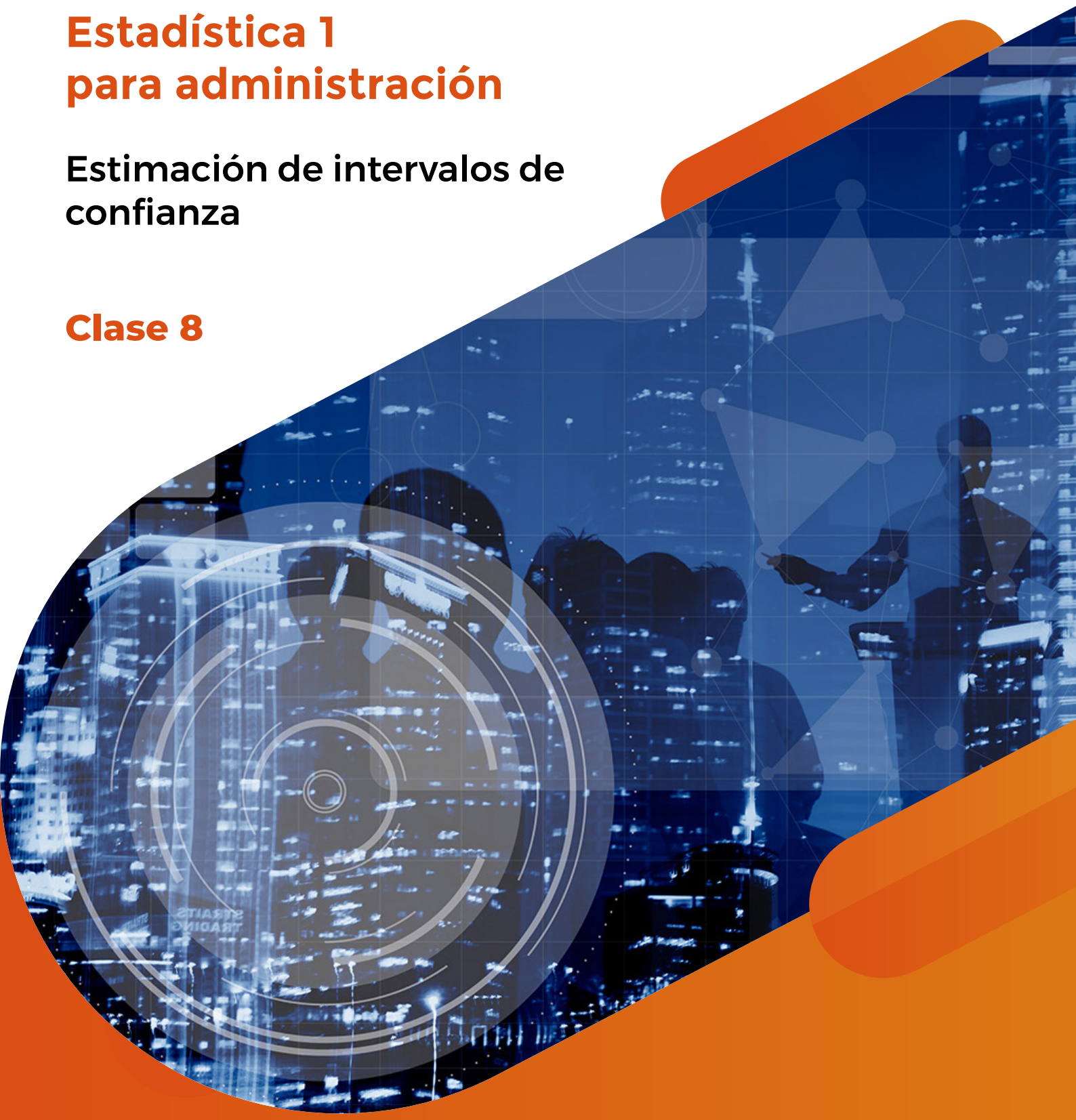
PUCE Virtual
EDUCACIÓN FLEXIBLE

Estadística 1 para administración

Estimación de intervalos de confianza

Clase 8



La excelencia no se improvisa

DESARROLLO

INTRODUCCIÓN DE LA CLASE

El análisis estadístico en investigaciones se basa en la idea clave de que hacemos observaciones sobre una muestra de sujetos y luego extraemos inferencias sobre la población de todos esos sujetos de los que se extrajo la muestra. Si la muestra del estudio no es representativa de la población, es muy posible que nos confundamos obteniendo resultados no verdaderos.

En la presente clase veremos cómo obtener los estadísticos que describen a la población tomando en cuenta la variación aleatoria en la muestra, debido a que los resultados de una sola muestra estarán sujetos a incertidumbre estadística, que está estrechamente relacionada con el tamaño de la muestra

Clase 8: Aplicar distribuciones de probabilidad con variables discretas y continuas en la solución de problemas, en distintas áreas del conocimiento.

Estimación de intervalos de confianza

¿Qué es inferencia?

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.

Los métodos paramétricos de la inferencia estadística se pueden dividir, básicamente, en dos:

- ✓ Métodos de estimación de parámetros
- ✓ Métodos de contraste de hipótesis.

Ambos métodos se basan en el conocimiento teórico de la distribución de probabilidad del estadístico muestral que se utiliza como estimador de un parámetro. (http://www.ub.edu/)

La inferencia nos permitirá trazar una línea lógica de causa-efecto, entre los diferentes puntos inferidos en la resolución del problema.

¿Qué es la estimación puntual?

Una estimación puntual es un valor único que describe mejor la población de interés, siendo la media de la muestra y la proporción de la muestra los más comunes.

¿Qué es un intervalo de confianza?

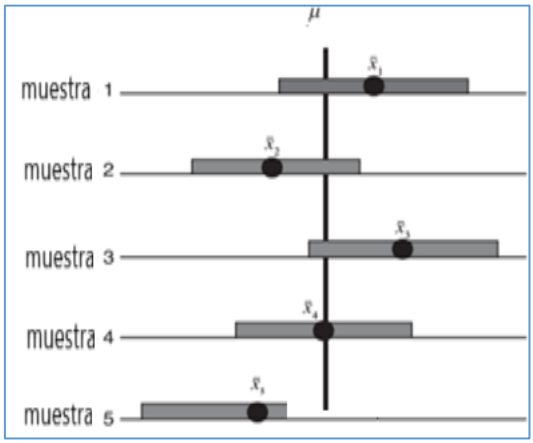
Existen muchas definiciones, de las cuales extraemos las más relevantes:

- ✓ Una estimación de intervalo combinada con una declaración de probabilidad
- ✓ Estimación de intervalo calculada a partir de estadísticas de los datos observados, que pueden contener el valor verdadero de un parámetro de población desconocido. El intervalo tiene un nivel de confianza asociado que califica el nivel de confianza de que un parámetro se encuentra en el intervalo
- ✓ Un rango de valores tan definido que existe una probabilidad específica de que el valor del parámetro se encuentre dentro de él.

De lo anterior podemos extraer las siguientes conclusiones: Es un intervalo de valores, es decir, este intervalo se encuentra entre un valor a y b. Dentro de este intervalo creemos que se encuentra una media poblacional de la variable bajo análisis. Y aunque no sabemos exactamente en que parte de este intervalo, podemos asegurar que existe una probabilidad concreta de que esto sea correcto.

l y sabemos que una muestra dada tendrá una media que se halla dentro de un nivel de significancia definido con la probabilidad de un error (el valor α)
Para el intervalo de confianza del 95 por ciento, el 95 por ciento de todos los intervalos contendrá la media de la población. El 5 por ciento restante podrían caer fuera del intervalo por temas de azar (muestra 5 en la figura 1).

Figura 1
Múltiples muestras y su media poblacional
Creación del autor Alfonso Prado

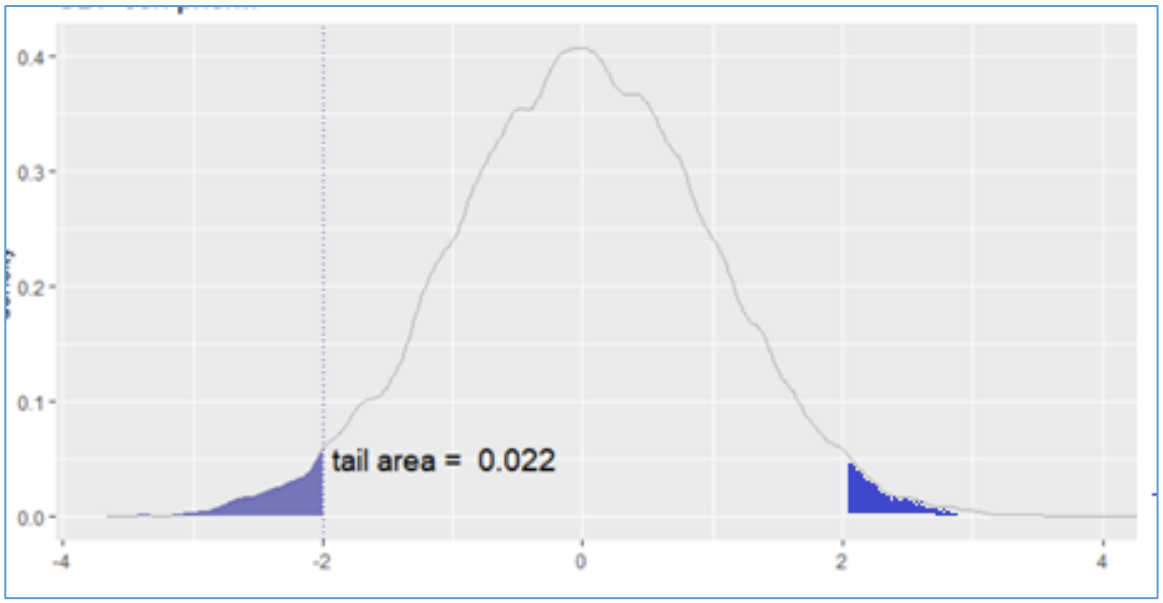


Notamos que la muestra 5 de los intervalos no contiene la media de la población. Estos caen fuera en lo que denomina valor alfa (α), el nivel de significancia α representa la frontera de colas de la distribución y como la curva es simétrica, tendremos α/2 a cada lado.

Figura 2

Distribución normal y tamaño de colas

Creación de autor: Alfonso Prado



Los niveles de confianza α están relacionados con los valores Z como se explica en la siguiente tabla:

Tabla 1

Relación entre valores Z y α

Creación del autor Alfonso Prado

1 - α	α/2	$z_{\alpha/2}$
0.90	0.05	1.645
0.95	0.025	1.96
0.99	0.005	2.575

Por otro lado, es importante entender la relación entre el valor Z y la precisión. Como habíamos visto en

clases anteriores mientras más alto es el valor de Z, el intervalo se hará más grande pero la precisión de nuestra inferencia será inferior.

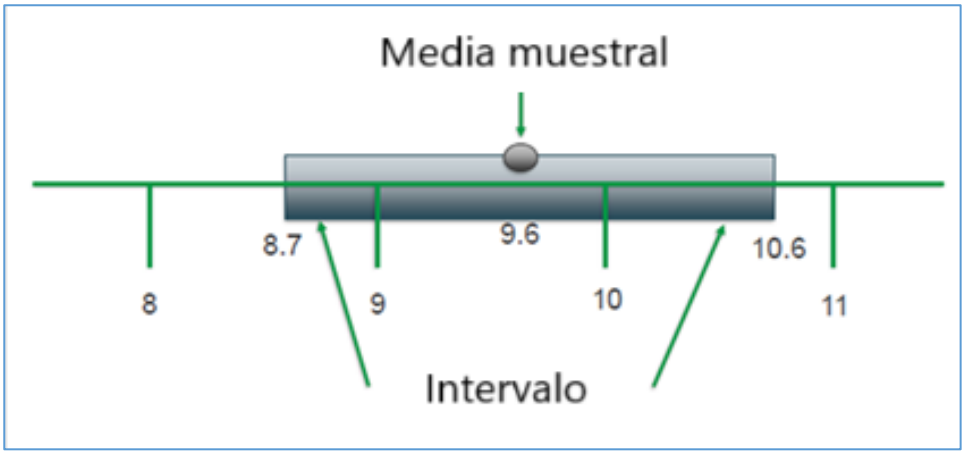
Veamos un ejemplo: Tenemos una población de personas. Esta población tiene un parámetro, vamos a decir la edad. Tenemos un rango de edades de interés, digamos 8.7 a 10.6 años. Como ya hemos visto en la distribución, si yo muestreo, tengo una probabilidad de que aparezcan sujetos con ciertas edades de lo cual se puede establecer el intervalo.

Si yo defino este intervalo de 0-100 años, tengo gran amplitud, pero no tengo precisión, si tengo poca amplitud tengo más precisión. La amplitud va a estar dada por los valores Z como se puede ver en la tabla 1.

Figura 3

Media muestral e intervalo

Creación de autor Alfonso Prado



Entonces definimos el intervalo de confianza para media (μ) con una confianza de (1-α) al intervalo (8.7 a 10.6 años) obtenido para una realización muestral dada, tal que: $P(8.7 \leq \mu \leq 10.6) = 1 - \alpha$.

El grado de confianza más habitual que se presenta es del 95%, pero no existe estandarización al respecto, y depende del estudio que se esté realizando.

Por tanto, un único estudio suele dar una estimación muestral imprecisa del valor total de la población en la que estamos interesados. Esta imprecisión se indica mediante la amplitud del intervalo de confianza:

Cuanto más amplio sea el intervalo, menor será la precisión. La amplitud depende esencialmente de tres factores.

- ✓ El tamaño de la muestra: los tamaños de muestra más grandes darán resultados más precisos con intervalos de confianza más estrechos. En particular, los intervalos de confianza amplios enfatizan la falta de fiabilidad de las conclusiones basadas en muestras pequeñas.
- ✓ La variabilidad de la característica que se estudia: cuanto menos variable sea (entre sujetos, dentro de los sujetos, a partir del error de medición y de otras fuentes), más precisa será la estimación muestral y más estrecho el intervalo de confianza.
- ✓ El grado de confianza requerido: cuanto mayor sea la confianza, más amplio será el intervalo.

Intervalo de confianza clásico

Se basa en la distribución y se forma utilizando la media muestral como una estimación puntual para el cual se adiciona y se resta un cierto valor para obtener los límites superior e inferior del intervalo de confianza respectivamente.

Figura 4
Fórmula Intervalo de Confianza clásico de la media poblacional (Prado A. 2025)
Adaptado de la fórmula intervalo de confianza según Webster(2000) capítulo 7

$$\text{CI for } \mu = \bar{x} \pm z * \underbrace{\sigma / \sqrt{(n)}}_{\text{SEM}}$$

Donde:

μ= media poblacional

σ=desviación poblacional

Z=Nivel de confianza requerido (ver tabla 1)

SEM=Error estándar de la media

$\bar{X}$ = media muestral

n=número de observaciones

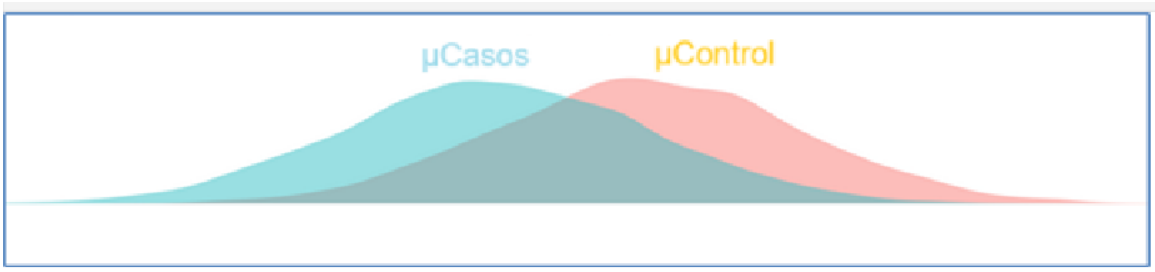
Por otro lado, muchas veces nos topamos con el problema de comparar dos medias poblacionales, esto puede darse porque, ya sea tenemos dos muestras de la misma población y queremos saber si existen diferencias, pero más comúnmente es porque

estamos tratando de encontrar diferencias entre sets que han sido sometidas a cierto tratamiento. En un experimento científico típico, nos interesan dos poblaciones (Çasos y Control) y si existe una diferencia entre sus medias (µCasos - µControl).

Figura 5

Diferencias entre medias en un estudios de Casos vs. Controles

Creación del autor Alfonso Prado



Aquí tenemos algunas situaciones diferentes que se pueden presentar dependiendo de las varianzas, los grados de libertad, si la muestra es pequeña o grande y del objeto de la prueba.

Cuando sigma es conocido muestras grandes

Tenemos 2 poblaciones cuyas muestras son grandes, las varianzas son conocidas, en este caso utilizaremos el estadístico Z mediante la fórmula:

Figura 6

Intervalo de confianza para diferencia entre muestras grandes con varianzas conocidas (Prado A. 2025)

Adaptado de Webster(2000) capítulo 7 9.1

$$(u1 - u2) = \left(\overline{x_1{}^2} - \overline{x_2{}^2}\right) \pm Z\sqrt{\frac{{\sigma_1}^2}{n1} + \frac{{\sigma_2}^2}{n2}}$$

Cuando sigma es desconocido muestras grandes

Tenemos 2 poblaciones cuyas muestras son grandes, las varianzas son desconocidas, en este caso utilizaremos el estadístico Z y sustituyendo σ por s mediante la fórmula

Figura 7

Intervalo de confianza para diferencia entre muestras, σ no conocidos(Prado A. 2025)
Adaptado de Webster(2000) capítulo 7

$$(u1 - u2) = \left(\overline{x_1{}^2} - \overline{x_2{}^2}\right) \pm Z\sqrt{\frac{{s_1}^2}{n1} + \frac{{s_2}^2}{n2}}$$

Cuando sigma son iguales pero desconocidos muestras pequeñas

En este caso calcularemos primero una varianza ponderada:

Figura 8
Fórmula de Varianzas ponderadas (Prado A. 2025)
Tomado de Webster(2000) capítulo 7

$${s_p}^2 = ({s_1}^2(n_1 - 1) + {S^2}_2\,(n_2 - 1))/(n_1 + n_2 - 2)$$

Varianza ponderada

Explica funciones y ejemplos
Enlace

El IC para la diferencia entre las medias poblacionales se halla entonces con una distribución t con grados de libertad n1+n2-2 de acuerdo a la siguiente fórmula

Figura 9
Cálculo de Intervalo de confianza para diferencia para muestras pequeñas, varianzas iguales pero desconocidas (Prado A. 2025)
Adaptado de Webster(2000) capítulo 9

$$(u1 - u2) = \left(\overline{x_1{}^2} - \overline{x_2{}^2}\right) \pm t\sqrt{\frac{{s_p}^2}{n1} + \frac{{s_p}^2}{n2}}$$

Cuando sigmas son desiguales y muestras pequeñas

En este caso no se ha encontrado una distribución exacta pero podemos aplicar una aproximación mediante el cálculo de grados de libertad ponderados como se muestra en la siguiente figura

Figura 10

Fórmula para cálculo de Grados de libertad ponderados (Prado A. 2025)

Adaptado de Webster (2000) capítulo 9

$$gl = \frac{\left(\frac{{s_1}^2}{(n_1)} + \frac{{S^2}_2}{(n_2)}\right)^2}{\frac{\left(\frac{{s^2}_1}{n_1}\right)^2}{(n_1 - 1)} + \frac{\left(\frac{{s^2}_2}{n_2}\right)^2}{n_2 - 1}}$$

Para luego calcular el estadístico t´ conseguido con los grados de libertad ponderados

Figura 11

Intervalo de confianza para diferencia para muestras pequeñas, varianzas desiguales y desconocidas (Prado A. 2025)

Adaptado de Webster(2000) capítulo 9

$$(u1 - u2) = \left(\overline{{x_1}^2} - \overline{{x_2}^2}\right) \pm t' \sqrt{\frac{{s_p}^2}{n1} + \frac{{s_p}^2}{n2}}$$

Veamos un ejemplo

Una empresa de construcción está interesada en comprar un bosque para explotacion de la madera. Tiene 2 opciones, para el primero se hace una muestral con 10 árboles, estimando que produce 69.5 p.t (Un pie tablar es 1 pie de largo x 1 pie de ancho x 1 pulgada de grueso) con una desviación de 2.2 pt, en el segundo se hace una muestra de 15 árboles y se obtiene una media de 72.2 on una desviación de 4.5.

Con un intervalo de confianza del 95% queremos estimar si hay diferencias entre los dos bosques.

En ese caso lo que están pidiendo es el IC para la diferencia en las medias poblacionales

xm1 <- 69.5

n1<- 10

sd1 <- 2.2

```
n2<- 15
xm2 <- 72.2
sd2 <- 4.5

 Dado que las desviaciones son diferentes calculamos los gl ponderados
gl <- ((sd1^2/n1 + sd2^2/n2)^2)  /(   (sd1^2/(n1))^2/(n1-1)  + (sd2^2/(n2))^2/(n2-1) )
gl <- round(gl,0)

tprima <- qt(.95,gl)
IC_medias_alto <- (xm1-xm2) + tprima* sqrt(sd1^2/n1 + sd2^2/n2)
[1] -0.37
IC_medias_bajo <- (xm1-xm2) - tprima* sqrt(sd1^2/n1 + sd2^2/n2)
[1] -5
 A un 95% de confianza podemos establecer que el bosque 1 tiene un producción
inferior
```

Sin embargo, de lo mencionado, muchas veces es complicado realizar muestreos con gran cantidad de observaciones o repetir muestreos de la misma población. Aquí entra la técnica de “Bootstrap” o intervalos de confianza basados en “Bootstrap”.

Técnicas de Remuestreo

De acuerdo a Webster 2020, el Bootstrap es un procedimiento de remuestreo que se puede utilizar para estimar la distribución de muestreo de casi cualquier estadístico, como la media, la mediana y los coeficientes de regresión, estos estadísticos toman un nombre genérico θ.

Una muestra Bootstrap es una muestra de los datos con reemplazo, esto quiere decir que cuando se realiza un remuestreo, el hecho de haber seleccionado un valor no impide que el mismo valor sea seleccionado en otro remuestreo.

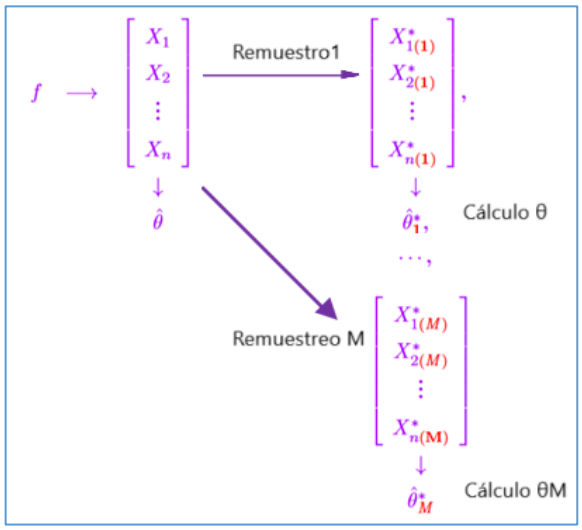
En comparación con los intervalos de confianza clásicos basados en ecuaciones, un intervalo de confianza Bootstrap no asume ninguna distribución específica. En cambio, supone que la distribución observada en la muestra es una buena representación de la

distribución de la población de la que se tomó la muestra. Por lo tanto, los intervalos de confianza Bootstrap se pueden utilizar como una alternativa si no cumplimos con los supuestos detrás de los intervalos de confianza clásicos o si no hay una ecuación para calcular dichos intervalos.

Bootstrap funciona de la siguiente forma. Asuma que tenemos una muestra de 10 observaciones, Bootstrap procede a realizar el primer remuestreo mediante la selección de 10 observaciones aleatorias, como resultado de este remuestreo es posible que algunos valores vayan a ser repetidos, así como otros valores de la muestra original podrían no aparecer. Sobre esta muestra se procede a realizar el cálculo del estadístico requerido (media, desviación etc.).
Por supuesto, la media de esta segunda muestra no tiene por qué coincidir con la media poblacional, entonces Bootstrap procede a hacer muestras adicionales, por ejemplo 1000 muestras. Mientras mayor sea la cantidad de remuestreos realizados, nos aproximaremos más a la media poblacional.

La siguiente figura explica el concepto

Figura 12
Proceso de remuestreo
Creación del autor Alfonso Prado



Luego sorteamos los valores $\theta_{1,2,3..M}$, y ubicamos los valores de 0.025 y 0.975% es decir los cuantiles 5% y 95% y habremos obtenidos el intervalo de confianza la media

poblacional. Es importante entender que dado este remuestreo, la técnica de Bootstrap no sirve para muestras pequeñas.

Recuerde el teorema del límite central, que indica que cuando la cantidad de observaciones es alta, la media muestral se aproximará a la media poblacional. En este caso mediante Bootstrap estamos generando una cantidad alta de remuestreos.
Bootstraping en investigación científica
Explica el uso de funciones para Bootstrap
Enlace

Determinación del tamaño apropiado de la muestra.

De acuerdo a Donnely: El tamaño de la muestra juega un papel importante al determinar la probabilidad de error, así como en la precisión de la estimación. Una vez que se ha seleccionado el nivel de confianza, dos factores importantes influyen en el tamaño muestral:

- ✓ La varianza de la población σ_x^2
- ✓ El tamaño del error tolerable que el investigador está dispuesto a aceptar.

Mientras que el primer factor está más allá del control del investigador (no hay nada que se pueda hacer sobre la varianza de la población), sin embargo, sí es posible limitar el tamaño del error.
Podemos determinar el tamaño n de una muestra aleatoria que conduce a un error estándar de la media que es menor que un valor de q deseado, suponiendo que conocemos la varianza (σ_x^2) . Sabemos que $\sigma_x=\sigma/\sqrt{n}$, por lo tanto, requerimos que $\sigma_x/\sqrt{(n)} < q$ despejando n , $n> (\sigma_x/q)^2$

Lo anterior puede reescribirse, de la siguiente forma:

Figura 13
Tamaño de la muestra requerido para un error dado

$$n = Z^2\,\sigma^2/(\bar{X}-\mu)^2$$

Ejemplo:

Uno de los problemas que tienen los grandes hoteles en el mundo es estimar la cantidad de habitaciones no utilizadas. Para esto deciden hacer una encuesta por noche para ver cuantas habitaciones están libres, obviamente la cantidad de habitaciones es muy grande, por lo que necesitan estimar el tamaño de la muestra. El hotel piensa que puede aceptar un error de 50 habitaciones. Se ha determinado que la desviación entre hoteles es de 165 habitaciones. ¿Cuál es el tamaño de la muestra requerida?

```
muestra <- function(err){
Z=qnorm(0.005)
Z
s=165
 aplicamos la fórmula de la figura 12
n<- (z^2)*(s^2)/err^2
n
}
muestra(50)
[1] 72.254
```

Podemos hacer el siguiente gráfico para entender esta relación

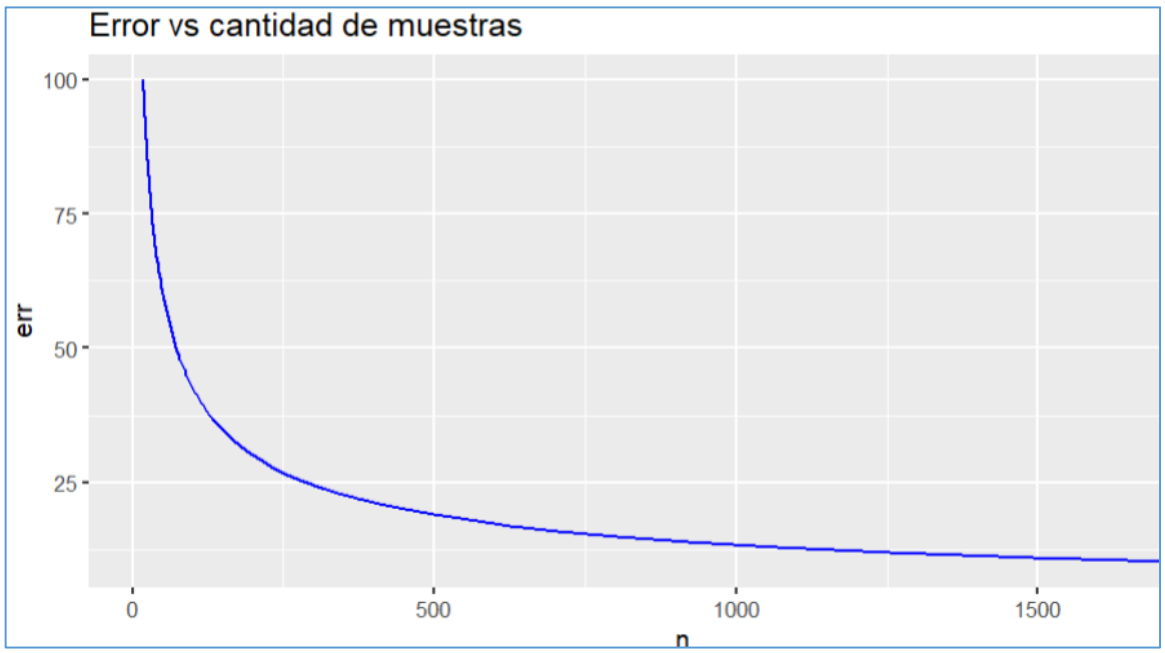
```
df <- data.frame(n=NULL, err=NULL)
for (e in seq(from=10 , to=100 , by=1))
{
   dfline<- data.frame(n=muestra(e),err=e )
 df <- rbind(df, dfline)
}
 View(df)

ggplot(data=df, aes(x=n, y=err)) + geom_line()+
   ggtitle("Error vs cantidad de muestras")
```

Notamos que en la medida que n crece la diferencia entre media muestral y poblacional disminuye

Figura 14

Relación entre el error ($\bar{\bar{X}}$ - μ) y la cantidad de muestras

Creación de autor Alfonso Prado



Referencias citadas en la Clase 8.

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

Donnelly R. , (2019), Business Statistics, Pearson Publishing

Estimación de parámetros,

https://diposit.ub.edu/dspace/bitstream/2445/125278/1/Estimaci%C3%B3n-apuntes1819-DD.pdf

(http://www.ub.edu/)

PUCE Virtual

La excelencia no se improvisa

síguenos