

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: you did not “run analyses” or “calculate models”; you “conducted analyses” and “estimated models”

Roles and Labels of Study Variables

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

- Reason (Explainer):

- In notation: x variable
 - Exogenous (is not explained)
- **Predictor**
 - My preferred generic term
- Independent variable (**IV**)
 - 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); also used for quantitative predictor in ANCOVA

- What is To Be Explained:

- In notation: y variable
 - Endogenous (is explained)
- **Outcome**
 - My preferred generic term
- Dependent variable (**DV**)
 - Used more often in experimental studies
- 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:
 - To what extent does positive feedback improve performance speed and accuracy more than neutral feedback?
 - Predictors:
 - Outcomes:
 - Is faster academic growth in elementary school related to more frequent reading to children when in preschool?
 - Predictors:
 - Outcomes:
 - How effective is teacher training for creating higher rates of positive feedback to a teacher's students?
 - Predictors:
 - Outcomes:

Types of Inferences: 2 possibilities in describing how x relates to y

- **x causes y \rightarrow causal inference** requires the following:
 - x variable had to come first (temporal precedence)
 - x variable was under complete experimental control during the study (best: through random assignment and experimental manipulation)
 - Study design eliminates all possible alternative explanations
- **x relates to y** (synonyms = **associative, correlational**)
 - We have observed a relationship, but we do not have the ability to infer cause given the design (i.e., it's an observational study with measures only)
 - In lieu of experimental control, we can attempt **statistical control**: include other predictors that represent alternative explanations for why x relates to y , and see if x is still related to y \rightarrow many research questions try to do this
- These 2 possibilities can only be distinguished by study design—they have nothing to do 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?
 - **“Generalized”** Linear Models using some other distribution and a transformed predicted outcome (called a “link function”) to address variable 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 distribution and logit link
 - Count outcome? Use Poisson-family distributions and log link
 - Come back in Spring 2023 to learn these *generalized* linear models 😊

What “Linear” in “Linear Models” Means

- Most predictive models have a “**linear**” form, which looks like this:
 - $y_i = (\text{constant} * 1) + (\text{constant} * X_{\text{pred}1_i}) + (\text{constant} * X_{\text{pred}2_i})...$
 - 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 predictors (the $X_{\text{pred}i}$ variables) as needed
 - Fortunately, this also means we can include categorical x_i predictors
- Historically, variants of the **general linear model** (for continu-ish outcomes) get siloed into different classes and called different names based on **what kind of x_i predictor variables are 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
 - We’ll use SAS GLM or REG, STATA REGRESS, and R LM for all of it!

Welcome to the GLM!

- Linear models use **new notation within one equation** to describe how all the x_i predictors relate to the y_i outcome(s) in your sample
 - 1 outcome? "**Univariate** GLM" 2+ outcomes? "**Multivariate** GLM"
- Starting point for univariate GLMs is always to represent central tendency and dispersion of the outcome variable (y_i)
 - We will use **mean and variance** to describe the outcome because the GLM uses the normal distribution (in which skewness should be 0)
- Your first GLM is the "**Empty**" model (=no predictors): $y_i = \beta_0 + e_i$
 - y_i = "y sub i": outcome variable for *each person* in your sample
 - β_0 = "**beta 0**" (sometimes called "beta not"—but not by me)
 - More generally, betas (β) will represent **values to be estimated** that will apply to the whole sample (i.e., betas are constants) = "**fixed effects**"
 - The beta **subscripts index each fixed effect** (starting at 0, then 1,2,...)

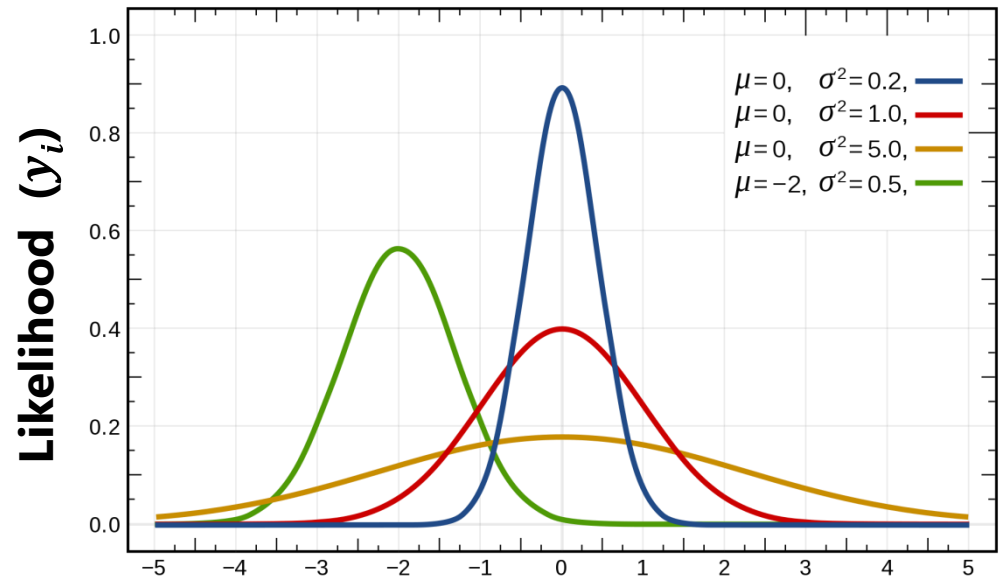
The “Empty” General Linear Model

- The “**Empty**” model (empty = no predictors): $y_i = \beta_0 + e_i$
 - β_0 = “beta 0” = “**the intercept**” (or “the constant”, ugh) and is defined as the predicted (expected) value for the y_i outcome when all x_i predictors = 0 (so the estimated value for β_0 will change as the predictors are changed)
 - 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, $\beta_0 = \bar{y}$)
- So what would β_0 be for:
 - The blue line? the red line?
 - But why do the red and blue lines differ????

Univariate Normal PDF

(Probability Density Function):

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



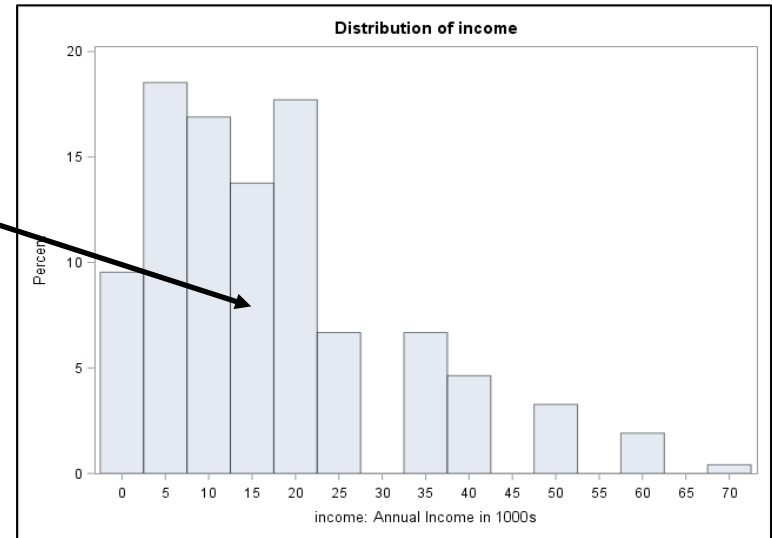
The “Empty” General Linear Model

- The “**Empty**” model (“no predictors”): $y_i = \beta_0 + e_i$ (in which $\beta_0 = \bar{y}$)
 - e_i = “e sub i” or “**residual**” = deviation between the actual y_i outcome for each person and y_i outcome predicted by the model (through the beta fixed effects)
 - Because the empty model predicts the same \bar{y} for all y_i values, the e_i residual for each person will just be the difference between y_i and β_0 : $e_i = y_i - \beta_0$
 - Rather than focusing on each individual e_i residual, we keep track of their **variance across persons** as the estimated model parameter, **denoted as σ_e^2**
 - You’ve seen this before: $Variance = s^2 = \frac{\sum_{i=1}^N (y_i - \bar{y})^2}{N-1} = \frac{\sum_{i=1}^N (e_i)^2}{N-1} = \text{now } \sigma_e^2$
 - In other words, the two parameters for the empty model outcome give us the y_i mean (as β_0) and y_i **variance** (as σ_e^2) → right now $\sigma_e^2 = \text{all the } y_i \text{ variance}$
- In writing linear models, the **notation refers to population parameters** instead of sample statistics or “estimates” (i.e., we use σ_e^2 instead of s^2)
 - Why? It’s understood that we only have one sample from which to estimate parameters used to make inferences about some (hypothetical) population

Quantifying Uncertainty of Sample Estimates

- Let's say we want to predict annual income in \$1000s in $N = 734$ using an empty model (= no predictors yet)

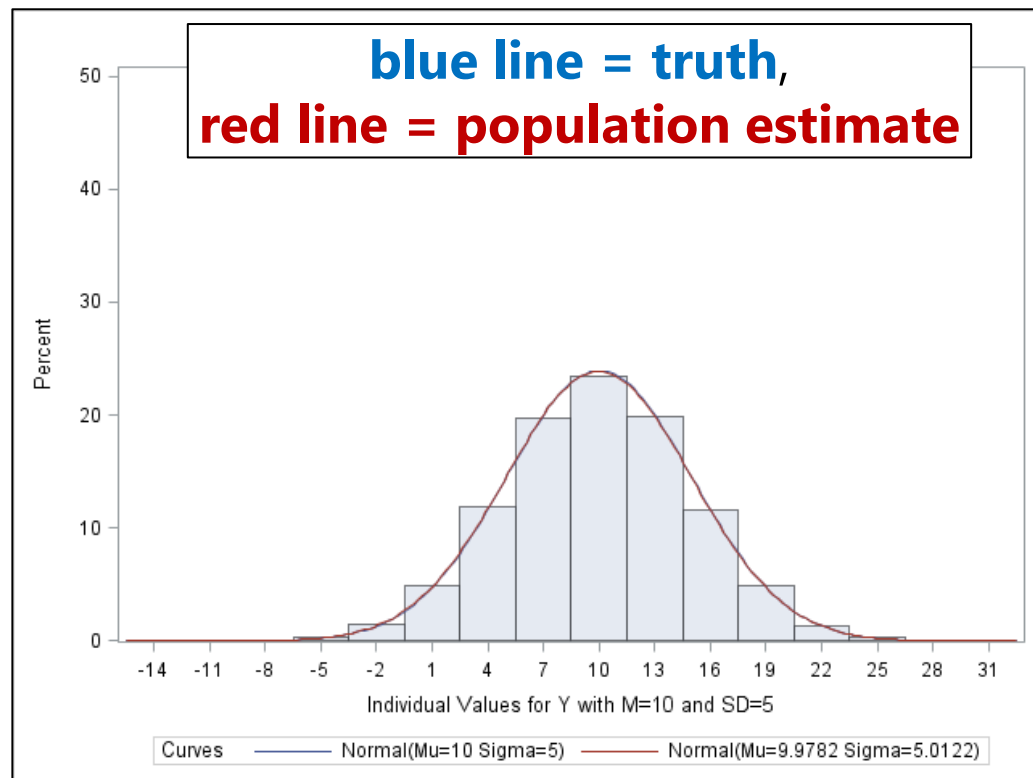
- $y_i = \beta_0 + e_i \rightarrow income_i = \beta_0 + e_i$
- Fixed intercept estimate $\beta_0 = 17.30$
(= sample \bar{y} because no predictors)
- Person-specific residual $e_i = y_i - \beta_0$
- Variance of e_i residuals $\rightarrow \sigma_e^2 = 190.21$
(= sample s^2 because no predictors)
- SD of $e_i \rightarrow \sigma_e = \sqrt{190.21} = 13.80$
(= sample s because no predictors)



- If the sample mean $\bar{y} = 17.30$ is supposed to be our best guess for the (unknown) population mean μ , **how good of an estimate** is it?
 - Said differently: If we had another sample from the same population, by how much would that 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 \rightarrow from the mean's "**sampling distribution**"
 - Note: This is NOT the same as a variable's distribution across persons (above)...!

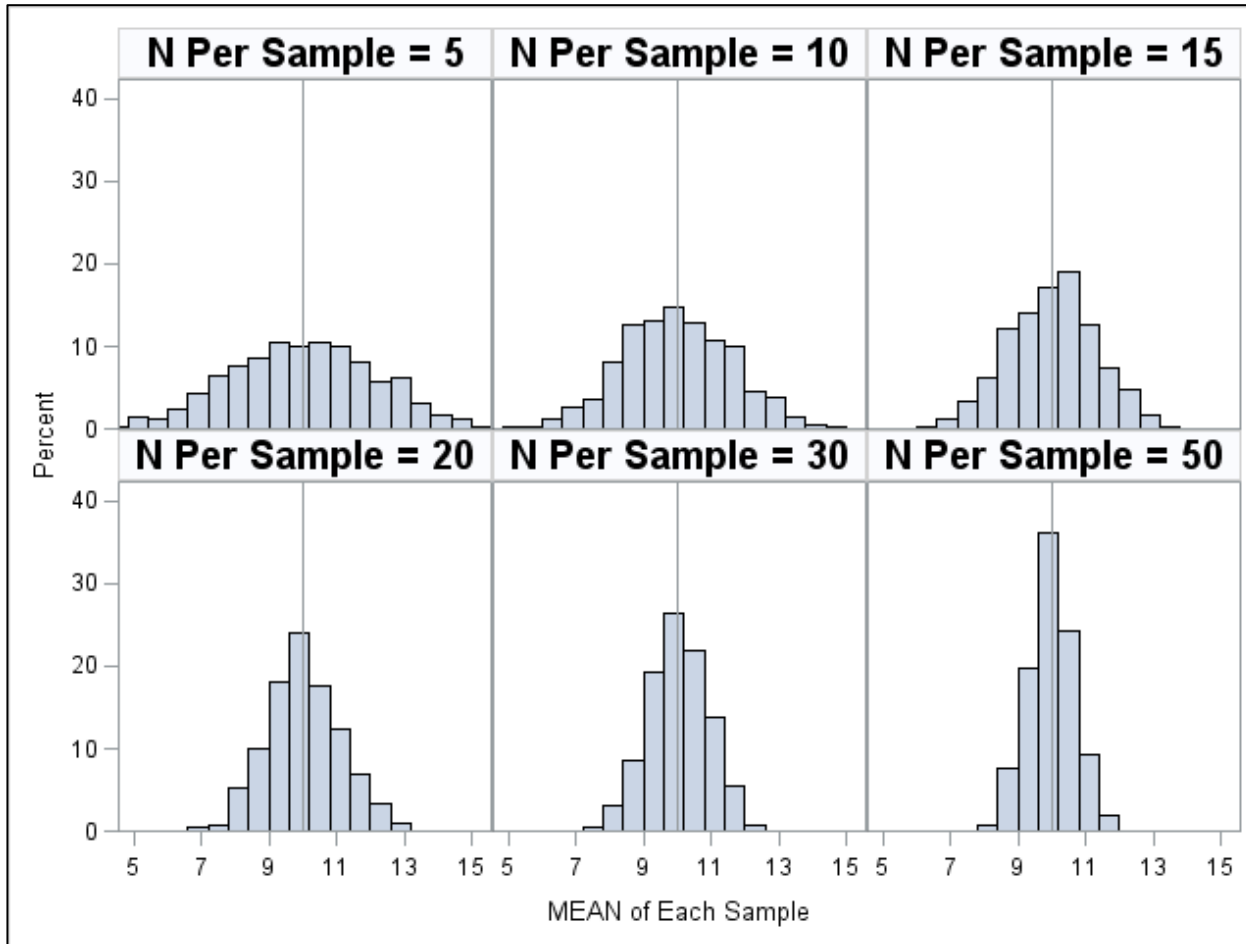
Building Intuition about Sampling Distributions of Estimates (the mean for now)

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



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

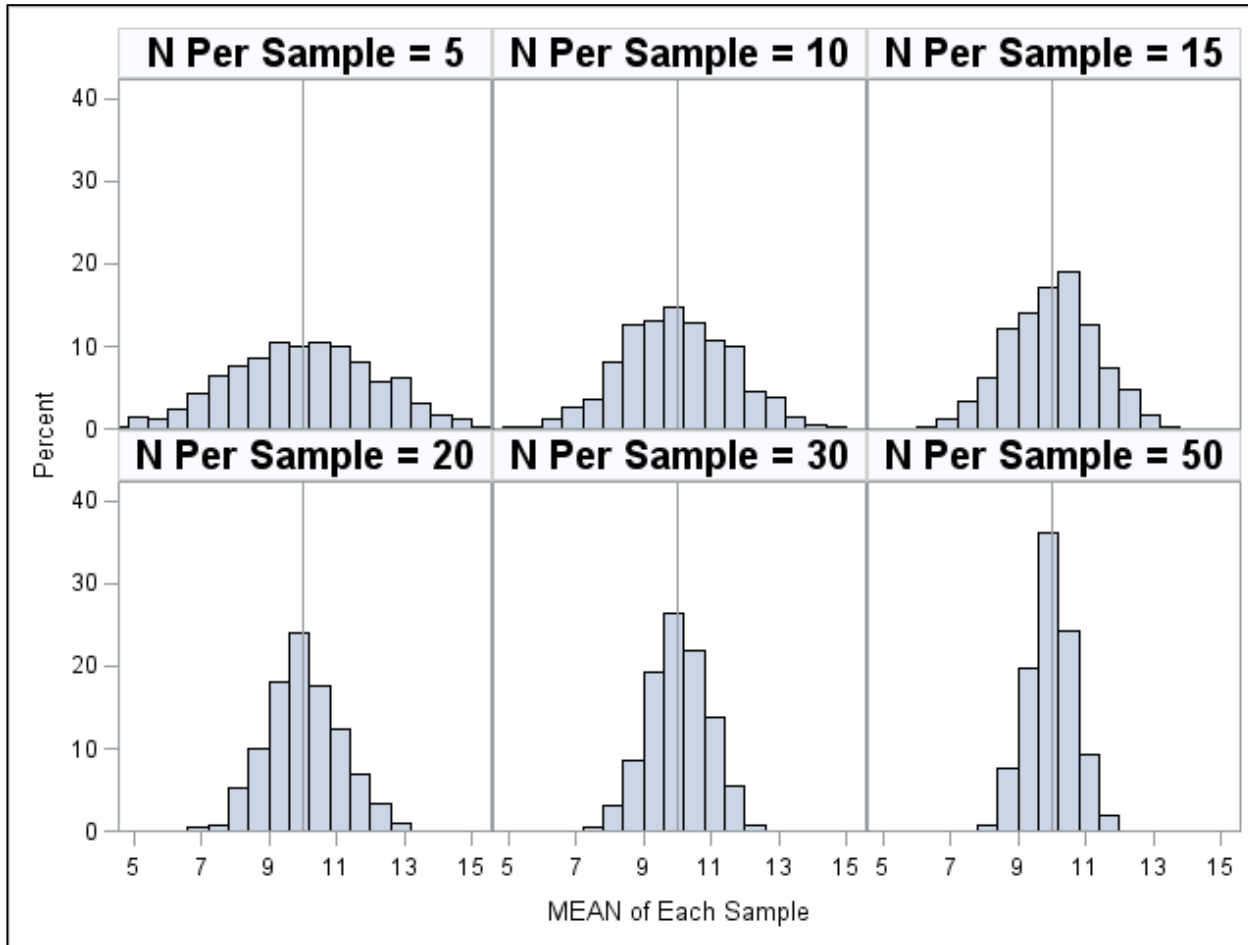
1 000 samples each for different $N \dots$



- Population values:
Mean $\mu = 10$
(SD $\sigma = 5$)
- 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 N** in each sample \rightarrow **less dispersion in \bar{y}_s** across samples (**more consistency**)

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

1000 samples each for different N ...



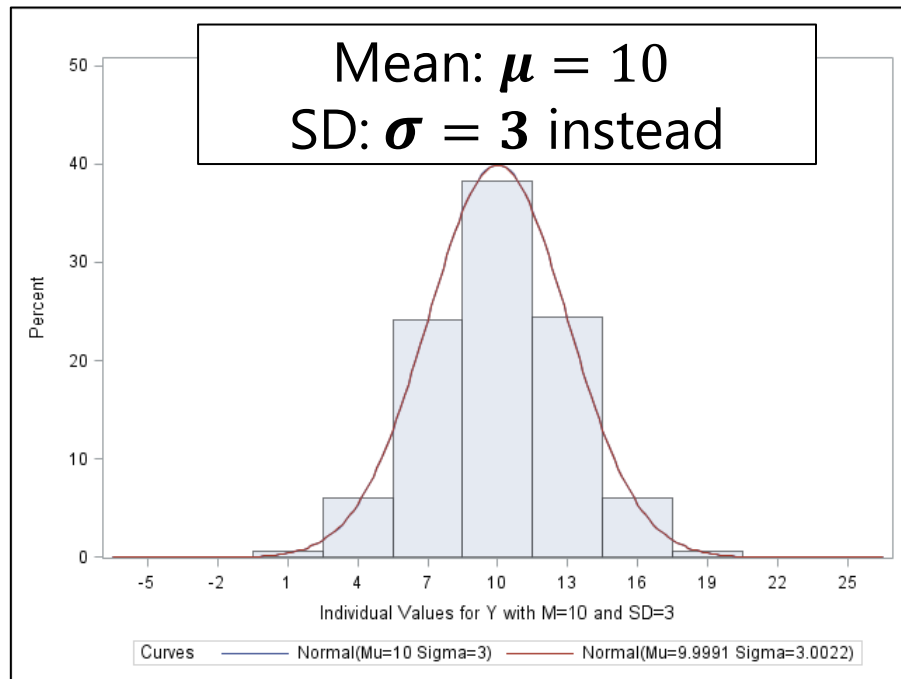
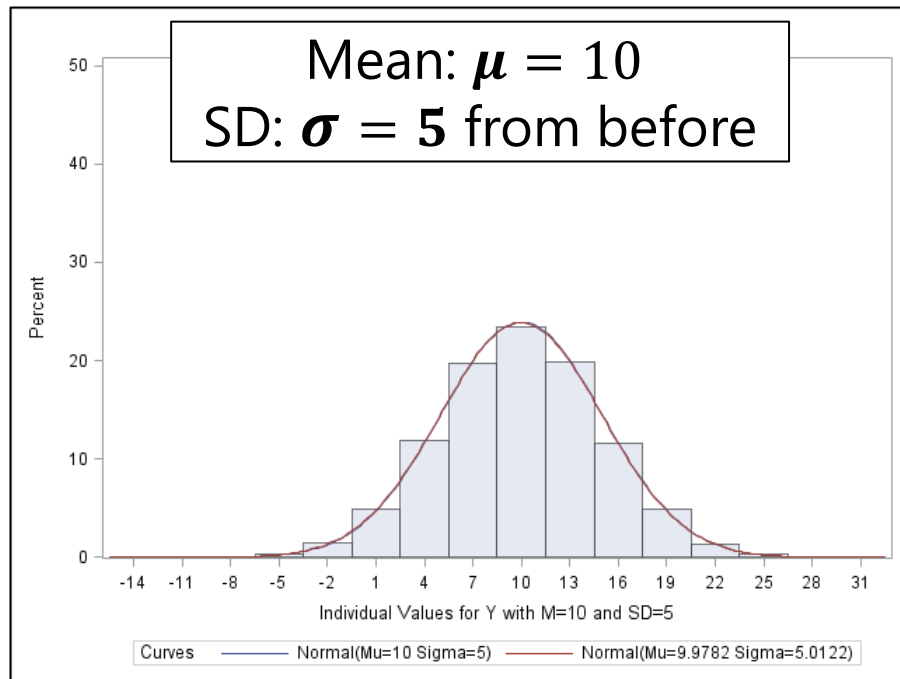
- Population values:
Mean $\mu = 10$
(SD $\sigma = 5$)
- **More $N \rightarrow$ less SD in \bar{y}_s across samples**

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

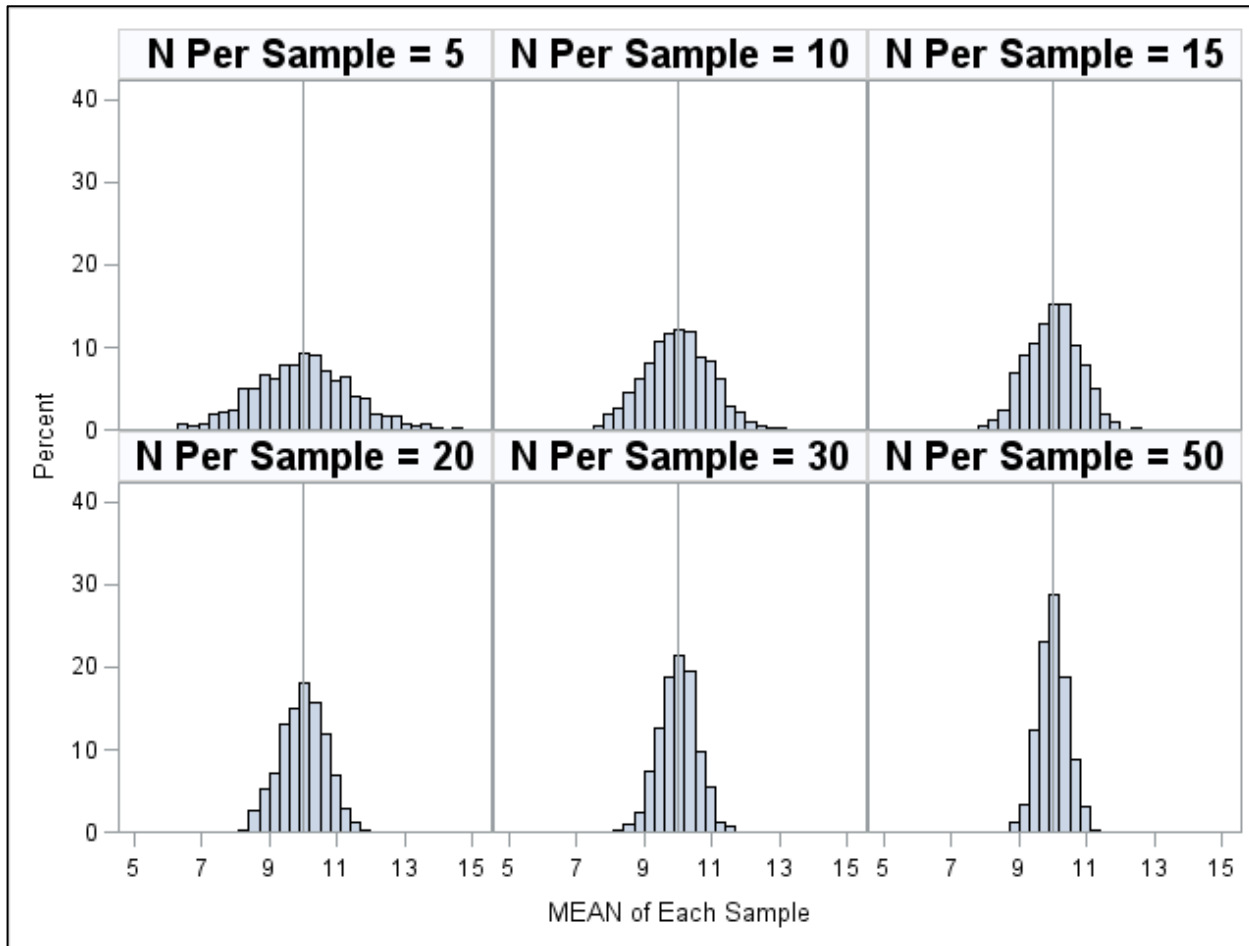
Note: These bars do not show individual people!
They are summaries for **distinct samples** of people.

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 on average**
- What else affects consistency of \bar{y}_s ? How **persons vary from each other!**



1 000 samples each for different $N \dots$

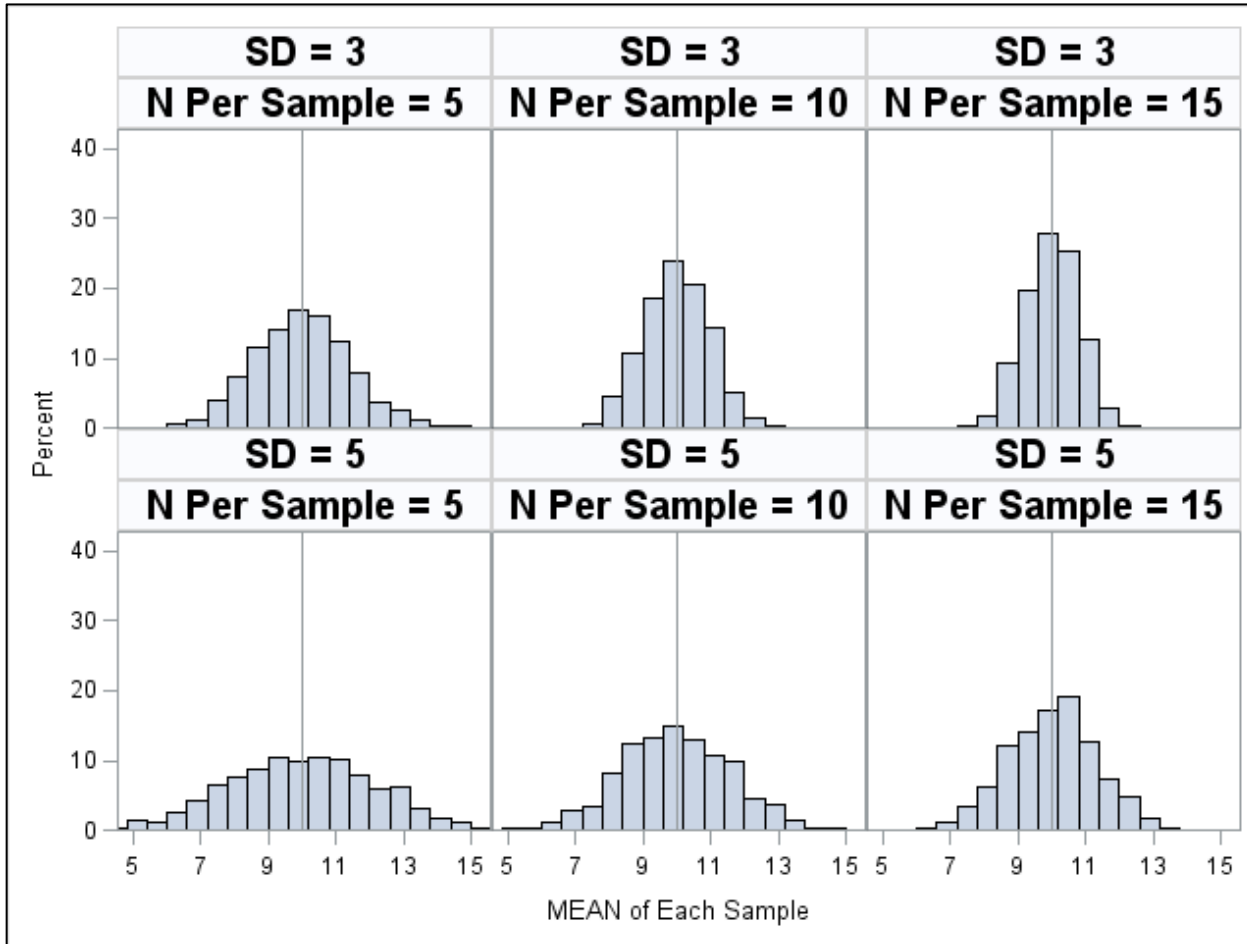


- Population values:
Mean $\mu = 10$
(SD $\sigma = 3$ now)
- **More $N \rightarrow$ less SD in \bar{y}_s across samples**

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

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

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 \rightarrow **less dispersion in \bar{y}_s** across samples

Top to bottom:

- **More SD** in each sample \rightarrow **more dispersion in \bar{y}_s** across samples
- **Dispersion \rightarrow More estimate inconsistency** across samples

Anticipating Inconsistency of Sample Mean \bar{y}_s

- In the example from the previous slides, we had a **simulated finite population** from which multiple different samples were selected
 - **Inconsistency of \bar{y}_s** could be indexed by standard deviation (*SD*) 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) = $SE = \frac{\sigma}{\sqrt{N}} = \sqrt{\frac{\sigma^2}{N}}$**
 - Note that **SE** includes the population SD σ , which must be replaced by the sample-estimated SD s when σ is unknown (i.e., most of the time)
 - **SE of the mean** is the expected average deviation of any given *sample mean* \bar{y} from the *population mean* μ (even if you do not know μ)
 - Is NOT the same as SD of y_i (s) which is the average deviation of any given *observation* (i.e., person) from the *sample mean* (that you can access)
 - 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 annual income in \$1000s in $N = 734$ using an empty model (= no predictors yet)

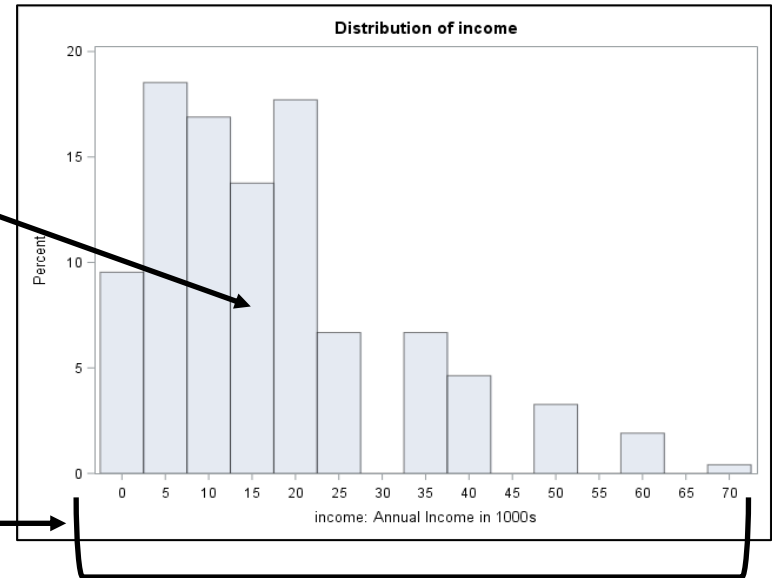
- $y_i = \beta_0 + e_i \rightarrow income_i = \beta_0 + e_i$
- Fixed intercept estimate $\beta_0 = 17.30$ (= sample \bar{y} because no predictors)
- Person-specific residual $e_i = y_i - \beta_0$
- Variance of e_i residuals $\rightarrow \sigma_e^2 = 190.21$ (= sample s^2 because no predictors)

- **SD** of $e_i \rightarrow \sigma_e = \sqrt{190.21} = 13.80$ (= sample s because no predictors)

- **SD is about the PEOPLE:** This tells us that individual income is expected to be off from the sample mean income \bar{y} by ± 13.80 on average

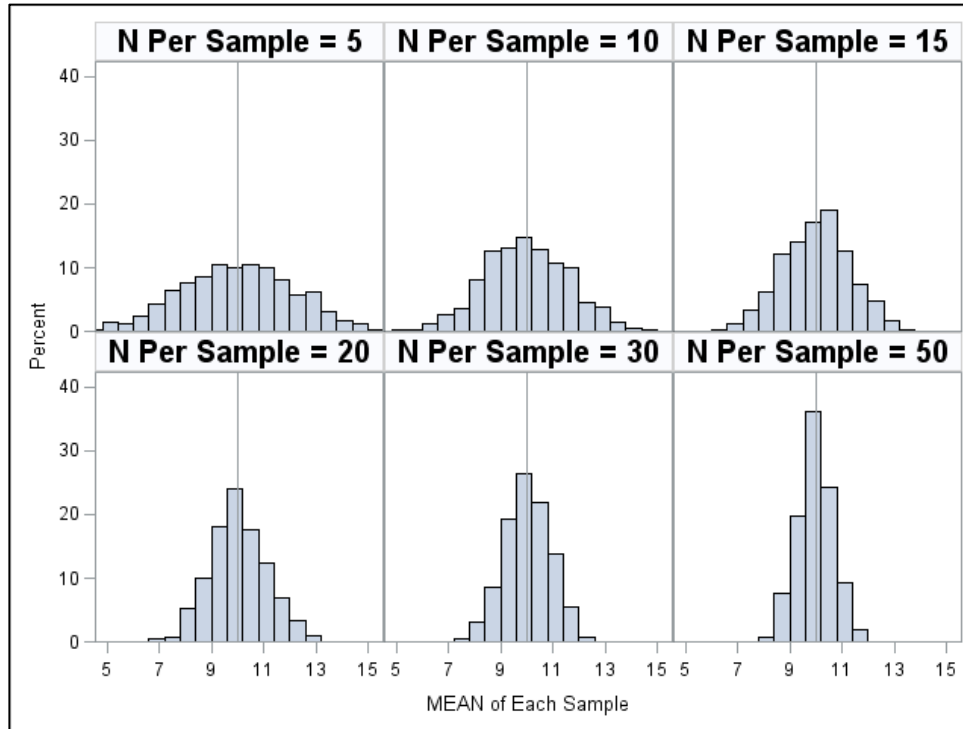
- Fixed intercept β_0 standard error: $SE = \frac{\sigma_e}{N} = \sqrt{\frac{\sigma_e^2}{N}} = 0.51$

- **SE is about the SAMPLES:** This tells us that sample mean income \bar{y} (as given by estimate for β_0 in an empty model) is expected to be off from the unknown population mean income μ by ± 0.51 on average



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

Population values for y_i variable: Mean $\mu = 10$, SD $\sigma = 5$



N	Mean \bar{y}_s	SD \bar{y}_s	Mean SE with:	
			σ	s
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

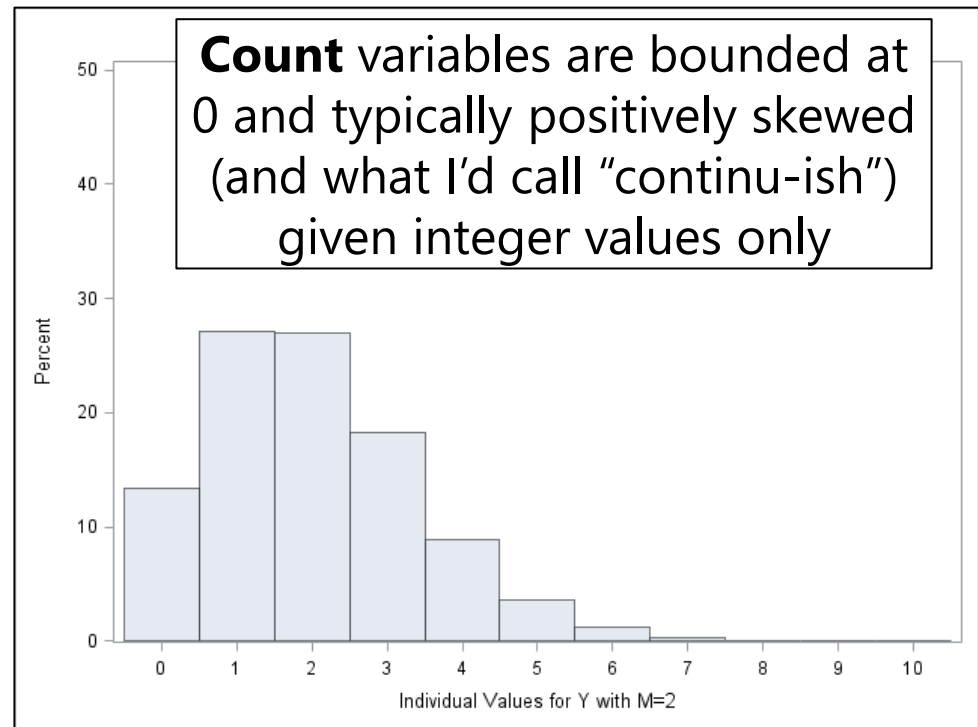
The greater the sample size N , the better the estimate of each sample's SD, and the less it matters that SE is formed with sample SD (s) instead of the population SD (σ). But this distinction will matter more in smaller samples....

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 $N \rightarrow$ more normal \bar{y}_s distribution \rightarrow "Central Limit Theorem"

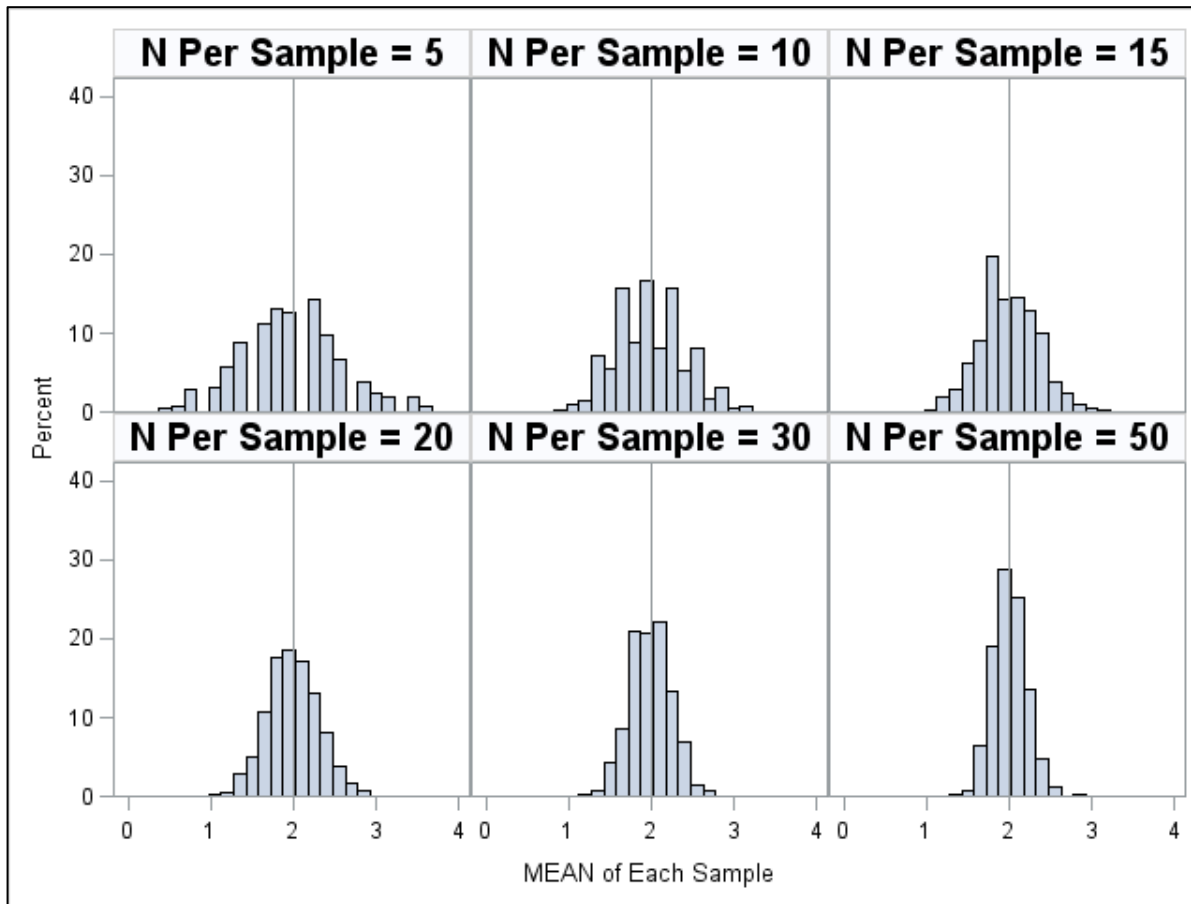
- Demo: I simulated a **count variable*** y_i in a population of 100,000 new fake people

- Population mean: $\mu = 2$
- Population VAR : $\sigma^2 = 2$
- So y_i is off the mean by $SD = \sqrt{2}$ on average



* Used a "Poisson" distribution here to generate y_i (in which $\mu = \sigma^2$)

1000 samples each for different N ... it works!



- Population values:
Mean $\mu = 2$
(VAR $\sigma^2 = 2$)
- **More $N \rightarrow$ less SD**
in \bar{y}_s across samples;
 \bar{y}_s is also more normal

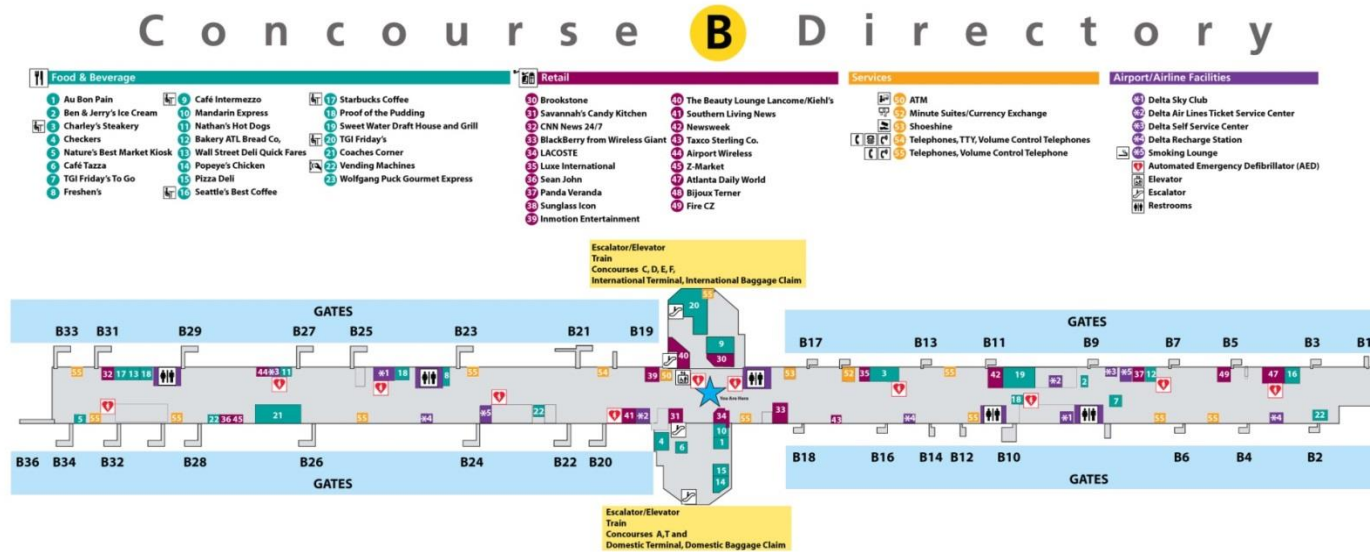
N	Mean \bar{y}_s	SD \bar{y}_s	Mean SE
5	1.98	0.61	0.59
10	1.99	0.42	0.43
15	1.99	0.35	0.36
20	1.99	0.31	0.31
30	1.99	0.25	0.25
50	2.01	0.20	0.20

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

Beyond Empty GLMs: 2 Fixed Effects

- Purpose of predictive linear models is to **customize** each person's expected outcome (beyond just the 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 move what 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?

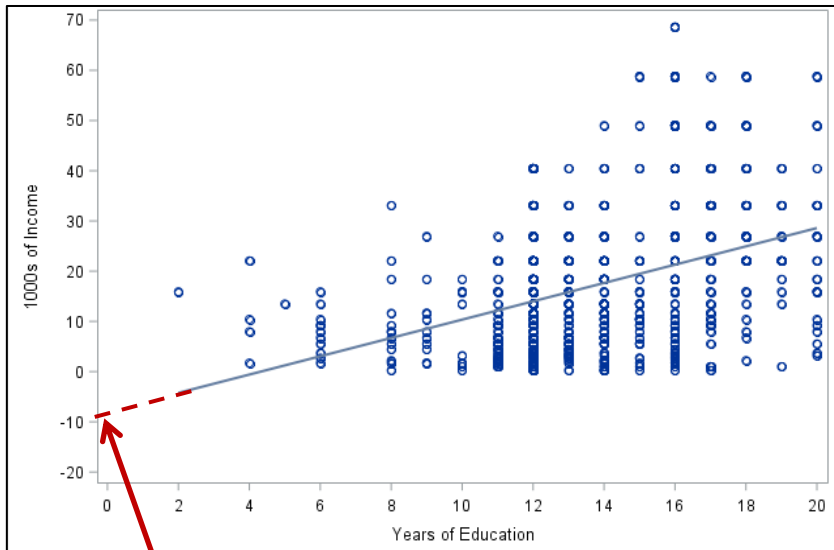
Intercept = “You are Here” Sign of Data

With **original** years of education:

$$x_i = \text{educ}, y_i = \text{income}$$

$$y_i = \beta_0 + \beta_1(x_i) + e_i$$

$$y_i = -7.89 + 1.82(x_i) + e_i$$



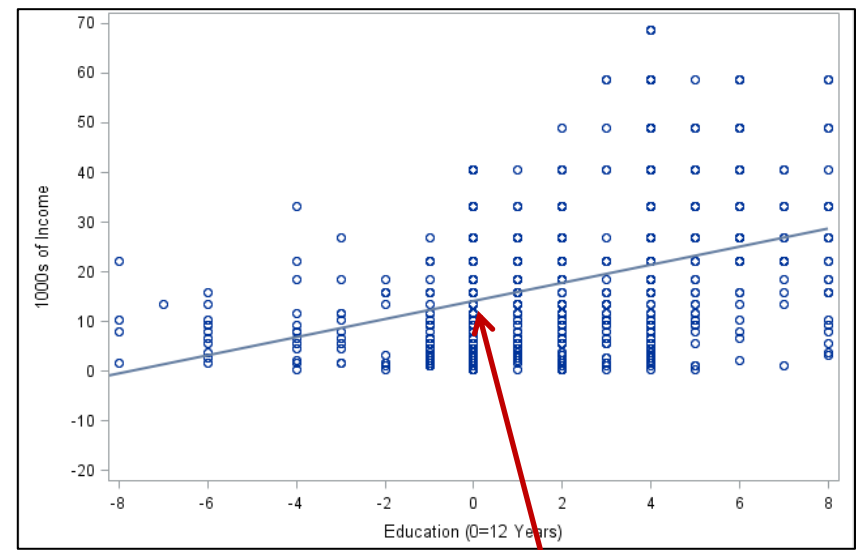
Intercept
 β_0

With **centered** education ($c = 12$):

$$x_i = \text{educ} - 12, y_i = \text{income}$$

$$y_i = \beta_0 + \beta_1(x_i) + e_i$$

$$y_i = 14.00 + 1.82(x_i) + e_i$$



Intercept
 β_0

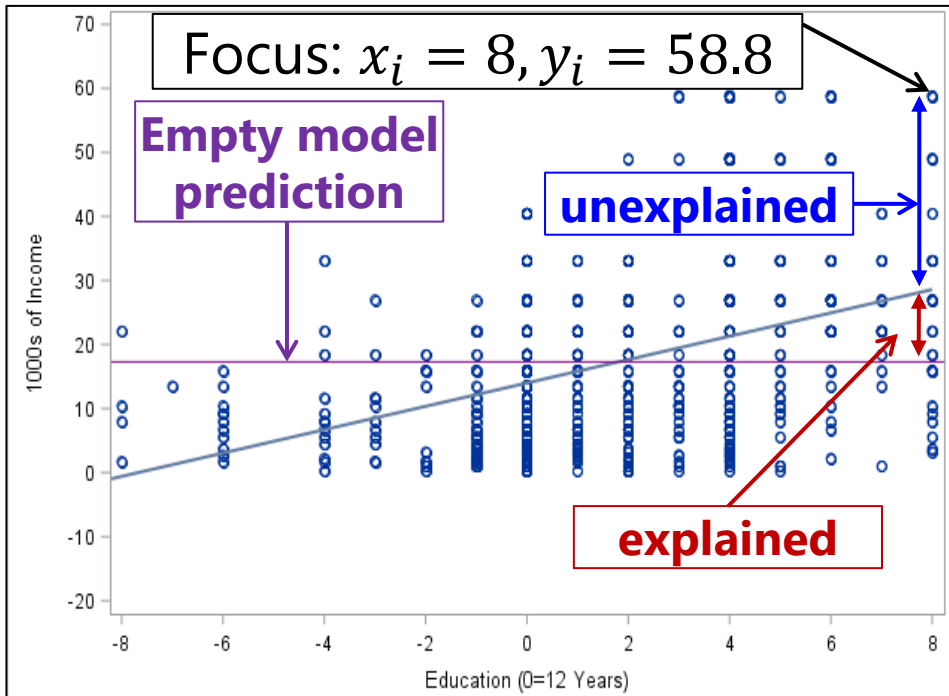
There is no *wrong* way to center, only *weird*. **Center so $x_i=0$ is meaningful.**

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 $x_i = \beta_1$
 - Total number of fixed effects = $k = 2$ here (used in later formulas)
- The y_i **created using the predictors** is called $\hat{y}_i = \text{"y hat"}$
 - $\hat{y}_i = \beta_0 + \beta_1(x_i) \rightarrow y_i = \hat{y}_i + e_i \rightarrow e_i = y_i - \hat{y}_i$
 - \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 will be **smaller than s^2 for y_i**

Visualizing GLM Residuals

Btw: β formulas result from the goal of minimizing the squared 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 $y_i = \text{income}$:

$$y_i = \beta_0 + e_i$$

$$\hat{y}_{Focus} = 17.3$$

$$y_{Focus} = 17.3 + 41.5$$

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

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

Add Education as Predictor:

$$y_i = \beta_0 + \beta_1 (\text{Educ}_i - 12) + e_i$$

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

$$y_{Focus} = 28.4 + 30.4$$

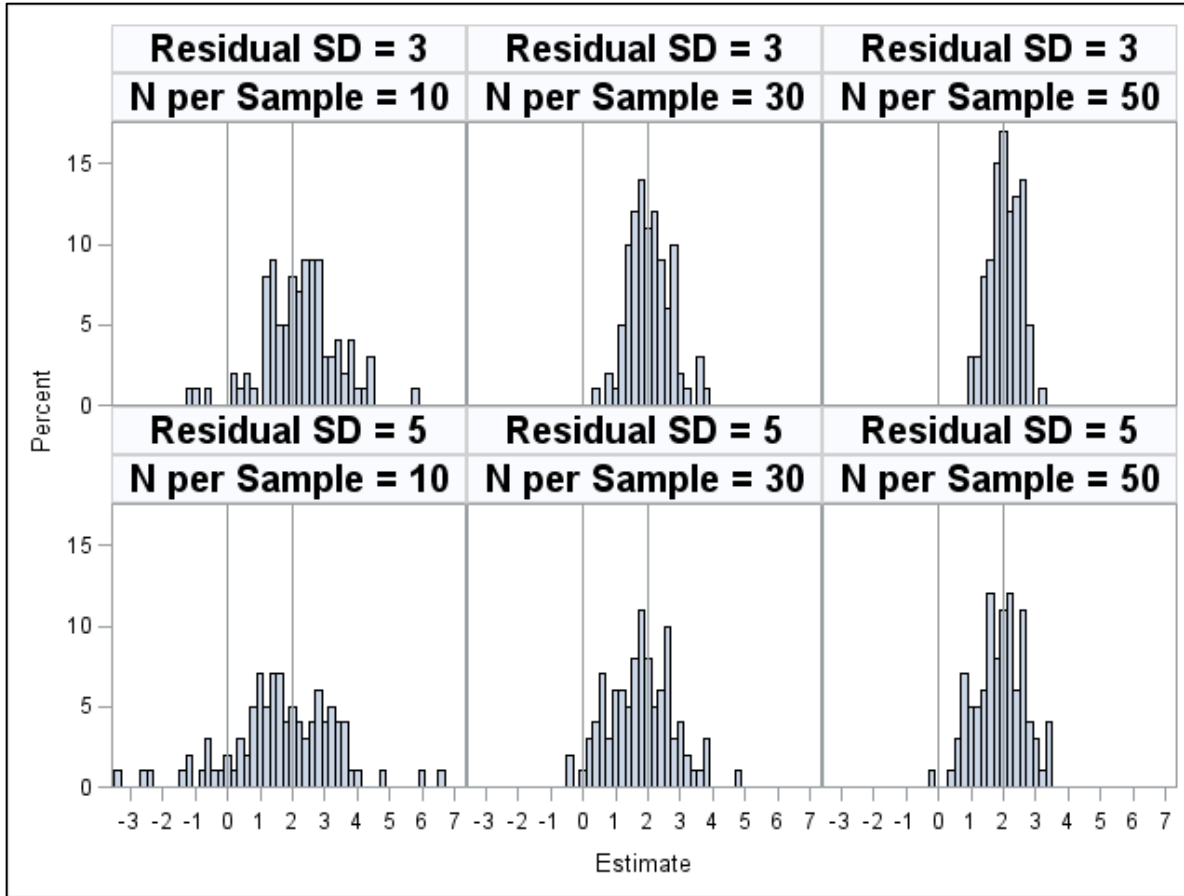
$$\text{Variance: } \sigma_e^2 = \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N-2} = 162.3$$

→ $162.3 = \text{leftover } y_i$ variance

Uncertainty of Sample Estimates

- Besides their estimates (i.e., answers using our sample that make σ_e^2 the smallest), we need to **quantify inconsistency** of both fixed effects: β_0 intercept and β_1 slope of predictor x_i
 - How much would my **sample intercept vary** across similar samples?
(just like SE of mean, but for the mean conditional on $x_i = 0$)
 - How much would my **sample slope vary** across similar samples?
(same idea in concept, but also taking into account the scale of x_i)
- Demo: I made my own (normally-distributed) quantitative variables x_i and y_i in a population of 100,000 fake people
 - **Predictor** x_i has mean = 0 and variance = 1
 - **Outcome** $y_i = \beta_0 + \beta_1(x_i) + e_i \rightarrow \hat{y}_i = 100 + 2(x_i)$
 - Residual variance is either $\sigma_e^2 = 9$ or $\sigma_e^2 = 25$
 - Let's see what happens to β_1 over 100 samples of varying N and σ_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:

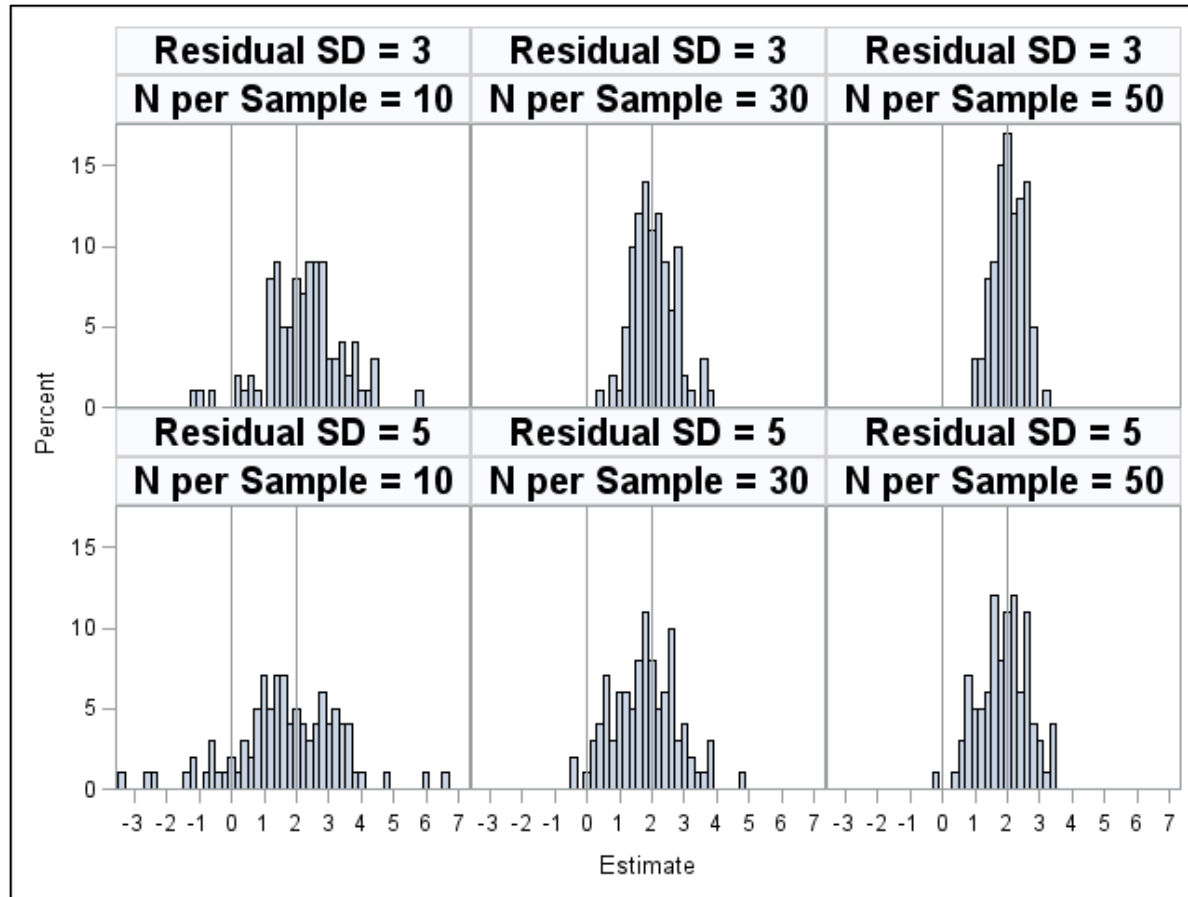
- **More N** in each sample \rightarrow **less dispersion in β_1** across samples

Top to bottom:

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

- **Dispersion \rightarrow Estimate inconsistency** across samples

Effects of N and SD on Dispersion of β_1



- Population values:
Slope $\beta_1 = 2$
- **More $N \rightarrow$ 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

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

Standard Errors in a One-Predictor GLM

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

- **Standard Error (SE)** for a slope estimate β_1 in a single-predictor GLM:

$$SE_{\beta_1} = \sqrt{\frac{\text{residual variance of Y}}{\text{variance of } x_i * (N - k)}} = \sqrt{\frac{\sigma_e^2}{s_x^2 * (N - k)}}$$

- Slope β_1 SE = $\sqrt{\frac{162.28}{8.46 * (734 - 2)}} = 0.16$

→ Sample slope is expected to be off from (unknown) population slope by ± 0.16 on average

- **SE of any predicted outcome \hat{y}_i** (including the outcome captured by β_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 $\hat{y}_i | x_i = \sqrt{\sigma_e^2} * \sqrt{\frac{1}{N} + \frac{(x_i - \bar{x})^2}{(N-1)s_x^2}} \rightarrow 0.55$ at $Educ_i = 12$ specifically

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
 - **SE**: Mean distance of sample (estimate) from population (true unknown) (**cough cough*, repeated homework question alert, *cough cough**)
 - Btw, **SE²** is known as “**sampling variance**” (but is not typically reported) —not to be confused with **SD²**, which is “**sample variance**” (ugh, I know)
- Moving forward: We can use an **SE to assess how far away** our sample estimate is **from a hypothesized** population value
 - By sampling only some persons, we expect some fluctuation in the sample-specific estimates across multiple similar samples, but how different 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"** (H_0) is a statement about the population parameter being equal to some stated (expected) value
 - e.g., for our GLM example education slope $\beta_1 = 1.82$, $H_0: \beta_1 = 0$
 - H_0 for a slope is usually 0, but it doesn't have to be!
 - Btw, you can also test \bar{y} against H_0 about $\mu \rightarrow$ "one-sample t -test"
- An **"alternative hypothesis"** (H_A) is a statement that contradicts the null hypothesis and **conveys allowed directionality of deviation** from stated H_0 value
 - **"One-tailed test"** is one-directional: $H_A: \beta_1 > 0$ OR $H_A: \beta_1 < 0$
 - **"Two-tailed test"** is "different than": $H_A: \beta_1 \neq 0$ (aka, $H_A: \beta_1 = !0$)
- Then how **far away is our sample estimate from H_0** ? We standardize this distance using its SE as a proxy for SD of its sampling distribution
 - 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"** = $\frac{Est - H_0}{SE} = \frac{\beta_1 - 0}{SE} = \frac{1.82 - 0}{0.16} = 11.28$

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
- 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 describe 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**” (and has a [great story](#))

Area Under Standard Normal Curve

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

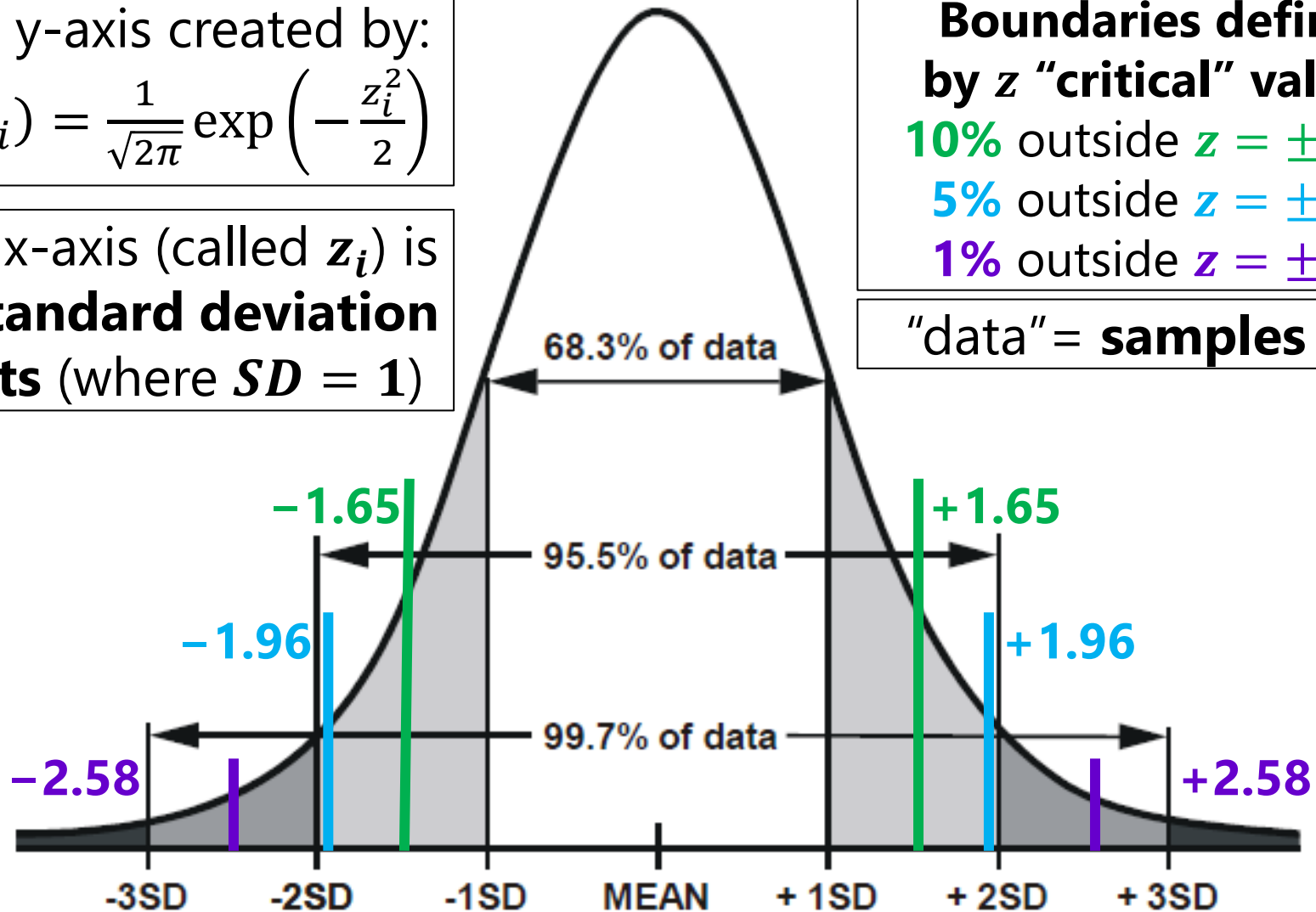


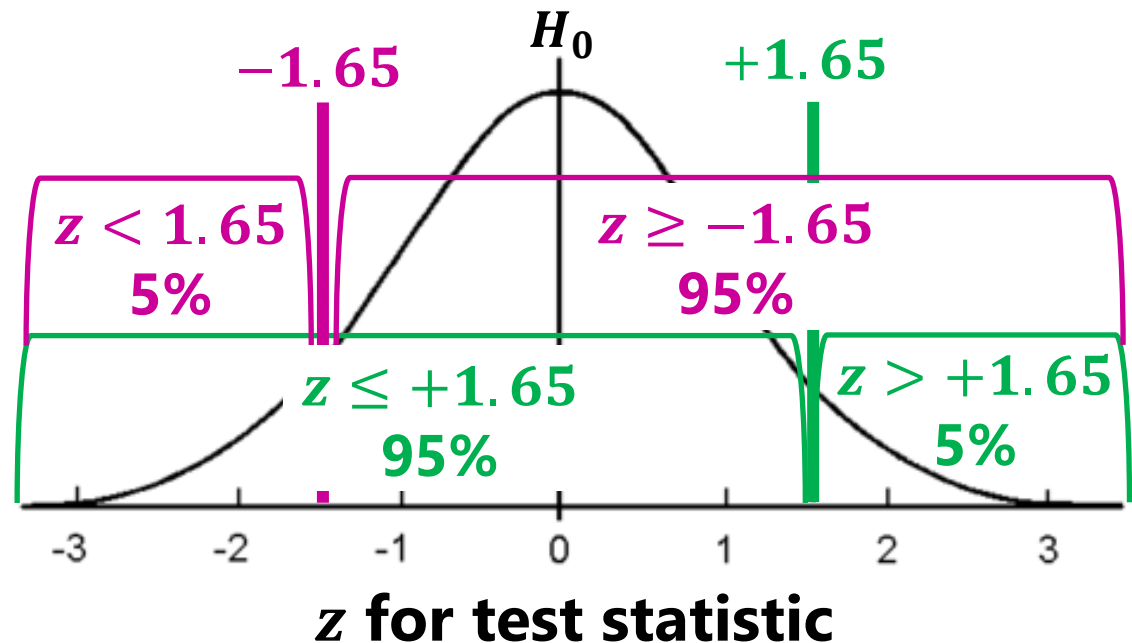
Image adapted from: <http://my.ilstu.edu/~gjin/hsc204-hed/Module-5-Summary-Measure-2/Module-5-Summary-Measure-28.html>

Denoting Expected vs. Unexpected

- “How unexpected” a test statistic is (e.g., **11.28** for $\beta_1 = 1.82$, $SE = 0.16$ here) requires **two more decisions made** ahead of time
- **How often** is “unexpected”? This is known as “**alpha level**”
 - Common is alpha = .05 (or .01 for conservative or .10 for lenient)
- **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 then create “**critical values**” for your PDF, which then **define boundaries** as to **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 **95%** “**expected**”) → two possible versions of one-tailed H_A :
- $H_A: \beta_1 > 0 \rightarrow$
 $Z_{critical} = +1.65$
 - Tests if β_1 is bigger or not bigger than H_0
 - If β_1 is actually smaller, conclude “not bigger”
- $H_A: \beta_1 < 0 \rightarrow$
 $Z_{critical} = -1.65$
 - Tests if β_1 is smaller or not smaller than H_0
 - If β_1 is actually bigger, conclude “not smaller”



Two-Tailed Test: \pm 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

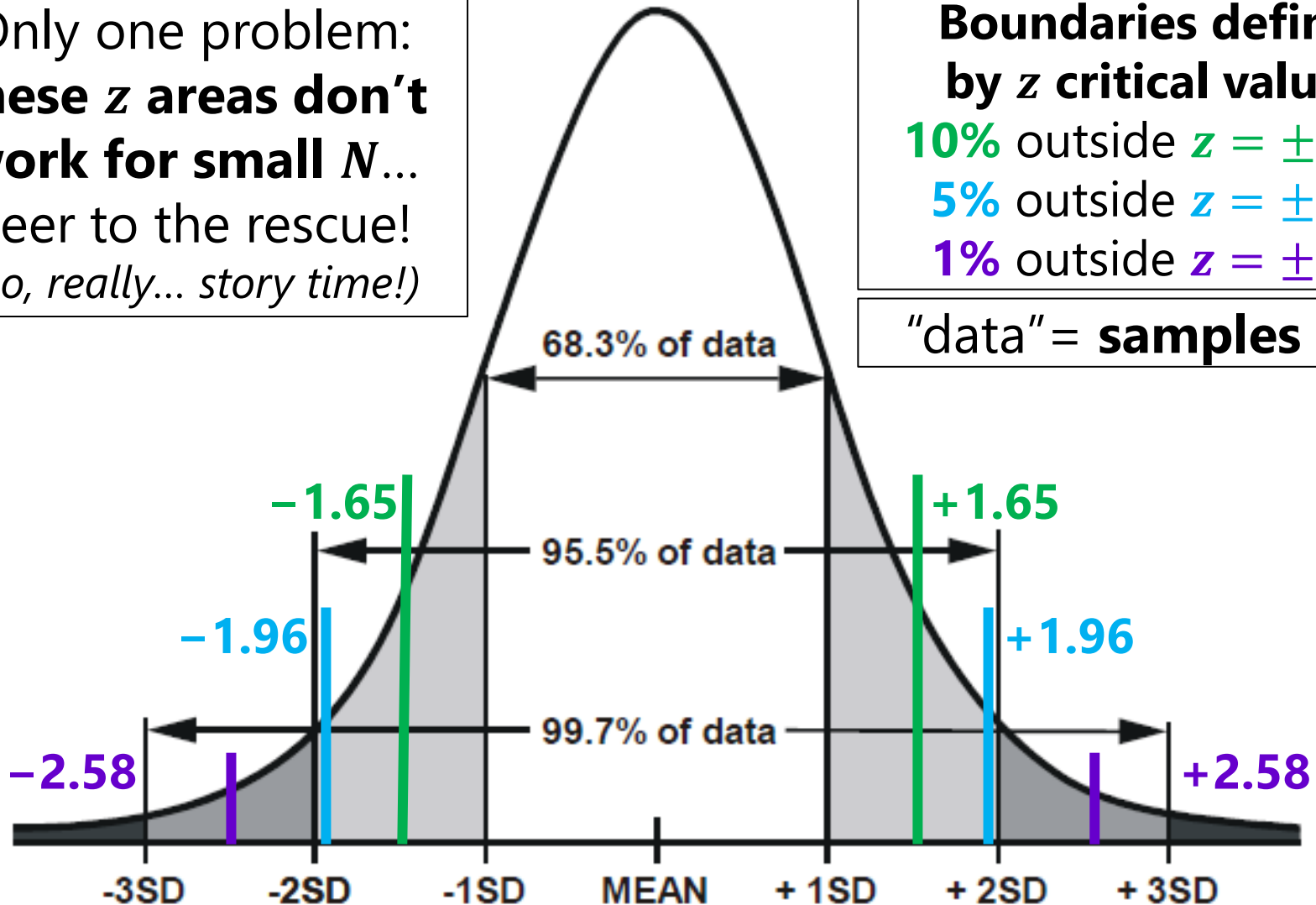
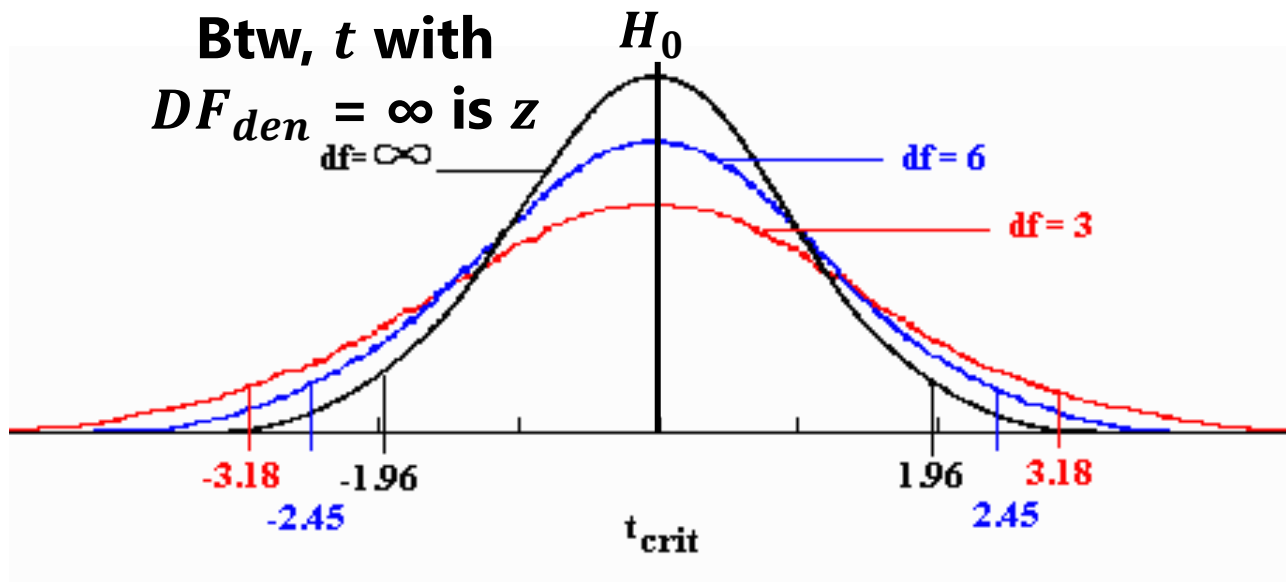


Image adapted from: <http://my.ilstu.edu/~gjin/hsc204-hed/Module-5-Summary-Measure-2/Module-5-Summary-Measure-28.html>

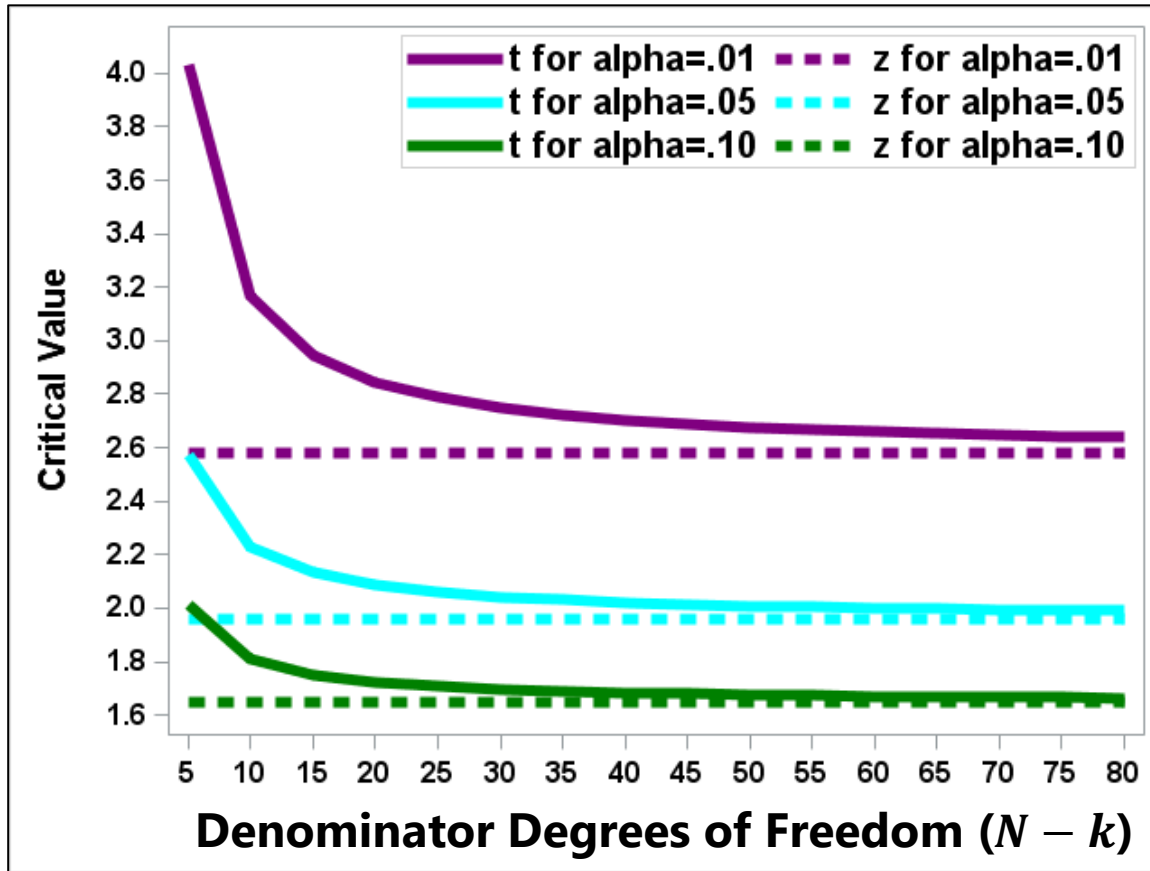
Meet Student's t Distribution

- 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 $DF_{den} \rightarrow$ reject $H_0 \rightarrow$ "**significant**" result



- $t_{critical}$ values for **alpha = .05** by DF_{den} shown here
- **With smaller N ,** have to go farther out to **get to 5%**

Critical Values for t versus z Distributions



With smaller $N - k$ (fewer DF_{den}), more extreme **test-statistics are needed when **using t** to say β_1 is “unexpectedly different” from H_0 (i.e., to cross the alpha-based boundary to be “**significant**”)**

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

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

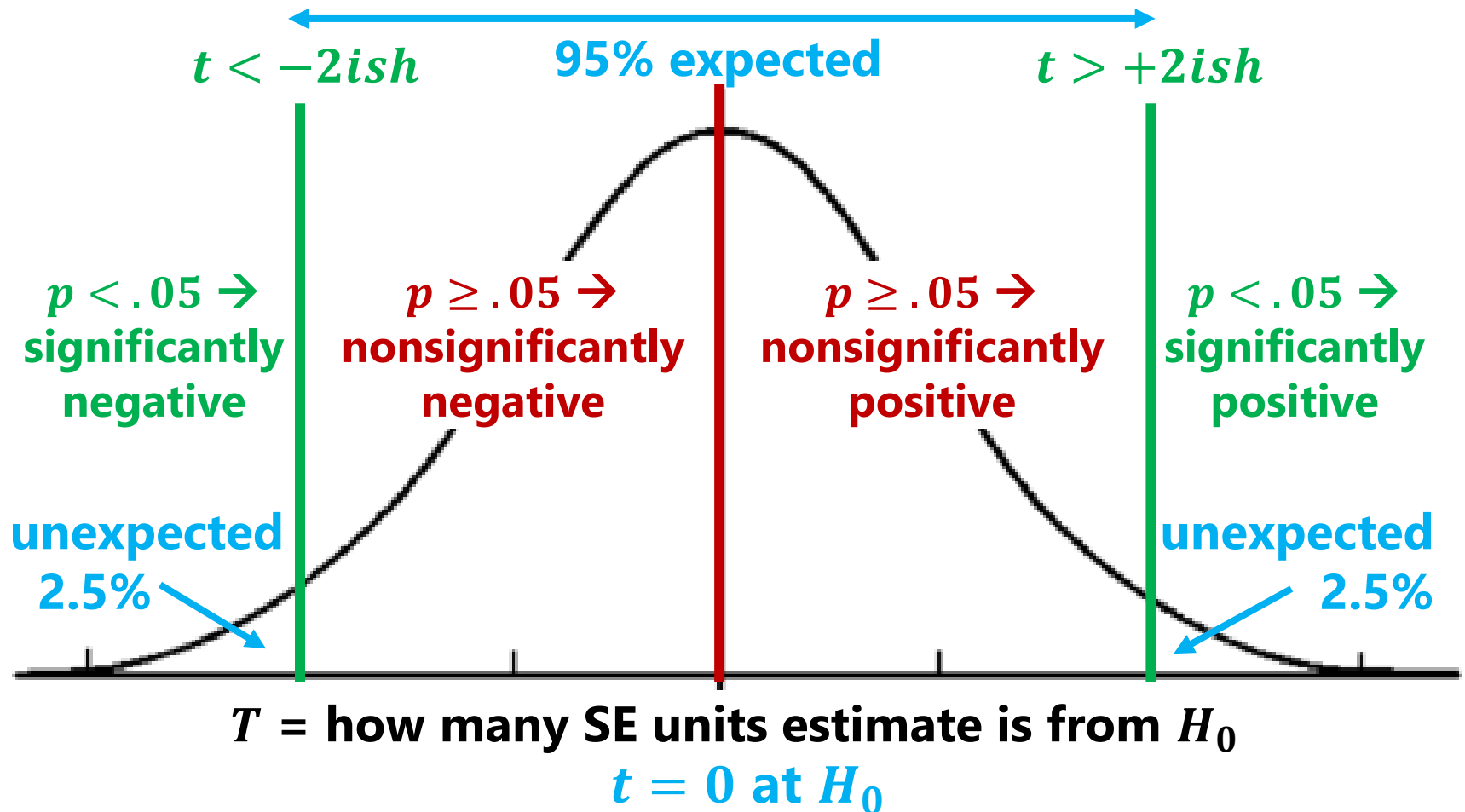


Image adapted from: <http://mathcenter.oxford.emory.edu/site/math117/normalDistribution/>

From Alpha to Confidence Intervals

- Critical values from a t -distribution can also be used to compute a “**confidence interval**” (CI) for your estimate
- **CI** = interval predicted to **contain the population value in (1 – alpha)%** of repeated samples (confidence = expected)
 - Conservative: $\alpha = .01 \rightarrow 1\%$ unexpected $\rightarrow 99\%$ CI
 - Common: $\alpha = .05 \rightarrow 5\%$ unexpected $\rightarrow 95\%$ CI
 - Lenient: $\alpha = .10 \rightarrow 10\%$ unexpected $\rightarrow 90\%$ CI
 - If **CI does not cross H_0** , then your result is “**significant**” at that alpha
- Confidence Interval: $CI = Estimate \pm (critical * SE)$
- e.g., for $\beta_1 = 1.82$, $SE = 0.16$, $DF_{den} = N - k = 732 \rightarrow z$ critical
 - **90%** CI for β_1 : $CI = 1.82 \pm (1.65 * 0.16) = 1.56 \text{ to } 2.08$
 - **95%** CI for β_1 : $CI = 1.82 \pm (1.96 * 0.16) = 1.51 \text{ to } 2.13$
 - **99%** CI for β_1 : $CI = 1.82 \pm (2.58 * 0.16) = 1.41 \text{ to } 2.23$

For reporting CI:
“lower bound” to
“upper bound”

Significance Tests: What's in Your Output

- Each **β fixed slope** has 6 relevant characteristics (*essential to report):
 - ***Estimate** = best guess for the fixed slope from our sample data
 - ***Standard Error** = SE = average distance of sample slope from population slope
→ expected *inconsistency* of slope across samples
 - **"t-value"** = $(\text{Estimate} - H_0) / SE = \text{test-statistic}$ for fixed slope against $H_0 (= 0)$
 - **Denominator DF** = $DF_{den} = N - k$ (where k = total number of fixed effects)
 - **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** = $CI = \text{Estimate} \pm t_{critical} * SE$ = range in which true (population) value of estimate is expected to fall across (95%) of samples
- Compare 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 (or are large enough), then t is treated as 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 β_1 in a single-predictor GLM:
$$SE_{\beta_1} = \sqrt{\frac{\text{residual variance of Y}}{\text{variance of } x_i * (N - k)}} = \sqrt{\frac{\sigma_e^2}{s_x^2 * (N - k)}}$$
- Example: $Income_i = \beta_0 + \beta_1(Educ_i - 12) + e_i$, $\sigma_e^2 = 162.28$,
 $N = 734$, $x_i = Educ_i - 12$: $M = 1.81$, $Var = 8.46$
 - Slope for education predictor: $H_0: \beta_1 = 0$, $H_A: \beta_1 \neq 0$
 $Est = 1.82$, $SE = \sqrt{\frac{162.28}{8.46 * (734 - 2)}} = 0.16$, $t = \frac{Est - 0}{SE} = \frac{1.82 - 0}{0.16} = 11.28$,
 $DF_{denominator} = N - k = 734 - 2 = 732$, so $p < .0001$,
 $95\% CI = Est \pm (t_{crit} * SE) = 1.82 \pm (1.96 * 0.16) = 1.51 \text{ to } 2.14$
 - Interpretation: Predicted income is **significantly higher** by 1.82k for each additional year of education (so reject H_0 that $\beta_1 = 0$)

SEs and CIs for Predicted Income

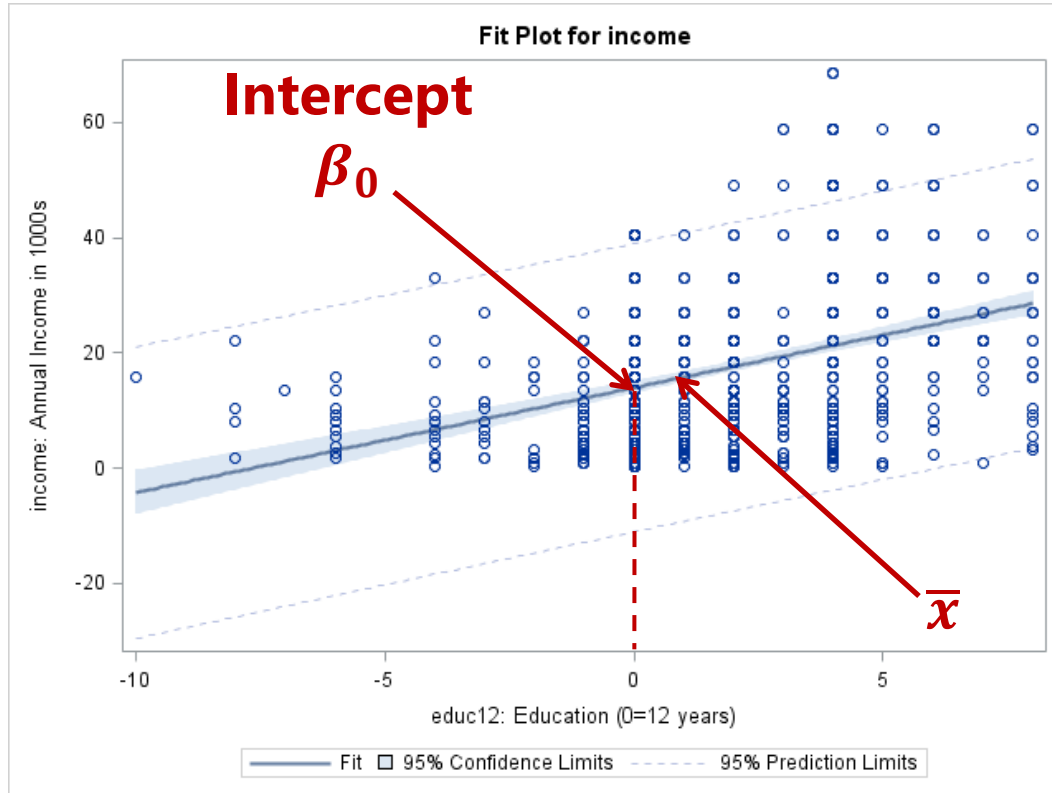
- **SE of any predicted outcome \hat{y}_i** (including the outcome captured by β_0 for $x_i = 0$) depends on the value of the predictor—the SE will increase as you move away from the predictor's mean:

$$\text{SE of } \hat{y}_i \mid x_i = \sqrt{\sigma_e^2} * \sqrt{\frac{1}{N} + \frac{(x_i - \bar{x})^2}{(N-1)s_x^2}}$$

SE (for β_0 or any \hat{y}_i) = average distance of sample predicted value from population value

- **$Income_i = \beta_0 + \beta_1(Educ_i - 12) + e_i$, $\sigma_e^2 = 162.28$,
 $N = 734, Educ_i - 12: M = 1.81, Var = 8.46$**
- SE and CI for predicted income when Education = 12?
 - Given by β_0 : $Est = 14.00$, $SE = \sqrt{162.28} * \sqrt{\frac{1}{734} + \frac{(0-1.81)^2}{(733)8.46}} = 0.55$,
 $95\% CI = Est \pm (t_{crit} * SE) = 14.00 \pm (1.96 * 0.55) = 12.91 \text{ to } 15.09$
- You can use SAS ESTIMATE, STATA LINCOM, or R GLHT to get predicted outcomes for any value of the model predictors...
 - Each has options to get predicted outcomes for each person in dataset!

Two CIs for Predicted Outcomes



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

- The **blue shading** shows the 95% range for the \hat{y}_i **outcomes from the fixed effects** (i.e., regression line)
 - They are narrowest at the predictor mean, and widen as moving away
- The **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 predictive equation uses the original variables as entered directly into the model—this is the “**unstandardized**” solution
- e.g., For our GLM with $y_i = \beta_0 + \beta_1(x_i) + e_i$
 - x_i is *Educ*_{*i*} − 12: $M = 1.81, Var = 8.46$
 - y_i is *Income*: $M = 17.30, Var = 190.21$
- **Unstandardized:** $Income_i = 14.00 + 1.82(Educ_i - 12) + e_i$
 - Unstandardized fixed slopes (β_{unstd}) can be standardized (β_{std}) as:

$\beta_{std} = \beta_{unstd} * \frac{SD_x}{SD_y}$	$std \beta_0$ will always be 0!
---	---------------------------------
- **Standardized (Std):** $y_i = 0 + 0.38(x_i) + e_i$
 - Standardized solution refers to variables that have been z-transformed into $M = 0, SD = 1 \rightarrow x_{std} = (x_i - \bar{x})/s_x$ and $y_{std} = (y_i - \bar{y})/s_y$
 - **For one predictor, $\beta_{std} = \text{Pearson correlation}$** (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**, x and y
 - To graph their relationship, we can request a **scatterplot**, in which values for x are shown on the x -axis and values for y are shown on the y -axis
 - Correspondence between x and y values will be captured by a general effect size called "**correlation**" (r); one specific type for *quantitative* variables is **Pearson**
 - Btw, Pearson's r for two binary variables is re-named "**phi**" r
 - Btw, Pearson's r for a binary and a quantitative variable is re-named "**point-biserial**" r
 - Correlations range continuously from **-1 to 1** (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 ($r = -1$):

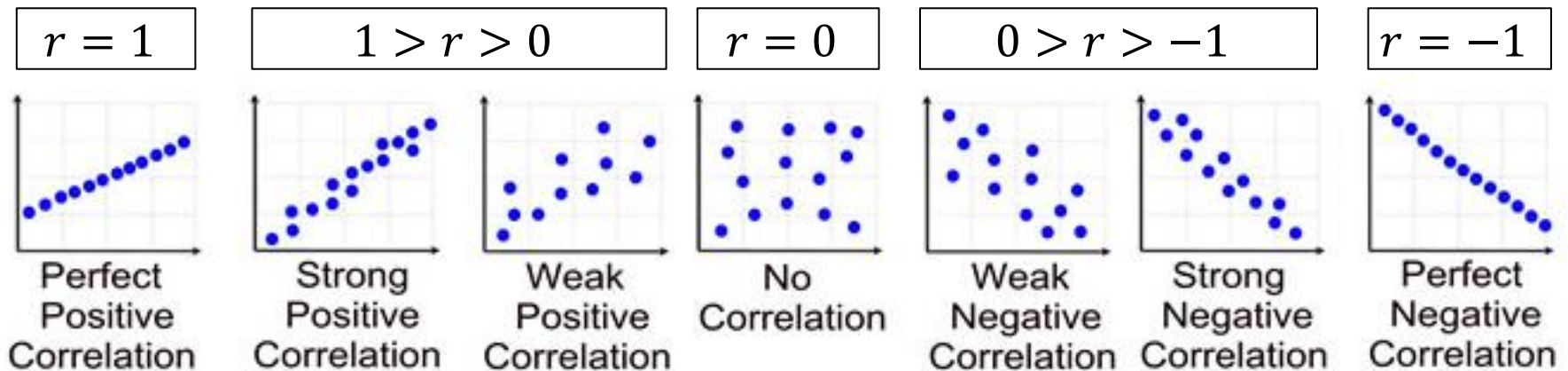


Image borrowed from: <https://mathbitsnotebook.com/AlgebraI/StatisticsReg/ST2CorrelationCoefficients.html>

What about Categorical Predictors?

- We just saw how a Pearson's r between two quantitative variables x_i and y_i can be represented equivalently with a general linear model (GLM) of x_i predicting y_i : $y_i = \beta_0 + \beta_1(x_i) + e_i$
 - Fixed slope β_1 captures a linear effect of x_i predicting y_i in an unstandardized metric (using x_i centered so intercept at 0 makes sense)
 - For how to model *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 C categories need C fixed effects** to distinguish the outcome means across all unique categories
 - After including the intercept β_0 , we still need $C - 1$ predictors, whose β_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** x_i predicting y_i : $y_i = \beta_0 + \beta_1(x_i) + e_i$
 - Create x_i so 0 = reference category, 1 = alternative category
 - Btw, this is also called an “**Independent** (or **two-sample**) **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
 - $marrygroup_i$: 0 = no, 1 = yes \rightarrow $Income_i = \beta_0 + \beta_1(Marry01_i) + e_i$
 - β_0 = **intercept** = expected income for unmarried persons
($Marry01_i = 0$)
 - β_1 = **slope** for $Marry01_i$ = expected mean difference for married persons **relative to unmarried persons**
 - e_i = **residual** = difference in model-predicted income (from \hat{y}_i) and actual income y_i , whose (residual) variance is estimated as σ_e^2

A GLM with a Binary Predictor

$$\text{Income}_i = \beta_0 + \beta_1(\text{Marry01}_i) + e_i \quad (\text{btw, } r = .23, p < .0001)$$

$$\text{Slope result: } \beta_1 = 6.22, SE = 0.996 \rightarrow t = 6.25, p < .0001$$

- Income predicted for unmarried:

$$\hat{y}_i = 14.45 + 6.22(0) = 14.45$$

- Income residual for unmarried:

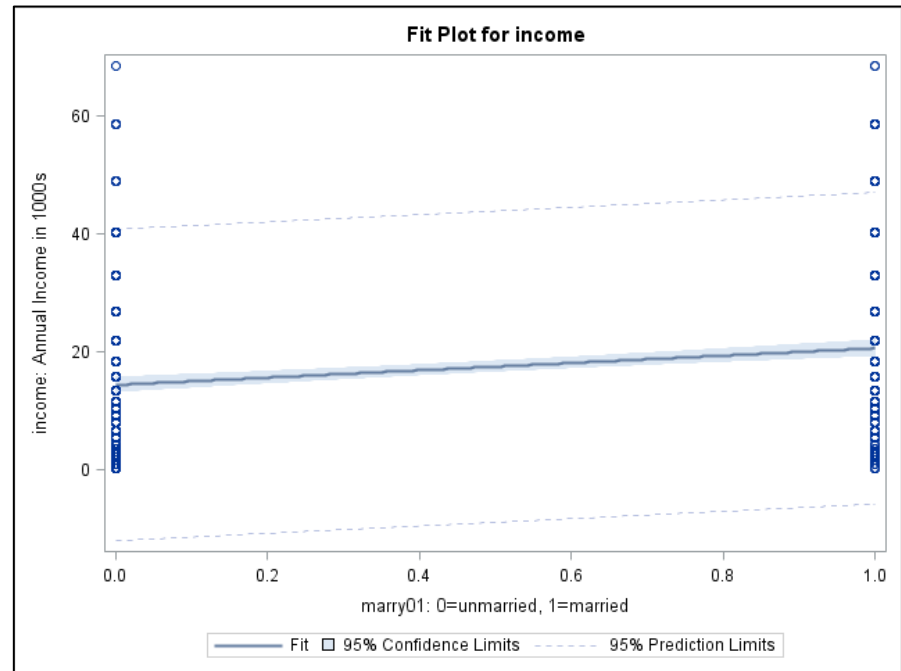
$$e_i = y_i - \hat{y}_i \rightarrow e_i = y_i - 14.45$$

- Predicted income for married:

$$\hat{y}_i = 14.45 + 6.22(1) = 20.67$$

- Income residual for married:

$$e_i = y_i - \hat{y}_i \rightarrow e_i = y_i - 20.67$$



- 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 the only kind possible for binary predictors (there is only one possible "unit difference" in a binary x_i from 0 to 1)

Effect Size for a Mean Difference: d

- For categorical predictors, an r effect size is less intuitive than a **Cohen's d** effect size, a **standardized mean difference** between two groups (labeled 0 and 1 here)

- 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 N in each group

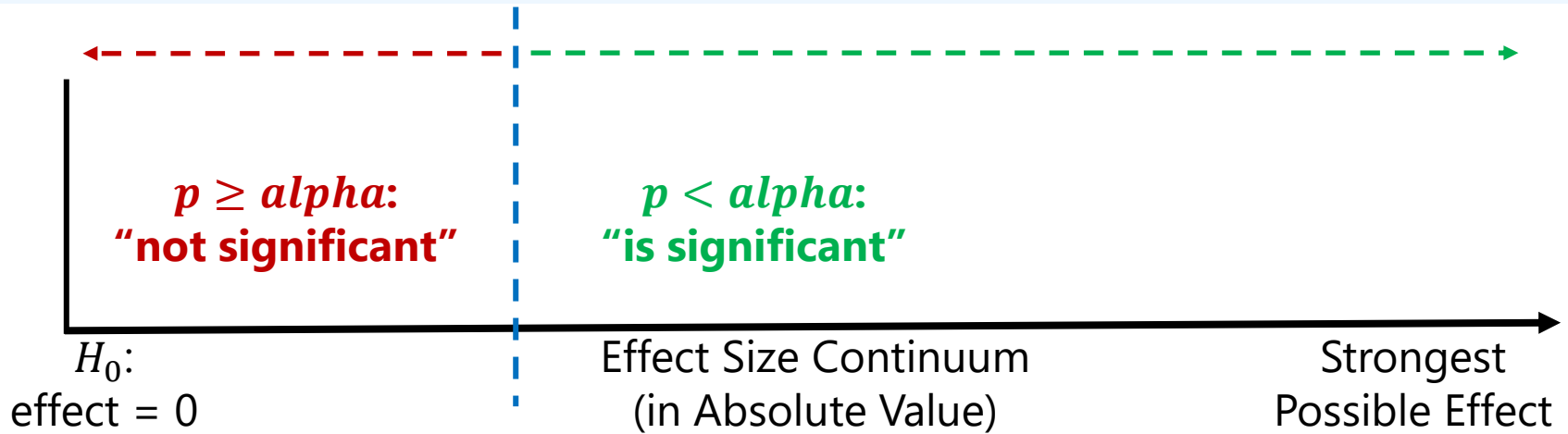
- If your GLM contains only one binary predictor, then the pooled SD is the same as the square root of GLM residual variance, $\sqrt{\sigma_e^2}$

- Otherwise, $\sqrt{\sigma_e^2}$ will be smaller because of the other predictors (*stay tuned for more about effect sizes given multiple predictors*)

- d or r can be found using the t test-statistic for a fixed effect:

- $d = \frac{2t}{\sqrt{DF_{den}}}$, $r = \frac{t}{\sqrt{t^2 + DF_{den}}}$, $d = \sqrt{\frac{4r^2}{1-r^2}}$, $r = \sqrt{\frac{d^2}{4+d^2}}$

Effect Size, Sample Size, and Test Statistics



- Role of **test statistics** (t when using denominator DF; z if not) is to **standardize** an estimate's deviation from the null hypothesis
 - When compared to reference distribution, they give you a p -value: probability of finding an effect \geq 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** \rightarrow Fewer people = **less power** (as discussed next!)

Decision Errors in Hypothesis Testing

- Usually, we test a null hypothesis against a two-sided alternative:
 - Typical null H_0 : effect (i.e., slope) = 0; alternative H_A : effect $\neq 0$
- **2 chances to get it right and 2 chances to get it wrong**, governed by:
 - **Alpha** (α) = expected percentage of **Type I errors** for a given H_0
 - Higher alpha \rightarrow less extreme boundaries for “significant” \rightarrow more Type I errors
 - **Beta** (β) = expected percentage of **Type II errors** for a given effect size
 - Usually expressed as $1 - \beta = \mathbf{Power}$: 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	<u>Correct:</u> Really NO Effect	<u>Miss:</u> Type II Error
Decision: Reject H_0	<u>False Alarm:</u> Type I Error	<u>Correct:</u> Really IS an Effect

Decision Errors in Hypothesis Testing

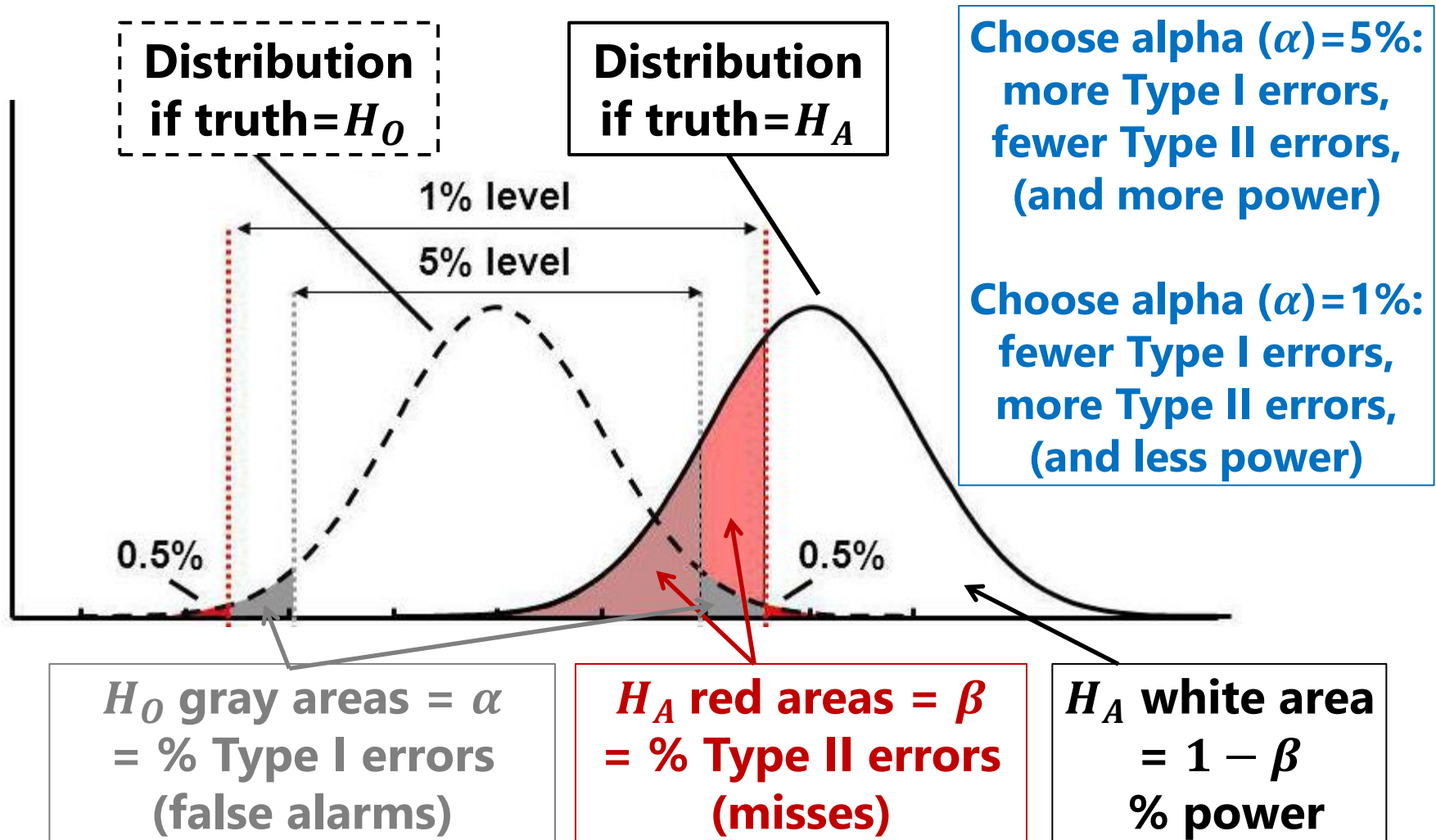
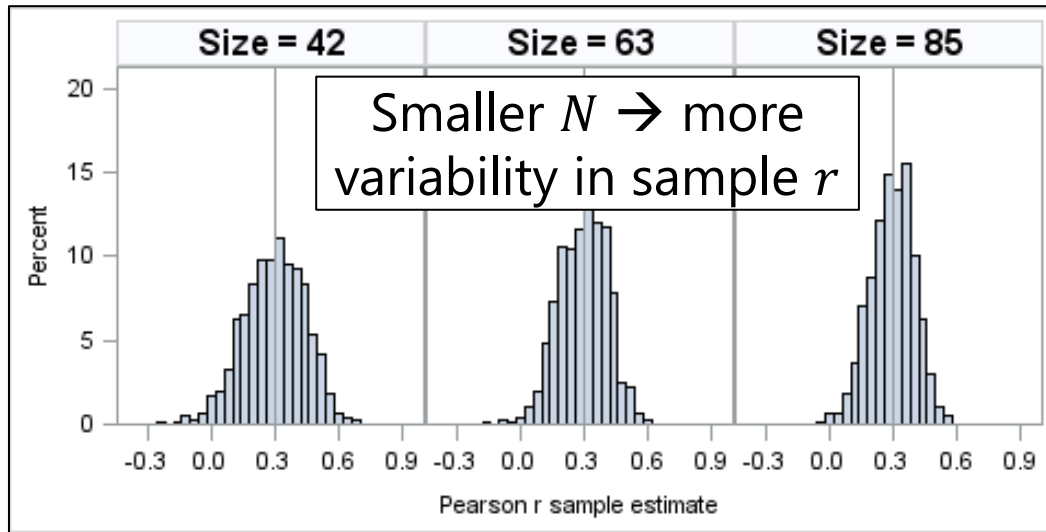
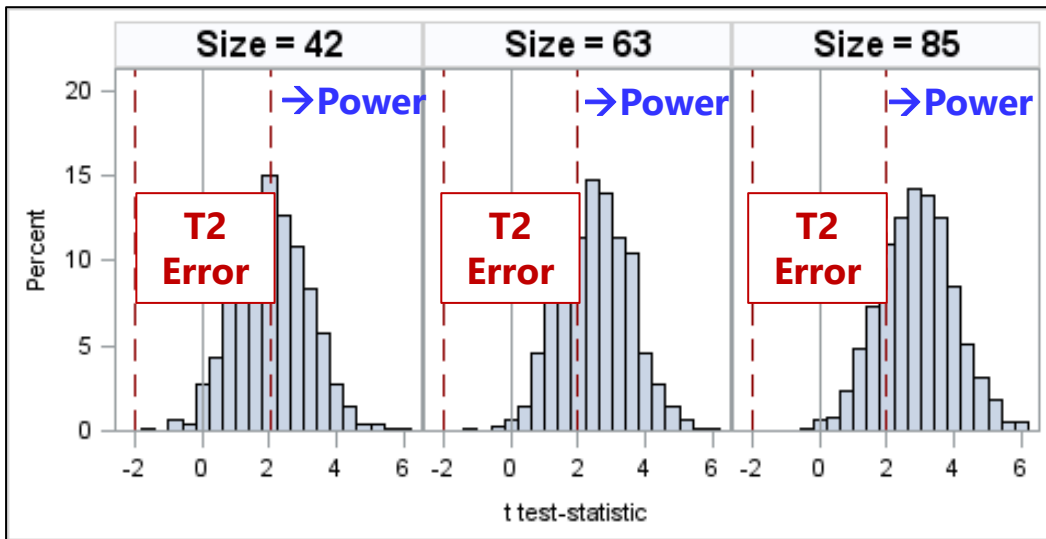


Image borrowed from: <https://images.app.goo.gl/eDuhatsiyKWjrUvcA>

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 (T2) Error: % not significant	Statistical Power: % significant
42	50%	50%
63	37%	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

- 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 $N = 194$
- 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 $r \geq .4$

Decisions and Decision Errors: Summary

- Given alpha (% unexpected), about the $(Est - H_0)/SE = t\text{-value}$:
- If ***t*-value falls outside** *t*-critical boundaries, then $p < \alpha$:
Result is sufficiently **unexpected** \rightarrow reject $H_0 \rightarrow$ "**significant**":
 - DO have to worry about a false alarm (Type I error \leftarrow 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 $p \geq \alpha$:
Result is sufficiently **expected** \rightarrow retain $H_0 \rightarrow$ "**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., form as outcome = constant*predictor + constant*predictor...) create expected outcomes from 1+ predictors
 - **General** linear models use a **normal** conditional distribution (***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)
 - $y_i = \beta_0 + e_i \rightarrow \beta_0 = \bar{y}$, variance of e_i residuals = $\sigma_e^2 \rightarrow$ all the y_i variance
 - **Single Predictor Model**: used to customize expected outcomes using a single predictor $\rightarrow y_i = \beta_0 + \beta_1(x_i - c) + e_i$ (C is centering constant)
 - $\beta_0 = \text{intercept}$ = expected y_i when $x_i = 0$
 - $\beta_1 = \text{slope}$ of x_i = difference in y_i per one-unit difference in x_i
 - $e_i = \text{residual}$ = deviation between actual y_i and predicted y_i (= \hat{y}_i)
 - x_i should have a meaningful 0 value \leftarrow center by subtracting constant c

Foreshadowing... please stay tuned!

- In a GLM with a **single predictor** (quantitative or binary), the effect size given by its **standardized slope** will 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 (coming in lecture 4)
 - Moderation of predictor effects (via interaction terms)
 - To test if predictor **slopes depend on** other predictors (lectures 5-6)