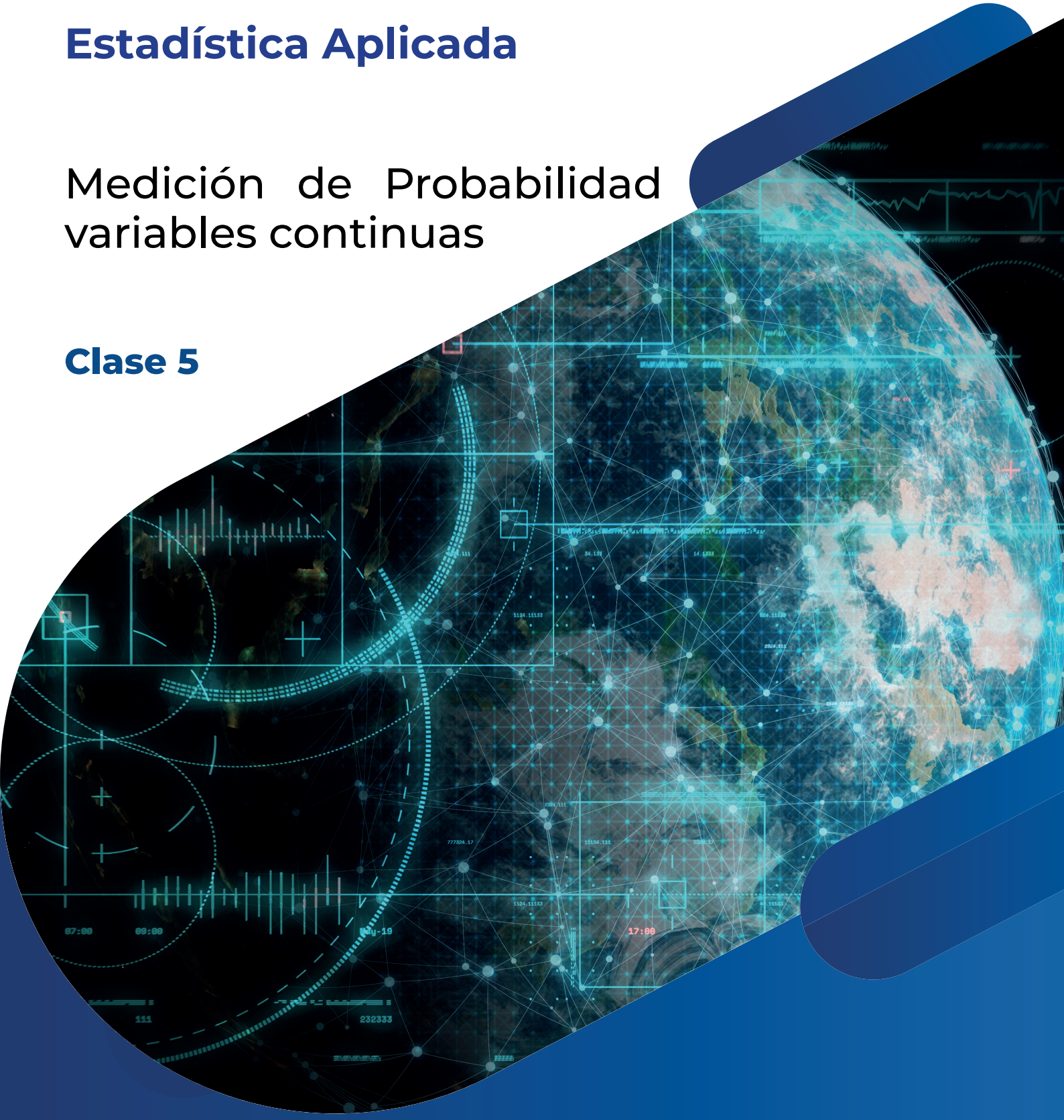
Estadística Aplicada

Medición de Probabilidad variables continuas

Clase 5



Ingeniería en ciberseguridad

La excelencia no se improvisa

1. INTRODUCCIÓN DE LA CLASE

En la clase anterior habíamos examinado la distribución normal y su importancia en la estadística inferencial. Sin embargo, cuando revisamos el concepto de muestra, mencionamos la condición de que la misma debe ser representativa con respecto a la población. Sin embargo, esto no siempre es posible, como explicaremos más adelante.

En esta clase veremos por qué no siempre podemos utilizar el intervalo de confianza de 1.96 cuando el tamaño de la muestra es pequeño. En su lugar, aprenderemos cuáles serían los valores apropiados en estos casos. Por otro lado, esta distribución está relacionada con las pruebas de hipótesis realizadas con la función t.test , que es un tema de vital importancia en la estadística.

Clase 1:

Resultados de aprendizaje que serán abordados con el contenido de la clase:

Calcular e interpretar estadísticos y parámetros para describir muestras y poblaciones

Reto # 2

5) Medición de Probabilidad variables continuas

En las clases anteriores habíamos mencionado el concepto de muestra y establecido que esta debe ser una parte representativa de la población que se selecciona para ser estudiada, ya que trabajar sobre la población generalmente es demasiado grande para hacer un trabajo de investigación viable. Por otro lado, si la población está estimada en cientos de miles o millones, entonces la muestra debería ser grande.

Pero trabajar con muestras grandes no siempre es posible; por ejemplo, un fabricante de autos debe certificar la resistencia de los modelos a choques frontales para salvaguardar la seguridad de los ocupantes, pero cada experimento que se realice será muy costoso debido a que terminará con la destrucción del vehículo. Esto puede ser muy costoso a largo plazo. En otros casos, puede ser imposible contar con suficientes muestras. Por ejemplo, un laboratorio farmacéutico desea probar un nuevo fármaco que ha desarrollado. La población objetivo puede ser muy grande, de modo que se necesitaría una muestra grande. Pero el laboratorio podría tener problemas para conseguir suficientes candidatos dispuestos a someterse como voluntarios para probar el fármaco.

De acuerdo con Webster, A. (2000), cuando estudiamos el proceso general de muestreo, estadístico e inferencia del parámetro, indicamos que, si la muestra es representativa, podríamos, por ejemplo, inferir el sigma de la población en base al s de la muestra. Pero con muestras pequeñas, la diferencia entre sigma y s puede ser significativa; más aún, si se realizaran 2 procesos de muestreo con pocas observaciones, los resultados de la inferencia poblacional serían diferentes. Aquí entra la distribución T.

La T es una familia de distribuciones de probabilidad continua utilizada al trabajar con poblaciones distribuidas normalmente, pero donde el tamaño de la muestra es pequeño y se desconoce la desviación estándar de la población.

5.1) Distribución T

La distribución t se utiliza cuando se cumplen tres condiciones:

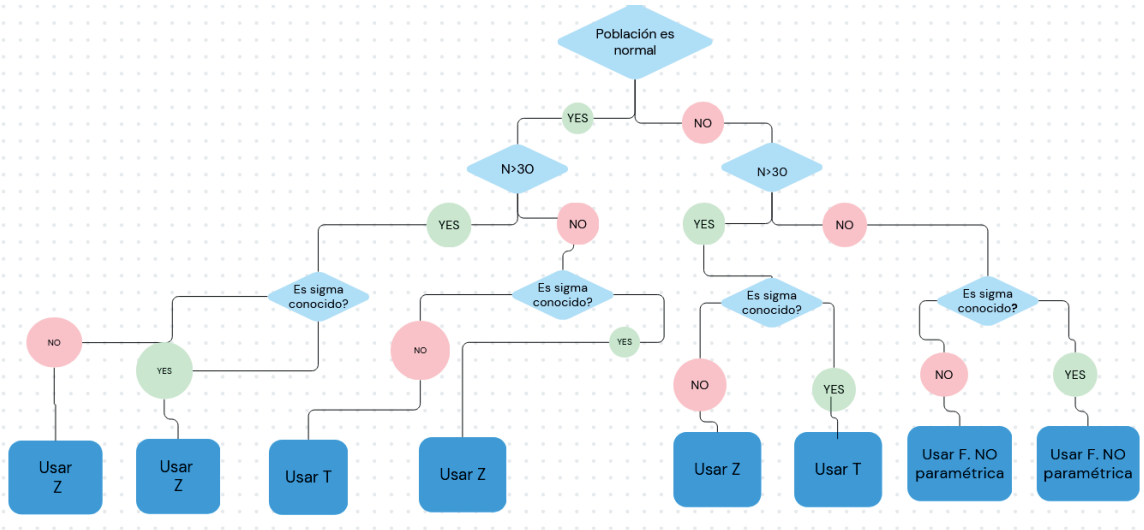
1. La muestra es pequeña. La pregunta que surge es qué se considera pequeño. Como veremos más adelante, los valores del estadístico t se aproximarán a los del estadístico Z cuando la cantidad de observaciones sea mayor a 30. Es decir, si, por ejemplo, estamos tratando con una muestra de 10 observaciones, las diferencias entre Z y t serán significativas.
2. Sigma (σ) es desconocida. Normalmente, no conocemos la varianza y la desviación de la población. En la clase anterior, dijimos que, si no se conoce sigma, esta puede ser reemplazada por $s/\sqrt{n}$, pero con muestras pequeñas la diferencia entre sigma (σ) y s puede ser importante. Sin embargo, si σ es conocida, la distribución Z se usa incluso si la muestra es pequeña.
3. La población presenta una distribución normal o casi normal. Esto es así dado que con muestras pequeñas es difícil asegurar que esta sigue una distribución normal. Como veremos más adelante, existen varias funciones que pueden indicar si una muestra es normal, pero en este caso, con muestras pequeñas, debemos recurrir a funciones no paramétricas.

En base a los conceptos mencionados, podemos establecer el siguiente mapa de decisión.

Figura 1

Mapa de decisión para usar distribución Z o T

Creación autor Alfonso Prado



TEOREMA DEL LÍMITE CENTRAL

Este teorema indica que, a medida que el tamaño de la muestra, n, se haga más grande, la distribución muestral de las medias de la muestra tenderá a seguir una distribución de probabilidad normal con una media igual a la media poblacional real, μ, y un error estándar de la media $\sigma x = \sigma/\sqrt{n}$.

Esto es así independientemente de la distribución, que podría ser Student T, Poisson, Binomial, etcétera, de la que se extrajo la muestra.

El teorema del límite central nos asegura que si tomamos una muestra lo suficientemente grande ($n \geq 30$), entonces la distribución muestral se distribuirá normalmente, independientemente de la distribución de la población en sí. Si la población de la que se extrajeron las muestras no es normal o si simplemente no sabemos si la población es normal o no, entonces el teorema del límite central se mantendrá siempre que tengamos una muestra grande de 30 o más.

En las clases, validaremos este concepto para las distribuciones de Poisson y Binomial

Estadístico T

Al igual que el concepto del estadístico Z, la distribución T cuenta con un estadístico llamado t que nos indica qué tan alejada de la media se halla un valor x.

Este indicador nos sirve para analizar la probabilidad en la distribución t. Al igual que el Z en la distribución gaussiana, la distribución T es una distribución normal con forma de campana, pero la distribución normal es más alta en el centro y más baja en ambos extremos o colas. El área debajo representa la probabilidad de las observaciones, por lo que es probable que la mayoría de los puntos de datos estén cerca del centro del gráfico, mientras que es probable que haya menos puntos en las secciones más altas o bajas, representadas por el área más pequeña hacia las colas del gráfico.

Debido a lo anterior, la probabilidad PDF y CDF de T no son iguales a la distribución gaussiana o normal en todos los casos.

De acuerdo con Rumsay (2009) En muchos casos, no se conoce la desviación estándar de la población, σ. Para estimar la media de la población utilizando un intervalo de confianza cuando se desconoce σ, se utiliza la fórmula siguiente.

Figura 2

Fórmula de estadístico t

$$t = \frac{(x - \mu)}{\left(\frac{S}{\sqrt{n}}\right)}$$

Donde:

μ= media de la población

x = media de la muestra

S= desviación estándar de la muestra

n=Tamaño de la muestra

GRADOS DE LIBERTAD EN DISTRIBUCIÓN t

Decimos que la distribución T es una "familia" debido a que su forma y probabilidades varían dependiendo del concepto de Grados de Libertad, que describimos a continuación:

En general, podemos decir que una variable puede tomar una cantidad infinita de valores. Pero cuando existen restricciones, entonces no todos los valores son posibles. Veamos un ejemplo: Se dispone de un conjunto de valores como {x1,x2,x3,x4}, para una variable, pero hay una restricción, digamos que la suma debe ser 20. Entonces, si en un momento determinado tengo {1,4,5,}, las primeras 3 variables pueden tomar cualquier valor, es decir, tengo 3 grados de libertad. Pero la cuarta está sujeta a la restricción (tiene que ser 10). Cuando se toma una muestra, de longitud n, n-1 vienen a ser los grados de libertad de la variable.

En general, definimos los grados de libertad como el número de observaciones menos el número de restricciones impuesta sobre tales observaciones.

CARACTERÍSTICAS DE LA DISTRIBUCIÓN T

Una distribución t es simétrica. Es una distribución en forma de campana que asume la forma de una distribución normal y tiene una media de cero.

Su distribución está parametrizada en base a los grados de libertad v =n-1

donde n es el tamaño de la muestra.

Su varianza=v/(v - 2), donde v representa el número de grados de libertad o en función de n como se muestra en la siguientes figura.

Figura 3

Varianza de la distribución T

$\sigma 2=(n-1)/(n-3)$

La varianza es mayor que 1 en todo momento. Sin embargo, tenga en cuenta que se acerca mucho a uno cuando hay muchos grados de libertad. Con una gran cantidad de grados de libertad, una distribución t se parece a una distribución normal.

Las colas de una distribución t son más gruesas y menos puntiagudas que las de una distribución normal, lo que indica una mayor probabilidad en las colas.

La forma de una distribución t cambia con el cambio en los grados de libertad. Cuanto mayor sean los grados de libertad, mayor será la probabilidad de que ocurra y cuanto mayor sea la magnitud, más se parecerá la forma de una distribución t a una distribución normal estándar.

La densidad de probabilidad PDF se calcula con la siguiente fórmula.

Figura 4

Función de densidad (PDF) de la distribución T

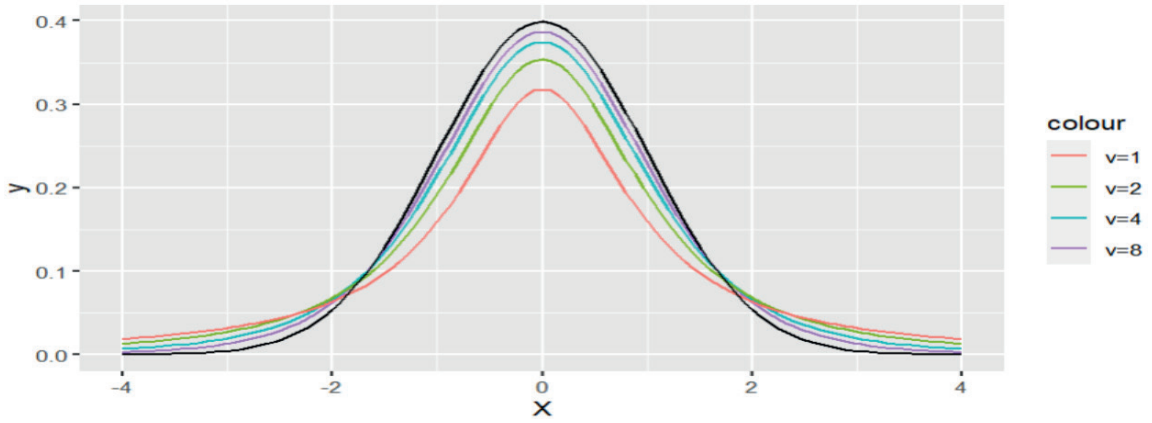
$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Mediante la formula anterior, podemos realizar el siguiente diagrama de la distribución en función de los grados de libertad.

Figura 5

Distribución T en función de grados de libertad

Creación de autor Alfonso Prado



Tema Relacionado

Descripción Grados de libertad

Enlace

5.2) funciones de distribución T

Funciones programáticas distribución t

Al igual que en la distribución normal, los prefijos usados para las funciones de distribución siguen siendo los mismos: **d** para densidad, **p** para probabilidad acumulada, q para obtener el inverso de **p** y **r** para crear un vector con determinados grados de libertad, y el sufijo será siempre **t**.

Figura 6

Funciones programáticas de distribución T

Función de Distribución	Distrib. Normal	
PDF	dt()	dt(x, v, ncp, log = FALSE)
CDF	pt()	pt(q, v, ncp, lower.tail = TRUE, log.p = FALSE)
Inverso CDF	qt()	qt(p, v, ncp, lower.tail = TRUE, log.p = FALSE)

Función para obtener un set de datos de la distribución	Distrib. Normal	
	rt()	rt(n, v, ncp)

Donde

v= grados de libertad

x=vector de valores

q=vector de cuantiles

p=vector de probabilidades

lower.tail=TRUE para calcular P[X≤x] , FALSE para calcular P[X>x]

Descripción de la Funciones

dt proporciona la densidad, pt proporciona la función de distribución, qt proporciona la función de cuantiles y rt genera desviaciones aleatorias.

Los argumentos no válidos generarán el valor de retorno NaN, con una advertencia.

Intervalo de confianza para la media poblacional

Una vez entendido el cálculo del estadístico **t**, podemos usarlo para inferir la media poblacional. Nótese que es similar a la fórmula utilizada en la distribución **Z**, pero en este caso sustituiremos el estadístico **Z** por **t**. Nuestra fórmula quedaría como se describe en la siguiente figura.

Figura 7

Intervalo de confianza para la media poblacional

$$IC \text{ para la media} = X \pm t^*\sigma/\sqrt{n}$$

Donde:

X es la media muestral

t es el estadístico calculado con los grados de libertad

¿Ahora nos preguntamos qué pasa si tengo 2 muestras de la misma población, pero con distinta cantidad de observaciones? Obviamente, los grados de libertad serán diferentes y posiblemente la desviación estándar también lo sea.

¿Cada una de estas muestras nos daría una media estimada para la población diferente? ¿Cuál vale?

Aquí tenemos dos casos: que las varianzas sean iguales entre las dos muestras y que las varianzas sean diferentes.

Si las varianzas son iguales, vamos a calcular una varianza ponderada, en donde los pesos son los grados de libertad n – 1 para cada muestra. Esta estimación ponderada se muestra en la siguiente figura:

Figura 8

Varianzas ponderadas

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

El intervalo de confianza para la diferencia entre las dos medias poblacionales se halla entonces con una distribución t con grados de libertad n1 + n2 - 2.

Temas Relacionados

Varianzas Ponderadas

Enlace

Figura 9

Intervalo de confianza para la diferencia entre medias poblacionales cuando $\sigma_1 = \sigma_2$

$$\text{I.C. para } (\mu_1 - \mu_2) = (\bar{X}_1 - \bar{X}_2) \pm t\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

Por otro lado, cuando las varianzas son diferentes se puede aproximar los grados de libertad según la siguiente figura

Figura 10

Grados de libertad ponderados cuando $\sigma_1 \neq \sigma_2$

$$\text{g.l.} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}$$

Y el intervalo de confianza para la diferencia entre medias poblacionales se calcularía de acuerdo con la sigui-

ente figura

Figura 11

Intervalo de confianza para la diferencia entre medias

$$\text{I.C. para } \mu_1 - \mu_2 = (\bar{X}_1 - \bar{X}_2) \pm t' \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Veamos algunos ejemplos de como se puede usar la distribución t.

Un fabricante de focos asegura que su producto dura 500 horas. Es imposible hacer un muestreo grande para validar, esta es una pista que debemos usar distribución T.

Se hace una muestra con 25 focos. La media obtenida es 505.36 y una s=12.07

Pregunta: ¿Está o no garantizada la calidad mencionada por el fabricante?

Para resolver este problema podemos hacerlo de dos maneras distintas:

a) Calculando la media poblacional

```
#Valores obtenidos del enunciado

n=25

media_muestral=505.36

s=12.07

mu=500

#Obtenemos el valor de t critico para IC=95%

tcritico<- qt(0.95,24)

tcritico

#aplicamos la fórmula

IC_bajo <- media_muestral - (tcritico*s/sqrt(25))

IC_alto <- media_muestral + (tcritico*s/sqrt(25))´

 IC_bajo

[1] 501.2299

 IC_alto

[1] 509.4901
```

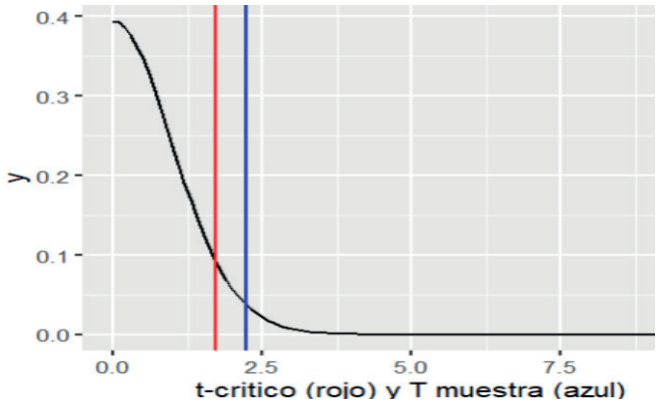
b) Comparando Tcalculado vs. Tcrítico. En este caso, vamos a calcular qué tan distanciado está la muestra con respecto a un Tcrítico obtenido para el 95% de intervalo.

```
#Calculo el valor de t

n=25

media_muestra=505.36

s=12.07

mu=500

#De acuerdo a la figura 2

tcalculado<- (media_muestra- mu)/(s /sqrt(n))

tcalculado

[1] 2.220381

#Obtenemos el valor de t-critico para 95% de confianza,

intconf =0.95

gl <-24 # 25 -1 grados de libertad

tcritico<-qt(intconf,gl, lower.tail=TRUE)

tcritico

1.710882
```

En base a este último código obtenemos la siguiente figura

Figura 12

Resolución analítica del problema anterior



En la gráfica podemos notar que la recta roja corresponde al Tcritico al 95%, es decir, que debajo de esa línea se ubicarán el 95% de las posibilidades. Y la recta azul corresponde al t-calculado de la muestra, que se encuentra hacia la derecha, esto indica que la calidad de los focos es todavía mejor que lo garantizado por el fabricante.

Ejercicio 2

En la cafetería de los estudiantes de la PUCE, existen máquinas expendedoras de café. Una muestra de 15 tazas arroja una media de 15.3 onzas, con una varianza de 3.5. Estudiantes indican que la cantidad de café dispensado no es la correcta. Se solicita al proveedor recalibrar la máquina, después de lo cual se realiza una nueva medición: una muestra de 10 tazas produce un promedio de Primero debemos darnos cuenta de que lo que solicita el problema es la diferencia entre medias, considerando que existen dos muestras pequeñas con varianza igual. Por lo tanto, las fórmulas que aplican son las de las figuras 8 y 9.

```
n1 <- 15

media1 <- 15.3

var1 <- 3.5

gl1 <- n1-1

n2<- 10

media2 <- 17.1

var2 <- 3.9

IC=0.975 #incluye cola izquierda

gl2 <- n2-1


#En primer lugar calcularemos la varianza ponderada al cuadrado

sp2 <- (var1*gl1 +var2*gl2)/(n1+n2-2)

sp2

#Segundo tenemos que calcular los grados de libertad ponderados

tcritico <- qt(IC,(n1+n2-2))

tcritico

#Por último calculamos el intervalo

ID_bajo <- (media1-media2) - tcritico*sqrt((sp2/15)+(sp2/10))

[1] -3.414905

ID_alto <- (media1-media2) + tcritico*sqrt((sp2/15)+(sp2/10))

[1] -0.1850947
```

Análisis: Revisando la respuesta de los intervalos de la diferencia (ID_bajo , ID_alto) notamos que esta diferencia no incluye el valor de 0 , notamos que esta diferencia no incluye el valor 0, por lo tanto, podemos concluir que sí existió una diferencia después de la recalibración.

Referencias citadas en la Clase 1.

Webster A., (2000), Estadística aplicada a los negocios, Irwin Professional Publishing

Rumsey D. (2009), Statistics II, Wiley Publishing Inc stats package - RDocumentation. (n.d.). https://www.rdocumentation.org/packages/stats/versions/3.6.2

Definición de los términos citados en la Clase 5.

Inferencia estadística	La inferencia estadística es el conjunto de métodos y técnicas que permiten inducir, a partir de la información empírica proporcionada por una muestra, cual es el comportamiento de una determinada población con un riesgo de error medible en términos de probabilidad.
Intervalo de confianza	El intervalo de confianza describe la variabilidad entre la medida obtenida en un estudio y la medida real de la población (el valor real). Corresponde a un rango de valores, cuya distribución es normal y en el cual se encuentra, con alta probabilidad, el valor real de una determinada variable. Esta «alta probabilidad» se ha establecido por consenso en 95%. Así, un intervalo de confianza de 95% nos indica que dentro del rango dado se encuentra el valor real de un parámetro con 95% de certeza

Profundización Clase 1.

Distribución T

Describe el uso de la función dt

Enlace

Distribución T

Describe el uso de las funciones pt y qt

Enlace

PUCE Virtual

La excelencia no se improvisa

síguenos