General Linear Models with One Predictor

- Topics:
  - Vocabulary and broad categories of predictive linear models
  - Special cases of GLMs (and review of hypothesis testing):
    - Empty model (with no predictors)
    - "(Simple) linear regression" (with one quantitative predictor)
    - "Independent (or two-sample) t-test" with a binary predictor
  - Relating effect size, Type I errors, Type II errors, and power
  - Foreshadowing uses of the GLM

Steps in Quantitative Data Analysis

- **Quantitative data analysis:** the process of applying statistical models to a sample of data to answer your research questions
  - Enter, download, or otherwise acquire quantitative data
  - Import data into statistical software and verify its accuracy of import
  - Ask for univariate descriptive statistics to describe variables (and especially min and max to double-check accuracy of data)

- **Select a family of statistical models** based on the characteristics of the variables of interest and the questions to be answered
  - Estimate statistical models, check results for potential problems…
  - Estimate more statistical models, check results again…
  - Estimate even more statistical models… interpret results!
  - Write up the results: Btw: In formal writing, you did not "run analyses" or "calculate models"; you "**conducted analyses**" and "**estimated models**"

When research questions are phrased as what is the role of x in explaining y, below are possible labels of x and y:

Reason (Explainer):

- In my notation: $\boldsymbol{x}$ **variable**
  - Exogenous (is not explained)
- **Predictor**
  - My preferred generic term
- Independent variable (IV) or explanatory variable
  - Used more often when variable is manipulated (like treatment)
- Covariate
  - Used for reasons the researcher is not interested in (but must include to keep others happy)

What is To Be Explained:

- In my notation: $\boldsymbol{y}$ **variable**
  - Endogenous (is explained)
- **Outcome**
  - My preferred generic term
- Dependent variable (DV) or response variable
  - Used more often in studies with experimental designs
- Criterion
  - Used in observational studies with "regression" models

Roles of Variables: Some Examples

In the following example research questions, identify which variables are **predictors or outcomes** and their likely types (**quantitative or categorical**):

1. To what extent does positive feedback improve performance speed and accuracy more than neutral feedback?
   - Predictors:
   - Outcomes:
2. Is faster academic growth in elementary school related to more frequent reading to children when in preschool?
   - Predictors:
   - Outcomes:
3. How effective is teacher training for creating higher rates of positive feedback to a teacher's students?
   - Predictors:
   - Outcomes:

Types of Inferences: 2 possibilities for how x relates to y

- **x causes y → causal inference** requires the following:
  - x variable had to come first (temporal precedence)
  - x variable was under complete experimental control (random assignment + manipulation)
  - Study design eliminates all possible alternative explanations

- **x relates to y → associative inference only** (i.e., correlational)
  - We have observed a relationship, but cannot infer cause given the observational design
  - In lieu of experimental control, we can attempt **statistical control**: include other predictors as alternative explanations for why x relates to y and see if x is still related to y → common!
- These 2 possibilities can only be **distinguished by study design**—they have nothing to with the type of variables collected (a common misconception)
- Because causal inference is rarely possible in studies of real people, we will **only use associative language** in describing model results in this class

Moving On to Predictive Linear Models

- Questions concerning more than 1 variable at a time are best answered using **predictive linear models**, in which one must designate which variables are predictors and which are outcomes
- Models come in different flavors based on type of outcome variable
  - Continu-ish quantitative outcome?
    - "**General**" Linear Models using the normal distribution—this semester!
  - Literally any other kind of outcome variable?
    - "**General*ized***" Linear Models using some other distribution and a transformed expected outcome (via a "link function") to address variables' possible values and boundaries—here are some examples:
      - Binary outcome? Use Bernoulli distribution and logit link
      - Ordinal outcome? Use multinomial distribution and cumulative logit link
      - Nominal outcome? Use multinomial distribution and baseline logit link
      - Binomial outcome? Use binomial-family distributions and logit link
      - Count outcome? Use Poisson-family distributions and log link

What "Linear" in "Linear Models" Means

- Most predictive models have a "**linear**" form, which looks like this:
  - $\boldsymbol{y_i = (\text{constant} * 1) + (\text{constant} * x1_i) + (\text{constant} * x2_i)}$...
  - Fortunately, this does NOT mean that we can ONLY predict linear relationships—we can STILL specify forms of nonlinear relationships with y_i for quantitative $\boldsymbol{x_i}$ predictors
  - Fortunately, this also means we can include categorical $\boldsymbol{x_i}$ predictors (stay tuned)

- Historically, variants of the **general linear model** (for continu-ish outcomes) get siloed into different classes and names based on the **kind of $\boldsymbol{x_i}$ predictors included**:
  - Called "(linear) (multiple) **regression**" if using quantitative predictors
  - Called "analysis of variance" (**ANOVA**) if using categorical predictors
  - Called "analysis of covariance" (**ANCOVA**) if using both predictor kinds
  - We are going to cover all of these as special cases of the General Linear Model ("the GLM")—separating them does way more harm than good (all using the base R `lm` function)

Welcome to the GLM!

- Linear models use **new notation within one equation** to describe how each of the x_i predictors relate to each y_i outcome in your sample
  - 1 outcome? "**Univariate** GLM" 2+ outcomes at once? "**Multivariate** GLM"
- GLMs start by recreating central tendency and dispersion of the outcome ($\boldsymbol{y_i}$)
  - We will use **mean and variance** to describe the outcome because the GLM uses the normal distribution (in which skewness and kurtosis should be ~0)

- Your first GLM is the "**Empty**" model (= no predictors): $\boldsymbol{y_i = \beta_0 + e_i}$
  - $\boldsymbol{y_i}$ = "**y sub i**": outcome variable for each person in your sample
  - $\boldsymbol{\beta_0}$ = "**beta 0**" (sometimes called "beta not"—but not by me)
    - More generally, betas ($\boldsymbol{\beta}$) will represent constant values to be estimated that will apply to the whole sample = "**fixed effects**" (as opposed to "random effects" that vary)
    - The beta **subscripts index each fixed effect** (starting at 0, then 1, 2, …)

The "Empty" General Linear Model

- The "**Empty**" model (empty = no predictors): $\boldsymbol{y_i = \beta_0 + e_i}$
  - $\boldsymbol{\beta_0}$ = "beta 0" = "**the fixed intercept**" (or "the constant", ugh) and is defined as the predicted (expected) value for the $\boldsymbol{y_i}$ outcome when all $\boldsymbol{x_i}$ predictors = 0 *(so the estimated value for $\boldsymbol{\beta_0}$ will change as the predictors change)*
  - We don't have any predictors yet, so the intercept takes on the single most likely value for everyone—the **sample** (or "**grand**") **mean** (so in this model, $\boldsymbol{\beta_0 = \bar{y}}$)
- So what would $\boldsymbol{\beta_0}$ be for:
  - The blue line? the red line?
  - But why do the red and blue lines differ???

Univariate Normal PDF

(Probability Density Function):

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma_y^2}} * \exp\left[-\frac{1}{2} * \frac{(y_i - \mu_y)^2}{\sigma_y^2}\right]$$

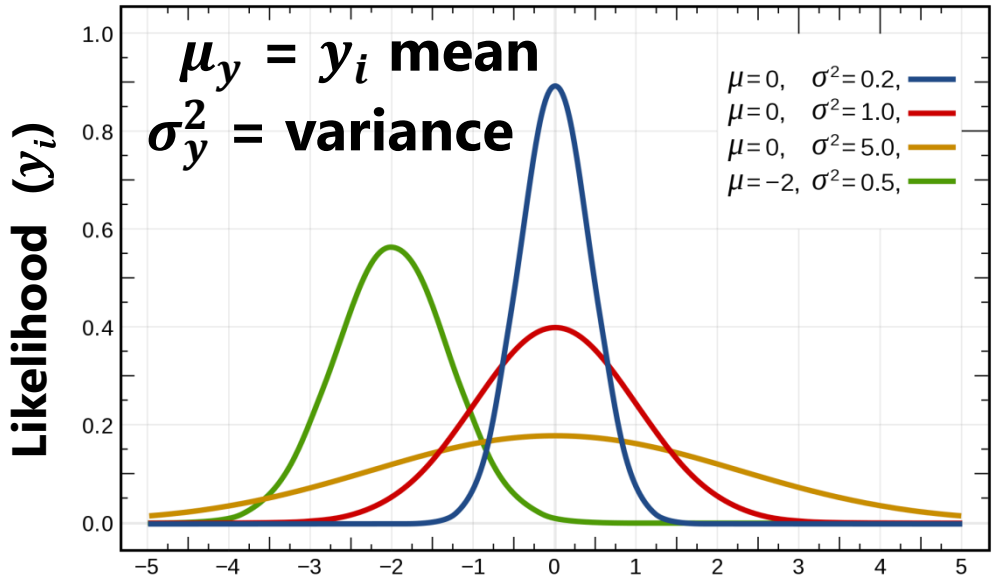


Image borrowed from: https://en.wikipedia.org/wiki/Normal_distribution

The "Empty" General Linear Model

- The "**Empty**" model ("**no predictors**"): $\boldsymbol{y_i = \beta_0 + e_i}$ (in which $\boldsymbol{\beta_0 = \bar{y}}$)
  - $\boldsymbol{e_i}$ = "**e sub i**" or "**residual**" = "**error**" = deviation of actual $\boldsymbol{y_i}$ for each person from $\boldsymbol{y_i}$ *as predicted by the model* (through the beta fixed effects)
  - Because the empty model predicts the same $\bar{y}$ for all $\boldsymbol{y_i}$ values, the $\boldsymbol{e_i}$ residual for each person will just be the difference between $\boldsymbol{y_i}$ and $\boldsymbol{\beta_0}$: $\boldsymbol{e_i = y_i - \beta_0}$
  - Rather than focusing on each individual $\boldsymbol{e_i}$ residual, we keep track of their **variance across persons** as the estimated model parameter, **denoted as** $\boldsymbol{\sigma_e^2}$
  - You've pry seen this before: $Variance(y_i) = \boldsymbol{s_y^2} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1} = \frac{\sum_{i=1}^{N}(e_i)^2}{N-1} =$ now $\boldsymbol{\sigma_e^2}$
  - In other words, the two parameters for the empty model outcome give us the $\boldsymbol{y_i}$ **mean** (as $\boldsymbol{\beta_0}$) and $\boldsymbol{y_i}$ **variance** (as $\boldsymbol{\sigma_e^2}$) → empty model $\boldsymbol{\sigma_e^2}$ = **all the** $\boldsymbol{y_i}$ **variance**

- In writing linear models, the **notation refers to population parameters** instead of sample statistics (i.e., we use $\boldsymbol{\sigma_e^2}$ instead of $\boldsymbol{s_y^2}$)
  - Why? It's understood that we only have one sample from which to estimate population parameters

R syntax for an "**empty**" model: $y_i = \beta_0(1) + e_i$

```
print("Empty GLM Predicting Income -- save as ModelEmpty")
print("1 represents fixed intercept")
ModelEmpty = lm(data=Lecture2, formula=income~1)
summary(ModelEmpty)          # Default output

Call:
lm(formula = income ~ 1, data = Lecture2)

Residuals:
   Min     1Q Median     3Q    Max
-17.06 -10.57  -3.83   4.75  51.30

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   17.303      0.509      34   <2e-16

Residual standard error: 13.8 on 733 degrees of freedom
```

In an empty model:

$$\beta_0 = \bar{y}$$

$$\sigma_e^2 = s_y^2 = var(y_i),$$

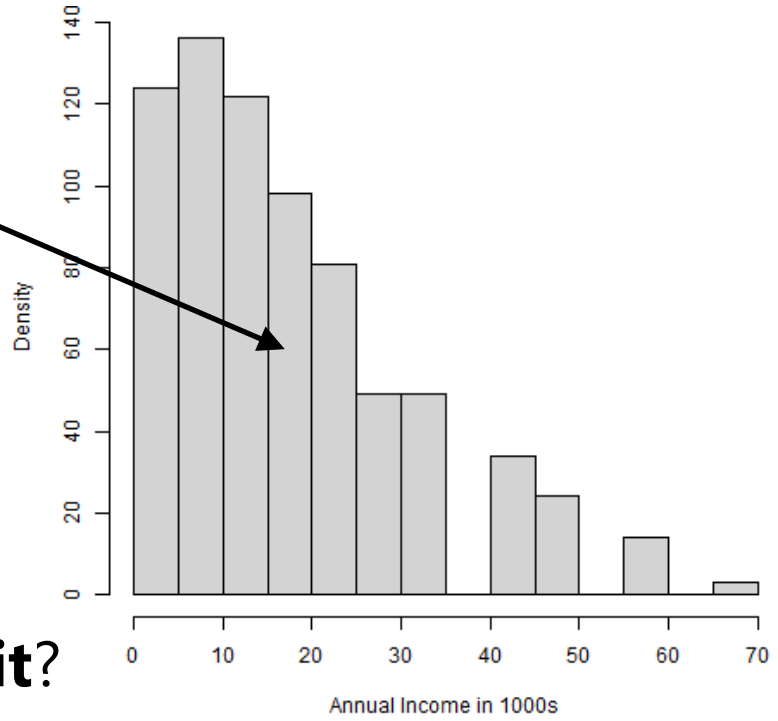
so $\sigma_e = s_y = SD(y_i)$

This default terminology of "**residual standard error**" is confusing because it really refers to the **standard deviation** of the e_i **residuals** (and standard error will now mean something else…!)

```
print("Descriptive Statistics for Quantitative or Binary Variables")
print(psych::describe(x=Lecture2[ , c("income","educ","marry")], fast=TRUE), digits=3)

        vars   n   mean    sd median   min  max range   skew kurtosis    se
income     1 734 17.303 13.792  13.47 0.245 68.6 68.36  1.156    1.075 0.509
educ       2 734 13.812  2.909  14.00 2.000 20.0 18.00 -0.230    0.777 0.107
marry      3 734  1.459  0.499   1.00 1.000  2.0  1.00  0.164   -1.976 0.018
```

Quantifying Uncertainty of Sample Estimates

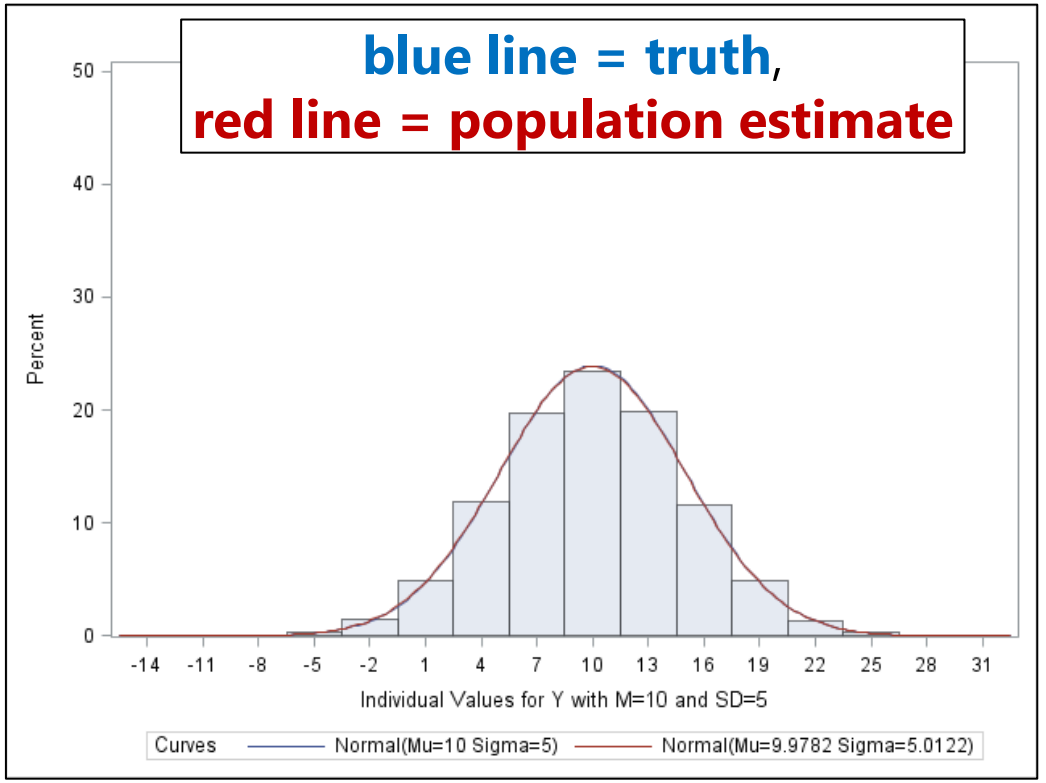
- Let's predict annual income in \$1000s in $N = 734$ using an empty model
  - $\boldsymbol{y_i = \beta_0 + e_i} \rightarrow \boldsymbol{income_i = \beta_0 + e_i}$
  - Fixed intercept estimate $\boldsymbol{\beta_0 = 17.30}$ (= sample mean $\bar{y}$)
  - Person-specific residual $\boldsymbol{e_i = y_i - \beta_0}$ (= how wrong)
  - Variance of $\boldsymbol{e_i}$ residuals → $\boldsymbol{\sigma_e^2 = 190.21}$ (= sample var $\boldsymbol{s_y^2}$)
  - SD of $\boldsymbol{e_i}$ → $\boldsymbol{\sigma_e = \sqrt{190.21} = 13.80}$ (= sample SD $\boldsymbol{s_y}$)



- If sample mean $\bar{y} = 17.30$ is our best guess for population mean $\boldsymbol{\mu_y}$, **how good of an estimate is it**?
  - How much would a new sample's mean differ from 17.30?
  - Need to know how inconsistent the sample mean is expected to be across repeated samples of same kind → from the mean's "sampling distribution"
  - **Note: This is NOT the same as a variable's distribution across persons (above)…!**

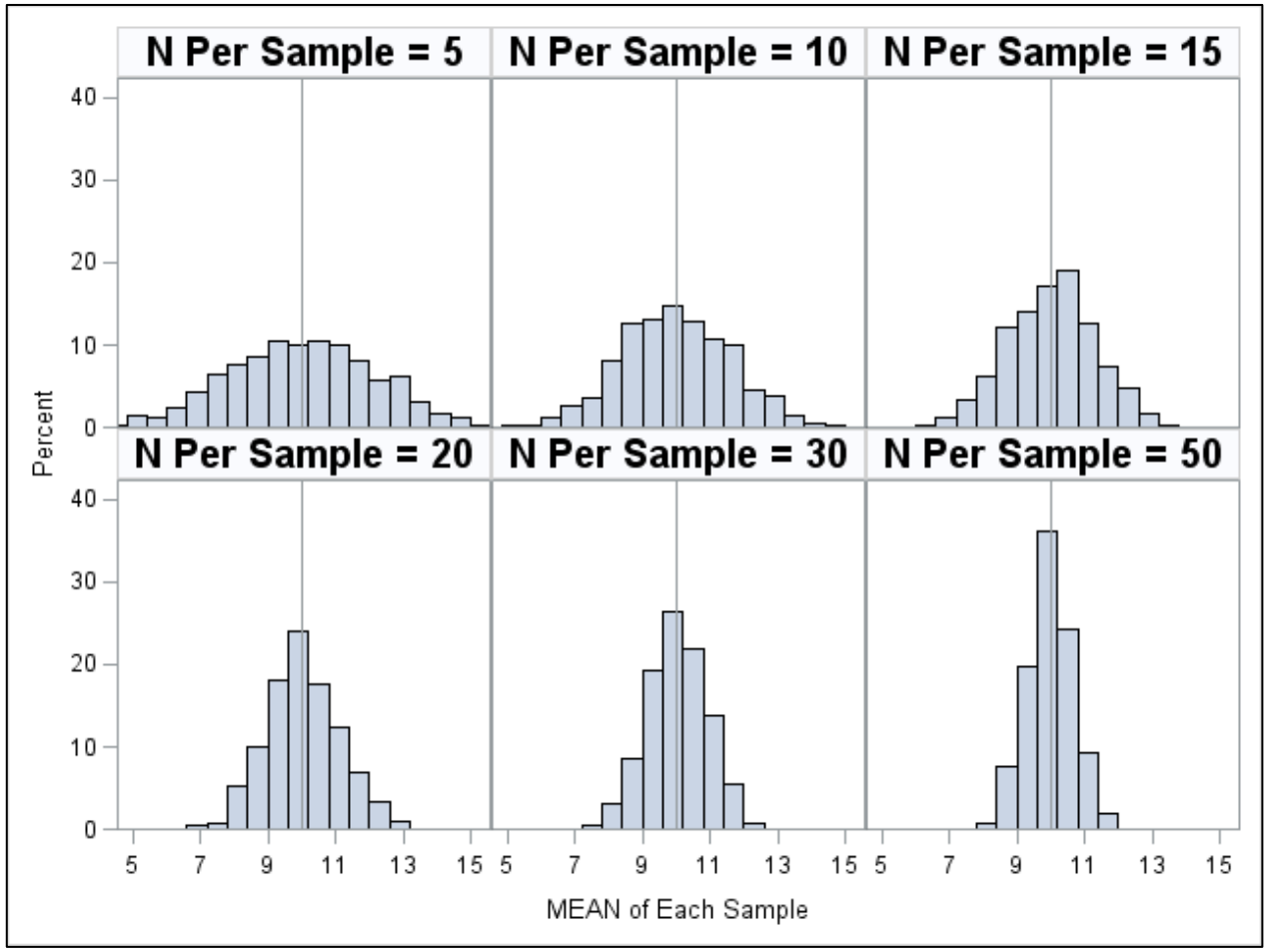
Building Intuition about Sampling Distributions of Estimates (the mean for now)

- What affects how close $\bar{y}$ is to "true" μ_y?
- Demo: I made my own quantitative variable* y_i in a population of 100,000 fake people
  - Population mean: $\mu_y = 10$
  - Population variance: $\sigma_y^2 = 25$
  - So y_i is off the mean by $\sigma_y = SD = 5$ on average



  - * Used a "**normal**" distribution here to generate y_i (as shown earlier)

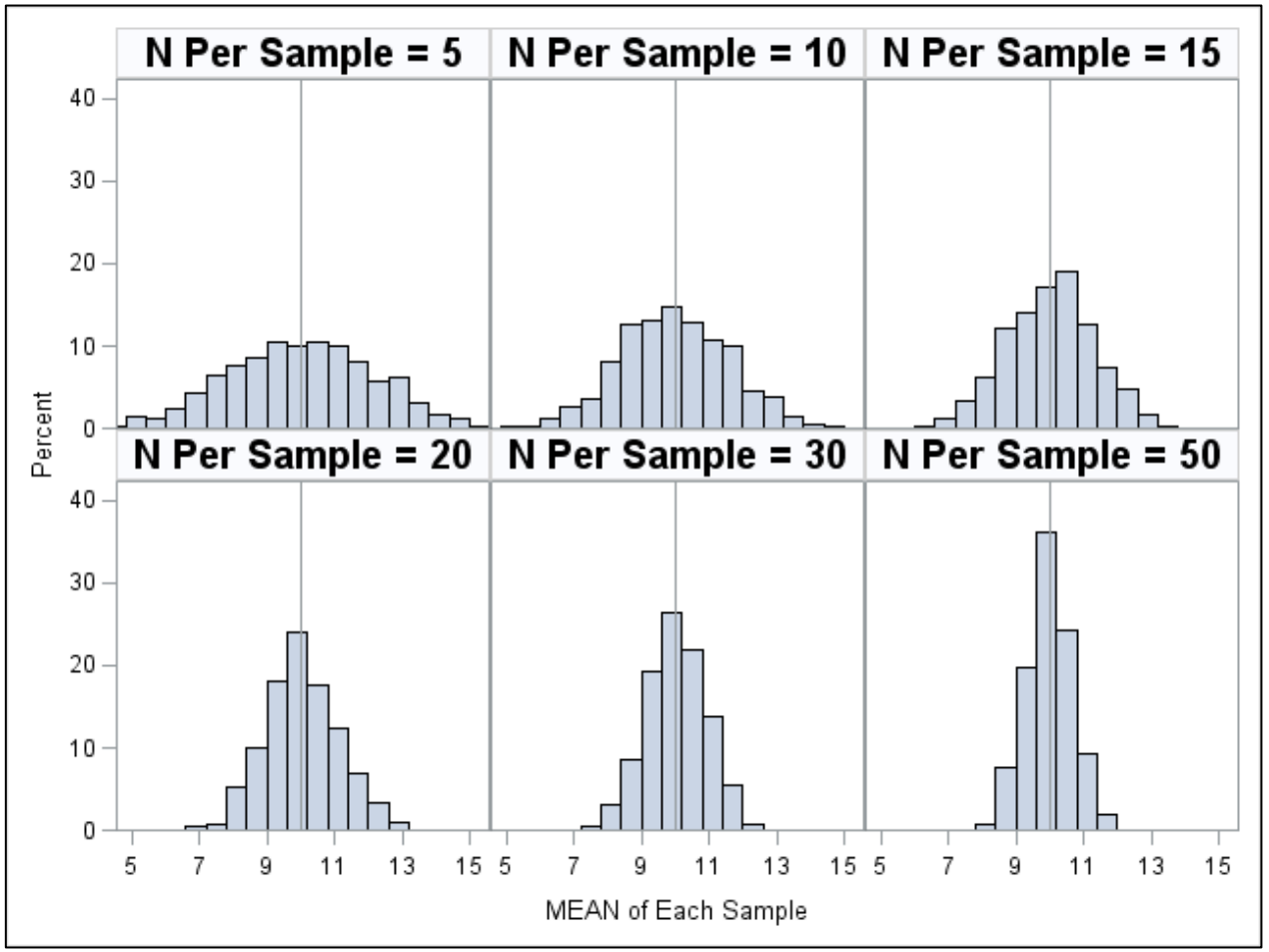
1000 samples each for different N…



Note: These bars do not show individual people!
They are summaries for **distinct samples** of people.

- Population values:
  Mean $\boldsymbol{\mu_y = 10}$, SD $\boldsymbol{\sigma_y = 5}$
- Histograms show **differences across samples** in each sample's **mean** ($\bar{y}_s$)
- These depict the N-specific "**sampling distribution**" of $\bar{y}_s$
- More $\boldsymbol{N}$ in each sample → less dispersion in $\bar{y}_s$ across samples (more consistency)

1000 samples each for different N…



Note: These bars do not show individual people!
They are summaries for **distinct samples** of people.

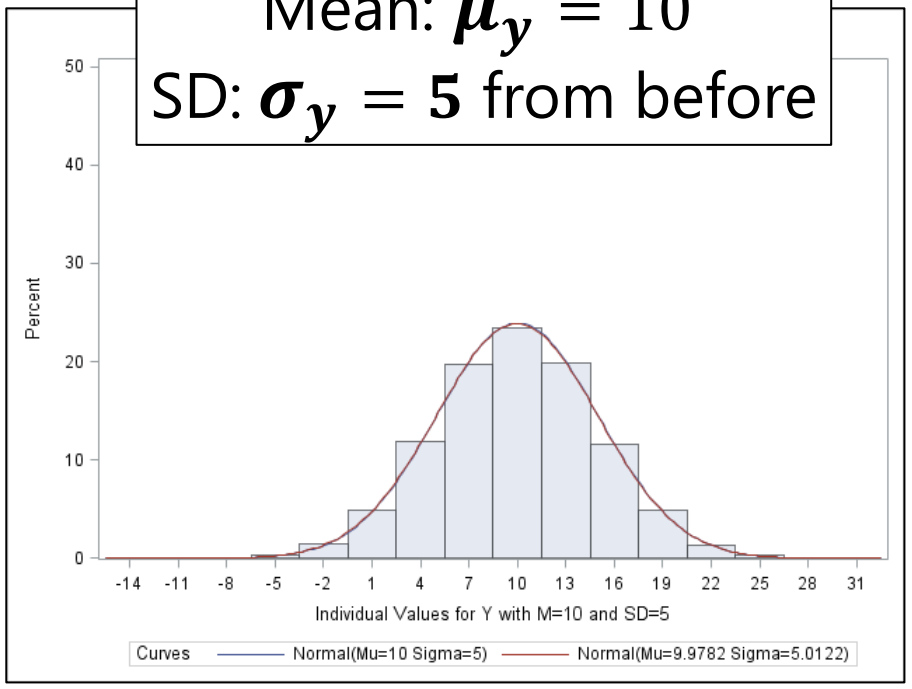
- <u>Population values:</u>
  Mean $\boldsymbol{\mu_y = 10}$, SD $\boldsymbol{\sigma_y = 5}$
- More $\boldsymbol{N}$ → **less** $\boldsymbol{SD}$ **in** $\boldsymbol{\bar{y}_s}$ across samples

N Per Sample	Mean $\bar{y}_s$	SD $\bar{y}_s$
5	9.97	2.17
10	9.98	1.60
15	10.00	1.28
20	10.03	1.08
30	10.03	0.89
50	9.97	0.69

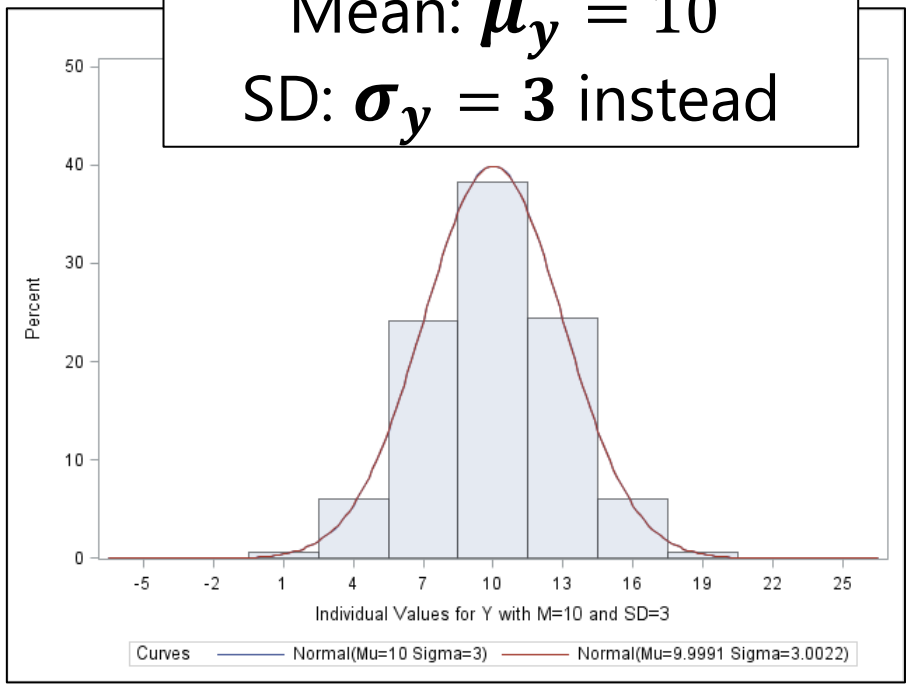
Building Intuition about Sampling Distributions of Estimates (the mean for now)

- As within-sample N **increases**, sample mean $\bar{y}$ will be **closer** to μ_y on average
- What else affects consistency of $\bar{y}_s$? How **persons vary from each other**!

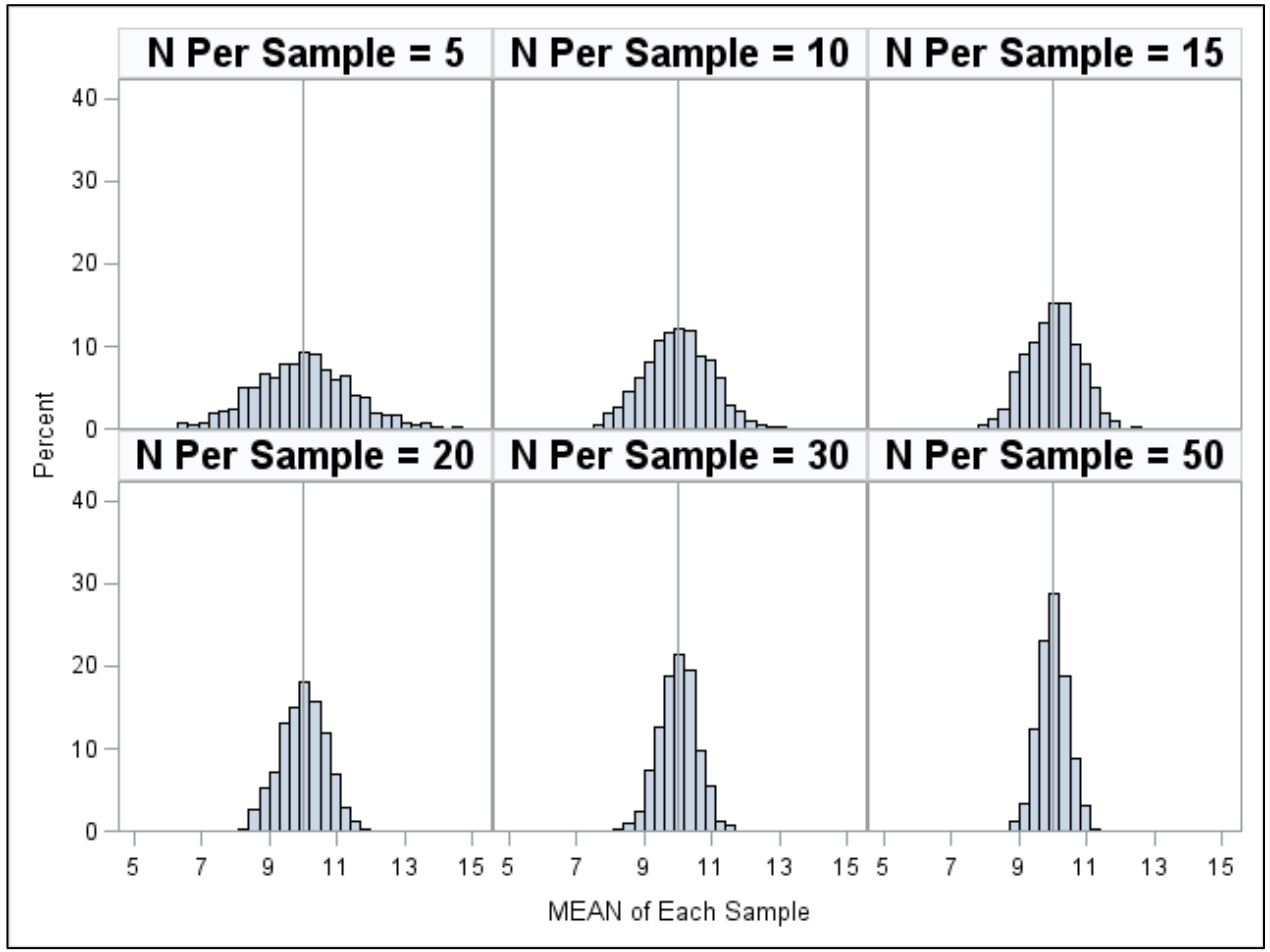
Mean: $\mu_y = 10$
SD: $\sigma_y = 5$ from before



Mean: $\mu_y = 10$
SD: $\sigma_y = 3$ instead



1000 samples each for different N…

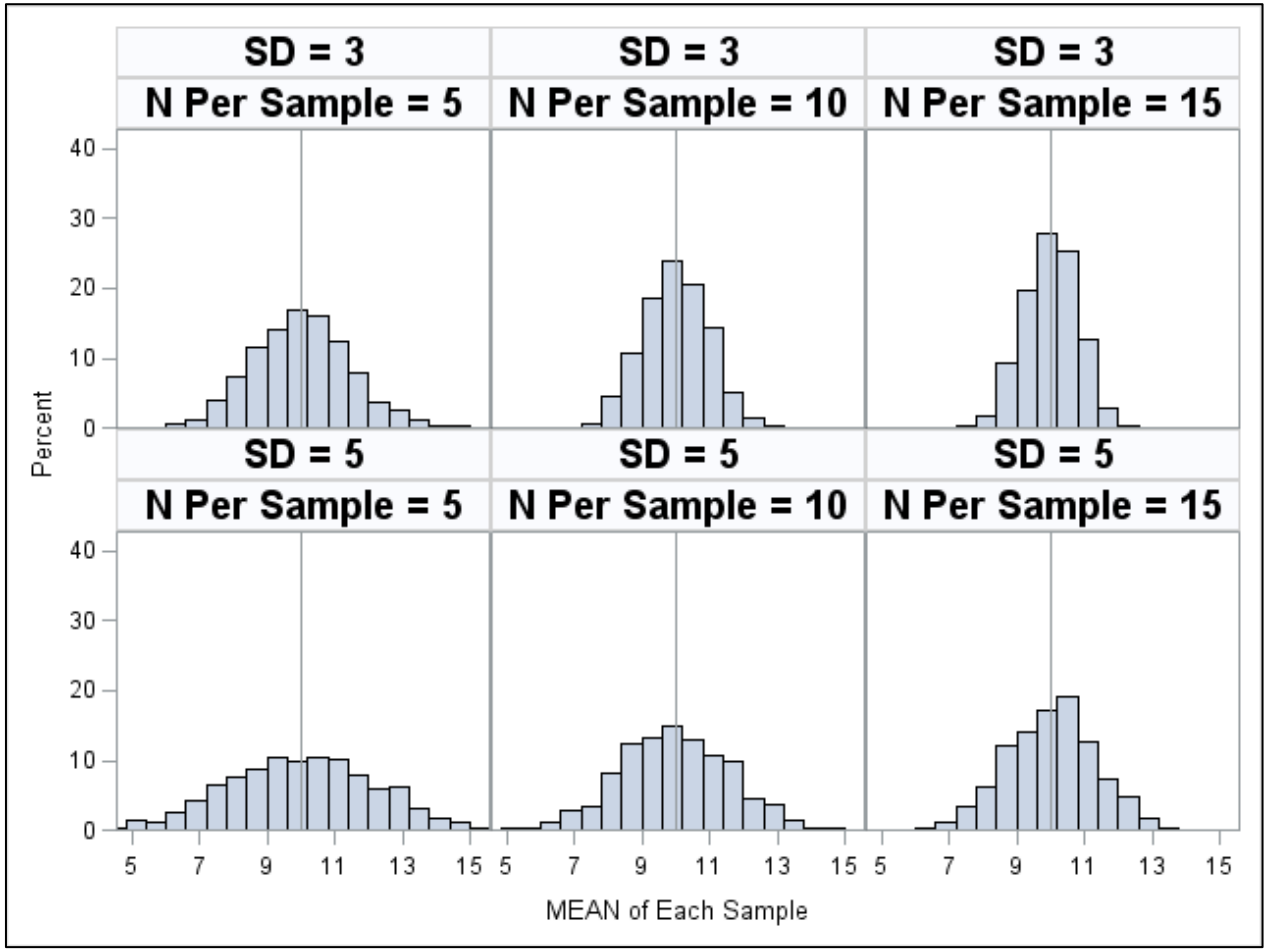


These bars still do not show individual people!
They are summaries for **distinct samples** of people.

- Population values:
  Mean $\boldsymbol{\mu_y = 10}$, SD $\boldsymbol{\sigma_y = 3}$ now
- More $\boldsymbol{N}$ → less SD in $\bar{\boldsymbol{y}}_{\boldsymbol{s}}$ across samples

N Per Sample	Mean $\bar{y}_s$	SD $\bar{y}_s$
5	10.01	1.42
10	10.00	0.96
15	10.01	0.78
20	9.99	0.67
30	10.00	0.56
50	10.00	0.42

Effects of N and SD on Dispersion of $\bar{y}_s$



These bars still do not show individual people!
They are summaries for **distinct samples** of people.

Left to right:

- **More** N in each sample → **less dispersion** in $\bar{y}_s$ across samples

Top to bottom:

- **More** SD in each sample → **more dispersion in** $\bar{y}_s$ across samples

- **Dispersion** → **More** estimate **inconsistency** across samples

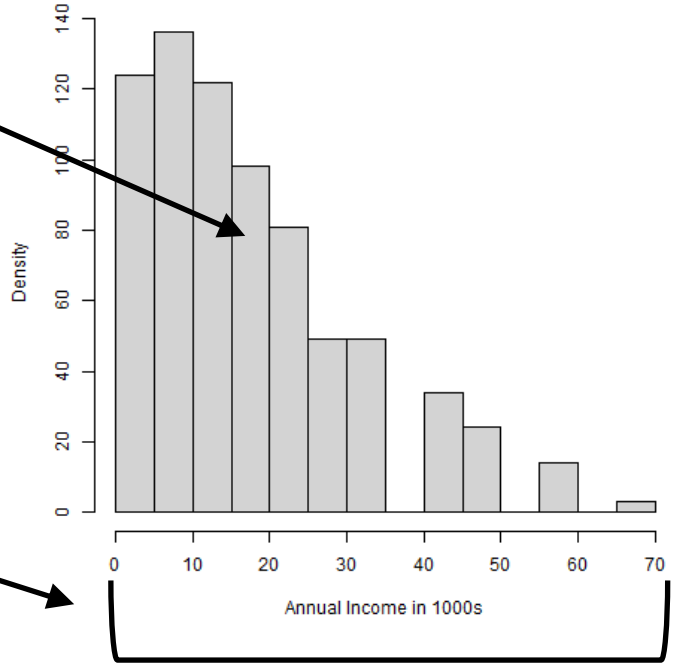
Anticipating Inconsistency of Sample Mean $\bar{y}_s$

- I had simulated a **known population** and selected multiple different samples
  - Inconsistency of $\bar{y}_s$ could be indexed by its standard deviation across samples → **more N, less variance → smaller SD** of $\bar{y}_s$ (= more consistent)

- Given only **one** sample, **we can still anticipate the SD** of $\bar{y}_s$:
  - SD of $\bar{y}_s$ across samples ← Standard Error (of Mean) = $SE = \frac{\sigma_y}{\sqrt{N}} = \sqrt{\frac{\sigma_y^2}{N}}$
    - SE includes the population SD σ_y, replaced by the sample SD s_y when σ_y is unknown
  - **SE of the mean** is the expected average deviation of any given sample mean $\bar{y}$ from the population mean μ_y (even if we do not know μ_y)
    - **Is NOT = SD of y_i** (s_y), the average deviation of any person from the sample mean
  - In general, the term **"SE" refers to the SD of an estimate's sampling distribution** (e.g., how the estimate of a sample's variance would differ across samples is also described by its own SE, but it's found differently)

Back to Our Example Empty Model

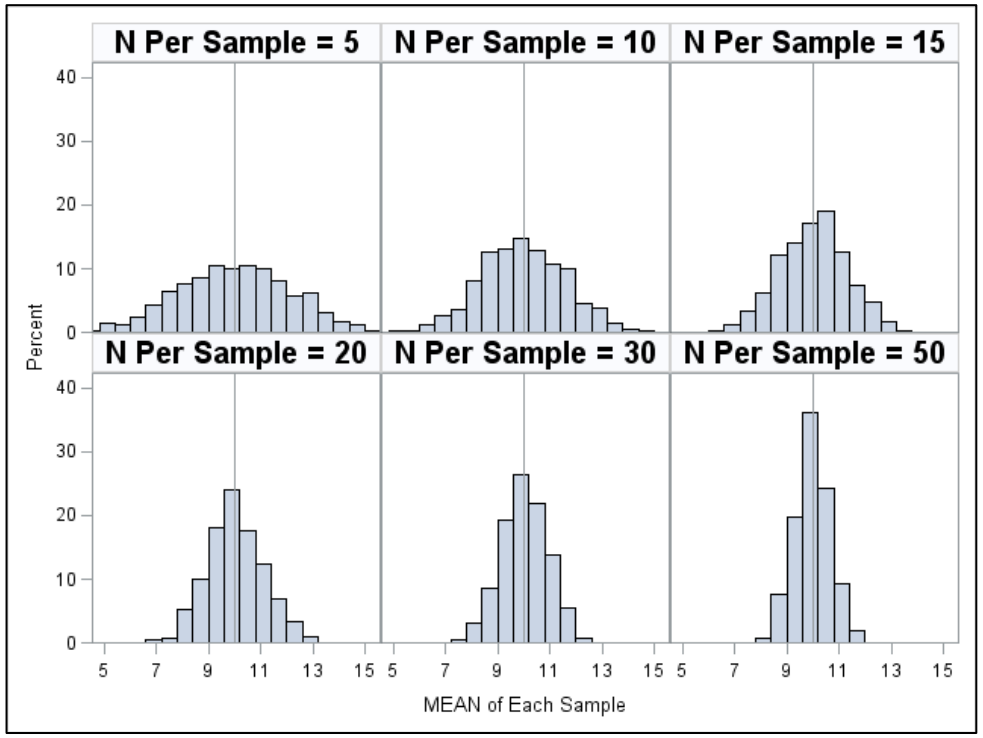
- Predict income in \$1000s in $N = 734$: $\boldsymbol{y_i = \beta_0 + e_i}$ → $\boldsymbol{income_i = \beta_0 + e_i}$
  - Fixed intercept estimate $\boldsymbol{\beta_0 = 17.30}$ (= sample $\bar{y}$)
  - Person-specific residual $\boldsymbol{e_i = y_i - \beta_0}$
  - Variance of $\boldsymbol{e_i}$ residuals → $\boldsymbol{\sigma_e^2 = 190.21}$(= sample $\boldsymbol{s_y^2}$)
- $\boldsymbol{SD}$ of $\boldsymbol{e_i}$ → $\boldsymbol{\sigma_e = \sqrt{190.21} = 13.80}$ (= sample $\boldsymbol{s_y}$)
  - $\boldsymbol{SD}$ **is about the PEOPLE**: This tells us that individual income is expected to be off from the sample mean income $\bar{y}$ by $\pm\,\mathbf{13.80}$ on average



- $\boldsymbol{\beta_0}$ standard error: $\boldsymbol{SE = \frac{\sigma_e}{\sqrt{N}} = \sqrt{\frac{\sigma_e^2}{N}} = 0.51}$
  - $\boldsymbol{SE}$ **is about the SAMPLES:** This tells us that sample mean income $\bar{y}$ (as given by estimate for $\boldsymbol{\beta_0}$ in an empty model) is expected to be off from the unknown population mean income $\boldsymbol{\mu}$ by $\pm\,\mathbf{0.51}$ on average

SE of Mean ≈ SD of Sampling Distribution for mean $\bar{y}_s$

Population values for y_i variable: Mean $\mu_y = 10$, SD $\sigma_y = 5$



N	Mean $\bar{y}_s$	SD $\bar{y}_s$	Mean SE with:	
			σ_y	s_y
5	9.97	**2.17**	2.24	2.13
10	9.98	**1.60**	1.58	1.55
15	10.00	**1.28**	1.29	1.28
20	10.03	**1.08**	1.12	1.11
30	10.03	**0.89**	0.91	0.91
50	9.97	**0.69**	0.71	0.71

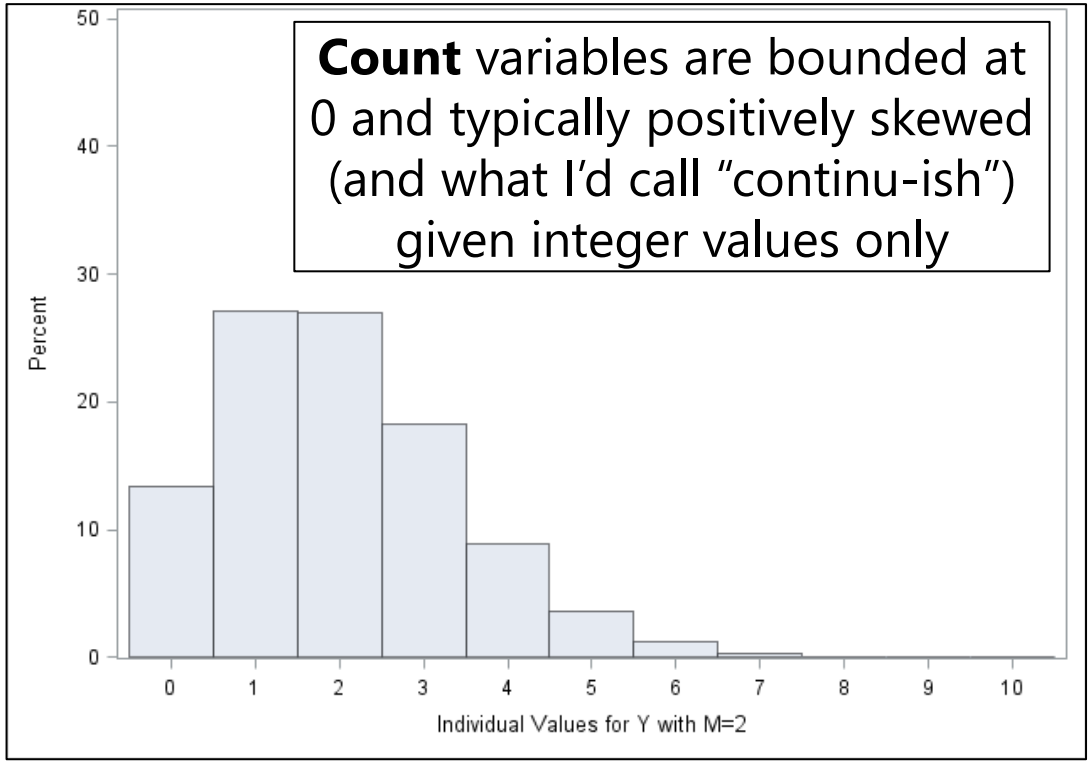
The greater the sample size N, the better the estimate of each sample's SD, and the less it matters that SE is formed with sample SD (s_y) instead of the population SD (σ_y). But this distinction will matter more in smaller samples….

What about Other Kinds of Variables?

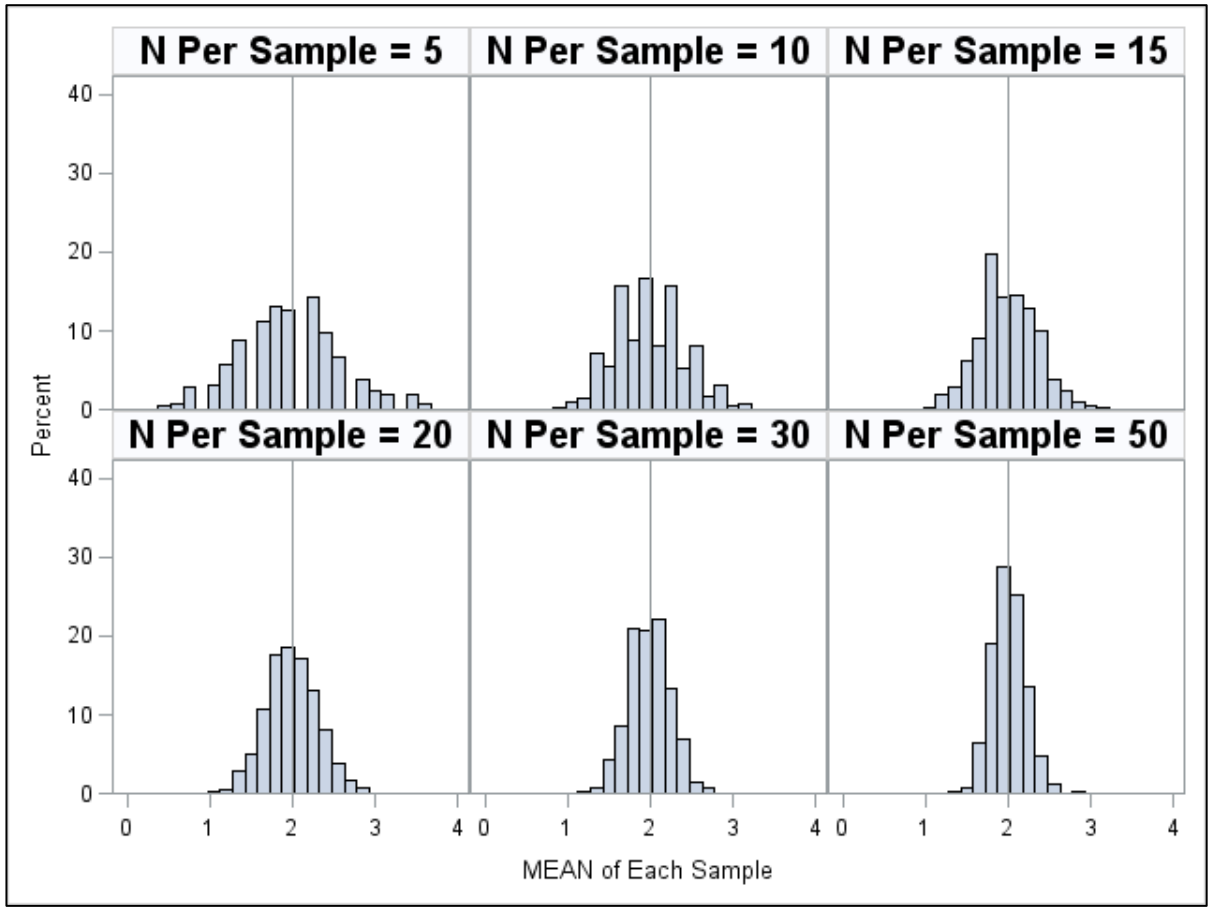
- It turns out **with more N the sampling distribution of $\bar{y}_s$ becomes more normal** *no matter what the observed variable's distribution is*
  - Btw: More N → more normal $\bar{y}_s$ distribution → "Central Limit Theorem"

- Demo: I simulated a **count variable*** y_i in a population of 100,000 new fake people
  - Population mean: $\mu_y = 2$
  - Population variance: $\sigma_y^2 = 2$
  - So y_i is off the mean by $SD = \sqrt{2}$ on average

  * Used a "**Poisson**" distribution to generate y_i (in which $\mu_y = \sigma_y^2$)



1000 samples each for different N… it works!



Note: The observed SD for the sampling distribution for $\bar{y}_s$: (a) is well-approximated by the mean SE for $\bar{y}_s$, and (b) appears normal, even for a count variable

- Population values:
  Mean $\boldsymbol{\mu_y = 2}$, VAR $\boldsymbol{\sigma_y^2 = 2}$)
- **More N → less SD in $\bar{y}_s$ across samples**; $\bar{y}_s$ is also more normal

N	Mean $\bar{y}_s$	SD $\bar{y}_s$	Mean SE
5	1.98	0.61	0.59
10	1.99	0.42	0.43
15	1.99	0.35	0.36
20	1.99	0.31	0.31
30	1.99	0.25	0.25
50	2.01	0.20	0.20

Section Review: Empty Model and Standard Errors

1. What does an empty model predict each person's outcome to be?

2. Would you expect more fluctuation in an outcome mean across repeated samples:

   a. If they each sampled 50 people or 100 people?

   b. If the variance of the outcome was 6 or 12?

3. What's the difference between the standard deviation of an outcome and the standard error of its mean?

4. The central limit theorem says, given a large enough sample size, all variables converge to a normal distribution. True or false?

Beyond Empty GLMs: 2 Fixed Effects

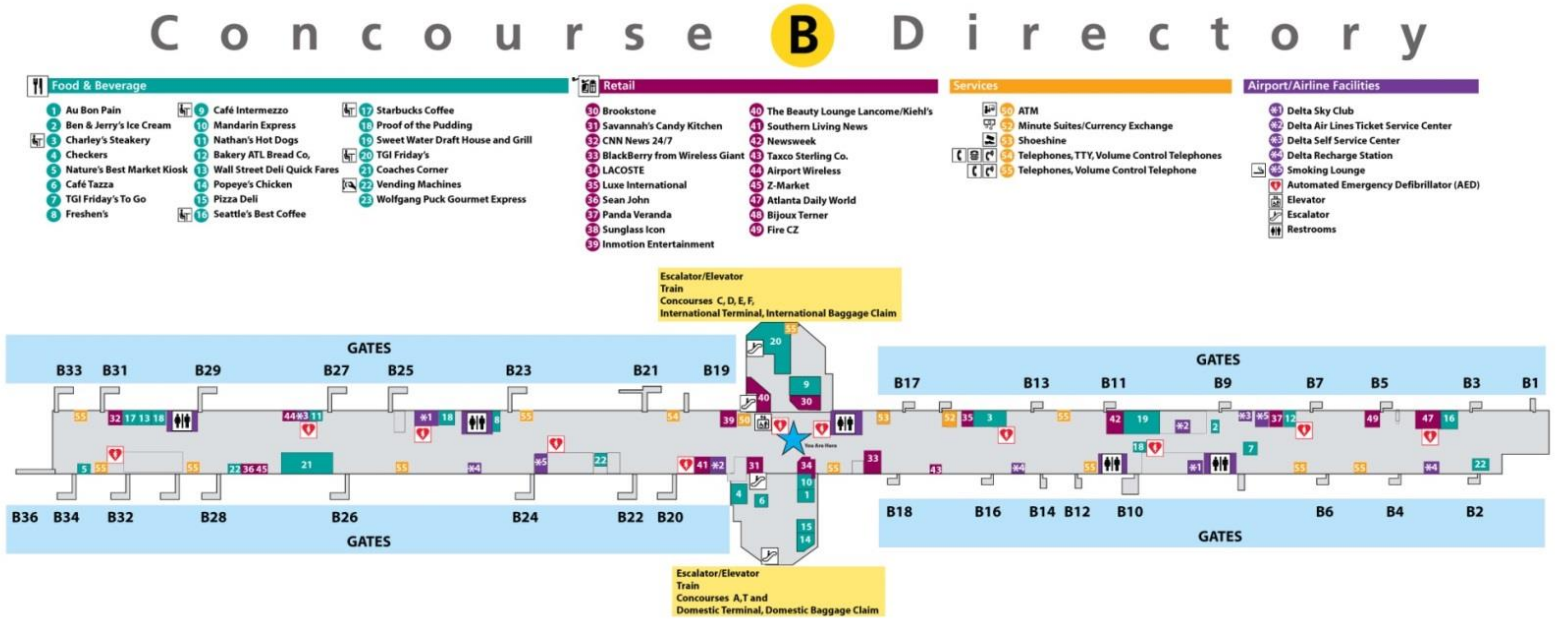
- Purpose of predictive linear models is to **customize** each person's expected outcome (beyond just the same mean) **by adding predictors**
  - Soon we will examine the unique effects of multiple predictors, but let's start with just **one quantitative predictor**: "(**simple**) **linear regression**"

- GLM to describe **how x_i predicts y_i**: $y_i = \beta_0 + \beta_1(x_i) + e_i$

  - **Fixed intercept β_0** = now is expected y_i specifically when $x_i = 0$
    - Changes from sample mean to a "**conditional mean**" (based on x_i)
    - Purpose is to adjust for any mean difference between x_i and y_i

  - **Fixed slope β_1** = difference in y_i per one-unit difference in x_i
    - Purpose is to capture a **linear relationship** between x_i and y_i
    - To create meaningful intercept when $x_i = 0$, you may need to rescale x_i by **centering**: **subtracting a constant c value to adjust what x_i = 0 means**

Why the Fixed Intercept β_0 *Should* Be Meaningful…



This is a very detailed map… but what do we need to know to be able to use the map at all?

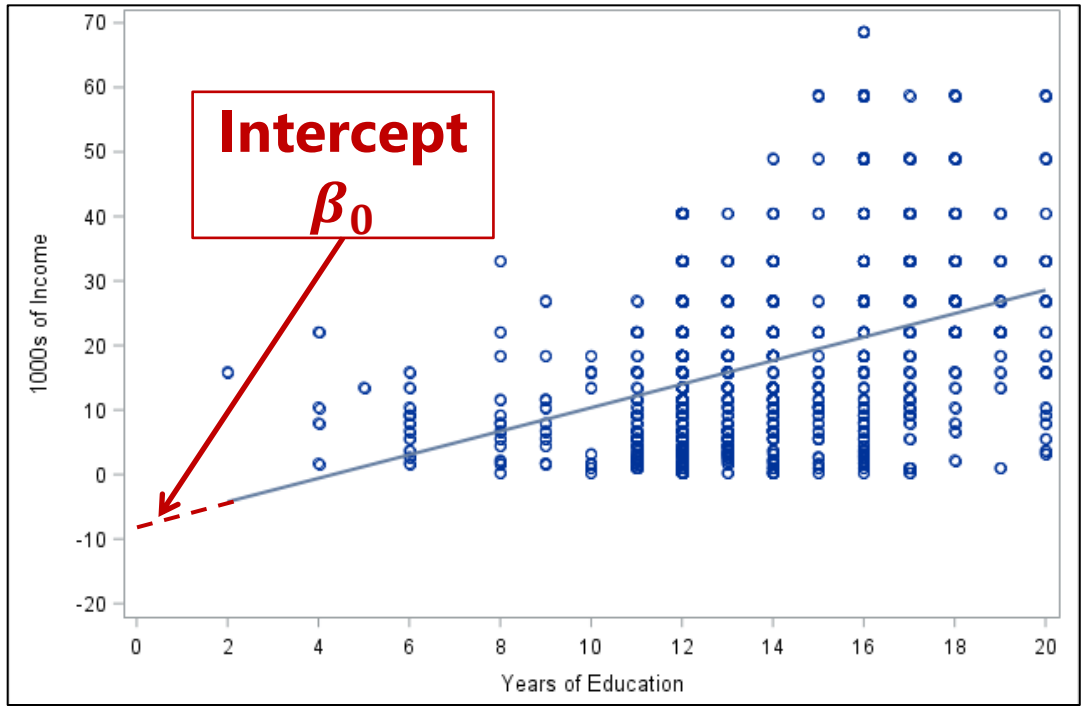
Intercept = "You are Here" Sign of Your Data

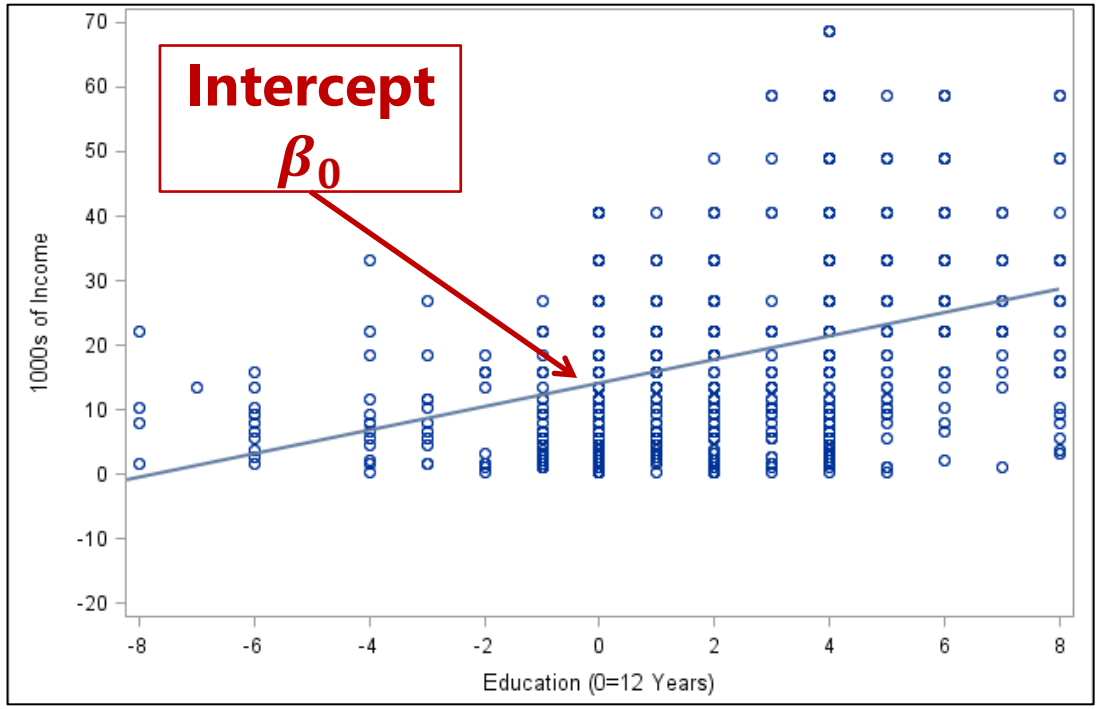
Using x_i = original years of education:

$$y_i = \beta_0 + \beta_1(x_i) + e_i$$
$$income_i = -7.89 + 1.82(educ_i) + e_i$$

After **centering x_i = education at 12**:

$$y_i = \beta_0 + \beta_1(x_i - 12) + e_i$$
$$income_i = 14.00 + 1.82(educ_i - 12) + e_i$$





There is no *wrong* way to center, only *weird*. **Center so x_i=0 is meaningful.**

Beyond Empty GLMs: Residual Variance

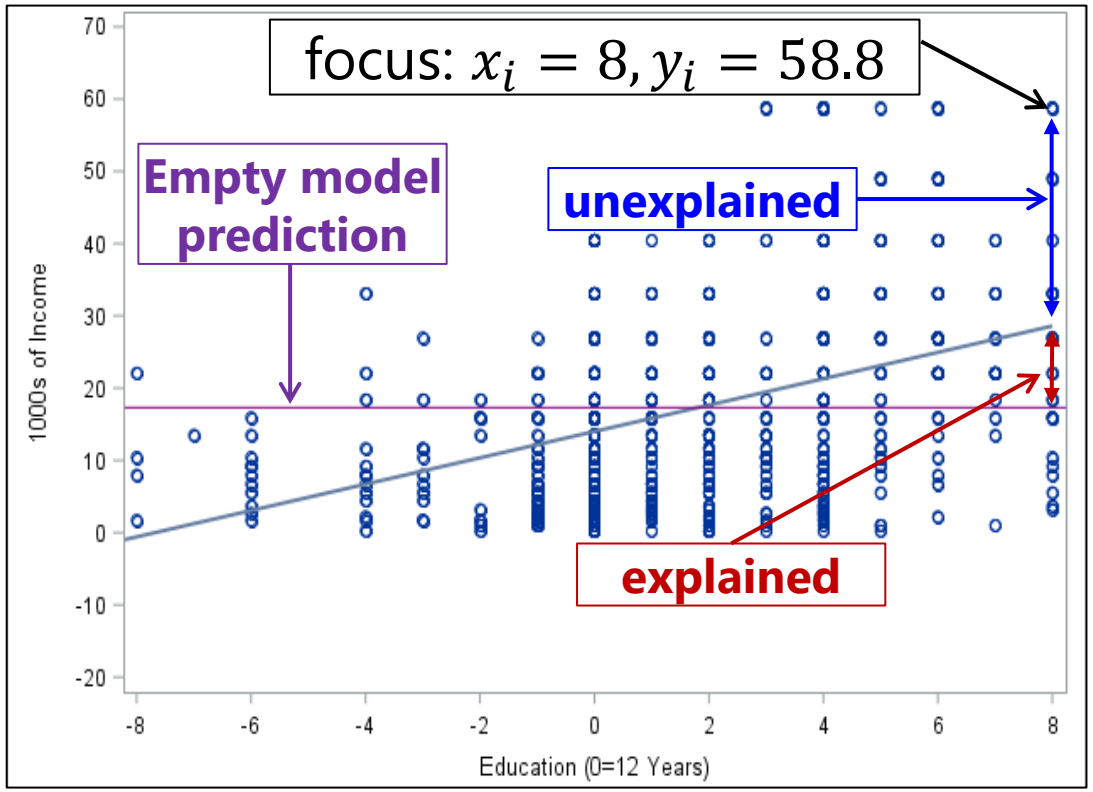
- GLM describes **how x_i predicts y_i**: $y_i = \beta_0 + \beta_1(x_i) + e_i$
  - Fixed intercept = β_0; fixed slope of predictor x_i = β_1
  - Total number of fixed effects = $k = 2$ here (note: k is used in later formulas)

- The y_i **created using the predictors** is abbreviated as $\hat{y}_i$ = "**y hat**"
  - $\hat{y}_i = \beta_0 + \beta_1(x_i)$ → $y_i = \hat{y}_i + e_i$ → $e_i = y_i - \hat{y}_i$
  - $\hat{y}_i$ is also called a "**conditional mean**" (because $\hat{y}_i$ will be the same for anyone with the same value of predictor x_i)

- Now we can find the new e_i **residual for each person**, and thus the **variance of the e_i residuals** across persons
  - "**residual variance**": $\sigma_e^2 = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N-2} = \frac{\sum_{i=1}^{N}(e_i)^2}{N-2}$
  - Once predictors are included, σ_e^2 should be **smaller than original s_y^2**

Visualizing GLM Residuals

Btw: $\boldsymbol{\beta}$ formulas result from the goal of minimizing the squared $\boldsymbol{e_i}$ residuals across the sample—this is called "**ordinary least squares estimation**"—let's see what happens for one example person below



Empty Model for y_i = income:

$$\boldsymbol{y_i = \beta_0 + e_i}$$

$$\boldsymbol{\hat{y}_{focus} = 17.3}$$

$$\boldsymbol{y_{focus} = 17.3 + 41.5}$$

Variance: $\boldsymbol{\sigma_e^2} = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N-1} = \boldsymbol{190.2}$

→ $190.2 = s_y^2$ = is **all** y_i variance

Add Education as Predictor:

$$\boldsymbol{y_i = \beta_0 + \beta_1(educ_i - 12) + e_i}$$

$$\boldsymbol{\hat{y}_{focus} = 14.0 + 1.8(8) = 28.4}$$

$$\boldsymbol{y_{focus} = 28.4 + 30.4}$$

Variance: $\boldsymbol{\sigma_e^2} = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N-2} = \boldsymbol{162.3}$

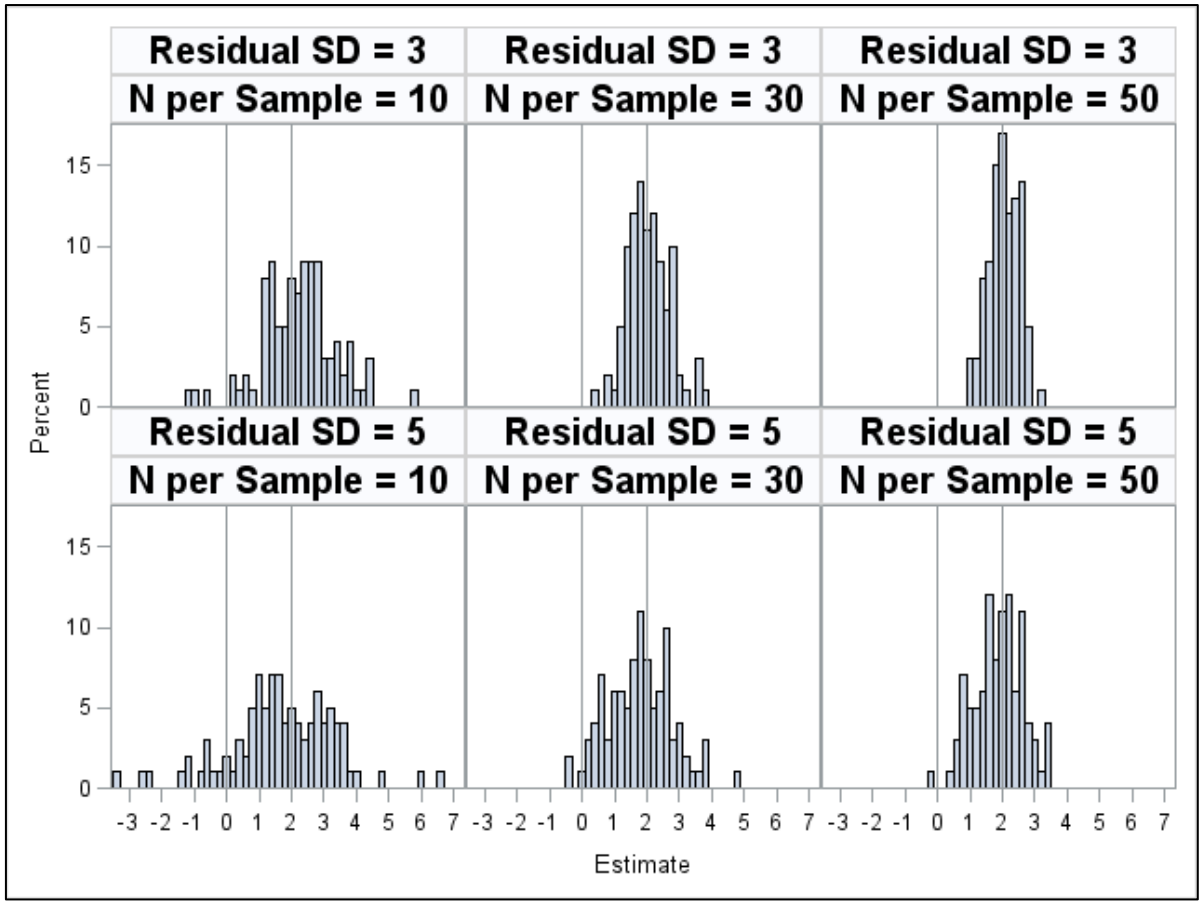
→ 162.3 = **leftover** y_i variance

Uncertainty of Sample Estimates

- Besides their estimates (i.e., answers in our sample), we need to **quantify the inconsistency** of both fixed effects: $\boldsymbol{\beta_0}$ intercept and $\boldsymbol{\beta_1}$ slope
  - How much would my **sample intercept vary** across repeated samples? *(just like SE of mean, but for the mean conditional on $\boldsymbol{x_i = 0}$)*
  - How much would my **sample slope vary** across repeated samples? *(same idea in concept, but also taking into account the scale of $\boldsymbol{x_i}$)*

- Demo: I made my own (normally-distributed) quantitative variables $\boldsymbol{x_i}$ and $\boldsymbol{y_i}$ in another population of 100,000 fake people
  - **Predictor** $\boldsymbol{x_i}$ has mean $= 0$ and variance $= 1$
  - **Outcome** $\boldsymbol{y_i = \beta_0 + \beta_1(x_i) + e_i}$ → $\boldsymbol{\hat{y}_i = 100 + 2(x_i)}$
  - Residual variance for $\boldsymbol{y_i}$ is either $\boldsymbol{\sigma_e^2 = 9}$ or $\boldsymbol{\sigma_e^2 = 25}$
  - Let's see what happens to $\boldsymbol{\beta_1}$ over 100 samples of varying $\boldsymbol{N}$ and $\boldsymbol{\sigma_e^2}$

Effects of N and SD on Dispersion of β_1



These bars still do not show individual people!
They are summaries for **distinct samples** of people.

Left to right:

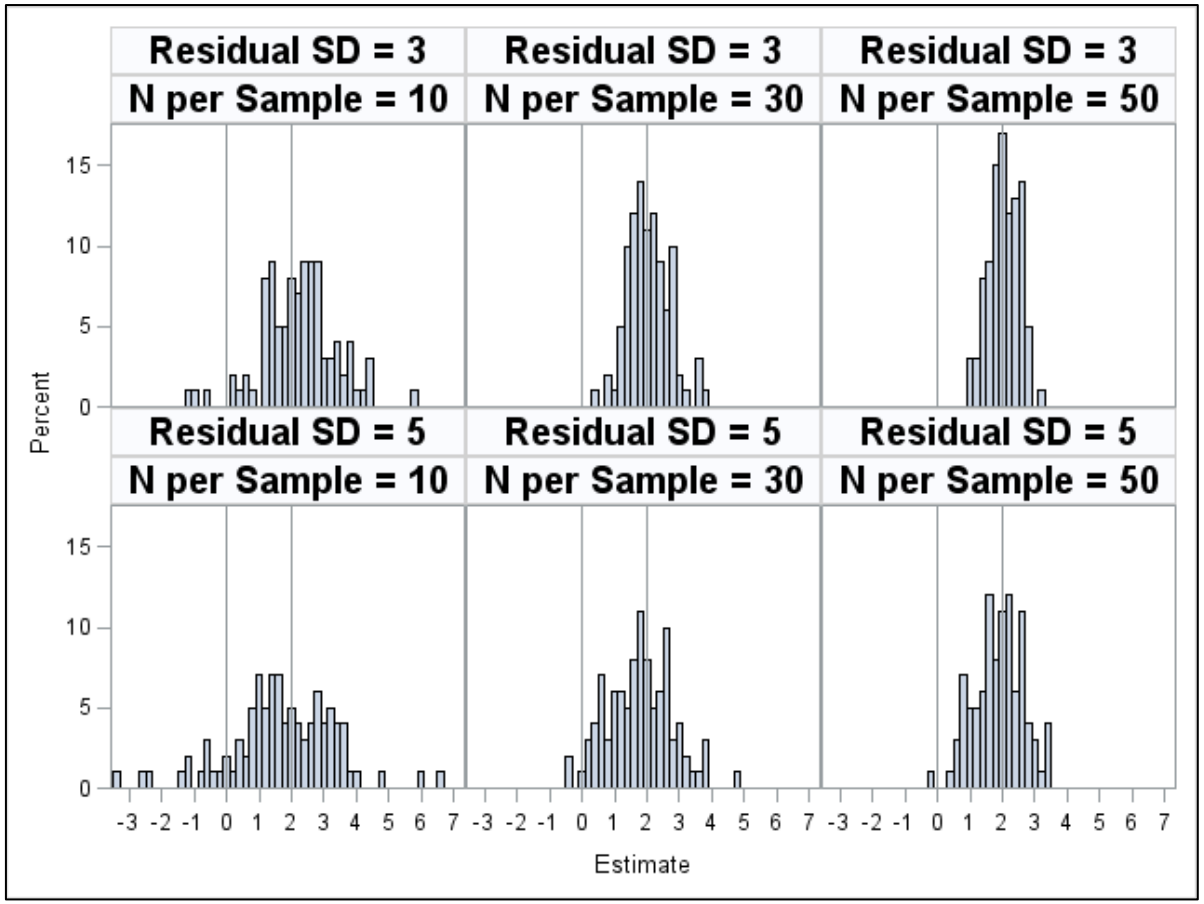
- **More N** in each sample → **less dispersion in β_1** across samples

Top to bottom:

- **More SD** in each sample → **more dispersion in β_1** across samples

- **Dispersion** → estimate **inconsistency** across samples

Effects of N and SD on Dispersion of β_1



These bars still do not show individual people!
They are summaries for **distinct samples** of people.

- Population values:
  Slope $\beta_1 = 2$
- **More N → less SD in β_1** across samples; β_1 is also more normal

N/ SD	Mean β_1	SD β_1	Mean SE
10/3	2.22	1.14	1.03
30/3	2.06	0.63	0.57
50/3	2.05	0.47	0.43
10/5	1.62	1.78	1.73
30/5	1.79	0.99	0.97
50/5	1.90	0.74	0.71

Standard Errors in a One-Predictor GLM

- Back to example GLM with $k = 2$ fixed effects
  - Model: $\boldsymbol{income_i = \beta_0 + \beta_1(educ_i - 12) + e_i}$
  - Result: $\boldsymbol{\hat{y}_i = 14.00 + 1.82(educ_i - 12)}$, $\boldsymbol{\sigma_e^2 = 162.28}$

- **Standard Error** (**SE**) for a slope $\boldsymbol{\beta_1}$ in a single-predictor GLM:

$$\text{SE}_{\beta_1} = \sqrt{\frac{\text{residual variance of } y_i}{\text{variance of } x_i * (N-k)}} = \sqrt{\frac{\boldsymbol{\sigma_e^2}}{s_x^2 * (N-k)}}$$

  - Slope $\boldsymbol{\beta_1}$ SE $= \sqrt{\frac{\mathbf{162.28}}{\mathbf{8.46}*(734-2)}} = 0.16$

→ Sample slope is expected to be off from population slope by $\pm$ **0.16** on average

---

- **SE of any predicted outcome** $\hat{y}_i$ (including that captured by $\boldsymbol{\beta_0}$ for $x_i = 0$) varies by predictor value—the SE will increase as you move away from predictor's mean:
  - SE of $\hat{y}_i \mid x_i = \sqrt{\boldsymbol{\sigma_e^2}} * \sqrt{\frac{1}{N} + \frac{(x_i - \bar{x})^2}{(N-1)s_x^2}}$ → 0.55 at $educ_i = 12$ specifically

"simple linear" regression in R (using my new output function)

```
print("Center education predictor so that 0 is meaningful for intercept")
Lecture2$educ12 = Lecture2$educ-12
# educ12: Education (0=12 years) # Label as a comment only

print("GLM Predicting Income from Centered Education 0=12 -- save as ModelEduc12")
ModelEduc12 = lm(data=Lecture2, formula=income~1+educ12)
obj=LMsummary(ModelEduc12, explain=TRUE)  # Full custom output with explanations
```

$$income_i = \beta_0 + \beta_1(educ_i - 12) + e_i$$
$$\hat{y}_i = 14.00 + 1.82(educ_i - 12),\ \sigma_e^2 = 162.28$$

```
Sums of Squares Table
               SS  DF        MS       F      p    R2
Model  20634.982   1 20634.982 127.157 <0.001 0.148
Error 118788.250 732   162.279
Total 139423.232 733   190.209

Explanation:
SS = Sum of Squares, MS = Mean Square, DF = Degrees of Freedom,
F = F test-statistic, p = two-sided p-value, R2 = R-square

Fixed Effects Table
            Est    SE      t      p    LCI    UCI
Intercept 13.998 0.554 25.265 <0.001 12.911 15.086
educ12     1.824 0.162 11.276 <0.001  1.506  2.141

Explanation:
Est = Estimate, SE = Standard Error, t = t test-statistic,
p = p-value, LCI = Lower Confidence Interval,
UCI = Upper Confidence Interval
```

Creating predicted outcomes using `glht` in R

$$income_i = \beta_0 + \beta_1(educ_i - 12) + e_i, \qquad \hat{y}_i = 14.00 + 1.82(educ_i - 12)$$

Predicted income for 8 years education: $\hat{y}_i = 14.00(1) + 1.82(-4) = 6.70$

Predicted income for 12 years education: $\hat{y}_i = 14.00(1) + 1.82(0) = 14.00$

Predicted income for 16 years education: $\hat{y}_i = 14.00(1) + 1.82(4) = 21.29$

Predicted income for 20 years education: $\hat{y}_i = 14.00(1) + 1.82(8) = 28.59$

```
print("Get predicted outcomes using glht from multcomp package -- save as PredEduc12")
print("In number lists below, the values are multipliers for each fixed effect in order")
PredEduc12 = glht(model=ModelEduc12, linfct=rbind(
  "Pred Income for  8 years (educ12=-4)" = c(1,-4),  # beta0(1) + beta1*(-4)
  "Pred Income for 12 years (educ12= 0)" = c(1, 0),  # beta0(1) + beta1*( 0)
  "Pred Income for 16 years (educ12= 4)" = c(1, 4),  # beta0(1) + beta1*( 4)
  "Pred Income for 20 years (educ12= 8)" = c(1, 8))) # beta0(1) + beta1*( 8)
obj=glhtSummary(glhtObject=PredEduc12) # Brief custom output
```

```
Linear Combinations Table

                                          Est    SE      t      p    LCI    UCI
Pred Income for  8 years (educ12=-4)   6.703 1.051  6.378 <0.001  4.640  8.767
Pred Income for 12 years (educ12= 0)  13.998 0.554 25.265 <0.001 12.911 15.086
Pred Income for 16 years (educ12= 4)  21.293 0.588 36.183 <0.001 20.138 22.449
Pred Income for 20 years (educ12= 8)  28.588 1.106 25.854 <0.001 26.417 30.759
```

Note how SE increases away from the center

Section Review: Single-Predictor GLMs

1. Given this generic model: $\boldsymbol{y_i = \beta_0 + \beta_1(x_i) + e_i}$

   a. Which terms are variables, and which terms are constants?

   b. Conceptually, what is the difference between $\boldsymbol{y_i}$ and $\boldsymbol{e_i}$?

2. Given this model with a linear slope of reading predicting math:

$$\boldsymbol{math_i = \beta_0 + \beta_1(reading_i) + e_i}$$
$$\boldsymbol{math_i = 10 + \ 2(reading_i) + e_i}$$

   What is the predicted math value for a student with reading = 3?

3. What is the difference between an estimate and a standard error?

From Descriptive to Inferential Statistics

- A **standard error** (**SE**) captures expected inconsistency of any kind of model estimate across repeated samples of the same kind
  - $\boldsymbol{SE}$: Mean distance of sample (estimate) from population (true unknown) *(*cough* HW question)*
  - Btw, $\boldsymbol{SE}^2$ is known as "**sampling** variance" (but is not typically reported) —not to be confused with $\boldsymbol{SD}^{\boldsymbol{2}}$, which is "**sample** variance" (ugh, I know)
  - Remember: $\boldsymbol{SE}$ → width across samples; $\boldsymbol{SD}$ → width across people in a sample

- Moving forward: We can use an $\boldsymbol{SE}$ **to assess how far away** any sample estimate is **from a hypothesized** population value
  - We expect **some fluctuation in the sample-specific estimates** across multiple samples
  - But **what is** "**too different**" for our sample to not likely be from the same population?
  - Said differently, if the population slope really were true, **how unexpected** (**what % of the time**) is it to have observed this sample's particular slope?
  - This process is known as "**null hypothesis significance testing**" and it requires us to make several decisions ahead of time: We need to operationally define "hypothesized value", as well as "too different" (or "how unexpected")…

Null Hypothesis Significance Testing

- A "**null hypothesis**" ($\boldsymbol{H_0}$) says the population parameter = some expected value
  - e.g., for our GLM example, education slope $\boldsymbol{\beta_1 = 1.82}$, $\boldsymbol{H_0}$: **population** $\boldsymbol{\beta_1 = 0}$
  - $\boldsymbol{H_0}$ for a slope is usually 0, but it doesn't have to be!

- An "**alternative hypothesis**" ($\boldsymbol{H_A}$) contradicts the null hypothesis to **convey allowed directionality of deviation** from stated H_0 value
  - "**One-tailed test**" is one-directional: $\boldsymbol{H_A}$: $\boldsymbol{\beta_1 > 0}$ OR $\boldsymbol{H_A}$: $\boldsymbol{\beta_1 < 0}$
  - "**Two-tailed test**" is "different than" : $\boldsymbol{H_A}$: $\boldsymbol{\beta_1 \neq 0}$ (aka, $\boldsymbol{H_A}$: $\boldsymbol{\beta_1 = !0}$)

- **How far away is our sample estimate from $\boldsymbol{H_0}$**? We standardize this distance using its $\boldsymbol{SE}$ as a proxy for the $\boldsymbol{SD}$ of its sampling distribution
  - For $\boldsymbol{\beta_1 = 1.82}$ and $\boldsymbol{H_0}$: $\boldsymbol{\beta_1 = 0}$, the standardized distance of $\boldsymbol{\beta_1}$ from $\boldsymbol{H_0}$

    is given by this "**test statistic**" $= \frac{\boldsymbol{Est - H_0}}{\boldsymbol{SE}} = \frac{\boldsymbol{\beta_1 - 0}}{\boldsymbol{SE}} = \frac{\boldsymbol{1.82 - 0}}{\boldsymbol{0.16}} = \boldsymbol{11.28}$

Null Hypothesis Significance Testing

- So how **far away is our sample estimate from H_0** (using SE)?
  - For $\beta_1 = 1.82$ and $H_0: \beta_1 = 0$, the standardized distance of β_1 from H_0 is given by this "**test statistic**" $= \frac{Est - H_0}{SE} = \frac{\beta_1 - 0}{SE} = \frac{1.82 - 0}{0.16} = 11.28$
  - So $\beta_1 = 1.82$ is 11.28 *sampling* standard deviations away from H_0
  - Test statistic = how far off from expected value / expected offness

- And **how unexpected** is that result—how often would we see a sample slope that far away from H_0, if H_0 really were true?
  - **Probability density functions** (**PDFs**) to the rescue! We use PDFs to describe the expected behavior of an estimate's sampling distribution
  - 2 PDF choices to evaluate a slope's test statistic (i.e., the 11.28 here)
    - "**Standard Normal**" distribution, which is known as "z"—that's next
    - "**Student's t**" distribution, which is known as "t"—stay tuned!

Area Under Standard Normal Curve (to express H_0)

Btw, y-axis created by:

$$f(z_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_i^2}{2}\right)$$

The x-axis (called z_i) is in **standard deviation units** (where $SD = 1$)

Boundaries defined by z "critical" values:

10% outside $z = \pm 1.65$

5% outside $z = \pm 1.96$

1% outside $z = \pm 2.58$

"data" = **samples** here!

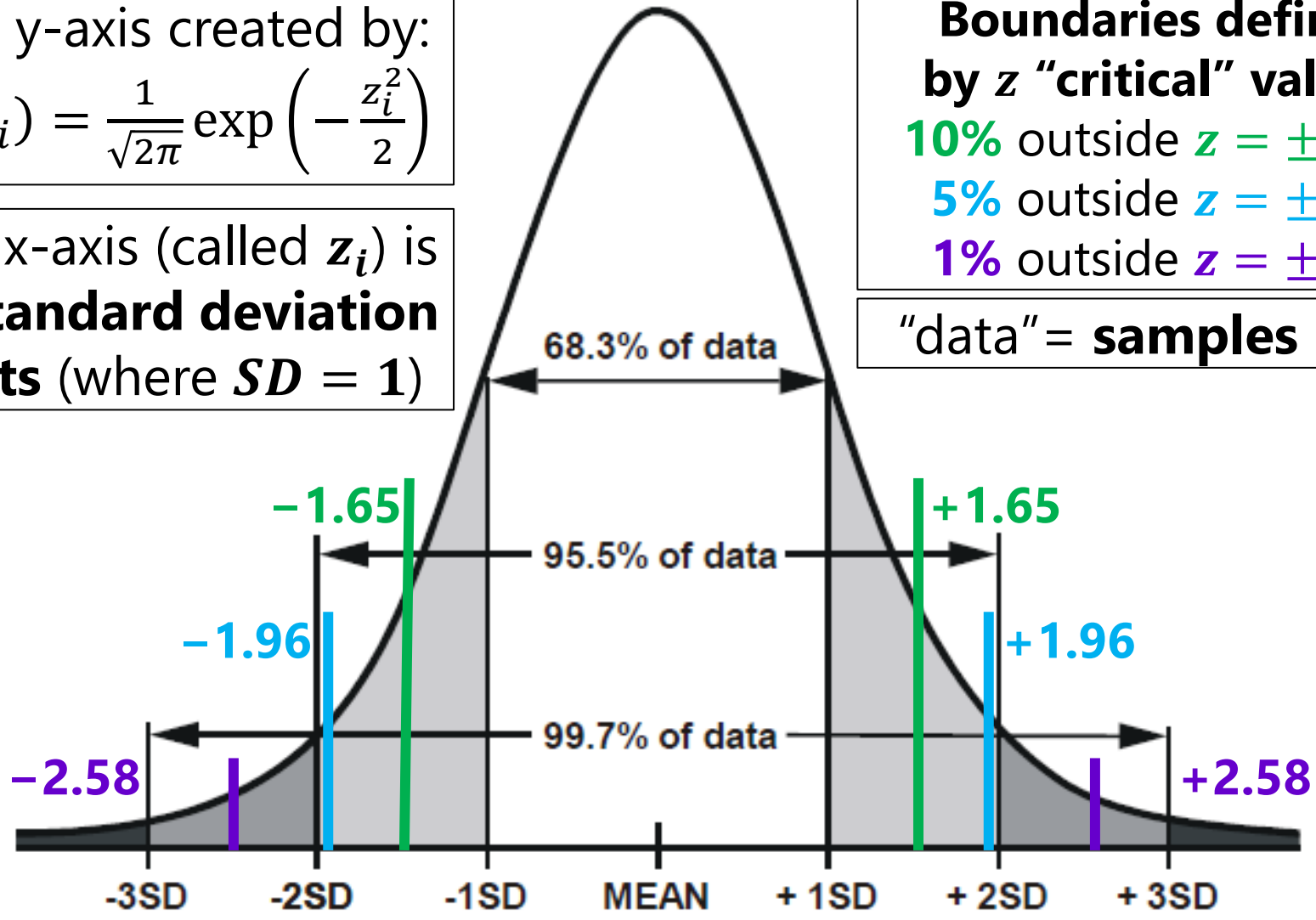


Image adapted from: https://medium.com/@harsharora0703/standard-normal-distribution-91f78a2b56e3

Denoting Expected vs. Unexpected (based on H_0)

- "How unexpected" (a test statistic is) requires **two more decisions made** ahead of time

- **How often** is "unexpected"? This is known as "**alpha level**" (α = alpha)
  - Common is **alpha = .05** (or **.01** for conservative or **.10** for lenient)
  - **Common "unexpected" means it would happen < 5% of the time**

- **Which directions** contain "unexpected" (how can you be wrong)?
  - "**One-tailed** test" allocates **ALL** of alpha % area to **one direction** (for H_A)
  - "**Two-tailed** test" allocates **HALF** of alpha % to **each possible direction**

- These decisions create "**critical values**" = **boundaries for where** "**unexpected**" **begins**
  - Test statistics that fall **inside boundaries** = sufficiently **expected**
    → retain (do not reject) H_0 = "statistically **nonsignificant**" result
  - Test statistics that fall **outside boundaries** = sufficiently **unexpected**
    → reject H_0 = "statistically **significant**" result

Allowed Directions of "Unexpected": One-Tailed Tests at Work

- Choices: H_0: $\beta_1 = 0$; probability declared "unexpected" is **alpha = .05** (so **5% is "unexpected"**) → two possible versions of **one-tailed** H_A:
- H_A: $\beta_1 > 0$ → $z_{critical} = +1.65$
  - Tests if β_1 is bigger or not bigger than H_0
  - If β_1 is actually smaller, conclude "not bigger" (= misleading)
- H_A: $\beta_1 < 0$ → $z_{critical} = -1.65$
  - Tests if β_1 is smaller or not smaller than H_0
  - If β_1 is actually bigger, conclude "not smaller" (= misleading)

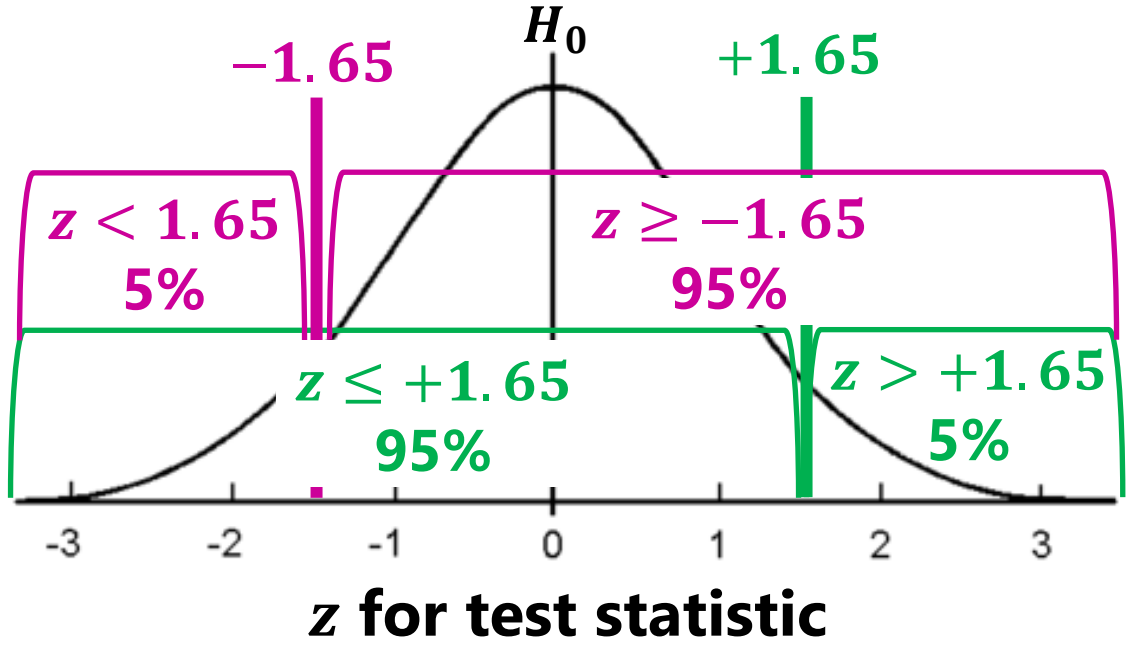


Image adapted from: https://discovery.cs.illinois.edu/learn/Simulation-and-Distributions/Normal-Distribution/

Two-Tailed Test: ± Critical Values Instead

Only one problem:
these z areas don't work for small N…
beer to the rescue!
(no, really… story time!)

Boundaries defined by z critical values:
10% outside $z = \pm 1.65$
5% outside $z = \pm 1.96$
1% outside $z = \pm 2.58$

"data" = **samples** here

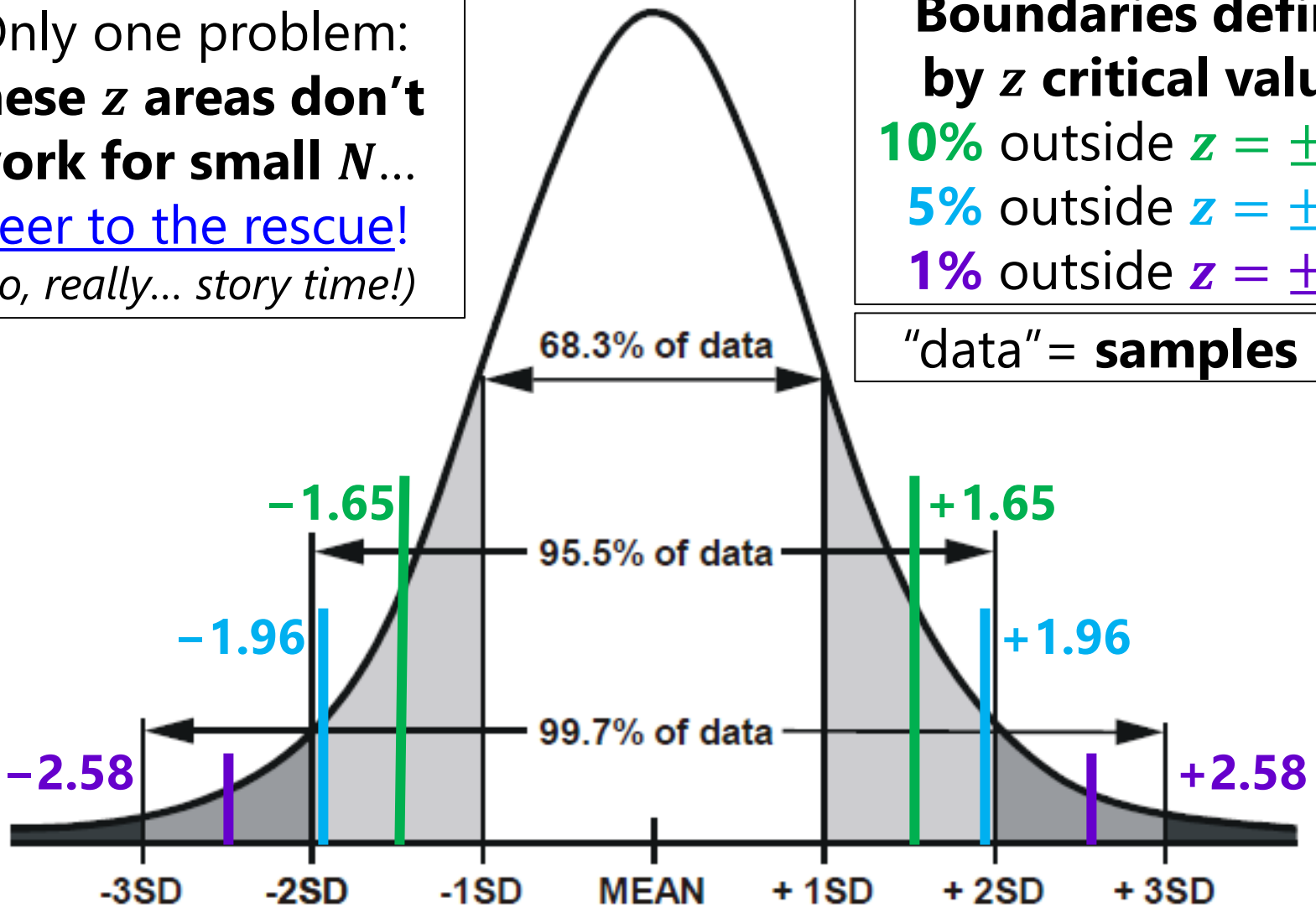
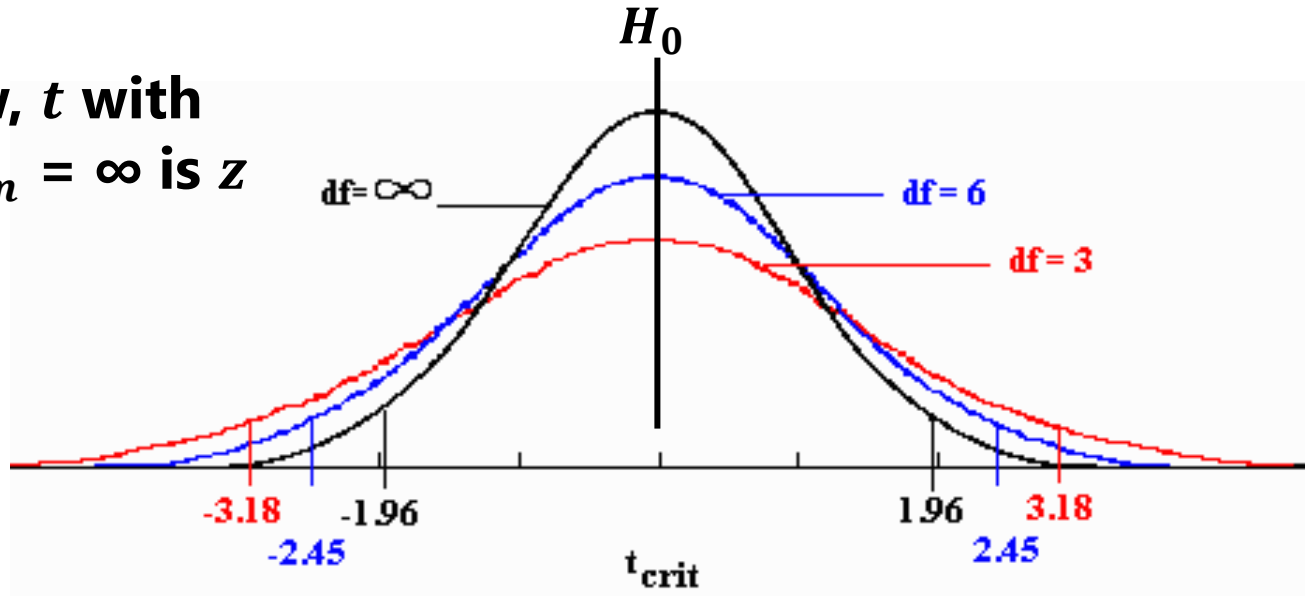


Image adapted from: https://medium.com/@harsharora0703/standard-normal-distribution-91f78a2b56e3

Meet Student's t Distribution (revised version of H_0)

- Both z (standard normal) and t distributions have the same metric: $M = 0$, $SD = 1$ (how far β_1 is from H_0 with sampling $SD \leftarrow SE$)
- But t is flatter than z, more so with fewer "**denominator degrees of freedom**": $DF_{den} = N - k$ (where k = fixed effects)
- Same rules: If t test statistic is outside boundaries set by t critical values for your DF_{den} → reject H_0 → "**significant**" result

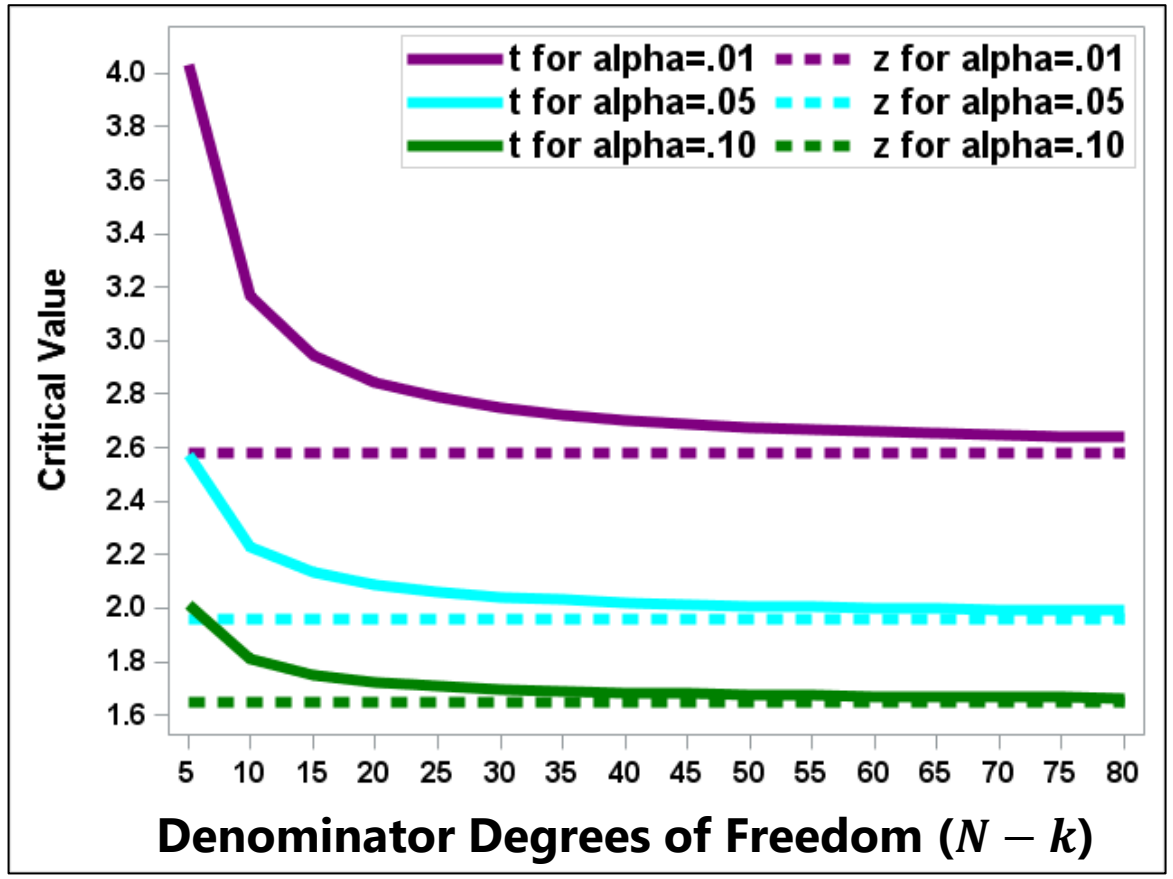
Btw, t with DF_{den} = ∞ is z



$t_{critical}$ values for **alpha = .05 by** DF_{den} shown here

With smaller N, we have to go farther out to **get to 5%**

Critical Values for t versus z Distributions



With smaller $\boldsymbol{N - k}$ (fewer $\boldsymbol{DF_{den}}$), more extreme test-statistics are needed when using $\boldsymbol{t}$ to say $\boldsymbol{\beta_1}$ is "unexpectedly different" from $\boldsymbol{H_0}$ (i.e., to cross the alpha-based boundary to be "significant")

$\boldsymbol{z}$ doesn't use DF_{den}

In the olden days, one needed to refer to tables of $t_{critical}$ values for a given alpha and DF, but now statistical software can give you the **exact p-value**: the probability of a more extreme t test-statistic than you found if the null hypothesis H_0 were true

Labeling t test-statistics and p-values: Two-tailed boundaries using alpha = .05

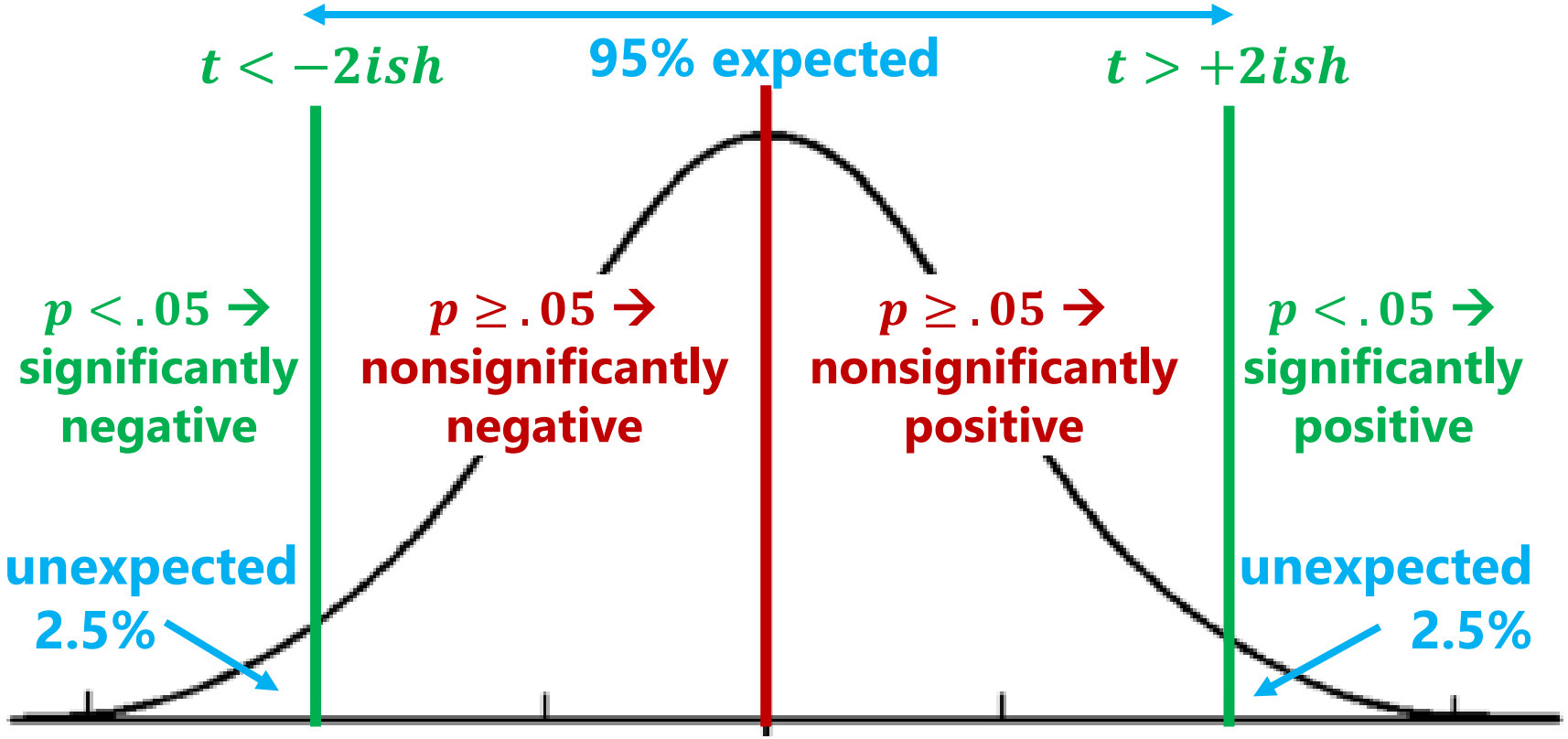


Image adapted from: http://mathcenter.oxford.emory.edu/site/math117/normalDistribution/

Mid-Section Review: NHST

1. What 3 things does the researcher need to decide ahead of time in planning a hypothesis test?

2. What purpose do test statistics (such as t or z) serve?

3. Why does Lesa say one-tailed tests are potentially misleading?

4. What does the term "statistically significant" mean?

From Alpha to Confidence Intervals

- Critical values from a t-distribution can also be used to compute a "**confidence interval**" ($\boldsymbol{CI}$) for your estimate
- $\boldsymbol{CI}$ = interval predicted to **contain the population value in (1−alpha)%** of repeated samples (confidence = expected)
  - Conservative: alpha = .01 → 1% unexpected → 99% CI
  - Common: alpha = .05 → 5% unexpected → 95% CI
  - Lenient: alpha = .10 → 10% unexpected → 90% CI
  - If $\boldsymbol{CI}$ **does not cross** $\boldsymbol{H_0}$, then your result is "**significant**" at that alpha level

- **Confidence Interval:** $\boldsymbol{CI = Estimate \pm (t_{critical} * SE)}$
- e.g., for $\boldsymbol{\beta_1 = 1.82}$, $\boldsymbol{SE = 0.16}$, $\boldsymbol{DF_{den} = N - k = 732}$ → $t_{critical}$
  - **90%** CI for $\boldsymbol{\beta_1}$: $CI = \mathbf{1.82} \pm (\mathbf{1.65} * \mathbf{0.16}) = 1.56\ to\ 2.08$
  - **95%** CI for $\boldsymbol{\beta_1}$: $CI = \mathbf{1.82} \pm (\mathbf{1.96} * \mathbf{0.16}) = 1.51\ to\ 2.13$
  - **99%** CI for $\boldsymbol{\beta_1}$: $CI = \mathbf{1.82} \pm (\mathbf{2.58} * \mathbf{0.16}) = 1.41\ to\ 2.23$

For reporting CI:
"lower bound" to "upper bound"

Significance Tests: What's in Your Output

- Each $\boldsymbol{\beta}$ **fixed slope** has 6 relevant characteristics (*essential to report):
  - ***Estimate** = best guess for the fixed slope from our sample data
  - ***Standard Error** = $\boldsymbol{SE}$ = average distance of sample slope from population slope
    → expected *inconsistency* of slope across samples
  - "$\boldsymbol{t}$**-value**" = (Estimate $- H_0$) / SE = **test-statistic** for fixed slope against H_0 (usually $H_0 = 0$)
  - **Denominator DF** = $\boldsymbol{DF_{den}} = N - k$ (where k = total number of fixed effects)
  - $\boldsymbol{p}$**-value** = (two-tailed) probability of fixed slope estimate *as or more extreme* if H_0 is true
    → how unexpected our result is on a t-distribution with 0=H_0, SD=SE
  - **(95%) Confidence Interval** = $\boldsymbol{CI} = Estimate \pm t_{critical} * SE$ = range in which true (population) value of estimate is expected to fall across (95%) of samples
- Compare $\boldsymbol{t}$ test-statistic to t critical-value at pre-chosen level of significance (where % unexpected = alpha level): this is a "**univariate Wald test**"
  - Btw, if denominator DF are not used (sample is "large enough"), then $\boldsymbol{t}$ is treated as $\boldsymbol{z}$ instead
    → same test-statistic, but different distribution for defining "unexpected"

Significance Test of Fixed Education Slope

- **Standard Error** (**SE**) for a fixed slope estimate $\boldsymbol{\beta_1}$ in a single-predictor GLM:

$$\mathrm{SE}_{\beta_1} = \sqrt{\frac{\text{residual variance of } y_i}{\text{variance of } x_i * (N-k)}} = \sqrt{\frac{\boldsymbol{\sigma_e^2}}{s_x^2 * (N-k)}}$$

- Example: $\boldsymbol{income_i = \beta_0 + \beta_1(educ_i - 12) + e_i}$, $\boldsymbol{\sigma_e^2 = 162.28}$,
$N = 734, x_i = educ_i - 12: M = 1.81, Var = 8.46$
  - Slope for education predictor: H_0: $\boldsymbol{\beta_1} = 0$, H_A: $\boldsymbol{\beta_1} \neq 0$

$$Est = \mathbf{1.82}, SE = \sqrt{\frac{\mathbf{162.28}}{\mathbf{8.46}*(734-2)}} = \mathbf{0.16}, \; t = \frac{Est-0}{SE} = \frac{\mathbf{1.82}-0}{\mathbf{0.16}} = 11.28,$$

$$DF_{den} = N - k = 734 - 2 = 732, \qquad t_{critical} = \mathbf{1.96}, \qquad \text{so } p < .0001,$$

$$95\%\ CI = Est \pm (t_{critical} * SE) = \mathbf{1.82} \pm (\mathbf{1.96} * \mathbf{0.16}) = 1.51 \text{ to } 2.14$$

  - Interpretation: Predicted income is **significantly higher by 1.82k per unit x_i**: for each additional year of education (so reject H_0 that $\boldsymbol{\beta_1} = 0$)

SEs and CIs for Predicted Income

- **SE of any predicted outcome $\hat{y}_i$** (including the outcome captured by $\boldsymbol{\beta_0}$ for $x_i = 0$) depends on the value of the predictor—the SE will increase as you move away from the predictor's mean:
  - SE of $\hat{y}_i \mid x_i = \sqrt{\sigma_e^2} * \sqrt{\frac{1}{N} + \frac{(x_i - \bar{x})^2}{(N-1)s_x^2}}$

SE (for $\boldsymbol{\beta_0}$ or any $\hat{y}_i$) = average distance of sample predicted value from population value

- $\boldsymbol{income_i = \beta_0 + \beta_1(educ_i - 12) + e_i}$, $\boldsymbol{\sigma_e^2 = 162.28}$,
  $N = 734, educ_i - 12: M = 1.81, Var = 8.46$
- SE and CI for predicted income when education = 12?
  - Given by $\boldsymbol{\beta_0}$: $Est = \mathbf{14.00}$, SE $= \sqrt{\mathbf{162.28}} * \sqrt{\frac{1}{734} + \frac{(0-1.81)^2}{(733)8.46}} = \mathbf{0.55}$,
    $95\%\ CI = Est \pm (t_{crit} * SE) = \mathbf{14.00} \pm (\mathbf{1.96} * \mathbf{0.55}) = 12.91$ to 15.09
- You can use any software package to get predicted outcomes, standard errors, and CIs for any value of the model predictors… and for any person in your dataset!

"simple linear" regression in R (using my new output function)

```
print("GLM Predicting Income from Centered Education 0=12 -- save as ModelEduc12")
ModelEduc12 = lm(data=Lecture2, formula=income~1+educ12)
obj=LMsummary(ModelEduc12, explain=TRUE)  # Full custom output with explanations
```

$$income_i = \beta_0 + \beta_1(educ_i - 12) + e_i$$
$$\hat{y}_i = 14.00 + 1.82(educ_i - 12),\ \sigma_e^2 = 162.28$$

```
Fixed Effects Table
             Est    SE      t      p    LCI    UCI
Intercept 13.998 0.554 25.265 <0.001 12.911 15.086
educ12     1.824 0.162 11.276 <0.001  1.506  2.141
```

```
Explanation:
Est = Estimate, SE = Standard Error, t = t test-statistic,
p = p-value, LCI = Lower Confidence Interval,
UCI = Upper Confidence Interval
```

Creating predicted outcomes using `glht` in R

$$income_i = \beta_0 + \beta_1(educ_i - 12) + e_i, \qquad \hat{y}_i = 14.00 + 1.82(educ_i - 12)$$

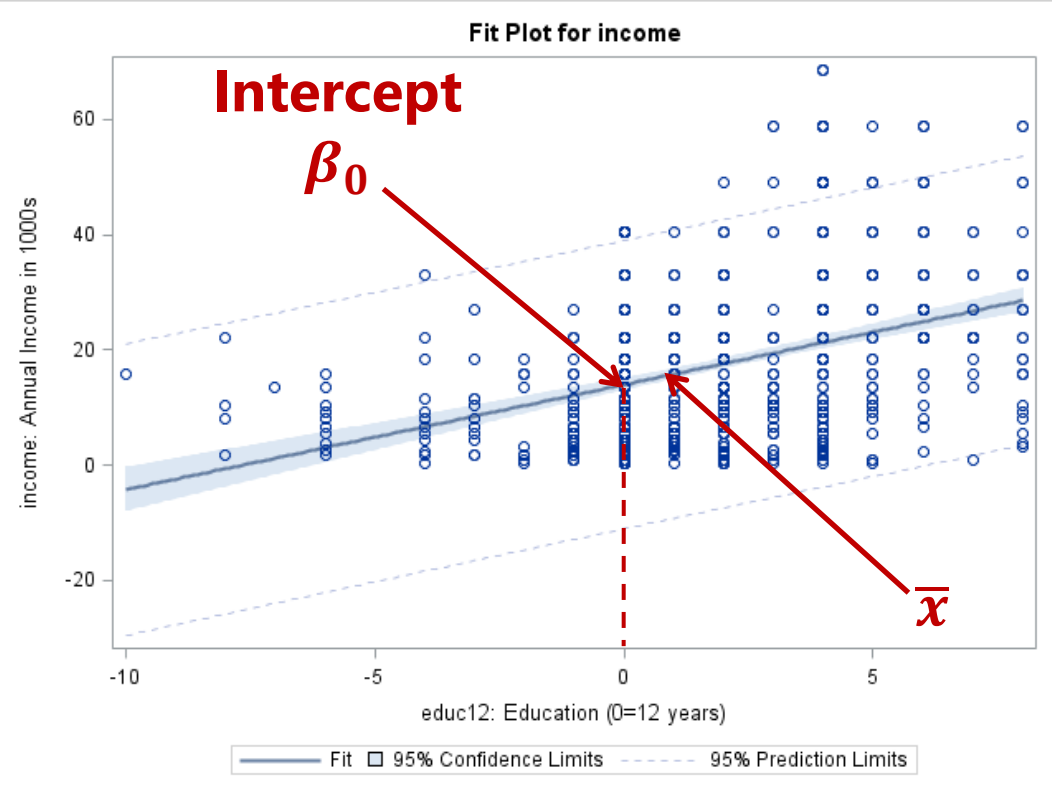
Predicted income for 8 years education: $\hat{y}_i = 14.00(1) + 1.82(-4) = 6.70$
Predicted income for 12 years education: $\hat{y}_i = 14.00(1) + 1.82(0) = 14.00$
Predicted income for 16 years education: $\hat{y}_i = 14.00(1) + 1.82(4) = 21.29$
Predicted income for 20 years education: $\hat{y}_i = 14.00(1) + 1.82(8) = 28.59$

```
print("Get predicted outcomes using glht from multcomp package -- save as PredEduc12")
print("In number lists below, the values are multipliers for each fixed effect in order")
PredEduc12 = glht(model=ModelEduc12, linfct=rbind(
  "Pred Income for  8 years (educ12=-4)" = c(1,-4),  # beta0(1) + beta1*(-4)
  "Pred Income for 12 years (educ12= 0)" = c(1, 0),  # beta0(1) + beta1*( 0)
  "Pred Income for 16 years (educ12= 4)" = c(1, 4),  # beta0(1) + beta1*( 4)
  "Pred Income for 20 years (educ12= 8)" = c(1, 8))) # beta0(1) + beta1*( 8)
obj=glhtSummary(glhtObject=PredEduc12) # Brief custom output
```

```
Linear Combinations Table

                                          Est    SE      t      p    LCI    UCI
Pred Income for  8 years (educ12=-4)  6.703 1.051  6.378 <0.001  4.640  8.767
Pred Income for 12 years (educ12= 0) 13.998 0.554 25.265 <0.001 12.911 15.086
Pred Income for 16 years (educ12= 4) 21.293 0.588 36.183 <0.001 20.138 22.449
Pred Income for 20 years (educ12= 8) 28.588 1.106 25.854 <0.001 26.417 30.759
```

Note how CI widens away from the center

Two Kinds of CIs for "Predicted" Outcomes



Blue shaded line is created by $t_{critical} * SE$; blue dotted line also adds in error from σ_e^2

- **Blue shading** shows the 95% range for the $\hat{y}_i$ **outcomes predicted from the fixed effects** (i.e., regression line)
  - They are narrowest at the predictor mean, and widen as moving away
- **Blue dashed lines** show the 95% range for the **actual** y_i **outcomes** including the residual variance (is way bigger!)

Effect Size via Standardized Slopes

- GLM uses the original variables—this is the "**unstandardized**" solution
- e.g., $\boldsymbol{y_i = \beta_0 + \beta_1(x_i) + e_i}$ → $\boldsymbol{income_i = 14.00 + 1.82(educ_i - 12) + e_i}$
  - Unstandardized fixed slopes ($\boldsymbol{\beta_{unstd}}$) can be standardized ($\boldsymbol{\beta_{std}}$) as:

$$\beta_{std} = \boldsymbol{\beta_{unstd}} * \frac{SD_x}{SD_y}$$

$std\ \beta_0$ will always be 0!

- **Standardized** (Std): $\boldsymbol{y_i = 0 + 0.38(x_i) + e_i}$
  - "Standardized" solution refers to variables that have been z-transformed into $M = 0, SD = 1$ → $x_{std} = (x_i - \bar{x})/s_x$ and $y_{std} = (y_i - \bar{y})/s_y$

```
print("Standardized fixed effect
      solution using lm.beta package")
lm.beta(ModelEduc12)
```

```
Standardized Coefficients::
(Intercept)       educ12
         NA       0.3847
```

  - **For one predictor, $\boldsymbol{\beta_{std}}$ = Pearson correlation** (range = −1 to 1)
- Why do this? To get an **effect size** that is independent of scaling and N!

Remember Pearson's Correlation r?

- For **two quantitative variables**, $\boldsymbol{x_i}$ and $\boldsymbol{y_i}$
  - To graph their relationship, we can request a **scatterplot** ($\boldsymbol{x_i}$ on the x-axis; $\boldsymbol{y_i}$ on the y-axis)
  - Relation will be captured by a general effect size called "**correlation**" (r); one type is **Pearson**
    - Btw, Pearson's r for two binary variables is re-named "**phi**" $\boldsymbol{r}$
    - Btw, Pearson's r for a binary and a quantitative variable is re-named "**point-biserial**" $\boldsymbol{r}$
  - Correlations range continuously from $\boldsymbol{-1}$ to $\boldsymbol{1}$ (with size indicated by absolute value)
- Here are some example scatterplots and the correlations they depict, ranging from perfectly positive ($r = 1$), to none ($r = 0$), to perfectly negative ($r = -1$):

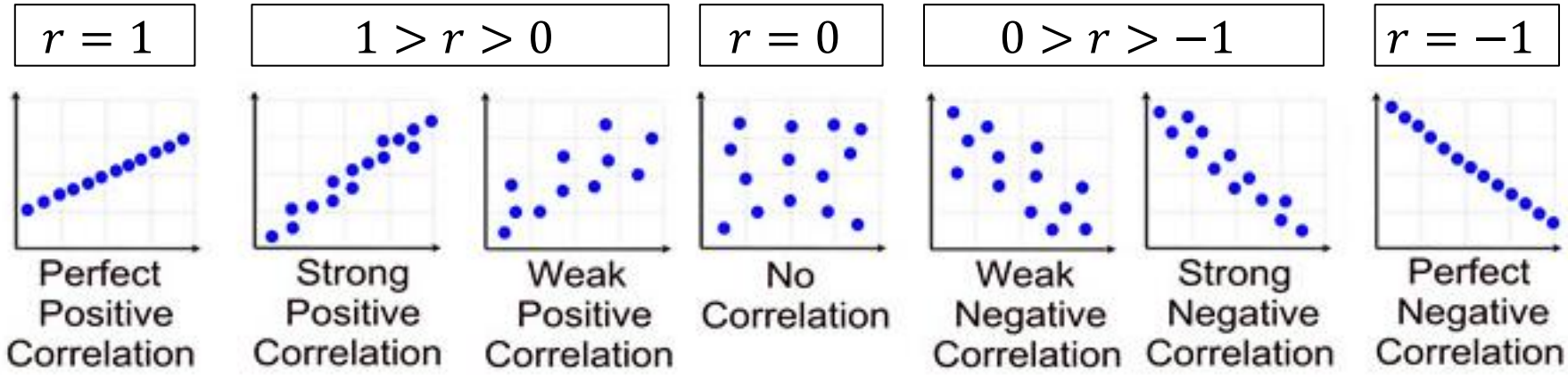


Image borrowed from: https://mathbitsnotebook.com/Algebra1/StatisticsReg/ST2CorrelationCoefficients.html

What about Categorical Predictors?

- A Pearson's r between two quantitative variables $\boldsymbol{x_i}$ and $\boldsymbol{y_i}$ can be represented equivalently with a GLM of $\boldsymbol{x_i}$ predicting $\boldsymbol{y_i}$: $\boldsymbol{y_i = \beta_0 + \beta_1(x_i) + e_i}$
  - Fixed slope $\boldsymbol{\beta_1}$ captures a **linear effect** of $\boldsymbol{x_i}$ predicting $\boldsymbol{y_i}$ in an unstandardized metric (via "centered" $\boldsymbol{x_i}$ so intercept at $x_i = 0$ makes sense)
  - For modeling *nonlinear* effects of quantitative predictors, stay tuned!

- Now we will see how to use the exact same type of GLM to predict a quantitative outcome from a single **categorical predictor**
  - General rule: **predictors with $\boldsymbol{C}$ categories need $\boldsymbol{C}$ fixed effects** to distinguish the outcome means across all unique categories
    - After including the intercept $\boldsymbol{\beta_0}$, we still need $C - 1$ predictors, whose $\boldsymbol{\beta_x}$ slopes then capture specific mean differences between categories
  - Let's start with a **binary variable**, which requires a single predictor slope

A GLM with a Binary Predictor

- GLM of **binary** $\boldsymbol{x_i}$ predicting y_i: $\boldsymbol{y_i = \beta_0 + \beta_1(x_i) + e_i}$
  - Create new $\boldsymbol{x_i}$ so 0 = reference category, 1 = alternative category
  - Btw, this is also called an "**independent** (or **two-sample**) $\boldsymbol{t}$**-test**" (even though all types of predictors use a t test-statistic to test significance)

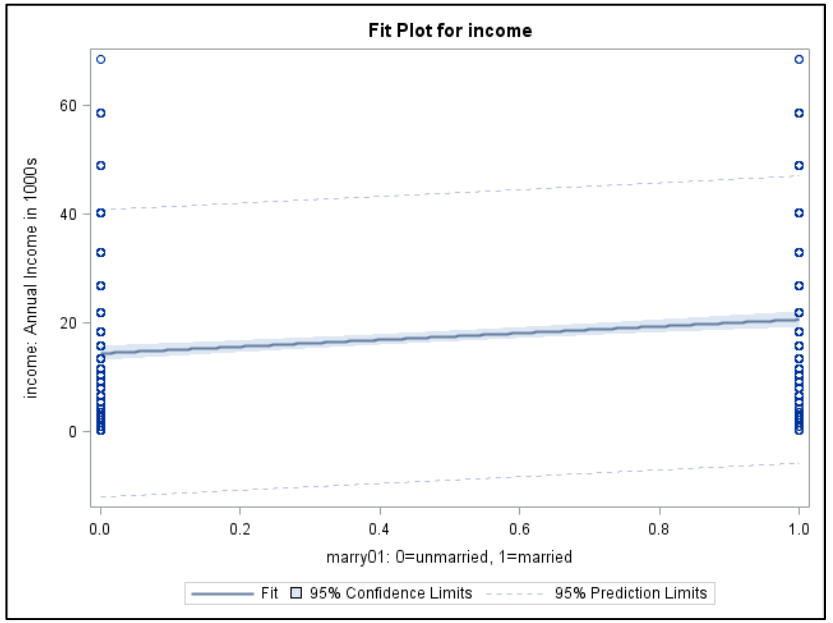
- e.g., annual income in $1000s predicted by binary marital status
  - $marry01$: 0 = no, 1 = yes → $\boldsymbol{income_i = \beta_0 + \beta_1(marry01_i) + e_i}$
  - $\boldsymbol{\beta_0}$ = **intercept** = expected income for unmarried persons ($marry01_i = 0$)
  - $\boldsymbol{\beta_1}$ = **slope** for $marry01_i$ = expected mean difference for married persons (=1) **relative to unmarried persons** (=0)
  - $\boldsymbol{e_i}$ = **residual** = difference in model-predicted income (from $\boldsymbol{\hat{y}_i}$) and actual income $\boldsymbol{y_i}$, whose (residual) variance is estimated as $\boldsymbol{\sigma_e^2}$

A GLM with a Binary Predictor

$\boldsymbol{income_i = \beta_0 + \beta_1(marry01_i) + e_i}$ (btw, $r = .23, p < .0001$)
Slope result: $\boldsymbol{\beta_1 = 6.22}, SE = \mathbf{0.996} \rightarrow t = 6.25, p < .0001$

- Income predicted for **unmarried**:
  $\boldsymbol{\hat{y}_i = 14.45 + 6.22(0) = 14.45}$
- Income residual for unmarried:
  $\boldsymbol{e_i = y_i - \hat{y}_i \rightarrow e_i = y_i - 14.45}$

- Predicted income for **married**:
  $\boldsymbol{\hat{y}_i = 14.45 + 6.22(1) = 20.67}$
- Income residual for married:
  $\boldsymbol{e_i = y_i - \hat{y}_i \rightarrow e_i = y_i - 20.67}$



- So married people ($x_i = 1$) are predicted to have **significantly higher** income by **$6.22 thousand dollars** on average than unmarried people ($x_i = 0$)
- A "linear" relationship is only kind possible for binary predictors (only one unit difference)

Effect Size for a Mean Difference: d

- For categorical predictors, an $\boldsymbol{r}$ effect size is less intuitive than a **Cohen's $\boldsymbol{d}$** effect size, a **standardized mean difference** between group 0 and group 1
  - Cohen's $d = \frac{\bar{y}_0 - \bar{y}_1}{SD_{pooled}}$, where $SD_{pooled} = \sqrt{\frac{s_0^2 + s_1^2}{2}}$
  - Other variants you might see: Glass' delta (δ) uses SD for only 1 group; Hedges' g weights by the relative sample size in each group

- If your GLM contains only one binary predictor, then the pooled SD is the same as the square root of GLM residual variance, $\sqrt{\boldsymbol{\sigma_e^2}}$
  - Otherwise, $\sqrt{\boldsymbol{\sigma_e^2}}$ will be smaller because of the other predictors *(stay tuned for more about effect sizes given multiple predictors)*
- **$\boldsymbol{d}$ or $\boldsymbol{r}$ can also be found directly using the $\boldsymbol{t}$ test-statistic for a fixed effect:**
  - $d = \frac{2t}{\sqrt{DF_{den}}}$, $r = \frac{t}{\sqrt{t^2 + DF_{den}}}$, $d = \sqrt{\frac{4r^2}{1-r^2}}$, $r = \sqrt{\frac{d^2}{4+d^2}}$

Formulas borrowed from: http://davidileitman.com/wp-content/uploads/2014/04/EffectSizeFormulas.pdf

GLM with binary predictor in R (using my new output function)

```
# Recode marry predictor so that 0 is meaningful
Lecture2$marry01=NA  # Create new empty variable
Lecture2$marry01[which(Lecture2$marry==1)]=0  # marry01=0 if marry=1
Lecture2$marry01[which(Lecture2$marry==2)]=1  # marry01=1 if marry=2
# marry01: 0=unmarried, 1=married             # Label as a comment only

print("GLM Predicting Income from Marry01 (0=Unmarried) -- save ModelMarry01")
ModelMarry01 = lm(data=Lecture2, formula=income~1+marry01)
obj=LMsummary(ModelMarry01, effectsizes=TRUE) # Custom output with effect sizes
```

```
Sums of Squares Table
               SS  DF        MS      F      p    R2
Model    7060.102   1  7060.102 39.044 <0.001 0.051
Error  132363.130 732   180.824
Total  139423.232 733   190.209

Fixed Effects Table
            Est    SE      t      p    LCI    UCI
Intercept 14.445 0.675 21.404 <0.001 13.120 15.770
marry01    6.224 0.996  6.249 <0.001  4.268  8.179

Effect Sizes for Fixed Effects Table
           Est      p     d    pr   sR2
marry01 6.224 <0.001 0.462 0.225 0.051
```

$$income_i = \beta_0 + \beta_1(marry01_i) + e_i$$
$$\hat{y}_i = 14.45 + 6.22(marry01_i),\ \sigma_e^2 = 180.82$$

Pearson correlation matrix of all example variables (for an r effect size instead of d)

```
        income educ marry
income    1.00 0.38  0.23
educ      0.38 1.00  0.05
marry     0.23 0.05  1.00
```

Creating predicted outcomes using `glht` in R

$$income_i = \beta_0 + \beta_1(marry01_i) + e_i, \qquad \hat{y}_i = 14.45(1) + 6.22(marry01_i)$$

Predicted income unmarried (marry01= 0): $\hat{y}_i = 14.45(1) + 6.22(0) = 14.45$

Predicted income married (marry01= 1): $\hat{y}_i = 14.45(1) + 6.22(1) = 20.67$

```
print("Get predicted outcomes using glht from multcomp package -- save as PredMarry01")
print("In number lists below, values are multiplier for each fixed effect in order")
PredMarry01 = glht(model=ModelMarry01, linfct=rbind(
  "Pred Income for Unmarried=0" = c(1,0),    # beta0(1) + beta1*(0)
  "Pred income for Married=1"   = c(1,1)))   # beta0(1) + beta1*(1)
obj=glhtSummary(glhtObject=PredMarry01)      # Brief custom output
```

```
Linear Combinations Table

                              Est    SE      t      p    LCI    UCI
Pred Income for Unmarried=0 14.445 0.675 21.404 <0.001 13.120 15.770
Pred income for Married=1   20.669 0.733 28.217 <0.001 19.231 22.107
```

0 = .54 of sample
1 = .46 of sample

This method of creating a binary (0/1) variable to represent two groups is my preferred strategy because then the reference for the intercept is one of the specific groups. Alternatively, you may see "**effect coding**"…

Section Review: r, CI, and Binary Predictors

1. Which is stronger, a correlation of .3 or -.6?

2. If a confidence interval for a slope overlaps 0, the relationship between the predictor and outcome is "significant" – true or false?

3. Which group has a higher outcome mean (0 or 1?) given:

   a. a positive slope?
   b. a negative slope?

4. If a binary predictor has a negative slope, which group is higher: 0 or 1?

2 reasons why I don't like effect coding (−1,1)

$$income_i = \beta_0 + \beta_1(marryEffect_i) + e_i, \qquad \hat{y}_i = 17.56 + 3.11(marryEffect_i)$$

Predicted income unmarried (marryEffect = −1): $\hat{y}_i = 17.56(1) + 3.11(-1) = 14.45$

Predicted income married (marryEffect = 1): $\hat{y}_i = 17.56(1) + 3.11(1) = 20.67$

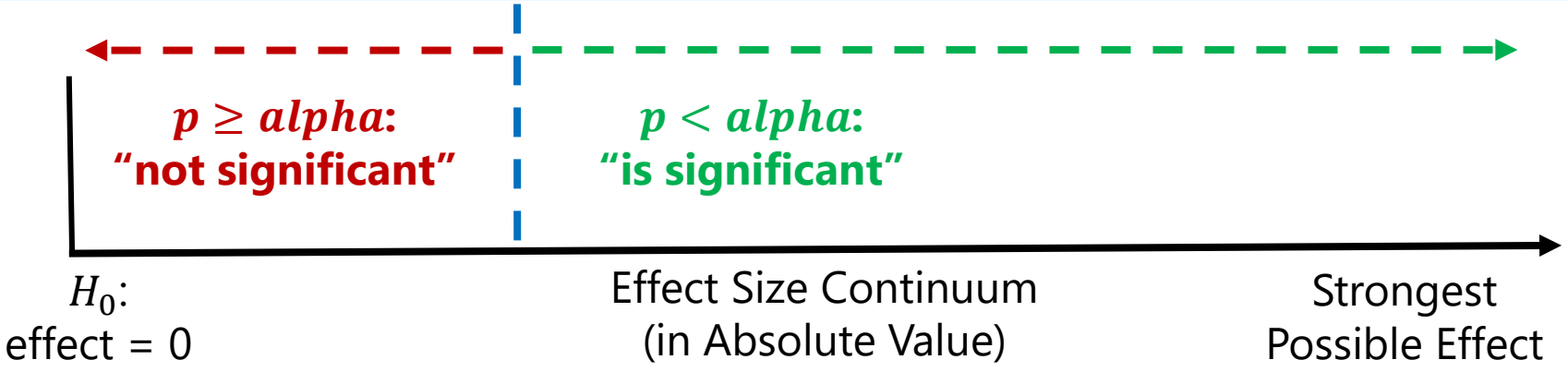
same $\hat{y}_i$

```
# Recode marry predictor to demonstrate why I don't like effect coding
Lecture2$maryEffect=NA  # Create new empty variable
Lecture2$maryEffect[which(Lecture2$marry==1)]=-1  # maryEffect= -1 if marry=1
Lecture2$maryEffect[which(Lecture2$marry==2)]= 1  # maryEffect=  1 if marry=2
# maryEffect: -1=unmarried, 1=married              # Label as a comment only

print("GLM Predicting Income from maryEffect (-1=Unmarried) -- save as ModelMaryEffect")
ModelMaryEffect = lm(data=Lecture2, formula=income~1+maryEffect)
obj=LMsummary(ModelMaryEffect, effectsizes=TRUE) # Custom output with effect sizes
```

```
Fixed Effects Table
             Est    SE      t      p    LCI    UCI
Intercept 17.557 0.498 35.255 <0.001 16.580 18.535
maryEffect  3.112 0.498  6.249 <0.001  2.134  4.090

Effect Sizes for Fixed Effects Table
              Est      p     d    pr   sR2
maryEffect 3.112 <0.001 0.462 0.225 0.051
```

1. **Who is the intercept for?**
2. **What does the slope now mean?**

Effect Size, Sample Size, and Test Statistics



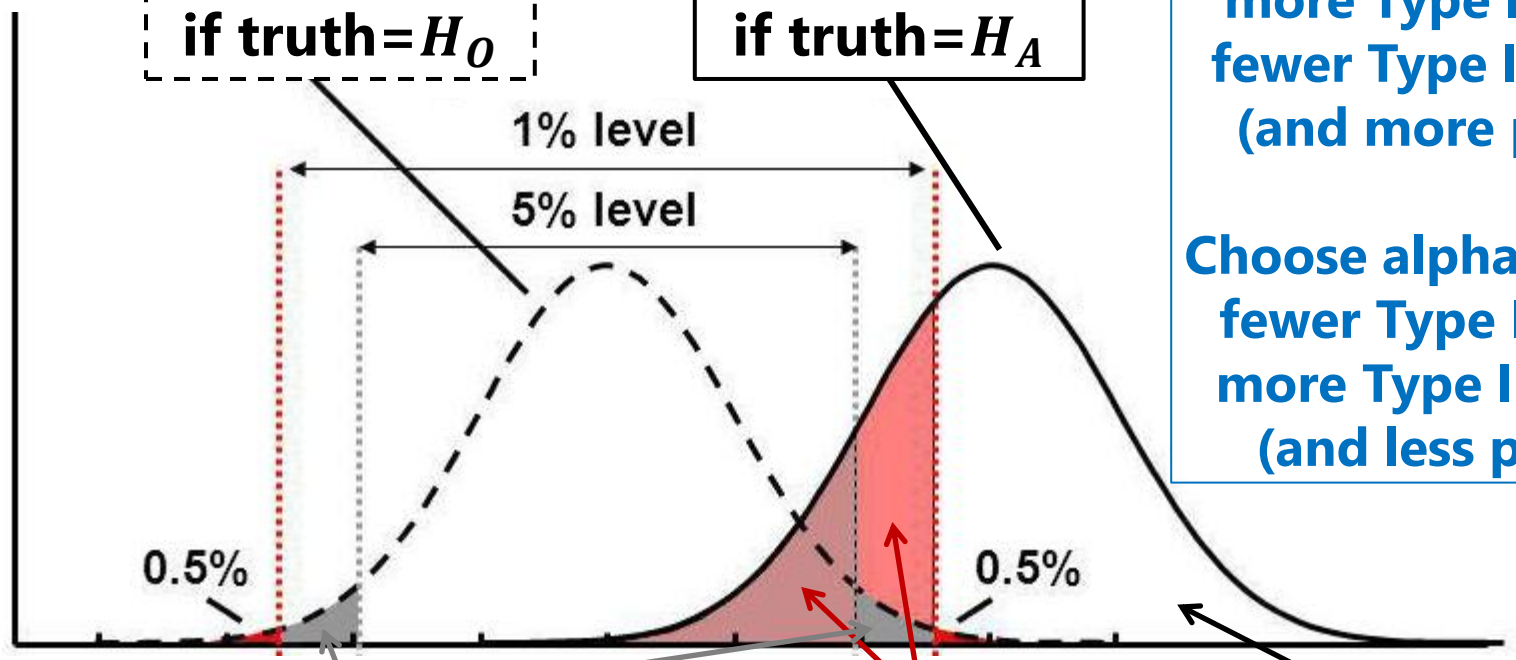
- **Test statistics** (e.g., t or z) **standardize** an estimate's deviation from the H_0 null hypothesis
  - When compared to "null" or H_0 reference distribution, they give you a p-value: probability of finding an effect ≥ the obtained effect **if H_0 is true**
  - **But test statistics** are a function of both **effect size** and **sample size N**!
- In other words, test statistics and alpha combine to locate the blue line above that divides effect sizes into "not significant" and "significant"
- **Blue line moves to the right** (is harder to "find" the same effect) given:
  - **Lower alpha level** (smaller % of distribution allowed for "unexpected")
  - **Smaller N** → Fewer people = **less "statistical power"** (as discussed next!)

Decision Errors in Hypothesis Testing

- Usually, we test a null hypothesis against a two-sided alternative:
  - Typical null H_0: slope = 0; alternative H_A: slope ≠ 0
- **2 chances to get it right and 2 chances to get it wrong**, governed by:
  - **Alpha** (α) = expected proportion of **Type I errors** *for a given* H_0
    - Higher alpha → less extreme boundaries for "significant" → more Type I errors
  - **Beta** (β) = expected proportion of **Type II errors** *for a given effect size*
    - Usually expressed as $1 - \beta$ = **statistical power**: Probability of finding a TRUE effect
    - More people N and/or greater effect size = more power (fewer Type II errors)!

	If Truth = H_0	**If Truth = H_A**
Decision: Retain H_0	**Correct: Really NO Effect**	**Miss: Type II Error**
Decision: Reject H_0	**False Alarm: Type I Error**	**Correct: Really IS an Effect**

Decision Errors in Hypothesis Testing



Choose alpha (α)=5%: more Type I errors, fewer Type II errors, (and more power)

Choose alpha (α)=1%: fewer Type I errors, more Type II errors, (and less power)

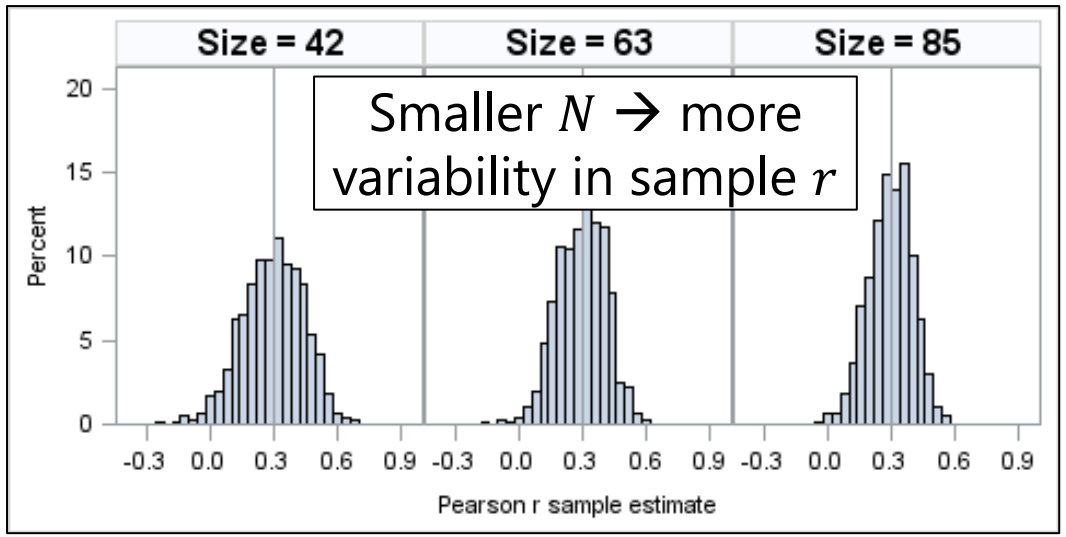
H_O gray areas = α = % Type I errors (false alarms)

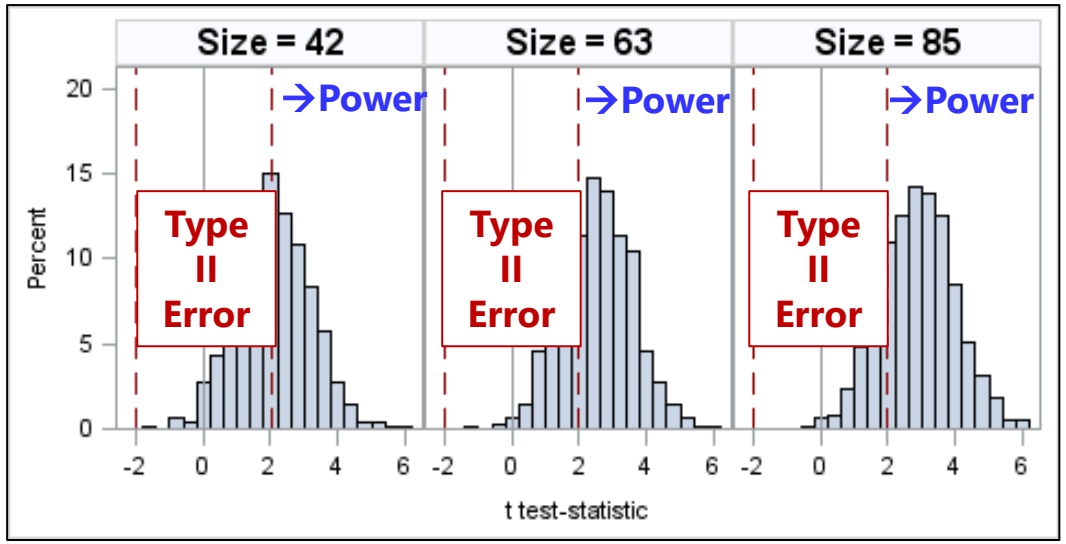
H_A red areas = β = % Type II errors (misses)

H_A white area = $1-\beta$ % power

Image was borrowed from: https://images.app.goo.gl/eDuhatsiyKWjrUvcA (now unknown)

Anticipating Statistical Power Tables…





- Demo: I simulated $r = .3$ for 100,000 fake persons
- Drew 1000 samples each of $N = 42, 63,$ or 85
- **Power = % area past $t_{critical}$** (is greater with more N)

N	Type II Error: % not significant	Statistical Power: % significant
42	50%	50%
63	34%	66%
85	21%	79%

Typical desired power = 80% (so Type II error rate = 20%)

Power Analysis for r Effect Size at $\alpha = .05$ (from Cohen, 1988, p. 102)

	r								
Power	.10	.20	.30	.40	.50	.60	.70	.80	.90
.25	167	42	20	12	8	6	5	4	3
.50	385	96	42	24	15	10	7	6	4
.60	490	122	53	29	18	12	9	6	5
2/3	570	142	63	34	21	14	10	7	5
.70	616	153	67	37	23	15	10	7	5
.75	692	172	75	41	25	17	11	8	6
.80	783	194	85	46	28	18	12	9	6
.85	895	221	97	52	32	21	14	10	6
.90	1047	259	113	62	37	24	16	11	7
.95	1294	319	139	75	46	30	19	13	8
.99	1828	450	195	105	64	40	27	18	11

- If you start with a target N, it's "**sensitivity analysis**" to find a "minimum detectable effect size" (MDES)
  - e.g., for $N = 30$, should have power > 80% for $r \geq .5$
  - e.g., for $N = 50$, should have power > 80% for $r \geq .4$

- Cells give N for row's power to find column's r
- If you start with target r to find N, it's "**a priori power analysis**"
  - e.g., for $r = .3$, 80% power is predicted for $N = 85$
  - e.g., for $r = .2$, 80% power is predicted for $N = 194$

Refer to these power tables for other common combinations

Decisions and Decision Errors: Summary

- Given alpha (% unexpected), about the $(Est - H_0)/SE$ = t-value:

- If ***t*-value falls outside** t-critical boundaries, then ***p* < alpha:**
  Result is sufficiently **unexpected** → reject H_0 → "**significant**":
  - DO have to worry about a false alarm (Type I error ← your p-value)
  - DO NOT have to worry about a Type II error (because you didn't miss!)
  - BUT—a significant result with low power is less likely to replicate!

- If ***t*-value falls inside** t-critical boundaries, then ***p* ≥ alpha:**
  Result is sufficiently **expected** → retain H_0 → "**nonsignificant**":
  - DO NOT have to worry about false alarm (Type I error not applicable)
  - DO have to worry about a miss (Type II error)
  - In planning studies, the conventional level of power (= 1 – Type II error) to aim for is .80 (which is much harder to do for smaller effects)

Summary: Introduction to GLMs

- Predictive linear models (i.e., formed as outcome = constant*predictor + constant*predictor…) create expected outcomes from 1+ predictors
  - **General** linear models use a **normal** conditional (residual) distribution *(btw, **Generalized** linear models use **some other** conditional distribution)*

- General linear models are called different names by type of predictor, but any kind of predictive model can be specified, for example:
  - **Empty Model**: no predictors; is used to recreate outcome mean and variance as unconditional starting point (sample mean is predicted for all persons)
    - $\boldsymbol{y_i} = \boldsymbol{\beta_0} + \boldsymbol{e_i}$ → $\boldsymbol{\beta_0} = \bar{\boldsymbol{y}}$, variance of $\boldsymbol{e_i}$ residuals = $\boldsymbol{\sigma_e^2}$ → all the $\boldsymbol{y_i}$ variance
  - **Single Predictor Model**: used to customize expected outcomes using a single predictor → $\boldsymbol{y_i} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}(\boldsymbol{x_i} - \boldsymbol{c}) + \boldsymbol{e_i}$ ($\boldsymbol{c}$ is centering constant)
    - $\boldsymbol{\beta_0}$ = **fixed intercept** = expected y_i when $x_i = 0$
    - $\boldsymbol{\beta_1}$ = **fixed slope** of x_i = difference in y_i per one-unit difference in x_i
    - $\boldsymbol{e_i}$ = **residual** = deviation between actual y_i and predicted $\hat{\boldsymbol{y}}_i$
    - $\boldsymbol{x_i}$ should have a meaningful 0 value ← center by subtracting constant $\boldsymbol{c}$

Foreshadowing… please stay tuned!

- In a GLM with a **single predictor** (quantitative or binary), the effect size given by its **standardized slope** will always be **equal to Pearson's r**
- So what's the point of estimating a GLM??? The real utility lies in **expanding the model** for at least one of these 3 reasons:

  - Multiple fixed slopes for a single predictor variable (in lecture 3)
    - To examine **nominal** or **ordinal predictors** of a quantitative outcome
    - To examine **nonlinear effects of a quantitative predictor** on a quantitative outcome (e.g., quadratic or piecewise spline predictors)
  - Multiple predictors (each potentially using 1+ fixed slopes)
    - To test the **unique effects** of correlated predictors after controlling for what information they have in common (in lecture 4)
  - Moderation of predictor effects (via interaction terms)
    - To test if predictor **slopes depend on** other predictors (in lecture 5)

Sample Results for Example Analyses

- The extent to which annual income in thousands of US dollars (*M* = 17.30, *SD* = 13.79) could be predicted from years of education (*M* = 13.81, *SD* = 2.91) and binary marital status (1 = unmarried 54.09%, 2 = married 45.91%) was examined in separate general linear models (i.e., simple linear regressions). All analyses were conducted using the lm function in R v. 4.5.2. Predicted outcomes were generated using the glht function within the multicomp R package v. 1.4-29.

- To create a meaningful model intercept, education was centered such that 0 = 12 years. Education was found to be a significant predictor of annual income: Relative to the reference expected income for a person with 12 years of education provided by the model intercept of 14.00k (SE = 0.55), for every additional year of education, annual income was expected to be higher by 1.82k (SE = 0.16, $p < .001$), resulting in a standardized coefficient = 0.38 (i.e., the Pearson correlation between annual income and education). For example, persons with only 8 years of education were predicted to have an annual income of only 6.70k (SE = 1.05), persons with 16 years of education were predicted to have an annual income of 21.29k (SE = 0.59), and persons with 20 years of education were predicted to have an annual income of 28.59k (SE = 1.11). *[Spoiler alert: we will test the adequacy of only a linear (constant) effect for years of education in Lecture 3.]*

Sample Results, Continued

- We then examined prediction of annual income by binary marital status. To create a meaningful model intercept, marital status was recoded so that 0 = unmarried persons and 1 = married persons. Marital status was also a significant predictor of annual income: Relative to the reference expected income for unmarried persons provided by the model intercept of 14.45k (SE = 0.67), married persons were expected to have significantly greater income by 6.22k (SE = 1.00, $p < .001$), resulting in a predicted income for married persons of 20.67k (SE = 0.73) and a standardized mean difference of Cohen's $d = 0.462$ (i.e., the groups differed by 0.462 standard deviation units).

- Note: because a GLM with a single binary predictor is also known as a "two-sample *t*-test" here is what the results would look like written from that angle…

  A two-sample *t*-test (i.e., assuming homogeneous variance across groups) was used to examine mean differences between unmarried and married persons in annual income. A significant mean difference was found, $t(732) = 6.25$, $p < .001$, such that annual income for married persons (M = 20.67k, SE = 0.73) was significantly higher than for unmarried persons (M = 14.45k, SE = 0.67).