ 325 y el ingreso por cada tonelada vendida es US$ 50. Los rendimientos obtenidos en cada lote fueron los siguientes (t/ha): 7.2 10.0 8.5 8.4 8.0 7.5 9.0 9.0 8.0 7.0 6.1 8.0

1. Identificar la poblacion, la muestra y las variables aleatorias consideradas.
2. Construir un diagrama de cajas y bigotes para los datos de rendimiento de los lotes.
3. Poner a prueba la hipótesis nula que dice que el margen bruto (la diferencia entre el ingreso y el costo) correspondiente a la media poblacional de los rendimientos que este nuevo híbrido produciría en los lotes del partido de Pergamino no supera los USS 50 por hectárea (Mostrar el desarrollo del análisis).
4. ¿Qué supuesto se dio por cierto para el análisis realizado en c?
5. Sobre la base del resultado obtenido, discutir la conveniencia de adoptar este nuevo hibrido en el partido de Pergamino.

## 4. Pasto para enfardar

En un establecimiento lechero se proyecta utilizar el pasto presente en una pastura de 20 ha para hacer una reserva de fardos de heno con la qual alimentar a las vacas lecheras durante el invierno. Por ello interesa determinar si los fardos a producir alcanzarán para cubrir la necesidad de 90 t de forraje para dicho período. Para concluir al respecto, se decide tomar una muestra de 25 marcos de 1 m² elegidos al azar dentro de la pastura. En cada marco, se cortará todo el forraje y se lo pesará luego de dejaro secar al aire del mismo modo que se hace para elaborar los fardos. Una vez realizadas las mediciones, se encuentra que el promedio de los pesos obtenidos es de 510 gramos/m² y el estimador del desvío estándar es de 100 g/m²

1. Identificar las unidades muestrales, la muestra y la población.
2. Formular hipótesis apropiadas para la evaluación propuesta.
3. Dando por cierto que el peso del forraje en un marco de 1 m² a elegir al azar en la pastura de 20 ha es una variable aleatoria con distribución Normal, calcular el valor-p y concluir con un nivel de significación ⍺ = 0,01.
4. Explicar la conclusión en términos de este problema.
5. Explicar el tipo de error que se puede haber cometido en este análisis y cuáles serían sus implicancias.

## 5. Titulación de arsénico

Para evaluar la exactitud de una nueva técnica de tulación para medir el contenido de arsénico en el agua una laboratorista prepara una solución acuosa que contiene exactamente 50.10^-3 mg de arsénico por litro. Luego, toma al azar 30 alicuctas de la solución y en cada una mide el contenido de arsénico con la técnica propuesta. Con los datos que genera, calcula la media muestral (52,18.10^-3 mg/1) y el estimador del desvío estándar (2,59.10^-3 mg/1).

1. Bajo los supuestos de distribución Normal y de independencia de las titulaciones ¿los datos permiten concluir, con a = 0.05, que la esperanza de las titulaciones difiere del contenido de arsénico de la solución?
2. Explicar qué es el nivel de significación a = 0.05 en términos de este problema.
3. Identificar la población de referencia y las variables aleatorias en este experimento. ¿Qué limitación importante tiene el procedimiento experimental seguido?
4. Construir e interpretar un intervalo de confianza 0.95 para la esperanza de las titulaciones.

## 6. Virus en duraznos

En una investigación sobre la susceptibilidad de plántulas de durazno a dos cepas diferentes de un virus, se tomaron de un vivero 15 plántulas al azar, en cada plántula se seleccionaron 2 hojas y cada una fue inoculada con una de las dos cepas virales. Al cabo de una semana se midió en cada hoja el tamaño de la lesión producida por el virus (en mm²). Los datos obtenidos figuran en la tabla:

Tamaño de lesión [mm²]

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Planta | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| Cepa A | 31 | 20 | 18 | 8 | 25 | 14 | 17 | 12 | 21 | 30 | 17 | 9 | 13 | 10 | 24 |
| Cepa B | 18 | 17 | 14 | 7 | 21 | 13 | 22 | 11 | 22 | 15 | 11 | 10 | 13 | 5 | 25 |

1. Estimar el promedio y la varianza de las diferencias entre los tamaños de las lesiones producidas por las dos cepas virales estudiadas.
2. Elegir un nivel de significación alfa y poner a prueba la siguiente hipótsis nula: “Las lesiones que producen las dos cepas virales tienen en promedio el mismo tamaño”.
3. Construir e interpretar un intervalo de confianza 1-alfa para la esperanza de la diferencia entre los tamaños de las lesiones producidas por las dos cepas virales.
4. ċQué supuesto condiciona la validez de la inferencia realizada en b y c?
5. Explicar qué representa el valor de alfa elegido en términos del problema.

## 7. Labranzas mecánicas vs. químicas

Un consorcio de productores agrícolas(CREA) lleva adelante una comparación de los rendimientos de maiz obtenidos con dos métodos de cultivo diferentes, labranza mecánica y labranza química. Para ello, cada socio del CREA elige al azar un lote que ha sido cultivado como una unidad al menos en los últimos 5 años, lo divide en dos y cultiva maíz aplicando uno de los dos tipos de labranza en cada mitad. Al final de la campaña, los productores reúnen el siguiente conjunto de datos de los rendimientos obtenidos:

Rendimiento de maíz [Uha]

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Productor | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Labranza mecánica | 8.9 | 7.8 | 10.1 | 9.7 | 9.2 | 9.1 | 9.9 | 8.4 | 9.0 | 7.2 |
| Labranza química | 8.8 | 6.8 | 12.9 | 11.9 | 8.0 | 12.2 | 9.1 | 11.2 | 10.5 | 10.1 |

1. Estimar el promedio y la varianza de las diferencias de rendimiento entre métodos de cultivo
2. Nombrar posibles causas de la varianza de las diferencias de rendimiento entre métodos de cultivo.
3. Formular hipótesis para evaluar si los dos métodos de cultivo producen en promedio igual rendimiento.

Dando por cierto que la diferencia de rendimientos en un lote a tomar al azar es una variables aleatoria con distribución Normal:

1. Calcular el valor p, concluir con un nivel de significación alfa = 0,05 y explicar la conclusión en términos del objetivo del estudio propuesto por el CREA
2. Construir e interpretar un intervalo de confianza 0,95 para la media poblacional de las diferencias entre los rendimientos obtenidos con los dos tipos de labranza en los lotes de los productores del CREA.

## 8. Profundidad de napas

Durante la última década, una importante superficie de los pastizales de la Región Pampeana ha sido reemplazada por forestaciones. Este cambio en el uso de la tierra puede producir consecuencias ambientales debidas a modificaciones de la hidrología local, como cambios en el caudal de los arroyos, en el contenido de sales del suelo o en la profundidad de la napa freática (agua subterránea). En un estudio orientado a evaluar el impacto de las forestaciones sobre el ciclo hidrológico, se seleccionaron al azar 10 forestaciones en el partido de Zárate y, en cada una ellas, se midió la profundidad de la napa freática (en metros) en el centro de la plantación y en el pastizal vecino a la misma. Los datos obtenidos figuran en la tabla:

Profundidad de napa [m]

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Forestación | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Centro plantación | 2.0 | 2.3 | 2.2 | 2.0 | 2.3 | 2.5 | 2.0 | 2.3 | 2.4 | 2.0 |
| Pastizal vecino | 1.5 | 1.6 | 1.6 | 1.5 | 1.8 | 1.8 | 1.5 | 1.6 | 1.7 | 1.4 |

1. Identificar la población, la muestra y las unidades muestrales y las variables aleatorias involucradas en este estudio.
2. Estimar la media y la varianza poblacionales de las diferencias en la profundidad de napa entre pastizales y forestaciones.
3. Dando por cierto que el desnivel de napa freática entre el centro de una forestación a elegir al azaren el partido de Zárate y su pastizal vecino es una variable aleatoria con distribución Normal, poner a prueba la hipótesis que dice que, en Zárate, las forestaciones han determinado un aumento en la profundidad promedio de la napa freática.
4. cQué tipo de error se puede haber cometido en la prueba de hipótesis anterior?. Explicar su significado en términos del problema.

## 9. Girasol

El asesor de una empresa agricola del Oeste de la provincia de Buenos Aires necesita recomendar acerca de la posible conveniencia de cambiar la variedad de girasol utilizada en los últimos años (Contiflor) por una alternativa (Paraiso). Para ello cuenta con un conjunto de datos de los rendimientos de dichas variedades en 5 establecimientos tomados al azar en la zona.

Rendimientos de girasol (kg/ha)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Establecimiento |  |  |  |  |
| Variedad | Don Juan | Alborada | Tobiano | El Lucero | La Martita |
| Paraiso | 1805 | 2200 | 1545 | 1875 | 1430 |
| Contiflor | 1724 | 1950 | 1325 | 1685 | 1148 |

1. ldentificar la población, la muestra y las variables aleatorias involucradas en este problema.
2. ¿Qué experimento aleatorio produce las variables aleatorias en cuestión?
3. Seleccionar una prueba de hipótesis para fundamentar la recomendación acerca de la posible conveniencia de cambiar la variedad de girasol. Justificar la selección.
4. Formular y poner a prueba la hipótesis para un nivel de significación apropiado. Concluir y emitir una recomendación.
5. Qué supuesto debería verificarse para que la inferencia realizada en d sea válida?

## 10. El Palmar

La vegetación característica del Parque Nacional “El Palmar” es una sabana con palmeras altas dispersas en una matriz de vegetación herbácea y arbustiva baja. Esta sabana es progresivamente invadida por árboles de especies que antes estaban restringidas a los bosquecitos que rodean a los rios y arroyos. Un estudiante de la Escuela para Graduados Alberto Soriano de la Facultad de Agronomía proyecta un experimento para determinar si el crecimiento promedio de las plántulas de Allophylus edulis (chal-chal), una de estas especies de arboles, difiere entre plántulas que están ubicadas bajo la copa de alguna palmera y plántulas que están ubicadas lejos de cualquier palmera. Para ello, seleccionará al azar 20 plántulas de Allophylus edulis establecidas bajo la copa de alguna palmera y 20 establecidas a más de 5 m de la palmera más cercana. Luego registrará el crecimiento de cada plántula durante un año, medido como la suma del crecimiento de todas sus ramas [cm].

1. Identificar las poblaciones, las muestras, las unidades muestrales y las variables aleatorias involucradas en esta investigación.
2. Explicar qué son los valores de crecimiento promedio que se intenta comparar.
3. Proponer dos causas posibles de la varianza de las variables aleatorias a registrar.

Luego de realizar el experimento propuesto, el estudiante registra los datos que figuran en siguiente la tabla:

Crecimiento total por plántula [cm]

Bajo copa de palmera: 15.1-16.0-17.0-17.6-16.3-14.4-14.6-16.4-17.5-19.6-15.4-18.0-14.2-15.8-15.0-17.7-19.1-15.7-18.4-18.3

Lejos de palmera: 18.2-19.5-21.1-16.1-16.0-16.4-18.6-16.9-16.6-15.5-18.5-13.2-18.3-18.1-20.3-14.8-20.5-14.8-15.9-16.9

1. A partir de los datos registrados, producir estimaciones puntales e intervalos de 95% de confianza para los valores de las medias poblacionales de los crecinmientos de las plántulas de Allophylus edulis ubicadas debajo de la copa de una palmera y de las ubicadas a más de 5 m de cualquier palmera. Presentarlas estimaciones en un gráfico de barras.
2. Formular y poner a prueba hipótesis apropiadas para realizar la evaluación propuesta. ćQué supuestos involucra la inferencia realizada en e?
3. Interpretar los resultados de la inferencia en términos del problema particular.
4. Explicar qué error se puede haber cometido. ¿Qué se hubiera podido hacer para disminuir la probabilidad de este tipo de error?

## 11. Cobre y GDP de terneros

Un técnico del INTA condujo un experimento para evaluar el efecto de la aplicación de cobre inyectable sobre la ganancia diaria de peso de terneros. Para ello, en un establecimiento de cría vacuna en el este del Chaco, tomó dos muestras aleatorias le terneros de un mes y medio de edad (pariciones tardías) y dos muestras aleatorias de terneros de dos meses y medio de edad (pariciones tempranas) hijos de vacas cebú inseminadas con semen de toros Polled Hereford. A fin de octubre pesó a todos los terneros seleccionados y luego sometió a los terneros y a las madres de una de las muestras de cada edad a un tratamiento de inyecciones de cobre y dejó a los restantes como controles sin tratar. Todos los terneros pastorearon con sus madres en campo natural hasta el destete cuatro meses después, momento en que se los volvió a pesar para calcular su ganancia de peso diario. A continuación se presenta un resumen de los datos generados en el experimento.

Ganancia de peso de los terneros [g/día]

|  |  |  |  |
| --- | --- | --- | --- |
|  | n | media | s |
| Pariciones tardias. Edad inicial: un mes y medio |  |  |  |
| + Tratados con Cu | 16 | 807.03 | 74.37 |
| + Control | 14 | 684,38 | 81,56 |
| Pariciones tempranas. Edad inicial: dos meses y medio |  |  |  |
| + Tratados con Cu | 16 | 768,00 | 63,44 |
| + Control | 18 | 736,72 | 78,55 |

1. Identificar las poblaciones, las unidades muestrales, las muestras y los tratamientos.
2. Formular y poner a prueba hipótesis para establecer si la aplicación de cobre resulta en mayor ganancia de peso diario promedio de los terneros de pariciones tardias (alfa = 0,01).
3. Formular y poner a prueba hipótesis para establecer si la aplicación de cobre resulta en mayor ganancia de peso diario promedio de los terneros de pariciones tempranas (alfa = 0,01).
4. ¿Qué error podría haber cometido en cada una de las dos pruebas realizadas? Describirlo en términos de este problema.
5. ¿Bajo qué supuestos son válidas las inferencia realizadas en b y c?

## 12. Gluten

La apiitud de la harina de trigo para panificación depende principalmente de su contenido de un complejo proteico denominado gluten. Para evaluar la posible influencia del sistema de secado del grano sobre su contenido de gluten, se seleccionaron al azar dentro de la provincia de Buenos Aires 12 plantas de acopio que utilizan un sistema de secado prolongado a baja temperatura y 15 plantas de acopio que utilizan un sistema de secado rapido con alta temperatura y se determinó el contenido de gluten del trigo (g/100g) procesado en cada una. Los datos obtenidos son los siguientes:

Contenido de gluten [g/100 g]

Baja temperatura: 25.9-25.3-25.1-26.6-25.5-24.7-24.6-27.0-25.2-25.2-25.0-28.6

Alta temperatura: 27.7-22.2-22.9-24.2-22.8-23.8-24.8-26.3-23.3-24.4-24.9-21.5-26.3-24.6-25.6

1. Identificar las unidades muəstrales, las muestras y las poblaciones.
2. Formular hipótesis apropiadas para evaluar si el contenido promedio de gluten es afectado por el sistema de secado.

Dando por cierto que los datos obtenidos son realizaciones de variables aleatorias con distribución Normal con idéntica varianza.

1. Poner a prueba la hipótesis nula con un nivel de significación alfa = 0,05.
2. Construir un intervalo de confianza 0,95 para la diferencia entre los promedios de contenido de gluten de trigo procesado con baja y con alta temperatura.
3. Explicar qué conclusión debería extraerse en relación con este fenómeno de interés y enunciar los supuestos bajo el cual dicha conclusión sería válida.