Introduction to Statistics I

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

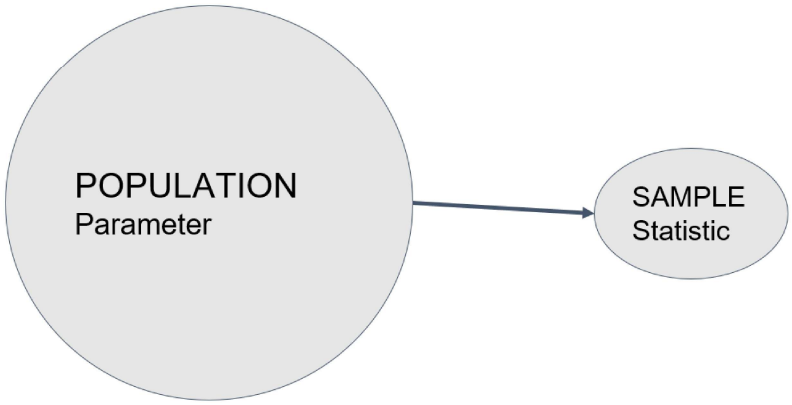
Previous Lecture

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

§7.3: Sampling Distributions for Central Limit Theorem

Previously: Using a sample distr and a mathematical model (normal distr) to estimate probabilities about the population.

Now: How to use stats ($\bar{x}$) to estimate the parameter (μ)?



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

Math Cat Says:

For our statistic, we'll focus on the sample **mean** ($\bar{x}$) (we could also look at other stats like the sample's: median, range, IQR, SD, etc).

To estimate μ, we first ask, "how do the stats $\bar{x}$ vary from **sample to sample**?"

Example: Your job at the Reese's pieces company is to ensure consistency of the candy. You sample 10 Reese's pieces and find their diameters (in millimeters):

9.62	9.49	9.57	9.38	9.5	9.53	9.55	9.47	9.47	9.52

. The average is $\bar{x} = 9.51$.

Questions to Explore:

- This is just one sample. What happens when we sample again?
- How does the sample mean $\bar{x}$ relate to population mean μ?
- And what role does the SD play in this relationship?

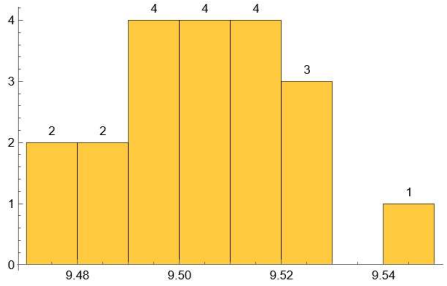
So, you take 20 samples, each consisting of 10 Reese's pieces from the "population," checking the diameter of each candy. You find the following results from your samples (each row is a sample of 10 Reese's).

9.62	9.49	9.57	9.38	9.5	9.53	9.55	9.47	9.47	9.52
9.43	9.53	9.43	9.45	9.45	9.5	9.49	9.5	9.41	9.55
9.53	9.45	9.45	9.49	9.55	9.54	9.5	9.48	9.57	9.44
9.54	9.57	9.46	9.48	9.55	9.55	9.59	9.47	9.53	9.44
9.45	9.45	9.46	9.47	9.5	9.52	9.44	9.62	9.45	9.49
9.48	9.53	9.5	9.48	9.57	9.6	9.5	9.6	9.62	9.47
9.45	9.53	9.48	9.48	9.52	9.51	9.59	9.52	9.56	9.51
9.58	9.47	9.5	9.54	9.43	9.54	9.54	9.48	9.38	9.54
9.48	9.48	9.42	9.47	9.52	9.51	9.48	9.5	9.39	9.49
9.45	9.51	9.53	9.48	9.56	9.49	9.48	9.41	9.49	9.51

9.41	9.48	9.44	9.47	9.52	9.46	9.45	9.55	9.45	9.49
9.52	9.54	9.54	9.48	9.46	9.48	9.58	9.46	9.48	9.59
9.44	9.52	9.55	9.52	9.52	9.42	9.45	9.55	9.51	9.47
9.5	9.48	9.49	9.53	9.48	9.47	9.47	9.43	9.42	9.5
9.57	9.46	9.51	9.53	9.44	9.45	9.45	9.47	9.53	9.55
9.57	9.47	9.5	9.61	9.49	9.51	9.52	9.5	9.56	9.41
9.5	9.49	9.57	9.51	9.6	9.61	9.42	9.5	9.51	9.54
9.61	9.47	9.43	9.45	9.54	9.54	9.4	9.58	9.53	9.47
9.56	9.45	9.54	9.49	9.53	9.48	9.48	9.53	9.52	9.54
9.54	9.46	9.48	9.52	9.53	9.42	9.52	9.56	9.55	9.48

Taking an average ($\bar{x}$) for each sample (row), we have:
$9.51,\ 9.47,\ 9.50,\ 9.52,\ 9.49,\ 9.54,\ 9.52,\ 9.50,\ 9.48,\ 9.49,\ 9.47,\ 9.51,\ 9.49,\ 9.48,\ 9.49,\ 9.51,\ 9.52,\ 9.50,\ 9.51,\ 9.50.$

Plotting these w/a histogram visualizes the **Sampling Distribution**:



Sample Means

Later on, you discover that the average **population** diameter is: $\mu = 9.5$, with a SD of $\sigma = 0.05$.

The **mean of the sampling distr** ($\mu_{\bar{x}}$) is defined as the average of your $\bar{x}$ values, giving us an approximation of μ:
$\mu \approx \mu_{\bar{x}} = \frac{\text{Sum of averages}}{\text{\# of samples}} = \frac{190.01}{20} \approx 9.5005.$ (pretty close to μ!)

We can also calculate the **SD of these sample means** $\bar{x}$ (using the methods of §3.2), giving us: 0.0178.
Note that this is much less than the population SD of $\sigma = 0.05$.

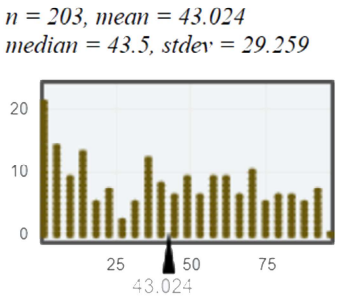
In fact, it turns out that you can calculate the sampling distr's SD as $\frac{\sigma}{\sqrt{n}}$, denoted *se* for "standard error",
where *n* is the sample size. So, for us, this would be: $se = \frac{\sigma}{\sqrt{n}} = \frac{0.05}{\sqrt{10}} \approx 0.01581$. (not too bad for just 20 samples!)
This estimate improves as the number of samples increases.

Summary

- Let $\bar{x}$ be the mean of a SRS of size n, drawn from a population with mean μ and SD σ.
- Repeated sampling results in a bunch of sample means ($\bar{x}$) which have a **sampling distr**.
- The **mean of the sampling distr** is: $\mu_{\bar{x}} \approx \mu$.
- The **SD of the sampling distr** is $se = \frac{\sigma}{\sqrt{n}}$.

Example: Internet Access

Suppose we let our population be the 203 countries of the world.
For each country, we can record the % of people w/internet access.
What are the parameters μ, σ from the given dot plot?

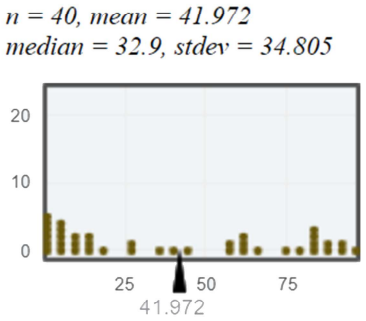


Population (each dot is a country)

$\mu = 43.02, \quad \sigma = 29.26.$

Sampling: We take a sample of 40 countries.
What symbols do we use for mean/SD from this sample?

Statistics: $\bar{x} = 41.97, \ s = 34.81.$



Sample

Let's experiment: On this applet choose (menu in upper left) "Percent w/Internet Access-2e,"
set $n = 40$, then click "Generate 1000 Samples."

What does the shape of the sample means look like?
Is it different from population distr above? SD? What if $n = 200$?

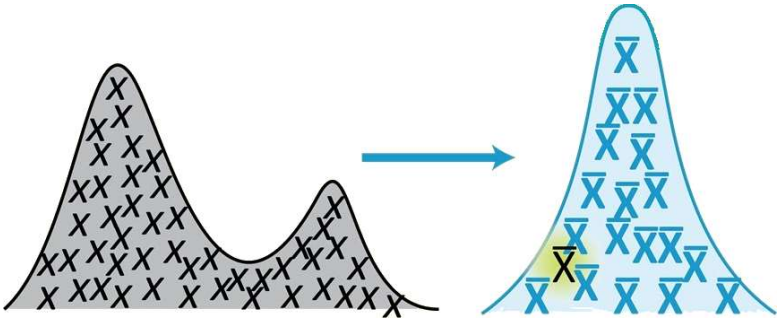
bit.ly/introstatsdata

Applet: Sampling Distr for a Mean

Sampling Distribution

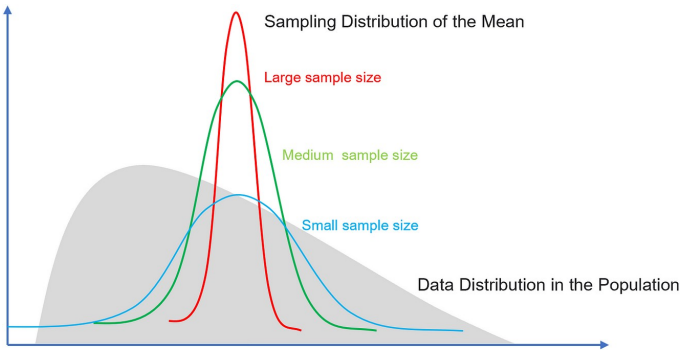
A **sampling distr** results when we take repeated samples of size n from a population w/mean μ and SD of σ.

Each time, we calculate each sample's mean $\bar{x}$.
What does the distr of these $\bar{x}$'s look like?



↑ Population → Distr of the sample means
μ

- **Shape**: Approximately normal (even if pop. distr is NOT).
- **Center**: at the population mean μ.
- **Spread**: SD decreases as n increases.



Sampling distr is approx normal, even if population distr isn't!

Central Limit Theorem (for quantitative vars)

Suppose samples of size n are taken from a population w/population mean μ and SD σ.

The distr of sample means $\bar{x}$ will have:

- **Shape** - approximately normal (exactly so if population distr is normal).
- **Mean** - at μ.
- **SD** (std error) - $se = \frac{\sigma}{\sqrt{n}}$.

Two Tech Conditions - CLT holds if:

- The samples are SRS, and
- The population you're sampling from is normal, **OR** as long as sample size is "big enough" ($n \geq 30$).

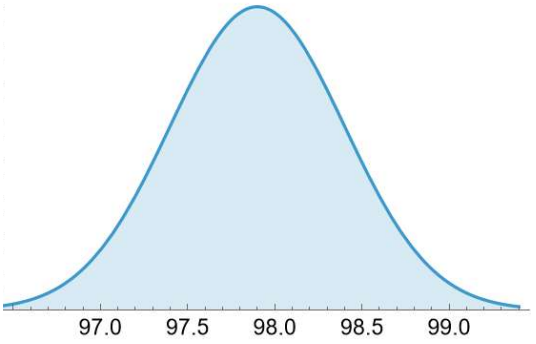
Body Temperatures

A study found that body temperatures of all healthy adults follow a normal distr w/mean of $97.9°F$ and SD of $0.5°F$. Are 97.9 & 0.5 parameters or statistics?

Symbols?

$\mu = 97.9, \qquad \sigma = 0.5$

What's the probability of a **single random adult** having a temp above 98°?



$z = \frac{\bar{x}-\mu}{\sigma} = \frac{98-97.9}{0.5} = 0.2.$

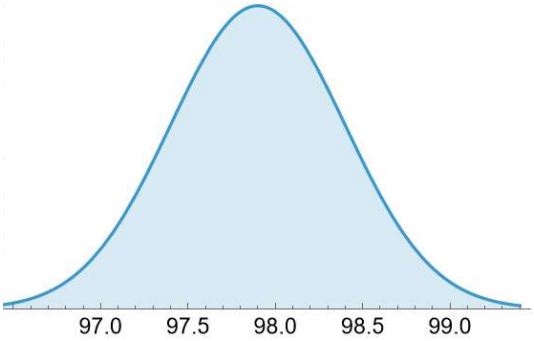
$P(x > 0.2) = 0.4207$ (from calculator.net: "z-score")

Now, let's take some samples. What if we took SRSs w/$n = 130$? Does CLT apply?

Yes, "SRS" and $n = 130 \geq 30$ (and "temps follow a normal distr").

What is the sampling distr's shape/center/spread?

Recall: $\sigma = 0.5,\ \mu = 97.9$ (from population)



Sample distr has:

- **Shape**: (exactly) Normal.
- **Center**: At population mean $\mu = 97.9$.
- **Spread**: SD (std error) is $se = \frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{130}} \approx 0.04385\,^\circ F$.

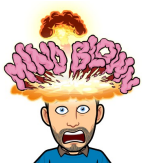
Instead of sampling one person, what's the prob of a SRS of 130 people having a sample mean above 98°?

$$z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} = \frac{98-97.9}{0.04385} \approx 2.281.$$

$P(x > 2.281) = 0.01127$ (from calculator.net: "z-score")

Recall the prob of a **single adult** having body temp above 98° is 0.4207.

But the prob of a **SRS of** 130 people having its sample mean above 98° is 0.01127.

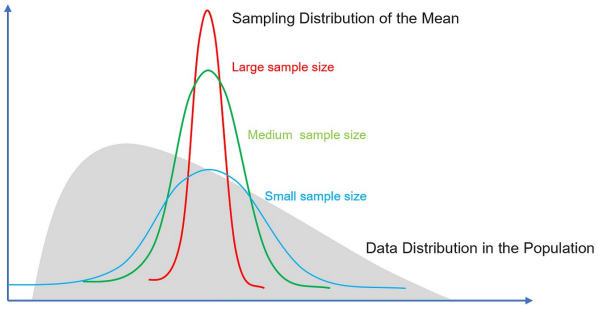


WHY?!?!

Sample Distr Spread vs n

Increasing sample size (n) ⇒ spread of statistics ($\bar{x}$) decreases.

Thus, the $\bar{x}$'s are more likely to be near μ as n increases.



Note that we are using the word "mean" in 3 different ways:

- **Population mean** μ
- **Sample mean** $\bar{x}$
- **Mean** ($\mu_{\bar{x}}$) **of the sample means**. Also known as the **mean of the sampling distr**.

Notation	Parameter (Population)		Statistic (Sample)		Sampling Distr	
Proportion (Qual)	p					
Mean (Quant)	μ	"mu"	$\bar{x}$	"x-bar"	$\mu_{\bar{x}}$	"mu sub x-bar"
Standard Deviation	σ	"sigma"	s		se	"standard error"

Activities: 7.3a

What did we learn?

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