Introduction to Statistics I

Instructor: Jodin Morey moreyj@lemoyne.edu

Previous Lecture

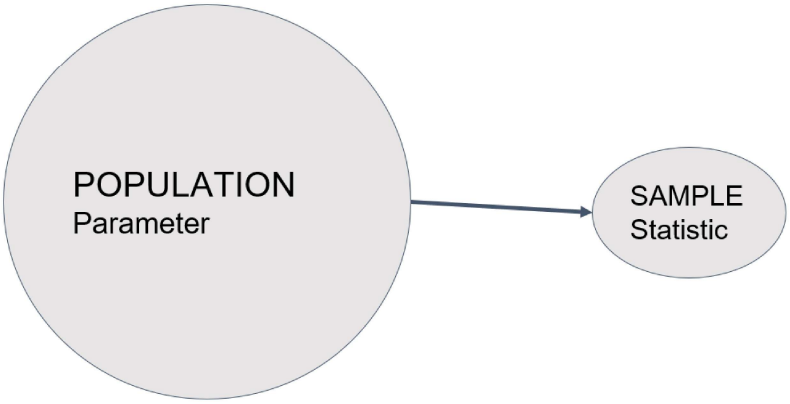
- Unique Normal distr for each mean and SD.
- Probability of observing data pt less/greater than some other value
- *z*-scores, areas under normal curves

Topic 13: Sampling Distributions

Proportions

Previously: How to use a sample distr to predict things about the population using a math model (normal distr).

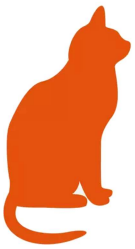
Now: How to use a statistic to estimate a parameter?



	(Population) Parameter		(Sample) Statistic	
Proportion	π	"pi"	$\hat{p}$	"p-hat"
Mean	μ	"mu"	$\bar{x}$	"x-bar"
Standard Deviation	σ	"sigma"	s	

Notation

Math Cat Says:



μ

Can repeated sampling tell us about the parameter?

Example: We want to know what proportion of Reese's Pieces are orange.

Collect 25 Reese's Pieces, count how many are orange.

Calculate proportion of candies that are orange. (recording this on the whiteboard)

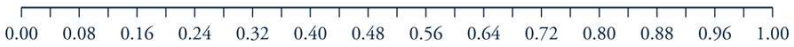
Observational unit? Variable? Variable type?

- Observational unit: **the candies**.
- Variable: **the color**.
- Variable type: **categorical** (binary if "orange or not").

Population? Sample? Statistic? Parameter? Symbols?

- Population: **All Reese's Pieces**.
- Sample: **The 25 we collected**.
- Statistic: **The proportion in our sample that were orange** - $\hat{p}$.
- Parameter: **The proportion of ALL Reese's Pieces that are orange** - π.

Experiment Results



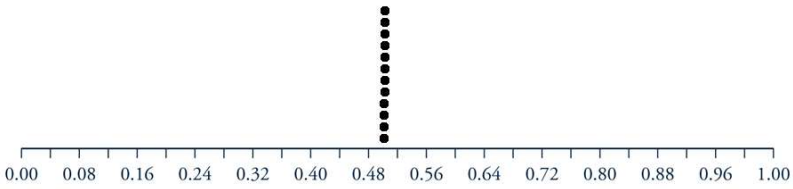
dotplot of students' sample proportions

bit.ly/introstatsdata

Applet: DotPlot

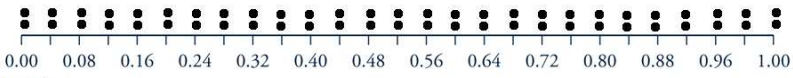
Now looking at everybody's proportions, what is the:

Observational unit? Variable? Variable type? Shape? Center? Spread?



Not Our Experiment Results. No spread.

Sampling Variability: The *fact* that the statistic $\hat{p}$ varies from sample to sample.



Also Not Our Experiment Results. Even spread.

Sampling Distribution: The *way* that the statistic $\hat{p}$ varies from sample to sample.

The samples gave different values of statistic $\hat{p}$.

But while $\hat{p}$ varies from sample to sample, it follows a pattern related to the parameter π.

The distr of the sample proportions $\hat{p}$ from sample to sample is the **Sampling Distr**.

Let's do a virtual version of this experiment. Do the following on the applet:

Sample of size (n): 25

Edit Proportion (π): 0.45

Then, "Generate 1000 Samples."

Does it look Normally Distr'd?

Where's the Center? What's the SD ("std. error")?

Repeat w/sample size: 75. Center? SD?

bit.ly/introstatsdata

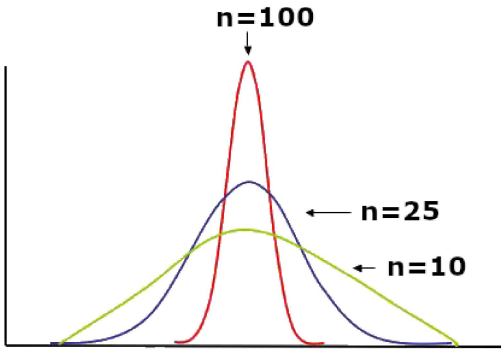
Applets: Sampling Distr

So we can describe the distr's as:

- Shape - Normal
- Center - Centered around parameter π
- Spread ??

Spread

What was the spread of the sampling distr as we changed sampling size?



As sample size increases, spread decreases

Central Limit Theorem (CLT)

CLT: Suppose a sample of size n is taken from a population w/parameter π.

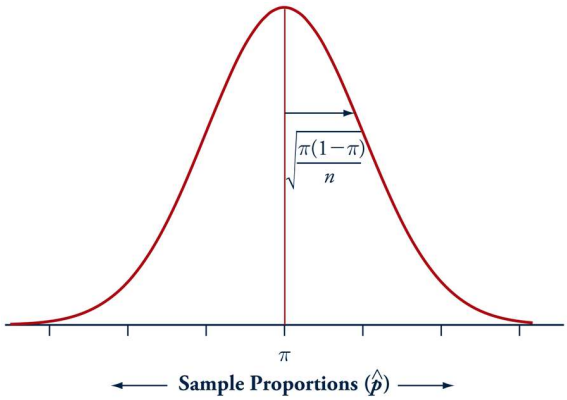
We can predict 3 things about the distr of sample proportions $\hat{p}$:

- Shape is approx. normal.

  (❗ will hold if $n\pi \geq 10$ **and** $n(1-\pi) \geq 10$.)

- Mean will be at π.

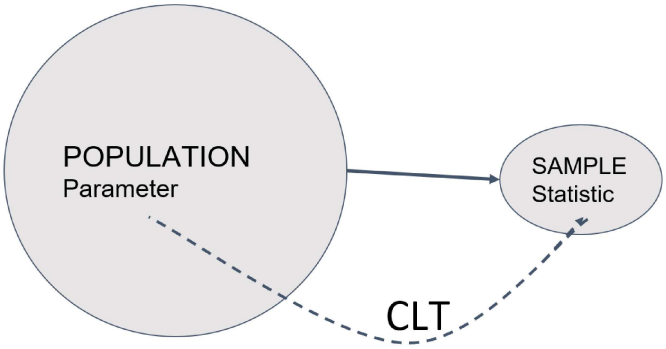
- $\hat{s} = \sqrt{\frac{\pi(1-\pi)}{n}}$.



CLT lets us predict the shape/center/spread of statistics $\hat{p}$ we'll observe given a certain parameter π. (three S's)

We can predict how far away from the parameter our statistics might get.

CLT describes relationship between parameter and statistic.



❗ Spread of sampling distr $\left(\sqrt{\frac{\pi(1-\pi)}{n}}\right)$ also depends on location (π) of the center of the distr. If the true parameter value π is near the "edges" of graph (0 or 1), there's less spread than if it's near the center.

$$\sqrt{\frac{1(1-1)}{n}} = \sqrt{\frac{0}{n}} = 0 = \sqrt{\frac{0}{n}} = \sqrt{\frac{0(1-0)}{n}}$$

Example: Le Moyne College advertises a 4-year graduation rate of 58%.

You take a sample of 50 alumni. You ask them whether they graduated in 4 yrs.

n, π, $\hat{p}$??

- $n = 50$
- $\pi = 0.58$ - proportion of Le Moyne students who graduate in 4 yrs.
- $\hat{p}$ is proportion in **our sample** who graduated within 4 yrs.

How to use CLT to describe distr of stats $\hat{p}$ we expect to see.

Step 1: Make sure CLT holds.

CLT holds if $n\pi \geq 10$ and $n(1-\pi) \geq 10$.

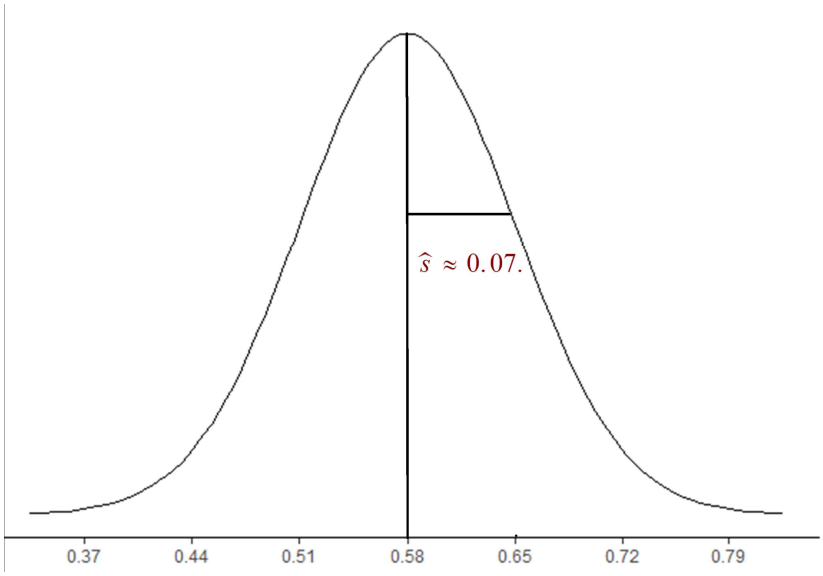
And, $50(0.58) = 29 \geq 10$ and $50(1-0.58) = 21 \geq 10$.

Step 2: Describe sampling distr

Upon repeatedly sampling, the distr of statistic $\hat{p}$ would be: Shape / Center / $\widehat{s}$?

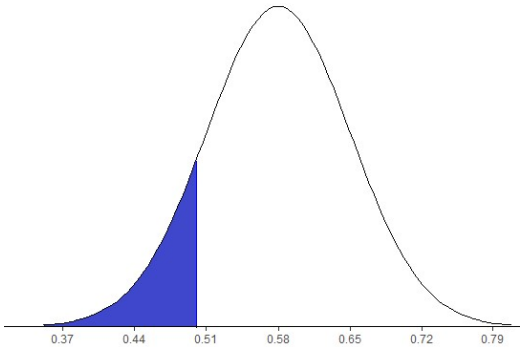
- Shape: normal,
- Center: centered at $\pi = 0.58$,
- $\widehat{s} = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.58(1-0.58)}{50}} \approx 0.07.$

Step 3: Sketch distr (shape/center/$\widehat{s}$)



Recall: we know how to learn things from normal curves: Normal probabilities answer questions like:

Probability that fewer than half of people in our sample have graduated in four years?



z-score for $x = 0.5$

The z-score:

$$z = \frac{\text{observation - mean}}{\hat{s}}$$

$$= \frac{0.50-0.58}{0.07} \approx -1.14.$$

Area to the left (table value) is 0.1271.

Interpretation (read carefully below) ??

For a random 50-person sample, there's a 0.1271 probability that:

fewer than half of those sampled will report they graduated in 4 yrs.

Equivalently: 12.71% of all 50-person-samples will have fewer than half report that they graduated in 4 yrs.

Activities: 13-2

What did we learn?

- Statistic → Sample, Parameter → Population
- Sampling Distr
- Spread (SD) of sampling distr: $\hat{s} = \sqrt{\frac{\pi(1-\pi)}{n}}$
- CLT: shape/center/$\hat{s}$. Holds if $n\pi \geq 10$ & $n(1-\pi) \geq 10$.