Introduction to Statistical Inference

Felipe José Bravo Márquez

December 10, 2020

Populations and Samples

- The main goal of statistical inference is investigate properties about a target **population**.
- Example: What is the average height of the Chilean people?
- 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**.
- A **sample staticic** or simply **statistic** is a quantitative measure calculated from the data. Examples: the mean, the standard deviation, the minimum, the maximum.
- Our goal in sampling is to determine the value of a statistic for an entire population of interest, using just a small subset of the population.
- We do this primarily to save time and effort.
- Idea: Why go to the trouble of measuring every individual in the population when just a small sample is sufficient to accurately estimate the statistic of interest? [Poldrack, 2019]

Statistical Inference (2)

- The process of drawing conclusions about a population from sample data is known as **statistical inference**.
- In statistical inference we try to **infer** the distribution that generates the observed data.
- Example: Given a sample $X_1, \ldots, X_n \sim F$, how do we infer F?
- However, in most 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 $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ 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.

Frequentist Aproaches

The satistical methods to be presented is this class are known as **frequentist (or classical)** methods. They are based on the following postulates [Wasserman, 2013]:

- Probability refers to limiting relative frequencies. Probabilities are objective properties of the real world.
- Parameters are fixed, unknown constants. Because they are not fluctuating, no useful probability statements can be made about parameters.
- Statistical procedures should be designed to have well-defined long run frequency properties. For example, a 95 percent confidence interval should trap the true value of the parameter with limiting frequency at least 95 percent.

There is another approach to inference called **Bayesian inference**, which is based on different posulated, to be discussed 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**.
- In a general sence, this 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.
- In this class we will consider this quantity of interest as a **population parameter** θ.
- By convention, we denote a point estimate of θ by $\hat{\theta}$ or $\hat{\theta}_n$.
- It is important to remark that while θ is an unknown fixed value, $\hat{\theta}$ depends on the data and is therefore a random variable.
- We need to bear in mind that the process of sampling is by definition a random experiment.

Point Estimation

Formal Definition

- Let $X_1, \ldots, X_n$ be n IID data points from some distribution F.
- A point estimator $\hat{\theta}_n$ of a parameter θ is some function of $X_1, \ldots, X_n$:

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

- The **bias** of an estimator is defined as:

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

- An estimator is unbiased if $\mathbb{E}(\hat{\theta}_n) = \theta$ or $\text{bias}(\hat{\theta}_n) = 0$

Sampling Distribution

- If we take multiple samples, the value of our statistical estimate $\hat{\theta}_n$ will also vary from sample to sample.
- We refer to this distribution of our estimator across samples as the **sampling distribution** [Poldrack, 2019].
- The sampling distribution may be considered as the distribution of $\hat{\theta}_n$ for all possible samples from the same population of size n[1].
- The sampling distribution describes the variability of the point estimate around the true population parameter from sample to sample.
- We need to bear in mind this is an imaginary concept, since we can't obtain all possible samples.

[1] https://courses.lumenlearning.com/boundless-statistics/chapter/sampling-distributions/

Standard Error

- The standard deviation of $\hat{\theta}_n$ is called the **standard error** *se*:

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

- The standard error tells us about the variability of the estimator between all possible samples of the same size.
- It is a measure of the uncertainty of the point estimate.

The Sample Mean

- Let $X_1, X_2, \ldots, X_n$ be a random sample of a population of mean μ and variance σ^2
- A sample statistic we can derive from the data is the **sample mean** $\overline{X_n}$

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

- The sample mean is a statitical estimate of the mean $\overline{X_n} = \hat{\mu}$.
- The sampling distribution of the sample mean:
- We can show that the sample mean is an unbiased estimator of μ:

$$\mathbb{E}(\overline{X_n}) = \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n} \times \mathbb{E}(\sum_{i=1}^{n} X_i) = \frac{1}{n}(n \times \mu) = \mu$$

Point Estimation (4)

- The standard error of the sample mean $se(\overline{X_n}) = \sqrt{\mathbb{V}(\overline{X_n})}$ can be calulated as:

$$\mathbb{V}(\overline{X_n}) = \mathbb{V}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\mathbb{V}(\sum_{i=1}^{n} X_i) = \frac{n}{n^2}\mathbb{V}(X_i) = \frac{\sigma^2}{n}$$

- Then, $se(\overline{X_n}) = \frac{\sigma}{\sqrt{n}}$

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

$$s_n^2 = \frac{1}{n}\sum_{i}^{n}(X_i - \overline{X_n})^2$$

- Una versión sin sesgo

$$s^2 = \frac{1}{n-1}\sum_{i}^{n}(X_i - \overline{X_n})^2$$

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

$$\hat{se}(\overline{X_n}) = \frac{s}{\sqrt{n}}$$

Estimación Puntual (6)

- Sean $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ 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.
- Unbiased ness used to receive much attention but these days is considered less important
- Many of the estimators we will use are biased.
- A reasonable requirement for an estimator is that it should converge to the true parameter value as we collect more and more data.
- 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 en las 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 $a = a(X_1, \ldots, X_n)$ y $b = b(X_1, \ldots, X_n)$ 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 $X_1, \ldots, X_n$ IID de distribución $N(\mu, \sigma^2)$
- Supongamos que μ es **desconocido** pero σ^2 es **conocido**.
- Sabemos que $\overline{X_n}$ es un estimador insesgado de μ
- Por la ley de los grandes números sabemos que la distribución de $\overline{X_n}$ se concentra alrededor de μ cuando n es grande.
- Por el CLT sabemos que

$$Z = \frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

  cuando n es grande
- Despejando, tenemos que $\mu = \overline{X_n} - \frac{\sigma}{\sqrt{n}} Z$

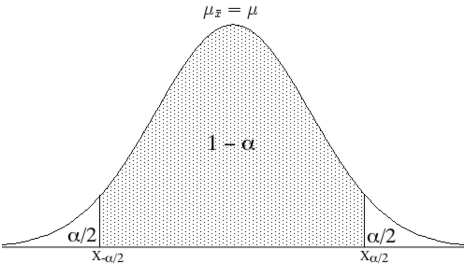
Intervalo de Confianza (5)

- Queremos encontrar un intervalo 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

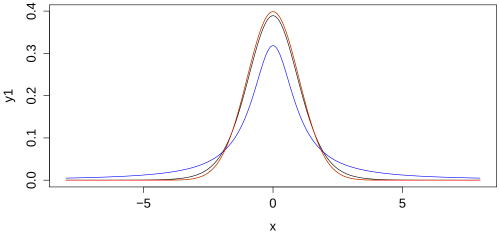
- Una V.A tiene distribución t con k grados de libertad cuando tiene la siguiente PDF:

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

- Cuando $k = 1$ se le llama distribución de **Cauchy**
- Cuando $k \to \infty$ 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 $s^2 = \frac{1}{n-1}\sum_i^n (X_i - \overline{X_n})^2$ tenemos:

$$T = \frac{\overline{X_n} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

- 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:

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

- 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
```

Maximum Likelihood Estimation

References I

Poldrack, R. A. (2019).
Statistical Thinking for the 21st Century.

Wasserman, L. (2013).
All of statistics: a concise course in statistical inference.
Springer Science & Business Media.