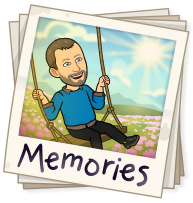
Introduction to Statistics I

Textbook: Elementary Statistics (4th Edition, by Navidi and Monk), and Workshop Statistics (4th Edition, by Rossman and Chance).

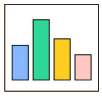
Previous Lecture



- Distr of Quantitative Vars
- Quant CLT: Shape/Center/SD. If SRS & pop. is normal or $n \geq 30$
- Symbols. Pop: p, μ, σ. Sample: $\bar{x}$, s. Sampling Distr: $\mu_{\bar{x}}$, se.

§7.4: Central Limit Theorem for Proportions

Previously we looked at sampling distr's and CLT with **quantitative** vars.
How about CLT w/**qualitative** vars?



Qual Distr

Can repeated sampling tell us something about the parameter?

yellow, brown, and orange candies

Example: We want to know what proportion of Reese's Pieces are orange.
You sample 25 Reese's Pieces and count how many are orange.
Then, you calculate the *proportion* of candies that are orange.
Observational unit? Variable? Variable type?

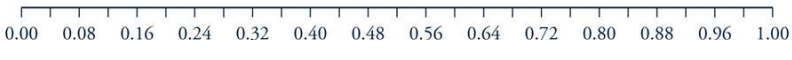
- Observational unit: **the candies**.
- Variable: **the color**.
- Variable type: **qualitative** (binary: "orange or not").

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** - p.

Let's do it! (allergies?)

Results



dotplot of students' sample proportions

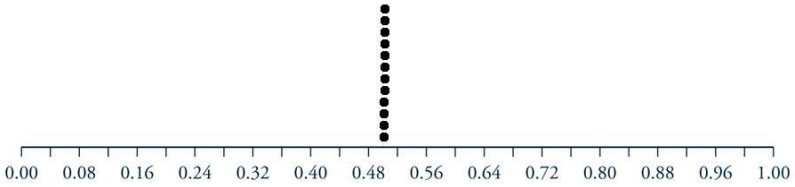
bit.ly/introstatsdata

Applet: DotPlot

Looking at the distr of everybody's *proportions in the dotplot*, what is the:

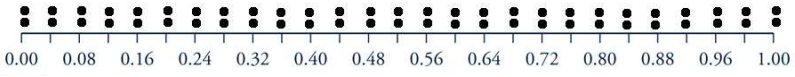
Observational unit? Variable? Variable type?

Shape? Center? Spread? How do these relate to the parameter?



This is a sample where there is no spread: it's unreasonable.

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



Sample where there is an even spread: also unreasonable.

Sampling Distr: The *way* the statistic $\hat{p}$ varies from sample to sample: Shape/Center/Spread.

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

Sample of size (n): 25

Edit Proportion (p): 0.45 (assumption)

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 for Proportions

So we can describe the distr's as:

- Shape - Normal
- Center - $\mu_{\hat{p}}$, centered around parameter p
- Spread ??

Example: Imagine you flip a coin. We know the prob of heads to be: $p = \frac{1}{2}$.

It turns out that the population SD is also: $\sigma = \frac{1}{2}$.

Let's pretend we **don't know** p, and instead we try to determine it **empirically**.

Head

Tail

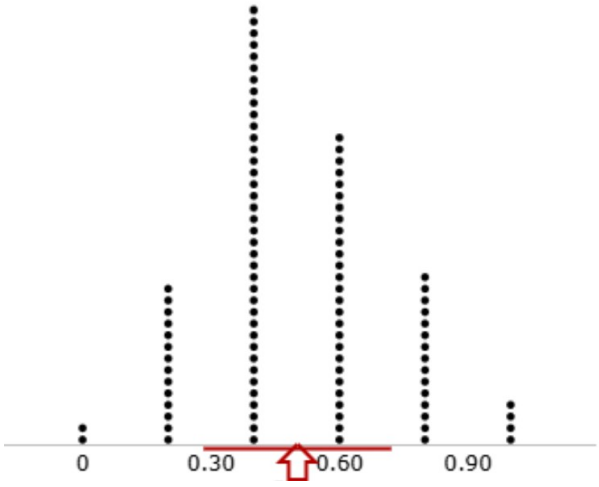
H *T*

Imagine flipping the coin 5 times, this will be a sample from the "population" with proportion heads $\hat{p}$.
Sampling 100 times, we get the following:

H,T,H,T,T	*H,T,T,T,T*	*H,T,H,H,H*	*T,H,T,H,T*	*H,H,H,H,H*	*T,T,T,H,H*	*H,T,H,T,T*	*H,H,H,T,T*	*H,H,T,T,T*	*T,T,H,H,T*
H,T,T,T,H	*H,H,H,T,T*	*H,H,H,H,T*	*H,H,H,T,T*	*T,H,T,H,T*	*T,T,H,H,T*	*T,T,H,T,T*	*T,T,T,H,H*	*H,T,H,T,T*	*T,H,H,H,T*
T,H,T,H,T	*T,T,H,T,H*	*T,T,H,H,T*	*T,H,H,T,H*	*T,H,T,H,H*	*H,T,T,H,T*	*H,T,H,T,T*	*T,H,H,T,H*	*T,T,T,H,T*	*T,H,H,T,T*
T,H,H,T,H	*T,H,T,H,H*	*H,H,H,H,T*	*H,H,T,T,T*	*T,T,H,T,T*	*H,H,H,T,H*	*H,H,H,T,H*	*H,H,H,T,T*	*H,H,T,T,T*	*T,T,H,T,H*
H,H,T,H,T	*H,T,T,H,H*	*H,T,H,T,H*	*H,H,T,H,H*	*H,T,T,T,T*	*T,T,T,H,T*	*T,H,H,H,T*	*H,H,H,H,H*	*H,H,H,T,T*	*H,T,T,H,H*
H,H,T,T,T	*H,T,T,T,H*	*T,H,T,T,H*	*H,T,T,T,T*	*H,H,T,H,H*	*T,T,H,H,H*	*H,T,T,H,T*	*H,H,T,H,T*	*T,H,T,H,T*	*H,T,H,H,T*
H,T,H,T,T	*T,H,H,H,H*	*T,T,H,H,T*	*T,H,T,T,H*	*T,H,H,H,T*	*T,T,H,T,T*	*H,H,H,T,T*	*T,H,H,H,T*	*H,H,H,H,T*	*H,H,T,T,H*
T,T,T,H,H	*H,H,T,H,H*	*T,T,T,T,H*	*H,T,T,T,T*	*T,T,T,T,T*	*T,T,T,T,T*	*H,T,H,H,H*	*H,T,H,H,H*	*T,T,T,T,T*	*T,T,T,H,T*
T,T,H,T,T	*H,H,H,T,T*	*H,T,T,H,T*	*T,T,H,H,H*	*H,T,T,T,H*	*H,T,T,T,H*	*H,T,H,T,H*	*T,T,H,T,H*	*H,H,H,H,H*	*T,T,T,H,T*
H,H,H,H,H	*T,H,H,H,T*	*H,H,T,H,T*	*H,H,T,H,T*	*T,H,H,T,T*	*H,T,H,H,H*	*T,H,T,H,T*	*H,T,T,T,H*	*T,H,T,T,H*	*H,H,H,H,H*

The proportions $\hat{p}$ of heads for each sample are:

0.4	0.2	0.8	0.4	1	0.4	0.4	0.6	0.4	0.4
0.4	0.6	0.8	0.6	0.4	0.4	0.2	0.4	0.4	0.8
0.4	0.2	0.6	0.4	0.6	0.6	0.4	0.4	0.4	0.4
0.4	0.6	1	0.4	0.2	0.6	0.8	0.8	0.4	0.2
0.8	0.4	0.6	0.8	0.4	0.2	0.6	0.8	0.8	0.4
0.6	0.2	0.4	0.4	0.6	0.6	0.6	0.6	0.4	0.6
0.4	0.6	0.6	0.2	0.8	0.2	0.6	0.6	0.8	0.4
0.4	0.8	0.2	0.4	0	0	0.6	0.8	0.2	0.2
0.2	0.6	0.4	0.4	0.4	0.4	0.6	0.4	1	0.4
0.8	0.8	0.6	0.6	0.4	0.6	0.6	0.2	0.4	1



The **mean of the sampling distr** is the average of these values, giving us an approximation for p: $\frac{\text{Sum of } \hat{p}\text{'s}}{\text{\# of samples}} = \frac{50.2}{100} = 0.502$. (pretty good!)

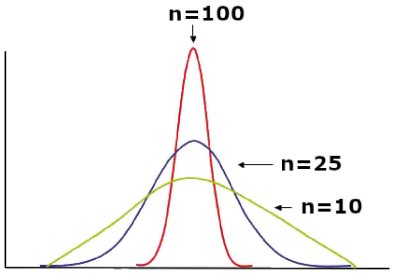
We can also calculate the SD of this sampling distr (using the methods of §3.2), giving us: 0.2193.
Note that this is less than the population SD of 0.5.

In fact, it turns out that the sampling distr SD is $\sqrt{\frac{p\,(1-p)}{n}}$, where n is the sample size.

So, for us this would be: $\sqrt{\frac{\frac{1}{2}(1-\frac{1}{2})}{5}} \approx 0.2236$. (pretty good!)

Spread vs Sample Size

Just as with quantitative vars, the spread of the sampling distr decreases as we increase the sampling size.

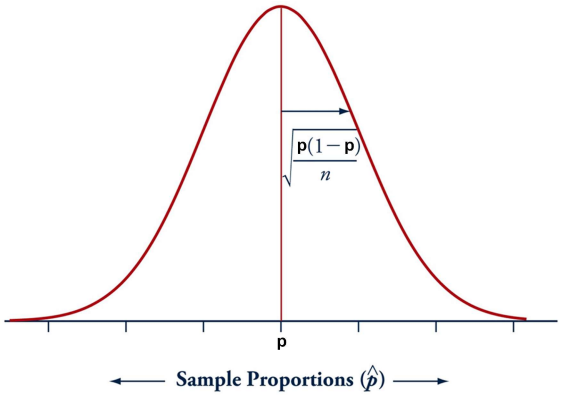


As sample size increases, spread decreases

Central Limit Theorem (CLT) (for qualitative vars)

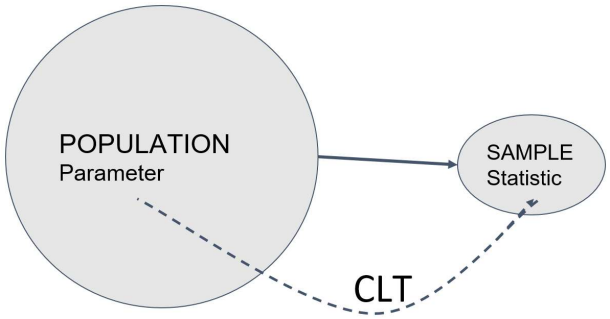
CLT: Suppose a sample of size n is taken from a population w/parameter p .
We can predict three things about the distr of sample proportions $\hat{p}$:

- Shape: approx. normal.
- Center: mean $\mu_{\hat{p}}$ will be at p .
- Spread (std error): $se = \sqrt{\frac{p\,(1-p)}{n}}$.



Two Tech Conditions - CLT holds if:

- The samples are SRS, and
- The sample size is "big enough":
  $np \geq 10$ **AND** $n(1-p) \geq 10$.



CLT describes the relationship between the parameters and statistics.

CLT predicts how far away from the parameter p our stats $\hat{p}$ tend to get.

Example: Le Moyne College advertises a 4-year graduation rate of 58%.
You take a SRS of 50 alumni. You ask them whether they graduated in 4 yrs.
n, p, $\hat{p}$??

- $n = 50$
- $p = 0.58$ - 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 the distr of stats $\hat{p}$ we expect to see.

Step 1: Make sure CLT holds.

CLT holds if SRS and $np \geq 10$ and $n(1-p) \geq 10$.

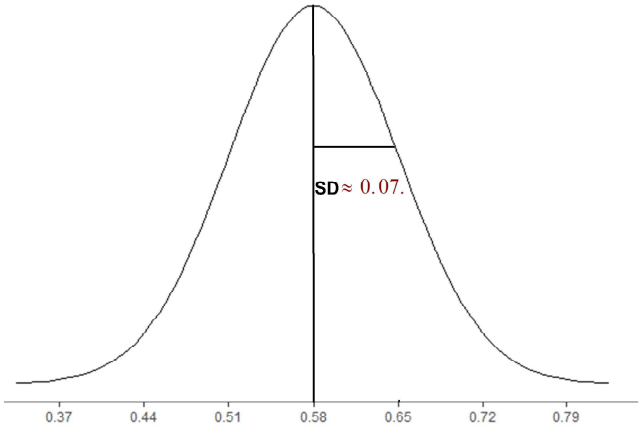
And, $50(0.58) = 29 \geq 10$ and $50(1-0.58) = 21 \geq 10$.

Step 2: Apply CLT.

Upon repeatedly sampling, the distr of statistics $\hat{p}$ will be: Shape / Center / $\sigma_{\hat{p}}$?

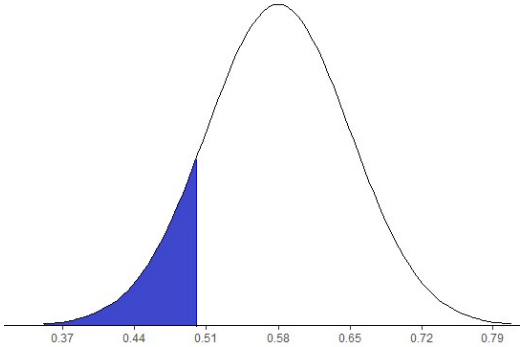
- Shape: normal,
- Center: $\mu_{\hat{p}}$, centered at $p = 0.58$,
- Spread (std error): $se = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.58(1-0.58)}{50}} \approx 0.06980.$

Step 3: Sketch distr (shape/center/spread)



Recall: we know how to learn things from normal distr's. So, w/proportion sample distr's, we can answer questions like:

For a random sample, what's the probability that ($\hat{p} < 0.5$) fewer than half of the people graduated in four years?



z-score for $\hat{p} = 0.5$

The z-score:

$$z = \frac{\text{observation - population proportion}}{\text{std error}} = \frac{\hat{p}-p}{se}$$

$$= \frac{0.50-0.58}{0.06980} \approx -1.146.$$

Area to the left is 0.1259. (using z-score calculator)

Interpretation?? (read carefully below)

Percent of Samples: 12.59% of all 50-person-samples will have fewer than half who graduated in 4 yrs.

Prob of a Single Sample: For a random 50-person sample, there's a 12.59% prob that fewer than half of your sampled alumni will have graduated in 4 yrs.

CLT Comparison

	Quantitative Var (mean)	**Qualitative Var** (proportion)
Sample Statistic	$\bar{x}$	$\hat{p}$
Tech Cond - SRS and:	When population is normal or $n \geq 30$	When $np \geq 10$ and $n(1-p) \geq 10$
Shape of Sampling Distr	Normal or Approximately Normal	Approximately Normal
Standard Deviation (std error)	$se = \frac{\sigma}{\sqrt{n}}$	$se = \sqrt{\frac{p(1-p)}{n}}$
Center (same as pop)	μ	p

Activity: 7.4a

What did we learn?

- Intuition behind Qualitative CLT
- Qualitative CLT results - sample distr has:

  Shape is ≈normal, center $\mu_{\hat{p}}$ is at parameter μ, spread is $se = \sqrt{\frac{p(1-p)}{n}}$

- CLT technical conditions: SRS & $np \geq 10$ & $n(1-p) \geq 10$.