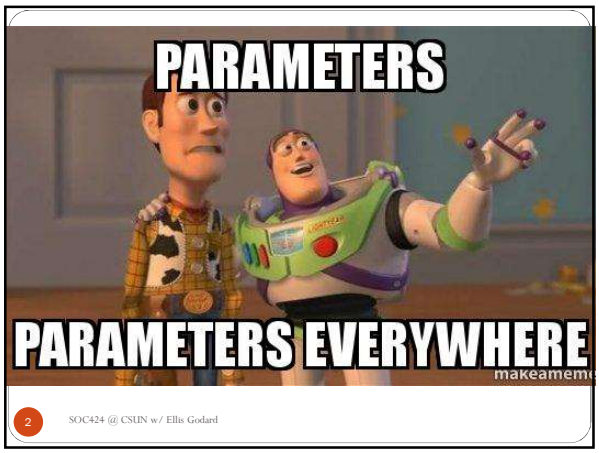


Reminders about Recoding

- **I gave 7 solutions for missing values**
  - Recoding wasn't one of them
- **I gave three options** (or patterns) **for recoding**
  - **Meaningful** division – logical/theoretical "cutoff points"
    - You combined Like & Like very much, Dislike & Dislike very much
    - Less than 9 years school is no HS; 9-11 is some HS; etc.
  - Two options involve data-derived "cutoff points"
    - **Graphical** division – look for clusters or "humps" in the histogram
    - **Statistical** division – e.g. even thirds, using cum.percent (will do in 497)
  - Note that none of those are about resolving missing values
- **Recoding & Missings are separate & unrelated**
  - Do NOT recode to try & address missing values – it WON'T

SOC424 – Statistics w/ Dr. Ellis Godard

Parameters & Estimation

Recoding Doesn't Fix Missings

- **You *might* choose to recode in this class**
  - e.g. if you can't find an ordinal, create one from an interval
  - But NEVER recode without a reason – & missing values isn't one!
    - **Recoding does not have anything to do with missing values**
      - Changing 99 to something else, doesn't tell SPSS to *ignore* 99
      - The *only* way to tell SPSS to do that, is put 99 in the missing column
      - That's what that column heading means, and why the column exists
    - You should always check for missing values & consider whether recoding is needed – but those are no more related than if I tell you to brush your teach in the morning and eat a good breakfast
      - Toothpaste is not breakfast, & breakfast doesn't clean your teeth

Outline for Today…

- **Misc. Pieces**
  - Reminders about Recoding (x2)
  - Definitions of Outliers
- **Estimates, & Qualities of Estimates**
- **Distributions, Errors, & Hypotheses**
- **The Central Limit Theorem & Sampling**
- **Estimates, Again**
- **Lab**

"Outlier" is Vague & Varies

- **Extreme value? If lots, "straggly tails"?**
- **But no *single* definition, so must *define* it**
  - One option: "any cases more than 1.5xIQR from mean"
    - IQR: the value of the 75th percentile minus value of the 25th
    - If mean age = 30, 25th percentile = 25, & 75th percentile = 35:
      - IQR = 75th – 25th, so 35 – 25 = 10
      - 1.5 x IQR = 1.5 x 10 = 15
      - Mean +/- (1.5 x IQR) = 30 +/- (15) = 30-15 & 30+15 = 15 & 45
    - Consider as outliers any values <15 or >45 (more than 1.5 IQR from Y)
      - If there are many cases with those values, call that "scraggly tails"
  - Another option (p.54): more than 1.5xIQR from *IQR*
    - 25th minus 1.5xIQR, and 75th *plus* 1.5xIQR

Statistical Inference

- Traditionally divided into two areas:
  1. **Problems of estimation**
     - Chapter Five
     - This and the next 4 lectures
     - Inferring parameters (population) from statistics (sample)
  2. **Testing of hypotheses**
     - Chapter Six
     - The remainder of the semester (lecture 16 to the end)
     - Assessing whether difference between 2 numbers is significant

Correspondence: Statistics & Parameters

Sample Statistics	Population Parameters
Mean ($\overline{Y}$)	**Mean (μ)**
Mode	Mode
Median	Median
Range	Range
IQR	IQR
Variance (s^2)	Variance (σ^2)
Std. Deviation (s)	**Std. Deviation (σ)**
Variation Ratio	Variation Ratio

Focus for now

Two *Kinds* of Estimation

- **1.A - Point Estimates**
  - Estimating parameters as a single number
  - Today & the next 2 lectures
- **1.B - Interval Estimates**
  - Estimating a range in which the parameter probably lies
  - 2 lectures after that (one interval, one categorical)

1st Property of a Good Estimator

- **An estimator is *unbiased* if the sampling distribution is centered around the parameter**
  - (e.g. sampling distribution of means has a mean that is the population mean)
  - Bias should diminish as sample size increases
- **A *biased* estimator would, on average, overestimate or underestimate the parameter**
  - Bias refers to what would happen over the long run, with repeated sampling.
- **The median can be biased by a skew in the data, whereas the mean will not be.**
  - Mean is susceptible to skew; median ignores it

Point Estimates of Parameters

A **sample statistic** is used to predict the value of the *corresponding* parameter (esp. a mean).

That prediction is associated with a level of confidence - **how confident we are** that the sample statistic is a *good* estimate of the population parameter.

2nd Property of a Good Estimator

- **An estimator with a small standard error relative to other estimators is said to be *efficient*.**
- **An *efficient* estimator tends to be relatively close to whatever parameter it estimates.**
- **The mean is more efficient than the median.**

For Our Purposes …

- **The sample mean $\bar{Y}$ is both unbiased and generally quite efficient estimator of the population mean**

and

- **the sample standard deviation s is an unbiased and quite efficient estimator of the population variance.**

Alternatives to the Sample Mean

- **One/some arbitrary observations**
  - "Person-in-the-street sample"
- **Sample median or mode**
- **"Trimmed mean"**
  - Delete highest & lowest measurements before averaging (& divide by the smaller sample size)

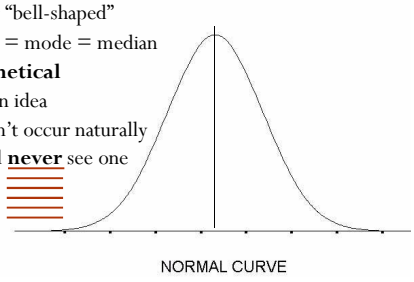
Distributions

- **Population Distribution** (real…ish?)
  - "Real" arrangement of a variable's data, in the population being studied.
- **Sample Distribution** (real, observed)
  - Arrangement (as illustrated in a histogram) of the data actually collected or observed
- **Normal Distribution** (unreal, hypothetical)
  - An arrangement w/ particular shape characteristics (similar to a "bell curve", but much more specific)
- **Now a 4th: Sampling Distribution** (completely hypothetical)

What is a Normal Distribution?

Symmetrical
- *Perfect* "bell-shaped"
- Mean = mode = median

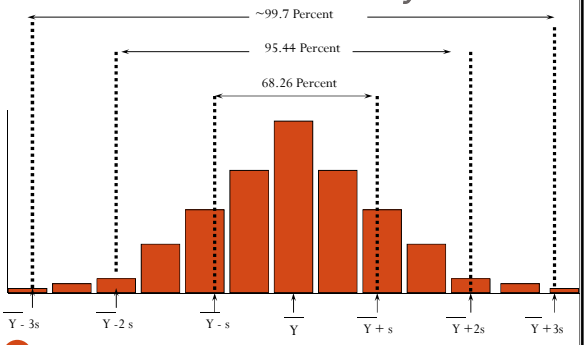
Hypothetical
- Just an idea
- Doesn't occur naturally
- You'll **never** see one



EMPIRICAL RULE *in words…*

If the distribution of a variable (as illustrated in its histogram) is normally distributed, then:
- ***About* 68 percent (68.26) of the measurements lie within one standard deviation from the mean (i.e. between Y - s and Y + s);**
- ***About* 95 percent (95.44) of the measurements lie two standard deviations from the mean;**
- **Almost *all* measurements (99.7%) lie within three standard deviations of the mean.**

EMPIRICAL RULE *visually…*



Sampling Error

- **Sample *statistics* are used to <u>estimate</u> population *parameters***
- **Statistics** (esp. mean & standard deviation) **provide *point estimates* of parameters**
- **They only *approximate* "actual" parameters**
  - We don't know them – & they might be unknowable
- **The difference between a statistic and its parameter is a *sampling error***
  - The error that results from sampling
  - For example, if the CSUN mean age is 20.1 but class mean age is 23.1, the sampling error is 3.0 (23.1-20.1)

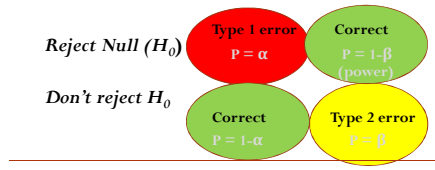
Two Kinds of Hypotheses

- **Science is <u>*not*</u> based on proof or confirmation**
  - It's based on testing ideas – about trying to *falsify* them
  - We *support* an idea by *rejecting its opposite*
- **<u>Research</u> Hypothesis**
  - The claim you want to make
    - e.g. the idea that there is a "significant difference" between 2 groups
  - Can't study that directly – instead, *all science (!!)* is about the….
- **<u>Null</u> Hypothesis**
  - Sorta the opposite of the research hypothesis
    - e.g. the idea that there is NOT a "significant difference"
    - Nothing going on – IV doesn't matter; doesn't explain the DV
  - Statistical tests address the null (not the research hypo) directly
    - Hoping to find something different enough from the null, that we can reject it
    - Science is always focused on the null – and on the risk of 2 errors….

Keep These Straight *(part A)*

- **Standard Deviation** (interval only)
  - Typical deviation from the mean
- ***Sampling* Error** (any variable)
  - Arithmetic difference btwn pop parameter & sample stat.
    - e.g. for mean education level (in years), if the parameter is 12 and the statistic is 13, the difference (the sampling error) is -1
  - *Want* sample to look like population (small sampling error)
    - Not always possible
    - Biggest problem = parameter not *known*
- ***Standard* Error** (any *statistic*)
  - Allows inferences even though sampling error unknown

Two Kinds of Errors



- **Type I is "worse"**
  - Want to reject Ho, but don't want to be wrong
- **Decide whether or not to reject based on "p"**
  - Hinges on standard errors – like z scores

Keep These Straight *(part B, preview)*

A. ***Population* distribution**
- What we *want* to say something about
- It has a standard deviation, that we don't know

B. ***Sample* distribution**
- The data we'll use to say something
- It has a standard deviation, that you can get from SPSS

C. ***Sampling* distribution**
- Hypothetical, but critical – doesn't exist, but I'll do an example
- Links A&B – *permits* inferences, w/ *confidence*
- It's standard deviation is called a *standard error* – the standard deviation of the sampling distribution

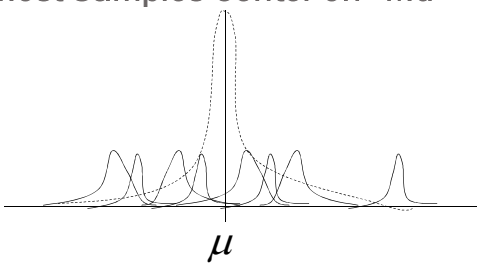
Errors are a Function of Samples

- **Samples vary**
  - But variation is somewhat predictable
  - We can estimate how much sample means vary
- **If we sample carefully, can reduce errors**
  - Esp., helps avoid Type I (rejecting a true null)
- **Key: The Central Limit Theorem…**

1st Implication of CLT: $\overline{Y}_{\overline{Y}} = \mu$

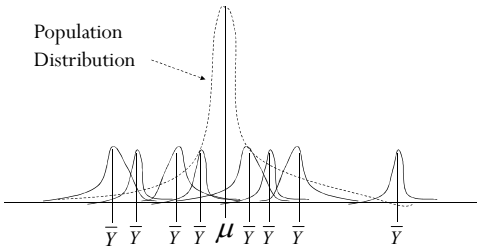
- **If we took every possible sample**
  - And they were random
  - And they were large enough
- **If we took the mean for every sample**
  - And put all of those means into a list
  - And treated them like a variable
  - W/ its own shape, central tendency, dispersion
- **That is a *sampling distribution***
  - A hypothetical collection of sample means
  - *Its* mean ($\bar{y}$ of $\bar{y}$'s) is the population mean (μ)

Most Samples Center on "Mu"



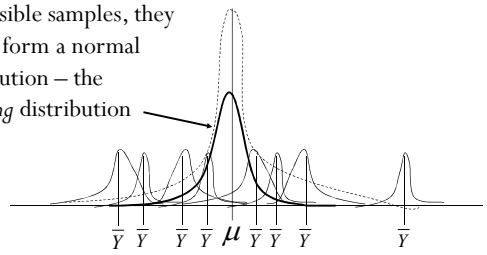
(first implication of the Central Limit Theorem)

$\bar{y}$'s all vary from "Mu" a bit



They together form a distribution

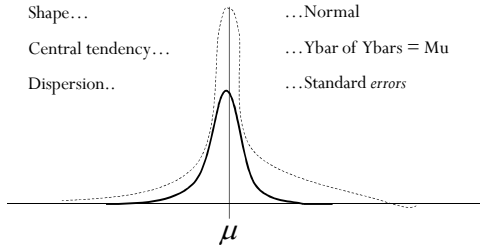
If we had *all* of the $\overline{Y}$s from *all* possible samples, they would form a normal distribution – the *sampling* distribution



Could Treat it Like Any Distribution

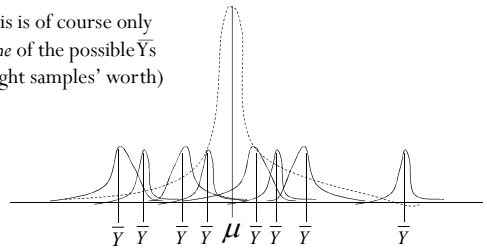
Can describe its:

Shape…	…Normal
Central tendency…	…Ybar of Ybars = Mu
Dispersion..	…Standard *errors*



Mean is (theoretically) *equal* to "Mu"

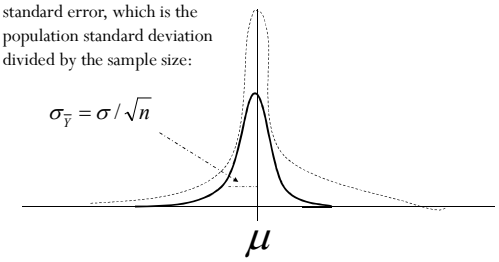
This is of course only *some* of the possible $\overline{Y}$s (eight samples' worth)



We can measure its dispersion

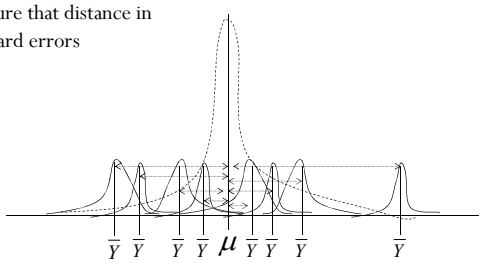
The standard deviation of that *sampling* distribution is the standard error, which is the population standard deviation divided by the sample size:

$\sigma_{\bar{Y}} = \sigma / \sqrt{n}$



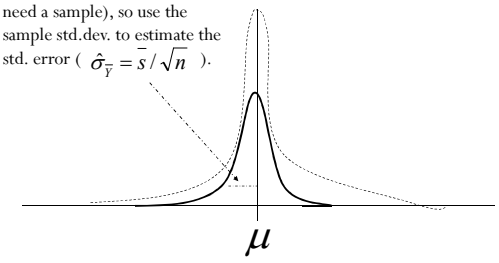
Each ȳ is a distance from Mu

If we knew Mu, we could measure that distance in standard errors



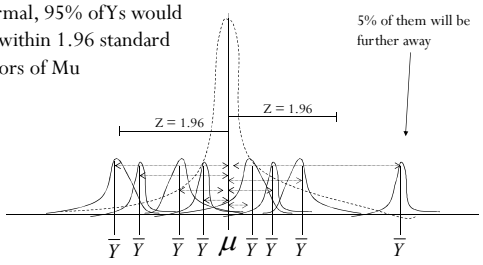
We estimate it w/ "s.e. mean"

We don't have the population std. dev. (else we wouldn't need a sample), so use the sample std.dev. to estimate the std. error ($\hat{\sigma}_{\bar{Y}} = s / \sqrt{n}$).



"Ybars" (sample means) distributed per z

Since the distribution is normal, 95% of $\bar{Y}$s would be within 1.96 standard errors of Mu



Keep These Straight *(part B, redux)*

A. ***Population* distribution**
- What we *want* to say something about
- It has a standard deviation, that we don't know

B. ***Sample* distribution**
- The data we'll use to say something
- It has a standard deviation, that you can get from SPSS

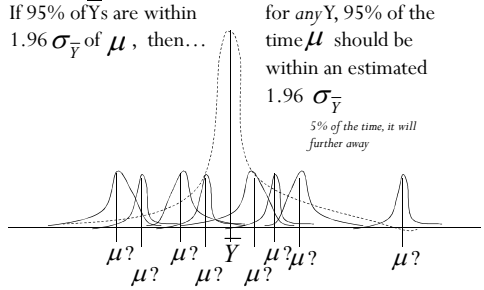
C. ***Sampling* distribution**
- Hypothetical, but critical – doesn't exist, but I'll do an example
- Links A&B – *permits* inferences, w/ *confidence*
- It's standard deviation is called a *standard error*

For estimation, invert that:

If 95% of $\bar{Y}$s are within $1.96\,\sigma_{\bar{Y}}$ of μ, then…

for *any* Y, 95% of the time μ should be within an estimated $1.96\,\sigma_{\bar{Y}}$

5% of the time, it will further away



CLT Inverted (Pts to Intervals)

***If* a sample is taken randomly and if it is sufficiently large…**

- An *estimated* 95% of the sample means will fall within 1.96 standard errors – that is, the population mean will be within 1.96 standard errors of the sample mean 95% of the time
- An *estimated* 99% of the sample means fall within 2.57 standard errors – that is, the population mean will be within 2.57 standard errors of the sample mean 99% of the time

Recent Lab: Music Index

- **Practical: Follow directions**
  - All of the steps (don't skip #1)
  - In order (don't try to do #4 before #1)
  - All of the instructions (e.g. freqs for "all your 13 new variables")
  - Use the data (esp. stats) to *tell a story*
- **Empirical: *Diversity* of tastes varies**
  - Original variables: *how much* each respondent likes each genre
  - Recoded variables: measure *whether* respondents like each
    - Some were most popular, some least
  - Index measures *how many* (not how much) genres each R likes
    - Diversity, not intensity
  - Most common (& median) is 6 of 12 – but mean higher (6.3)

Next Lab: Estimation

- **Comparing PAEDUC & MAEDUC**
  - Start w/ frequencies, histograms, stats (center & spread)
  - 1. Which is more dispersed?
    - Use appropriate statistical terms *and statistics!*
  - 2. Within what range (of years of education) do we estimate that 68.26% of the cases are for each variable?
    - Assume that the empirical rule applies.
  - 3. Are there outliers / scraggly tails?
    - Use the *first* definition from *this* lecture (in a yellow box – slide #7)
  - 4. What are the two sampling errors?
    - Assume that the population mean for both is 10 years.