

— class: inverse, middle

## Chapter 15: Parameters and Statistics

— Parameters and Statistics We have discussed using sample data to make inference about the population. In particular, we will use sample **statistics** to make inference about population **parameters**.

### Definition (Parameter)

A **parameter** is a number that describes the population. In practice, parameters are unknown because we cannot examine the entire population.

### Definition (Statistic)

A **statistic** is a number that can be calculated from sample data without using any unknown parameters. In practice, we use statistics to estimate parameters.

## — Greek Letters and Statistics

### Greek Letters (Parameters)

- Greek letters like  $\mu$  and  $\sigma^2$  represent the truth about the population.

### Latin Letters (Statistic)

- Latin letters like  $\bar{x}$  and  $s^2$  are calculations that represent guesses (estimates) at the population values.

The goal for the class is for the latin letters to be good guesses for the greek letters:

$$\text{Data} \longrightarrow \text{Calculation} \longrightarrow \text{Estimates} \xrightarrow{\text{hopefully!}} \text{Truth}$$
$$X \longrightarrow 1/n \sum_{i=1}^n X_i \longrightarrow \bar{x} \xrightarrow{\text{hopefully!}} \mu$$

— Examples of Parameters Some parameters of distributions

we've encountered are

- $n$  and  $p$  in  $X \sim B(n, p)$  where

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

- $a$  and  $b$  in  $X \sim U(a, b)$

$$f(x) = \frac{1}{b - a}$$

- $\mu$  and  $\sigma^2$  in  $X \sim N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2}$$

— Mean and Variance Two population parameters of particular interest are

- the mean, denoted  $\mu$ , defined by  $E(X)$

- the variance, denoted  $\sigma^2$ , defined by  $E(X^2) - E(X)^2$

We **do not** observe these. Therefore, we guess using

- the sample mean,  $\bar{X}$
- the sample variance,  $s^2$

Why do we use these as our guess?

— Getting the right sample Before we talk about the properties of sample statistics, we need to make sure we have the right sample. We talked about good ways to generate a sample.

**The right sample is the most important part of any data analysis.**

A Simple Random Sample has no bias, and has observations that are from the same population.

In this case, we say all of the observations in our sample are

identically distributed. In math, this means for any two observations  $X_i$  and  $X_j$ ,

- $Pr(X_i < x) = Pr(X_j < x)$

— Independent Observations Does observing  $X_i$  impact the value of  $X_j$ ? Sometimes yes (time series, spatial dependence), but hopefully not.

To simplify things, we need to assume independent sample observations, meaning  $Pr(X_i = a \mid X_j = b) = Pr(X_i = a)$ .

Intuitively, this means that *observing* one outcome doesn't help you *predict* any other outcome.

To summarize, we want an *i.i.d.* sample, i.e. sample observations that are independent and identically distributed.

— Sample Statistics are Random Variables For a sample  $X_1, \dots, X_n$

of the random variable  $X$ , any function of that sample,  $\hat{\theta} = g(X_1, \dots, X_n)$ , is a **sample statistic**. For example,

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

Because  $X_1, \dots, X_n$  are random variables, any sample statistic  $\hat{\theta} = g(X_1, \dots, X_n)$  is itself a random variable! That means,

- There is some distribution for the values of  $\hat{\theta}$

— **Sampling Distributions** This is one of the most important concepts in the course. One **trial** would consist of the following:

- **Random Sample** - Grab a group of observations from the population

- **Sample Statistic** - Take your particular random sample and calculate a sample statistic (e.g. sample mean)

**Sampling Distribution** - Imagine repeatedly grabbing a different group of observations from the population and calculating the sample mean. This is performing many **trials**.

The sample means themselves will have a distribution. This is what we call the **Sampling Distribution**.

— Sampling Distribution Example

```
“r, echo = F, out.width =  
"90knitr::include_graphics("samplingdistnofxbar.png")”
```

— Sampling Distributions We will only observe 1 sample in the world though.

How does the concept of sampling distribution help us?

- Since we don't know the true population parameter, Our sample statistic will be our best guess at the possible true value.

- If we know the sampling distribution, then we can consider uncertainty about our sample statistic.

— Law of Large Numbers Under certain conditions, we can use the Law of Large Numbers (LLN) to guarantee that  $\bar{X}$  approaches  $\mu$  as the sample size grows large.

— Law of Large Numbers



## Theorem (Law of Large Numbers)

*Let  $X_1, X_2, \dots, X_n$  be an i.i.d. set of observations with  $E(X_1) = E(X_2) = \dots = E(X_n) = \mu$ . Define the sample mean of size  $n$  as  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ . Then*

$$\bar{X}_n \rightarrow \mu \quad \text{as} \quad n \rightarrow \infty.$$

Intuitively, as we observe a larger and larger sample, we average over randomness and our sample mean approaches the true population mean.

— Properties of

Let  $X_1, X_2, \dots, X_n$  be an i.i.d. sample with  $E(X_i) = \mu$  and  $\text{Var}(X_i) = \sigma^2 < \infty$ . Then

- $E(\bar{X}) = \mu$ ,
- $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$ .

Intuitively, we grab many samples from a population. The variance of our sample averages shrinks as we observe more observations per sample.

— Clicker Question Suppose we sample 100 observations from a distribution with  $\mu = 15$  and  $\sigma^2 = 25$ . What are  $E(\bar{X})$  and  $\text{Var}(\bar{X})$ ?

- (a)  $E(\bar{X}) = 15$ ,  $\text{Var}(\bar{X}) = 25$ .
- (b)  $E(\bar{X}) = 0.15$ ,  $\text{Var}(\bar{X}) = 0.25$ .

(c)  $E(\bar{X}) = 15$ ,  $Var(\bar{X}) = 5$ .

(d)  $E(\bar{X}) = 15$ ,  $Var(\bar{X}) = 0.25$ .

— When is

*Although we know the mean and variance of, we generally don't know its distribution.*

## Theorem

*Let  $X_1, X_2, \dots, X_n$  be an i.i.d. sample with  $X_i \sim N(\mu, \sigma^2)$  for  $i = 1, 2, \dots, n$ .*

*Then*

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Intuitively, if all the observations come from the same normal distribution then the sample average is also normally distributed and centered at the true mean.

— Central Limit Theorem If the number of observation per sample is large (we will discuss this more later), then we know, that the distribution of  $X_i$  doesn't matter, we will always have

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}).$$