Sampling Distributions

Grinnell College

October 11, 2024

Review

We have so far spent time on:

- Looking at how data is collected
  - Sampling methods
  - Biases
  - Experiment vs. Obs. Study
- Making displays of our data
  - Bar/Histogram/Boxplot/Scatter
  - Tables
- Describing what we see in our displays
  - Quant. → shape/center/spread
  - Associations
  - Correlations
  - Proportions / Percentages / Probabilities / Odds

We have mostly just been working with our sample and describing what we see. We need to go one step further.

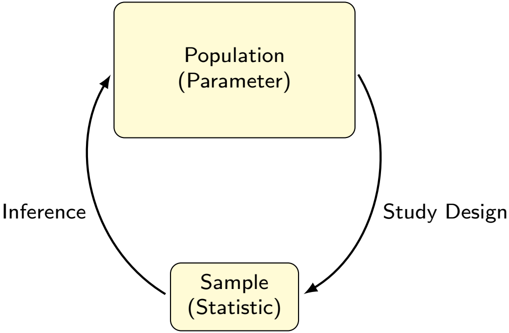
Review – Statistical Framework

Population is a big group of subjects/events/things about which we wish to learn about

Sample is a subgroup of pop. that we collect data from

Parameter is a *quantifiable* attribute of the pop. (most of the time this value is unknown)

Statistic is a numerical summary of the sample that we calculate from our sample data



BIG IDEA: Parameter value is unknown → we use the statistic to estimate it

More on Parameters & Statistics

Statistics are numerical summaries of the sample
Parameters are numerical summaries of the population

- **Inference**: use statistics (known) to estimate parameters (unknown)

Typically we will use special notation to differentiate *population parameters* (things we wish to know) from *statistics* computed from our sample:

	Population Parameter	Sample Statistic
Mean	μ	$\overline{x}$
Standard Deviation	σ	s
Proportion	p	$\hat{p}$
Correlation	ρ	r
Regression	β	b's or $\hat{\beta}$'s

Examples – Covid Vaccine

According to the U.S. Census Bureau, as of October 11, 2021, 83.3% of U.S. adults 18 years and older had received at least one dose of a COVID-19 vaccine. This is based on a representative sample of civilians aged 18 and over.

What are we trying to learn about?

- p = population proportion (%) of US adults that received at least one dose of a COVID-19 Vaccine = ?

What are we using to learn about it?

- $\widehat{p}$ = sample proportion (%) of US adults that received at least one dose of a COVID-19 Vaccine = .833 = 83.3%

Examples – Florida Fish

Data was collected from 53 lakes in Florida. For each lake, the mercury level (parts per million) was computed for a large mouth bass. Based on this sample of fish, the mean mercury level was 0.527 ppm

Research question: What is the average mercury level of fish (Large Mouth Bass) in all Florida lakes?

Parameter we are trying to learn about:

- $\mu =$ pop. average mercury level of Large Mouth Bass in Florida lakes

Statistic we are using to answer the research question

- sample mean mercury level of fish from 53 lakes $= 0.527$ppm

Sources of Error

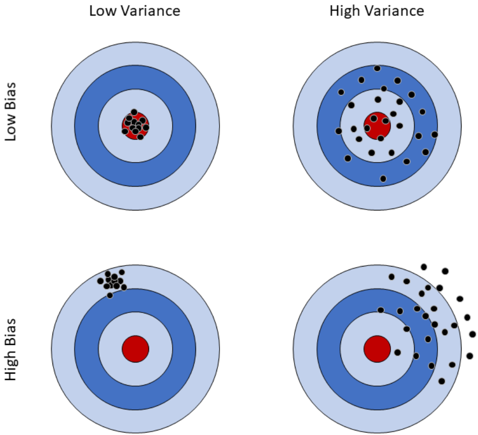
The statistic value will probably not *exactly* equal the parameter

- If the sample is *representative*, our estimate should be *close* to the parameter we wish to know

There are two main reasons why our sample statistic may differ from the population parameter:

1. **Sampling Bias** – A systemic flaw in how the sample was collected
2. **Sampling Variability** – Differences between samples due to *random chance*

Bias and Variance



Bias and Variance

Next step: We want to find out how different we can expect the statistic to be from the parameter

There is a **big problem** when we have a *biased* sample.

- We often cannot quantify how the sample differs from the pop.
- If the sample is 'messed up,' we don't know how wrong the statistic is
- It is super difficult if not impossible to correct for this

For **sampling variability** – maybe we just do some simulations of taking a whole bunch of samples and looking at the behavior of the corresponding statistics?

- We are going to give names to a few things to help us do this

Review – Distributions

Recall that a **distribution** describes:

- What values our variable can take
- How frequently they occur

For the rest of today we will mostly focus on looking at quantitative variables.

Population Distribution

A **population distribution** is quite literally the distribution of a variable for the entire population

- If we had complete info we could find the value for any parameter
- Almost always we don't *actually* know what this looks like
- If we took a *census* we could construct this

Sample Distribution

A **sample distribution** is the distribution of a variable for a single sample (data we've collected)

- We can use the data to calculate statistics (ex: $\overline{x}$, s)
- The statistics give us an idea of how the pop. may look

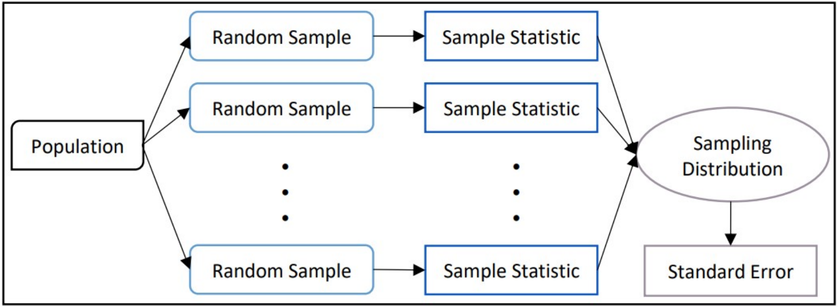
Sampling Distribution

What if I actually had many samples? I could plot the distribution of all of the resulting statistics.

A **sampling distribution** is the distribution of a whole bunch of statistics from *many* samples (all with same sample size)

- This will allow us to see how much the statistic itself varies from sample to sample
- This is a theoretical tool – in practice we only ever take one sample

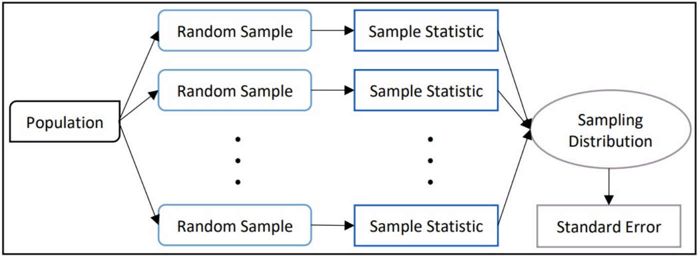
Sampling Distribution



How to construct?

1. Start with population.
2. Take a sample and compute the statistic of choice
3. Take another sample and compute the statistics
4. Continue to take more samples and compute statistic each time
5. Plot all of the statistics in a histogram or dotplot

Sampling Distribution



Population Variability

- Cases within the pop. vary → can describe with pop. standard deviation (σ)

Sample Variability

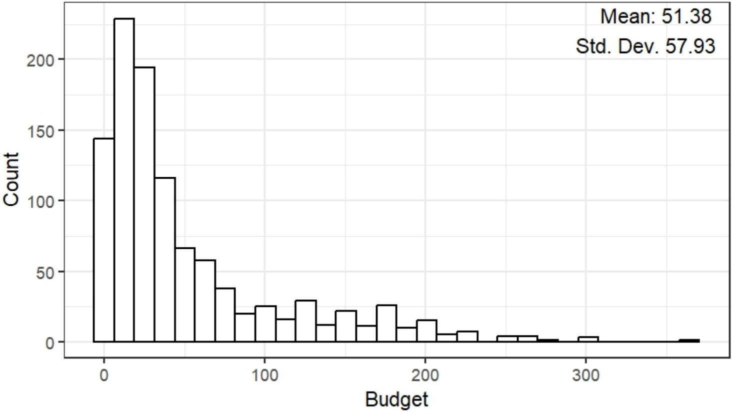
- Cases within a sample vary → can describe with sample standard deviation (s)

Sampling Variability

- Statistics from many samples vary → describe using the std. dev. of sampling distribution
- **Standard Error (SE)** = std. dev. of sampling distribution

Example – Movie Budgets

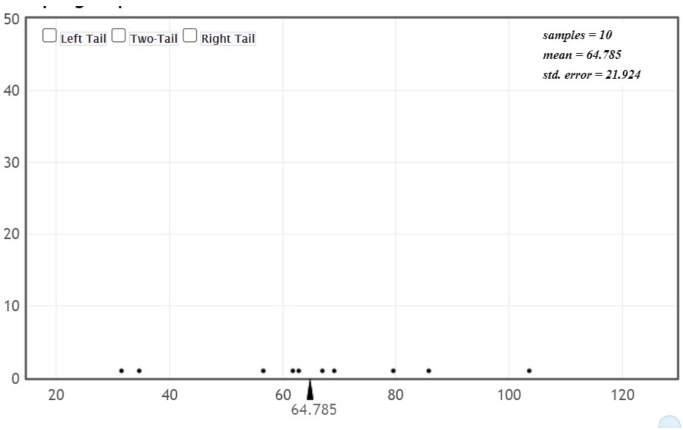
Hollywood movies released between 2012 and 2018. Recorded in millions of dollars. (n=1056)



Example – Movie Budgets

Sampling Distribution (sample size = 10 for each sample)

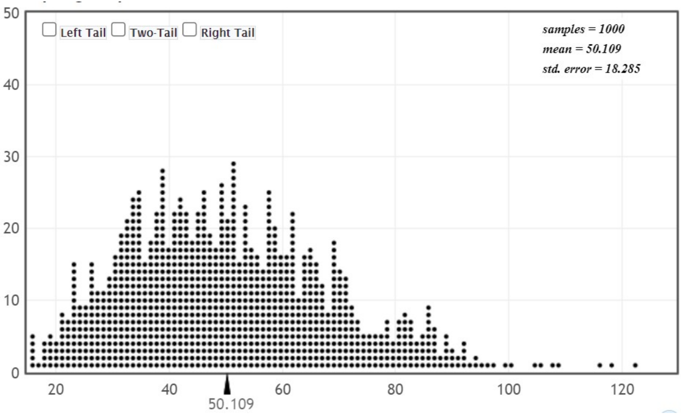
- 10 samples
- Each dot represents the mean budget for 1 random sample



Example – Movie Budgets

Sampling Distribution (sample size = 10 for each sample)

- 1000 samples



Pop: $\mu = 51.38$

- Is the sampling distribution mean similar?

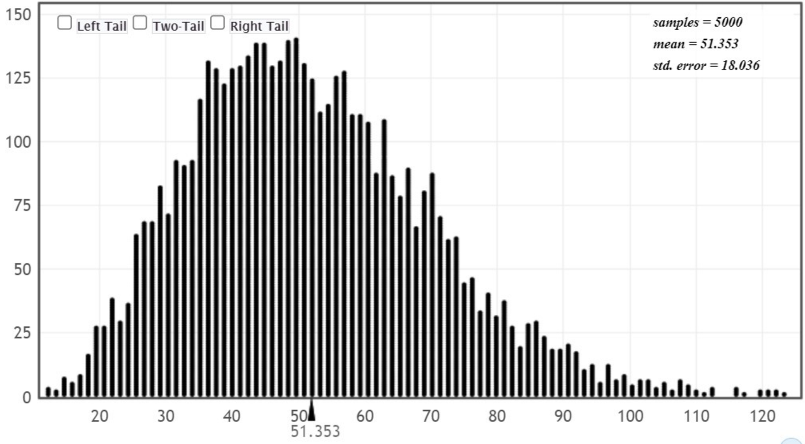
Sample Size

What happens the sampling distribution when we change the sample size?

Example – Movie Budgets

Sampling distribution (n=10 for each sample)

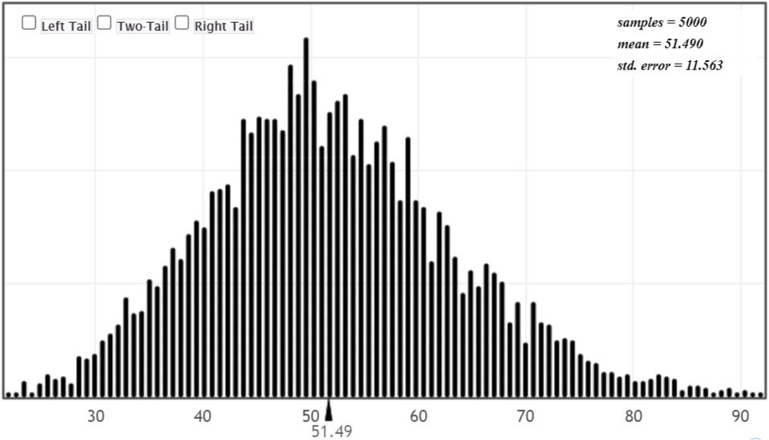
- Shape?
- SE = 18.036



Example – Movie Budgets

Sampling distribution (n=25 for each sample)

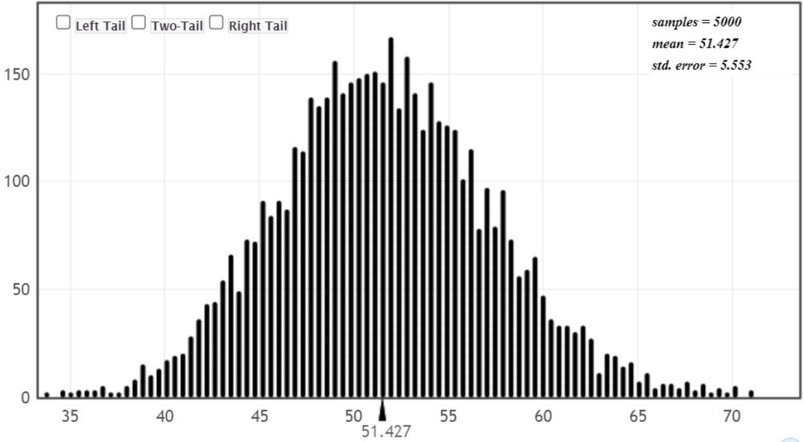
- Shape?
- SE = 11.536



Example – Movie Budgets

Sampling distribution (n=100 for each sample)

- Shape?
- SE = 5.553



Summary of Sampling Distributions

Purpose: Sampling distributions let us see how much variability there is for our estimates (the statistics) from many samples

Center?
As long as we are using *random samples*, then the mean of the sampling distribution will be close to the true parameter value.

Shape?
As the sample size of each sample increases:

- the distribution becomes more and more bell-shaped

Standard Error (SE) – variability of statistics from our many samples
As the sample size of each sample increases:

- The standard error of the sampling distribution decreases
- our estimates become more *precise*

Next Time

We are going to keep working with sampling distributions.

We will use Standard Error to help us quantify how far off we think our estimates might be.

We can use sampling distributions for all kinds of statistics like medians, proportions, correlations (not just means)

Check Your Understanding

- What does **Inference** mean?
- What was the purpose of constructing sampling distributions in terms of estimation?
- How does sample size affect sampling distributions?