Intervalos de Confianza

Marzo 4, 2024

Prof. Sergio Béjar

Departamento de Estudios Políticos, CIDE

Objetivo(s)

1. Revisar la lógica del muestreo aleatorio (infinito) de una población conocida.
2. Introducir a los estudiantes a los "intervalos de confianza".
3. Reafirmar cómo hacemos inferencia de una muestra a una población.

Un Breve Desvío

En las próximas clases vamos a aprender "pruebas de hipótesis" vis a vis la muestra y la población.

- i.e. ¿Cuál es la probabilidad de nuestra muestra estadística dado el parámetro poblacional?

¿Qué es una Hipótesis?

Las hipótesis son declaraciones sobre la relación que existe entre una variable independiente y una variable dependiente.

- Variable dependiente: lo que queremos explicar.
- Variable independiente: lo que creemos que explica la variación en la variable dependiente.

¿Qué Deben Decir las Hipótesis?

Las hipótesis deben comunicar lo siguiente:

1. La causa y el efecto.
2. El tipo de relación esperada entre las variables.
3. La unidad de análisis.
4. Claridad en el tipo de medición que se usa para ambas variables.

Tipos de Relaciones Propuestas

- Positiva.
- Negativa.
- Cero.
- Curvilínea.

Haciendo Pronósticos Sobre la Población

Asumamos lo siguiente:

- Tenemos un político que es más odiado que querido.
- La población (n = 250,000) está evaluando al político usando un termómetro [0:100]
- Nosotros asignamos los parámetros poblacionales.

Queremos hacer pronósticos sobre la Población en base a muestras de la Población.

Creamos los Datos

Revisemos como creamos los datos.

```
# rbnorm() de {stevemisc}
Poblacion <- rbnorm(250000, mean = 42.42, sd = 38.84,
                    lowerbound = 0,
                    upperbound = 100,
                    round = TRUE,
                    seed = 8675309)
```

Vemos cuál es la media y la desviación estándar..

```
mean(Poblacion)
```

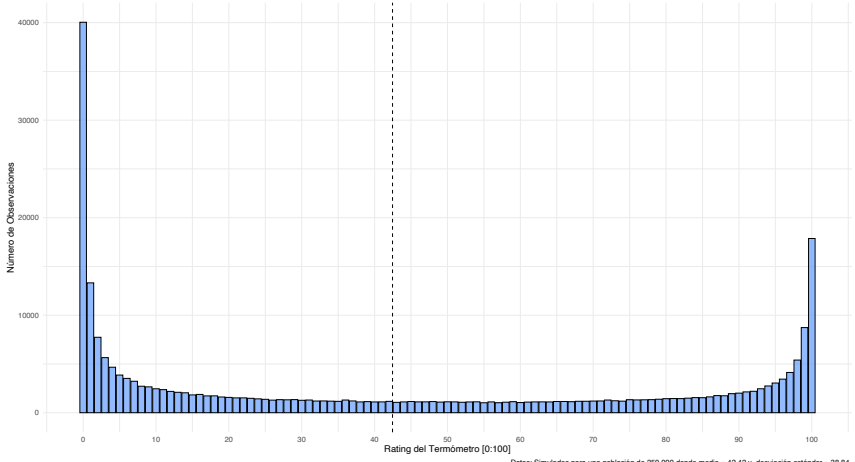
```
## [1] 42.45977
```

```
sd(Poblacion)
```

```
## [1] 38.88818
```

La Distribución de Ratings de Nuestra Población

Estos datos son una aproximación de la forma y distribución de de los ratings de políticos polarizantes.



Datos: Simulados para una población de 250,000 donde media = 42.42 y desviación estándar = 38.84.
La linea vertical es la media de la población.

Teorema de Límite Central

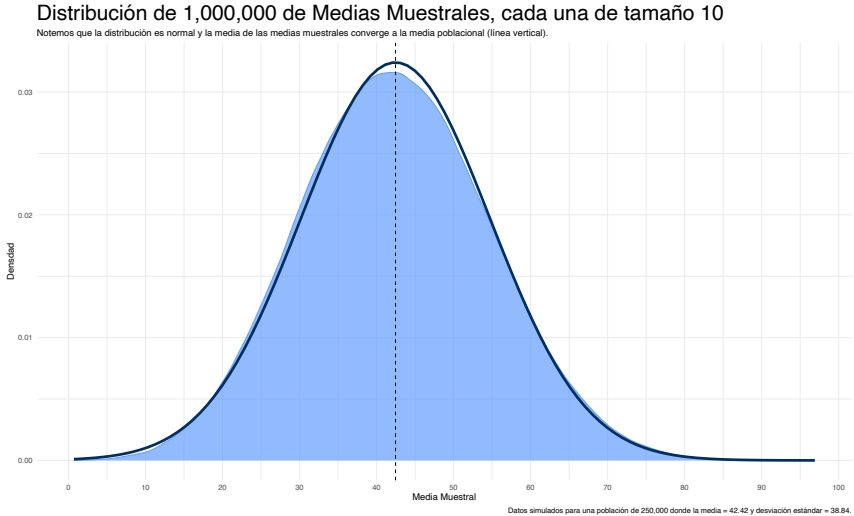
El **Terema de Límite Central** sugiere que con un número infinito de muestras de tamaño n, extraídas de una población de N unidades, las medias muestrales están **distribuídas normalmente**.

Por lo tanto:

- La media de las medias muestrales es igual a μ.
- El E.M.A. será igual al error estándar de la media muestral. $(\frac{\sigma}{\sqrt{n}})$

Código en R

```
set.seed(8675309)
# Usamos {dqrng} que es más eficiente (i.e. rápido) para hacer muestreo
# Esta es la función dqsample(), paquete {dqrng}
Popsamples <- tibble(
  samplemean=sapply(1:1000000,
           function(i){ x <- mean(
             dqsample(Poblacion, 10,
                    replace = FALSE))
           }))
```


Distribución de 1,000,000 de Medias Muestrales, cada una de tamaño 10
Notemos que la distribución es normal y la media de las medias muestrales converge a la media poblacional (línea vertical).
Densidad
0.03
0.02
0.01
0.00
0
10
20
30
40
50
60
70
80
90
100
Media Muestral
Datos simulados para una población de 250,000 donde la media = 42.42 y desviación estándar = 38.84.

Estandarización

La distribución normal que acabamos de ver no nos permite hacer mucha inferencia.

- Pero la **estandarización de una distribución normal** (o equivaler valores de una distribución normal a una distribución normal estándar) la hará más útil.

$$z = \frac{\text{Desviación de la Media}}{\text{Unidad Estándar}} \tag{1}$$

La unidad estándar varía de acuerdo a lo que queramos.

- Si tenemos una sola muestra aleatoria es igual a la desviación estándar.
- Si estamos comparando medias de varios muestreos aleatorios, entonces es el error estándar.

Estandarización

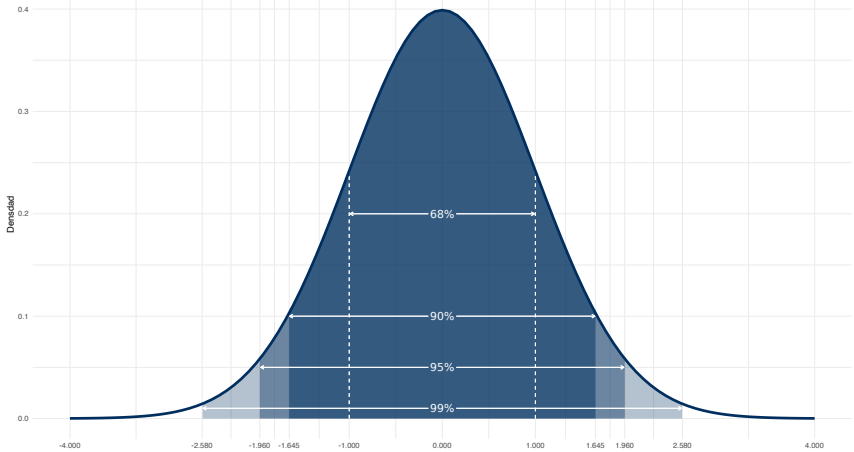
Mientras mas grande sea el valor de Z, mayor es la diferencia con respecto a la media.

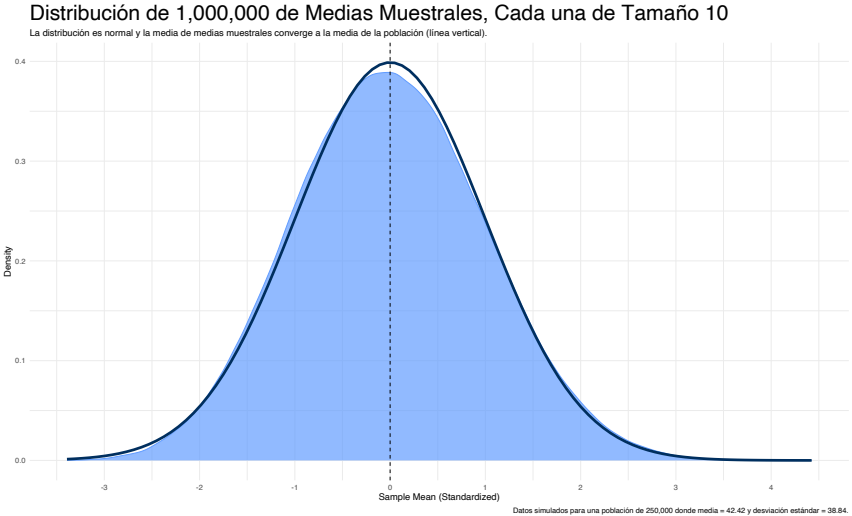
- Cuándo $Z = 0$, no hay desviación de la media (obviamente).

La estandarización nos permite hacer un mejor resúmen de la función normal.

El Area Debajo de la Distribución Normal

Las colas se extienden hacia infinito y tienden asintóticamente a 0, pero toda el área suma 1. 95% de todos los posibles valores están a 1.96 unidades estándar de la media.




Distribución de 1,000,000 de Medias Muestrales, Cada una de Tamaño 10
La distribución es normal y la media de medias muestrales converge a la media de la población (línea vertical).
Density
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3
4
Sample Mean (Standardized)
Datos simulados para una población de 250,000 donde media = 42.42 y desviación estándar = 38.84.

Inferencia Usando la Distribución Normal

¿Qué sigue? Vamos a asumir el siguiente escenario para ilustrar lo que sigue.

- Tenemos una muestra de 100 personas de la Población que hemos hablado toda la clase (i.e. la de 250,000 personas)

```
set.seed(8675309)
nuestramuestra <- sample(Poblacion, 100, replace = FALSE)
mean(nuestramuestra)
```

```
## [1] 43.64
```

- No conocemos μ (sabemos que es 42.46).
- Y asumimos que conocemos σ (aunque sabemos que es 38.89), no muy realista pero nos sirve para este ejemplo...
- Tenemos una n de 100 y $\overline{x}$ de 43.64 .

Nos interesa sacar conclusiones sobre el lugar en el que se encuntra la media poblacional.

Inferencia Usando la Distribución Normal

Para sacar conclusiones sobre el parámetro poblacional usando la muestra:

- Debemos tomar en cuenta el ruido generado por el muestreo aleatorio.
- Pero, nunca vamos a poder estar 100% seguros de que esa conclusión es cierta.

Un **intervalo de confianza del 95 por ciento** es informativo.

- Podemos estar seguros de que nuestros estimadores muestra estarán (o caerán) en ese intervalo con un 95% de probabilidad.
- Lo operacionalizamos de la siguiente forma $\overline{x} \pm (1.96)$*(error estándar).

Inferencia Usando la Distribución Normal

Saquemos el intervalo de confianza para nuestro ejemplo.

- Tenemos $\overline{x}$.
- También conocemos n y asumimos una σ conocida.
- Error Estándar $= 3.889$ ($\frac{\sigma}{\sqrt{n}} = \frac{38.88}{\sqrt{100}} = 3.88$)

Código en R para estimar Error Estándar

```
round(sd(Poblacion)/sqrt(length(nuestramuestra)), 3)
```

Inferencia Usando la Distribución Normal

Ahora debemos sacar los límites (superior/inferior) de nuestro intervalo de confianza al 95% de probabilidad.

$$\text{Límite Inferior} = \overline{x} - (1.96) * (e.s.) \tag{2}$$

$$\text{Límite Superior} = \overline{x} + (1.96) * (e.s.) \tag{3}$$

Estimación en R

```r
#cálculo del error estándar de la media
esm <- sd(Poblacion)/sqrt(length(nuestramuestra))

#intervalo de confianza de la media con 95% de probabilidad
c(mean(nuestramuestra) - 1.96*esm, mean(nuestramuestra) + 1.96*esm)
```

```
## [1] 36.01792 51.26208
```

Inferencia Usando la Distribución Normal

Lo que nuestro cálculo anterior significa es:

- que si tomamos 100 muestras de 100 personas cada una ($n = 100$), 95 de esas 100 muestras aleatorias van a tener en promedio una media entre 36.02 y 51.26.

Hay que notar que por ahora no estamos sacando ninguna conclusión sobre la media poblacional real.

Ejemplo de Inferencia

Asumamos que un seguidor del político no nos cree que $\overline{x}$ es correcto.

- Dice que es mucho más alto. Digamos: 56.61.

¿Qué podemos hacer para probarle que lo que nos dice no es correcto?

Ejemplo de Inferencia

Esta es una pregunta probabilística

- i.e. ¿Cuál es la probabilidad de $\overline{x} = 43.64$ si $\mu = 56.61$?

Y esto lo podemos contestar usando los valores de Z.

$$z = \frac{\overline{x} - \mu}{e.s.} \tag{4}$$

En R

```
(mean(nuestramuestra) - 56.61)/esm
```

```
## [1] -3.335204
```

Encontramos valor de Z

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842

Figure 1: Encontramos valor de Z

...o en R

```r
# de una cola (i.e. Nosotros asumimos la dirección)
pnorm(abs((mean(nuestramuestra) - 56.61)/esm), lower.tail=FALSE)
```

```
## [1] 0.0004261848
```

```r
# de dos colas (i.e. No sabemos la dirección)
# "dos colas" es usualmente el default.
2*pnorm(abs((mean(nuestramuestra) - 56.61)/esm), lower.tail=FALSE)
```

```
## [1] 0.0008523696
```

Ejemplo de Inferencia

¿Cuál es la probabilidad de que una muestra aleatoria produzca un valor de Z de -3.3352?

- Respuesta: 0.00043

En otras palabras: si μ fuera 56.61, observaríamos $\overline{x}$ aproximadamente 4 veces en 10,000 intentos, en promedio.

- Algo bastante improbable.

Ejemplo de Inferencia

¿Cuál es la conclusión?

- Podemos decir que ese seguidor está sugiriendo algo que MUY probablemente no es correcto.
- En realidad nuestra media muestral es mucho más cercana a μ.

Sin emabrgo,

- Este procedimiento no necesariamente nos dice cuál es el valor de μ.
- Más bien estamos comunicando lo que nosotros pensamos que es altamenete improbable que suceda.

¿Qué Sabemos de la Media Poblacional?

¿Qué tan probable era nuestra $\overline{x}$ de 43.64 dado que μ era 42.46? Mismo procedimiento.

```
(mean(nuestramuestra) - mean(Poblacion))/esm
```

```
## [1] 0.3034927
```

```
# Una cola (i.e. Nosotros asumimos dirección)
pnorm(abs((mean(nuestramuestra) - mean(Poblacion))/esm),
        lower.tail=FALSE)
```

```
## [1] 0.3807572
```

```
# Dos colas (i.e. No hacemos pronóstico sobre la dirección)
2*pnorm(abs((mean(nuestramuestra) - mean(Poblacion))/esm),
        lower.tail=FALSE)
```

```
## [1] 0.7615144
```

La probabilidad de nuestra media muestral, dada la media poblacional (que conocemos), es 0.38.

- Esto SÍ es posible. No podemos descartar la media poblacional de nuestra muestra aleatoria como lo hacíamos en el ejemplo en donde la media

Algunas Derivaciones

Asumimos que sabíamos σ, aunque no supieramos μ. ¿Y si no conocemos ningua?

- Usamos la desviación estándar de la muestra (s).
- Y hacemos el mismo procedimiento pero ahora usando una **distribución *t* de Student**.
- Es casi idéntica a una distribución normal, pero con colas más anchas cuando tenemos menos **grados de libertad**.
  - Grados de libertad = n - k (i.e. número de observaciones - número de parámetros [aquí: 1])

El grado de incertidumbre se incrementa con menos grados de libertad.

Intervalo de Confianza (t-student)

La fórmula para calcular el intervalo de confianza cuando usamos un distribución *t-student* es:

$$\text{Límite Inferior} = \overline{x} - (t) * (e.s.) \tag{5}$$

$$\text{Límite Superior} = \overline{x} + (t) * (e.s.) \tag{6}$$

Distribución *t* de Student

Table of Probabilities for Student's t-Distribution

df	0.600	0.700	0.800	0.900	0.950	0.975	0.990	0.995
1	0.325	0.727	1.376	3.078	6.314	12.706	31.821	63.657
2	0.289	0.617	1.061	1.886	2.920	4.303	6.965	9.925
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032
6	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707
7	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499
8	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355
9	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250
10	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169
11	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106
12	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055
13	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012
14	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977
15	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947
16	0.258	0.535	0.865	1.337	1.746	2.120	2.583	2.921
17	0.257	0.534	0.863	1.333	1.740	2.110	2.567	2.898
18	0.257	0.534	0.862	1.330	1.734	2.101	2.552	2.878
19	0.257	0.533	0.861	1.328	1.729	2.093	2.539	2.861
20	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845
21	0.257	0.532	0.859	1.323	1.721	2.080	2.518	2.831
22	0.256	0.532	0.858	1.321	1.717	2.074	2.508	2.819
23	0.256	0.532	0.858	1.319	1.714	2.069	2.500	2.807
24	0.256	0.531	0.857	1.318	1.711	2.064	2.492	2.797
25	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787
26	0.256	0.531	0.856	1.315	1.706	2.056	2.479	2.779
27	0.256	0.531	0.855	1.314	1.703	2.052	2.473	2.771
28	0.256	0.530	0.855	1.313	1.701	2.048	2.467	2.763
29	0.256	0.530	0.854	1.311	1.699	2.045	2.462	2.756
30	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750
40	0.255	0.529	0.851	1.303	1.684	2.021	2.423	2.704
60	0.254	0.527	0.848	1.296	1.671	2.000	2.390	2.660
120	0.254	0.526	0.845	1.289	1.658	1.980	2.358	2.617

df (degrees of freedom) = number of samples - 1
1 - alpha (for one tail) or 1 - alpha/2 (for two tails)

Figure 2: Probabilidades Distribución *t* Student

...o en R

```r
# media propuesta
(mean(nuestramuestra) - 56.61)/
  (sd(nuestramuestra)/sqrt(100))  -> tstat1
tstat1
```

```
## [1] -3.311028
```

```r
 # media actual
(mean(nuestramuestra) - mean(Poblacion))/
  (sd(nuestramuestra)/sqrt(100)) -> tstat2
tstat2
```

```
## [1] 0.3012929
```

. . . siguiendo en R

```r
# probabilidad si el seguidor está en lo correcto
pt(-abs(tstat1), df = 100-1) # una cola
```

```
## [1] 0.0006489941
```

```r
2*pt(-abs(tstat1), df = 100-1) # dos colas
```

```
## [1] 0.001297988
```

```r
# probabilidad sabiendo cuál es la media poblacional
pt(-abs(tstat2), df = 100-1) # una cola
```

```
## [1] 0.3819116
```

```r
2*pt(-abs(tstat2), df = 100-1) # dos colas
```

```
## [1] 0.7638231
```

Pruebas de Hipótesis (Breve Recapitulación)

Todas las pruebas de hipótesis tienen 5 partes:

- Supuestos (Tipo de los datos, randomización, distribución de la población, tamaño de la muestra).
- Hipótesis (Hipótesis nula (H_0) e hipótesis alternativa H_a).
- Prueba estadística (i.e. nos permite saber que tan lejos estamos del parámetro poblacional).
- P-value (la probabilidad de que el valor de nuestra prueba sea igual al valor observado o a un valor más extremo H_a).
- Conclusión.

Hipótesis Nula e Hipótesis Alternativa

La hipótesis nula sobre una media poblacional μ se define como:

$$H_0 = \mu = \mu_0 \tag{7}$$

La hipótesis alternativa se define así:

$$H_a = \mu \neq \mu_0, \text{dado que} H_a : \mu \neq 0 \tag{8}$$

Ejemplo 1

Por experiencia se ha comprobado que las calificaciones de los estudiantes del primer exámen de métodos cuantitativos aplicados en el CIDE están normalmente distribuídas con una media de 75 y una varianza de 36. El Director de la licenciatura en CPRI quiere saber si el grupo del año en curso, de 16 estudiantes, es típico con una probabilidad del 10%. Al hacer el exámen, el promedio de calificación fue 82. ¿Cuál sería la conclusión?

1. Definir H_0 y H_a
2. Encontrar el valor z de p $= .1$ (prueba de dos colas, 5% en cada lado)
3. El valor de la media muestral es 82
4. Calcular z que es igual a 4.67
5. $4.67 > 1.645$
6. Rechazamos H_0 y aceptamos H_a

Ejemplo 2

Si tenemos una muestra al azar que consiste de 9 individuos, la cual nos da $\overline{x} = 23$ y $\sigma = 4$. Probar H_0: $\mu = 21$ contra H_a: $\mu > 21$. Asumir p = .01
Resultado: $1.5 < 2.33 ==>$ Aceptamos H_0

Conclusión: El Proceso de Inferencia

Va de nuevo el proceso de inferencia.

1. Asumimos la media hipotética como correcta (digamos que es nuestra hipótesis).
2. Probamos la aseveración sobre la media hipotética con la información de la muestra aleatoria.
3. Inferimos sobre nuestra aseveración usando inferencia probabilística.

Algo así como: $p(\bar{x}|\mu)$.

- Nótese que **no** es: $p(\mu|\bar{x})$.

Conclusión: El Proceso de Inferencia

Nunca sabremos μ.

- Pero el procedimiento descrito aquí no ayuda a dar una respuesta aunque sea indirecta.
- Entre un determinado intervalo de confianza: "No puedo desechar esto."
- *Fuera* del intervalo de confianza deseado: "lo que sabemos es altamente improbable."

Table of Contents