

Introduction to Statistical Inference

Felipe José Bravo Márquez

October 1, 2020

Statistical Inference

- In order to draw conclusions about a **population**, it is generally not feasible to gather all the data about it.
- We must make reasonable conclusions about a population based on the evidence provided by **sample data**.
- The process of drawing conclusions about a population from sample data is known as **statistical inference**.

Statistical Inference (2)

- In statistical inference we try to **infer** the distribution that generates the observed data
- Example: Given a sample $X_1, \dots, X_n \sim F$, how do we infer F ?
- In some cases we are only interested in inferring some property of F (e.g., its **mean** value).
- Statistical models that assume that the distribution can be modeled with a finite set of parameters $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ are called **parametric models**.
- Example: if we assume that the data comes from a normal distribution $N(\mu, \sigma^2)$, μ and σ would be the parameters of the model.
- A **sample statistic** is a quantitative measure calculated from the data.

Frequentist and Bayesian Approaches

- Freq 1: Probability is restricted to relative counts.
- Freq 2: Uncertainty comes from sample variation.
- Freq 3: Parameters are unknown quantities not random variables.
- Bayesian 1: Probability describes degree of belief.
- Bayesian 2: We can make probability statements about parameters.
- Bayesian 3: We make inferences about a parameter by producing a probability distribution of it.
- For the remainder of this class we will stay with the frequentist approach. The Bayesian approach will be adopted later in the course.

Point Estimation

- Point estimation is the process of finding the **best guess** for some quantity of interest from a **statistical sample**.
- The quantity of interest could be a parameter in a parametric model, a CDF F , a probability density function f , a regression function r , or a prediction for a future value Y of some random variable.
- By convention, we denote a point estimate of θ by $\hat{\theta}$ or $\hat{\theta}_n$
- Es importante remarcar que mientras θ es un valor fijo desconocido, $\hat{\theta}$ depende de los datos y por ende es una variable aleatoria.

Estimación Puntual (2)

Definición Formal

- Sean X_1, \dots, X_n n observaciones IID de una distribución F
- Un estimador puntual $\hat{\theta}_n$ de un parámetro θ es una función de X_1, \dots, X_n :

$$\hat{\theta}_n = g(X_1, \dots, X_n)$$

- El **sesgo** (bias) de un estimador se define como:

$$\text{bias}(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n) - \theta$$

- Un estimador es insesgado si $\mathbb{E}(\hat{\theta}_n) = \theta$ o $\text{bias}(\hat{\theta}_n) = 0$

Estimación Puntual (3)

- La distribución de $\hat{\theta}_n$ se conoce como la **distribución muestral**
- La desviación estándar de $\hat{\theta}_n$ se conoce como **error estándar** *se*:

$$se(\hat{\theta}_n) = \sqrt{\mathbb{V}(\hat{\theta}_n)}$$

- El error estándar nos habla sobre la variabilidad del estimador entre todas las posibles muestras de un mismo tamaño.

Estimación Puntual (4)

- Sea X_1, X_2, \dots, X_n una muestra aleatoria de una población de media μ y varianza σ^2
- Se define la **media muestral** \overline{X}_n o $\hat{\mu}$ como:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

- Es un estimador insesgado:

$$\mathbb{E}(\overline{X}_n) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \times \mathbb{E}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n}(n \times \mu) = \mu$$

- Su error estándar sería $se(\overline{X}_n) = \sqrt{\mathbb{V}(\overline{X}_n)}$ donde

$$\mathbb{V}(\overline{X}_n) = \mathbb{V}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \mathbb{V}\left(\sum_{i=1}^n X_i\right) = \frac{n}{n^2} \mathbb{V}(X_i) = \frac{\sigma^2}{n}$$

- Entonces $se(\overline{X}_n) = \frac{\sigma}{\sqrt{n}}$

Ejemplos de Estimación Puntual (5)

- Por lo general no sabemos σ de la población.
- Cuando queremos estimar la varianza de una población a partir de una muestra hablamos de la **varianza muestral**:
- Existen dos estimadores comunes, una versión sesgada

$$s_n^2 = \frac{1}{n} \sum_i^n (X_i - \bar{X}_n)^2$$

- Una versión sin sesgo

$$s^2 = \frac{1}{n-1} \sum_i^n (X_i - \bar{X}_n)^2$$

- Cuando no sabemos la varianza de la población y queremos estimar la media, el error estándar es estimado:

$$\hat{se}(\bar{X}_n) = \frac{s}{\sqrt{n}}$$

Estimación Puntual (6)

- Sean $X_1, \dots, X_n \sim \text{Bernoulli}(p)$ y sea $\hat{p}_n = \frac{1}{n} \sum_i X_i$
- Luego $\mathbb{E}(\hat{p}_n) = \frac{1}{n} \sum_i \mathbb{E}(X_i) = p$, entonces \hat{p}_n es insesgado.
- El error estándar *se* sería

$$se = \sqrt{\mathbb{V}(\hat{p}_n)} = \sqrt{p(1-p)/n}$$

- El error estándar estimado \hat{se} :

$$\hat{se} = \sqrt{\hat{p}(1-\hat{p})/n}$$

Estimación Puntual (7)

- Se espera que un buen estimador sea insesgado y de mínima varianza.
- Un estimador puntual $\hat{\theta}_n$ de un parámetro θ es **consistente** si converge al valor verdadero cuando el número de datos de la muestra tiende a infinito.
- La calidad de un estimador se puede medir usando el **error cuadrático medio** (MSE)

$$MSE = \mathbb{E}_{\theta}(\hat{\theta}_n - \theta)^2$$

Estimación Puntual (8)

- Si para un estimador $\hat{\theta}_n$, su $bias \rightarrow 0$ y su $se \rightarrow 0$ cuando $n \rightarrow \infty$, $\hat{\theta}_n$ es un estimador consistente de θ .
- Por ejemplo, para la media muestral $\mathbb{E}(\overline{X}_n) = \mu$ lo que implica que el $bias = 0$ y $se(\overline{X}_n) = \frac{\sigma}{\sqrt{n}}$ que tiende a cero cuando $n \rightarrow \infty$. Entonces \overline{X}_n es un estimador consistente de la media.
- Para el caso del experimento Bernoulli se tiene que $\mathbb{E}(\hat{p}) = p \Rightarrow bias = 0$ y $se = \sqrt{p(1-p)/n} \rightarrow 0$ cuando $n \rightarrow \infty$. Entonces \hat{p} es un estimador consistente de p .

Intervalo de Confianza

- Sabemos que el valor de un estimador puntual **varía** entre una muestra y otra
- Es más razonable encontrar un **intervalo** donde sepamos que valor **real del parámetro** se encuentra dentro del intervalo con una cierta **probabilidad**.
- La forma general de un intervalo de confianza es la siguiente:

$$\text{Intervalo de Confianza} = \text{Estadístico Muestral} \pm \text{Margen de Error}$$

- Entre más ancho el intervalo mayor incertidumbre existe sobre el valor del parámetro.

Intervalo de Confianza (2)

Definición

- Un **intervalo de confianza** para un parámetro poblacional desconocido θ con un **nivel de confianza** $1 - \alpha$, es un intervalo $C_n = (a, b)$ donde:

$$\mathbb{P}(\theta \in C_n) = 1 - \alpha$$

- Además $a = a(X_1, \dots, X_n)$ y $b = b(X_1, \dots, X_n)$ son funciones de los datos
- El valor α se conoce como el nivel de **significancia**, generalmente se toma como 0.05 lo que equivale a trabajar con un nivel de confianza de 95%
- La significancia se puede interpretar como la probabilidad de equivocarnos.

Intervalo de Confianza (3)

Interpretación

- Existe mucha **confusión** de como interpretar un intervalo de confianza
- Una forma de interpretarlos es decir que si repetimos **un mismo experimento** muchas veces, el intervalo contendrá el valor del parámetro el $(1 - \alpha)\%$ de las veces.
- Esta interpretación es correcta, pero rara vez repetimos un mismo experimento varias veces.
- Una interpretación mejor: un día recolecto datos creo un intervalo de 95% de confianza para un parámetro θ_1 . Luego, en el día 2 hago lo mismo para un parámetro θ_2 y así reiteradamente n veces. El 95% de mis intervalos **contendrá** los valores reales de los parámetros.

Intervalo de Confianza (4)

- Se tienen n observaciones independientes X_1, \dots, X_n IID de distribución $N(\mu, \sigma^2)$
- Supongamos que μ es **desconocido** pero σ^2 es **conocido**.
- Sabemos que \bar{X}_n es un estimador insesgado de μ
- Por la ley de los grandes números sabemos que la distribución de \bar{X}_n se concentra alrededor de μ cuando n es grande.
- Por el CLT sabemos que

$$Z = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

cuando n es grande

- Despejando, tenemos que $\mu = \bar{X}_n - \frac{\sigma}{\sqrt{n}} Z$

Intervalo de Confianza (5)

- Queremos encontrar un intervalo $C_n = (\mu_1, \mu_2)$ con un nivel de confianza $1 - \alpha$:

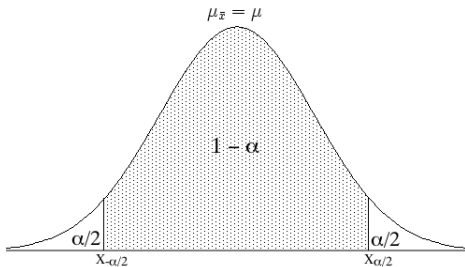
$$\mathbb{P}(\mu_1 \leq \mu \leq \mu_2) = 1 - \alpha$$

- Sea $z_a = \Phi^{-1}(1 - a)$, con $a \in [0, 1]$ donde Φ^{-1} es la función cuantía de una normal estandarizada
- Esto es equivalente a decir que z_a es el valor tal que $1 - \Phi(z_a) = \mathbb{P}(Z \geq z_a) = a$
- Por simetría de la normal $z_{\alpha/2} = -z_{(1-\alpha/2)}$

Intervalo de Confianza (6)

- Se tiene que

$$\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$



Intervalo de Confianza (7)

- El intervalo de confianza para μ es:

$$C_n = (\overline{X}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}})$$

- Entonces $z_{\alpha/2}$ nos dice cuantas veces tenemos que multiplicar el **error estándar** en el intervalo.
- Mientras menor sea α mayor será $z_{\alpha/2}$ y por ende más ancho será el intervalo.
- Demostración:

$$\begin{aligned}\mathbb{P}(\mu \in C_n) &= \mathbb{P}(\overline{X}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \overline{X}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}) \\ &= \mathbb{P}(-z_{\alpha/2} < \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{\alpha/2}) \\ &= \mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) \\ &= 1 - \alpha\end{aligned}$$

Intervalo de Confianza (8)

- Como $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ podemos usar la función cuantía de la normal para calcular intervalos de confianza en R

```
> alpha <- 0.05
> xbar <- 5
> sigma <- 2
> n <- 20
> se <- sigma/sqrt(n)
> error <- qnorm(1-alpha/2)*se
> left <- xbar-error
> right <- xbar+error
> left
[1] 4.123477
> right
[1] 5.876523
>
```

Distribución T

- En la práctica, si no conocemos μ es poco probable que conozcamos σ
- Si estimamos σ usando s , los intervalos de confianza se construyen usando la distribución **T-student**

Distribución T

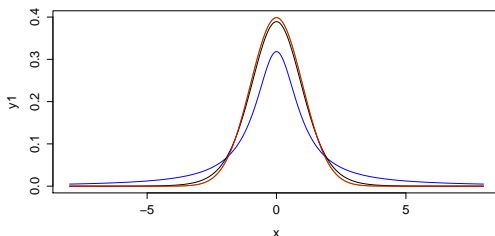
- Una V.A tiene distribución t con k grados de libertad cuando tiene la siguiente PDF:

$$f(t) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})(1 + \frac{t^2}{k})^{(k+1)/2}}$$

- Cuando $k = 1$ se le llama distribución de **Cauchy**
- Cuando $k \rightarrow \infty$ converge a una distribución normal estandarizada
- La distribución t tiene colas más anchas que la normal cuando tiene pocos grados de libertad

Distribución T (2)

```
x<-seq(-8,8,length=400)
y1<-dnorm(x)
y2<-dt(x=x,df=1)
y3<-dt(x=x,df=10)
y4<-dt(x=x,df=350)
plot(y1~x,type="l",col="green")
lines(y2~x,type="l",col="blue")
lines(y3~x,type="l",col="black")
lines(y4~x,type="l",col="red")
```



Intervalo de Confianza (9)

- Sea $s^2 = \frac{1}{n-1} \sum_i^n (X_i - \bar{X}_n)^2$ tenemos:

$$T = \frac{\bar{X}_n - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

- Sea $t_{n-1,a} = \mathbb{P}(T > a)$, equivalente a la función cuantía qt evaluada en $(1 - a)$
- El intervalo de confianza resultante es:

$$C_n = (\bar{X}_n - t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}, \bar{X}_n + t_{n-1,\alpha/2} \frac{s}{\sqrt{n}})$$

- Como las colas de la distribución t son más anchos cuando n es pequeño, los intervalos de confianza resultantes son más anchos

Intervalo de Confianza (10)

- Calculemos un intervalo de confianza para la media de `Petal.Length` de los datos del **Iris** con 95% de confianza

```
>data(iris)
>alpha<-0.05
>n<-length(iris$Petal.Length)
>xbar<-mean(iris$Petal.Length)
>xbar
[1] 3.758
>s<-sd(iris$Petal.Length)
>se<-s/sqrt(n)
>error<-qt(p=1-alpha/2,df=n-1)*se
>left<-xbar-error
>left
[1] 3.473185
>right<-xbar+error
>right
[1] 4.042815
```

- Otra forma:

```
>test<-t.test(iris$Petal.Length,conf.level=0.95)
>test$conf.int
[1] 3.473185 4.042815
```


Test de Hipótesis

- Cuando queremos probar si alguna **propiedad** asumida sobre una población se contrasta con una muestra estadística usamos un **Test de Hipótesis**
- El test se compone de las siguientes hipótesis:
 - **Hipótesis Nula** H_0 : Simboliza la situación actual. Lo que se ha considerado real hasta el presente.
 - **Hipótesis Alternativa** H_a : es el modelo alternativo que queremos considerar.
- La idea es encontrar suficiente **evidencia estadística** para rechazar H_0 y poder concluir H_a
- Si no tenemos suficiente evidencia estadística **fallamos en rechazar** H_0

Test de Hipótesis (2)

Metodología para Realizar un Test de Hipótesis

- Elegir una hipótesis nula H_0 y alternativa H_a
- Fijar un nivel de significancia α del test
- Calcular un estadístico T a partir de los datos
- El estadístico T es generalmente un valor estandarizado que podemos chequear en una tabla de distribución
- Definir un criterio de rechazo para la hipótesis nula. Generalmente es un valor crítico c .

Test de Hipótesis (3)

- Ejemplo: Se sabe que la cantidad de horas promedio de uso de Internet mensual en Chile país es de 30 horas
- Supongamos que queremos demostrar que el promedio es distinto a ese valor.
- Tendríamos que $H_0 : \mu = 30$ y $H_a : \mu \neq 30$
- Fijamos $\alpha = 0.05$ y recolectamos 100 observaciones
- Supongamos que obtenemos $\bar{X}_n = 28$ y $s = 10$
- Una forma de hacer el test es construir un intervalo de confianza para μ y ver si H_0 está en el intervalo.

```
> 28-qt (p=0.975, 99) *10/sqrt (100)
```

```
[1] 26.01578
```

```
> 28+qt (p=0.975, 99) *10/sqrt (100)
```

```
[1] 29.98422
```

- El intervalo sería la zona de aceptación de H_0 y todo lo que esté fuera de éste será mi región de rechazo.
- Como 30 está en la región de rechazo, rechazo mi hipótesis nula con un 5% de confianza.

Test de Hipótesis (4)

- Otra forma de realizar el test es calcular el estadístico $T = \frac{\overline{X}_n - \mu_0}{\frac{s}{\sqrt{n}}}$

- En este caso sería

$$T = \frac{28 - 30}{\frac{10}{\sqrt{100}}} = -2$$

- Como $H_a : \mu \neq 30$, tenemos un test de dos lados, donde la región de aceptación es

$$t_{n-1, 1-\alpha/2} < T < t_{n-1, \alpha/2}$$

```
> qt(0.025, 99)
[1] -1.984217
> qt(0.975, 99)
[1] 1.984217
```

- Como T está en la región de rechazo, rechazamos la hipótesis nula.

Test de Hipótesis (5)

- Generalmente, además de saber si rechazamos o fallamos en rechazar una hipótesis nula queremos saber la evidencia que tenemos en contra de ella.
- Se define un **p-valor** como la probabilidad de obtener un resultado al menos tan extremo como el observado en los datos dado que la hipótesis nula es verdadera.
- “Extremo” significa lejos de la hipótesis nula.
- Si el **p-valor** es menor que el nivel de significancia α , rechazamos H_0
- Ejemplo:

```
> data(iris)
> mu<-3 # La hipótesis nula
> alpha<-0.05
> n<-length(iris$Petal.Length)
> xbar<-mean(iris$Petal.Length)
> s<-sd(iris$Petal.Length)
> se<-s/sqrt(n)
> t<-(xbar-mu)/(s/sqrt(n))
> pvalue<-2*pt(-abs(t),df=n-1)
> pvalue
[1] 4.94568e-07 # es menor que 0.05 entonces rechazamos H0
```

Test de Hipótesis (6)

- La forma elegante de hacerlo en R:

```
> t.test(x=iris$Petal.Length,mu=3)
```

One Sample t-test

```
data:  iris$Petal.Length
t = 5.2589, df = 149, p-value = 4.946e-07
alternative hypothesis: true mean is not equal to 3
95 percent confidence interval:
 3.473185 4.042815
sample estimates:
mean of x
 3.758
```

Test de Hipótesis (7)

- Tenemos dos tipos de errores cuando realizamos un test de hipótesis
- Error tipo I: es cuando rechazamos la hipótesis nula cuando ésta es cierta.
- Este error es equivalente al nivel de significancia α
- Error tipo II: es cuando la hipótesis nula es falsa pero no tenemos evidencia estadística para rechazarla.
- Para mitigar los errores tipo I generalmente usamos valores de α más pequeños.
- Para mitigar los errores tipo II generalmente trabajamos con muestras más grandes.
- Existe un trade-off entre los errores tipo I y tipo II.

	Retener H_0	Rechazar H_0
H_0 es verdadera	✓	error tipo I
H_1 es verdadera	error tipo II	✓



L. Wasserman *All of Statistics: A Concise Course in Statistical Inference*, Springer Texts in Statistics, 2005.