PUCE Virtual
EDUCACIÓN FLEXIBLE

Estadística 1 para administración

Medición de Probabilidad variables continuas 2

Clase 5

La excelencia no se improvisa

INTRODUCCIÓN DE LA CLASE 5

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. Pero 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 pequeña, en su lugar aprenderemos cuales 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.

Medición de Probabilidad variables continuas 2

Distribución T
Empezaremos con algunas definiciones básicas

TEOREMA DEL LÍMITE CENTRAL

Este teorema indica que a medida que el tamaño de la muestra, n, se hace más grande, la distribución muestral de las medias de la muestra tiende 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 etc. 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 la clase 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 que 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 más 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.
De acuerdo a 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 utilizará la fórmula siguiente.

Figura 1
Fórmula de estadístico t Prado A. (2025)
Adaptación de la fórmula para la resolución intervalos de confianza según Webster (2000) capitulo 7

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

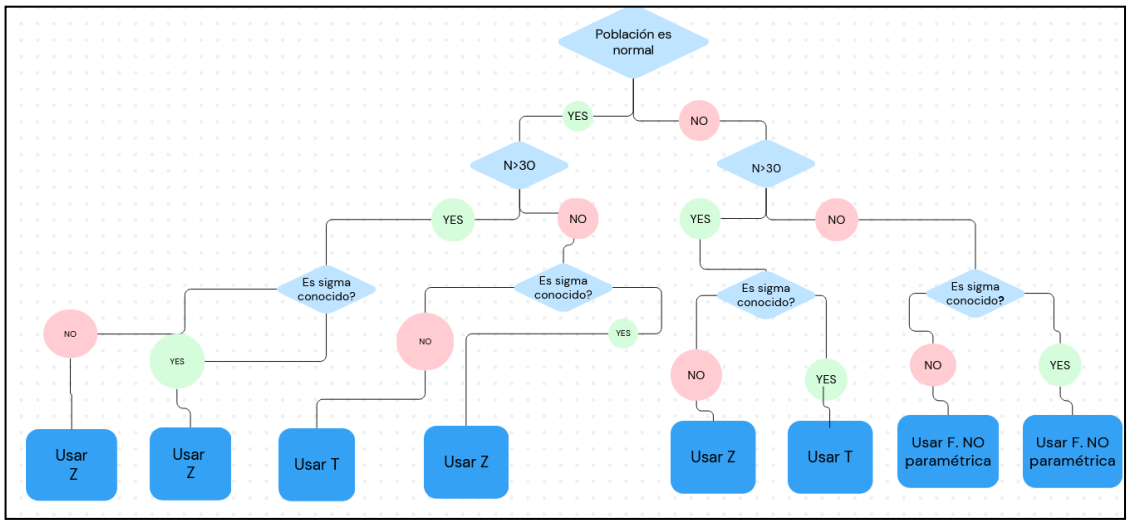
Donde:
μ= media de la población
x = media de la muestra
S= desviación estándar de la muestra
n=Tamaño de la muestra

La distribución t se utiliza cuando se cumplen las tres condiciones:

1. **La muestra es pequeña**. La pregunta que surge es que se considera pequeño. Como veremos más adelante, los valores del estadístico t se aproximarán a los de estadístico Z cuando la cantidad de observaciones se encuentre sobre 30 observaciones. 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 desviación de la población. En la clase anterior habíamos dicho que, si no se conoce sigma, esta puede ser reemplazado por s/√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 inclusive si la muestra es pequeña
3. **La población presenta una distribución normal o casi normal.** Esto es obvio dado que con muestras pequeñas es difícil asegurar que ésta 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 deberemos recurrir a funciones no paramétricas.

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

Figura 2
Mapa de decisión para usar distribución Z o T
Creación autor Alfonso Prado



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 tengo una restricción, digamos que la suma de los elementos 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.

Esto se debe a que la suma de las desviaciones de las observaciones con respecto a la media ($x_1 - \bar{x}$, $x_2 - \bar{x}$, $x_3 - \bar{x}$, $x_4 - \bar{x}$ y $x_5 - \bar{x}$) debe ser igual a cero. Si utilizásemos n para calcular la varianza, estaríamos subestimando la variabilidad de la población. Por ello, recurrimos a n – 1 para conseguir una estimación no sesgada de la varianza de la población.
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 se semeja a la forma de una distribución normal y tiene una media de cero.
- ✓ Su distribución está parametrizada en base al concepto de 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.
- ✓ La distribución t tiene una media de cero.

- ✓ La distribución normal supone que la desviación estándar de la población es conocida. La distribución t no hace esta suposición.

Figura 3
Fórmula para el cálculo de la varianza de la distribución T. Prado A. (2025)
Adaptación de la fórmula de la varianza de Webster (2000) capitulo 7

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

La varianza es mayor que 1 en todo momento. Sin embargo, tenga en cuenta que se acerca mucho a 1 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 sea 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 Prado A. (2025)
Adaptación de la fórmula de la probabilidad para muestras pequeñas de Webster (2000) capitulo 7

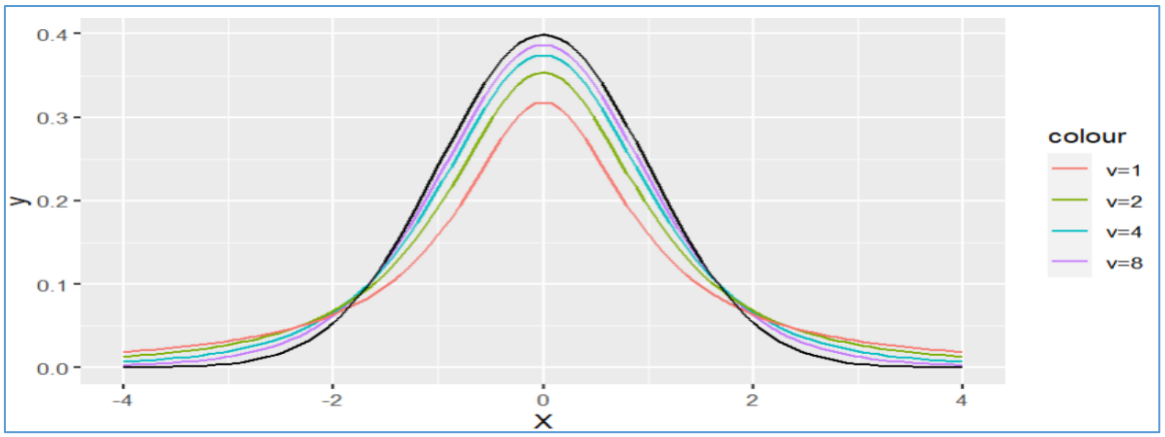
$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Donde
ν representa los grados de libertad
Γ La función gamma

Función gamma
Presenta definición y casos de uso

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



Grados de libertad

La presente revisión tiene como objetivo dar una idea intuitiva acerca de los grados de libertad y sus fundamentos

Enlace

Funciones Programáticas de la distribución T

Al igual que la distribución normal, los prefijos usados para las funciones de distribución siguen siendo las mismas: 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

Creación autor Alfonso Prado

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

La función pt() sirve para encontrar la probabilidad acumulada entre -∞ y el valor de la variable x , para un distribución con grados de libertad V dada, y el argumentos lower.tail que indica en que sentido debe acumular las probabilidades, por defecto utilizaremos el argumento lower.tail=FALSE.

La función qt() es el inverso de pt(). En pt la función toma un valor q y retorna su probabilidad acumulada, mientras que en qt toma como argumento la probabilidad acumulada y retorna el valor q debajo del cual se logra dicha probabilidad

Por último, la función rt() nos permite crear un vector de n valores cuya distribución es T y tiene grados de libertad v.

Intervalo de confianza para la media poblacional

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

Figura 7
Intervalo de confianza para la media poblacional Prado A. (2025)
Adaptación de la fórmula de Webster (2000) capitulo 7

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

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 será distinta.
¿Cada una de estas muestras nos daría una media estimada para la población diferente, cual 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
Fórmula para el cálculo de Varianzas ponderadas Prado A. (2025)
Adaptación de la fórmula de la varianza de Webster (2000) capitulo 7

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

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. Esto se debe a que para demostrar que existe diferencia entre las medias de las dos muestras, deberemos calcular el valor p usando $(n_1 - 1) + (n_2 - 1)$ grados de libertad. Esto porque hemos tenido que calcular dos medias y por lo tanto se han reducido en 2 los grados de libertad.

Varianza ponderada
Presenta definiciones y casos de uso
Enlace

Figura 9
Intervalo de confianza para la diferencia entre medias poblacionales cuando $\sigma_1 = \sigma_2$. Prado A. (2025)
Adaptación de la fórmula del Intervalo de confianza de Webster (2000) capitulo 7

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

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

Figura 10
Fórmula para el cálculo de Grados de libertad ponderados cuando $\sigma_1 \neq \sigma_2$, Prado A. (2025)
Adaptación de la fórmula según Webster (2000) capitulo 7

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

Y el intervalo de confianza para la diferencia entre medias poblacionales se calcularía de acuerdo a la siguiente figura

Figura 11
Intervalo de confianza para la diferencia entre medias con grados de libertad ponderados . Prado A. (2025)

Adaptación de la fórmula de intervalos de confianza según de Webster (2000) capitulo 7

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

Veamos algunos ejemplos de cómo se puede usar la distribución t

Un fabricante de focos asegura que su producto dura 500 horas. Es imposible hacer un muestreo significativo 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: Esta o no garantizada la calidad mencionada por el fabricante.

Para resolver este problema podemos hacerlo de dos maneras distintas:

a) Calculando la media poblacional

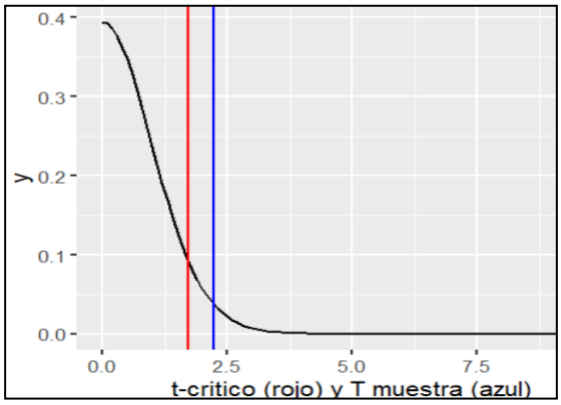
```
#Valores obtenidos del enunciado
n=25
media_muestral=505.36
s=12.07
mu=500
#Obtenemos el valor de t critico para IC=95%
tcritico<- qt(0.95,24)
tcritico
#aplicamos la fórmula
IC_bajo <- media_muestral - (tcritico*s/sqrt(25))
IC_alto <- media_muestral + (tcritico*s/sqrt(25))´
 IC_bajo
[1] 501.2299
 IC_alto
[1] 509.4901
```

b) Comparando Tcalculado vs. Tcrítico. En este caso vamos a calcular que 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
Comparación tcritico vs tcalculado
Creación autor Alfonso Prado



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 el tcalculado contiene más del 95% de los casos, 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 da una media de 15.3 onzas con una varianza de 3.5. Estudiantes indican que la cantidad de café dispensado no es correcto. Se solicita al proveedor recalibrar la máquina Después de lo cual se hace una nueva medición que muestrea con 10 tazas produce un promedio de 17.1 onzas con una varianza de 3.9. Se asume que la varianza es constante antes y después del ajuste. Queremos validar si la recalibración produjo diferencias.

Primero debemos darnos cuenta que, lo que pide el problema es la diferencia entre las medias considerando que existen 2 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, por lo tanto, podemos concluir que existió una diferencia después de la recalibración.

Referencias citadas en la Clase 5

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
The Data Visualisation Catalogue, : https://datavizcatalogue.com/

PUCE Virtual

La excelencia no se improvisa

síguenos