Bivariate Association and Significance Testing

- Topics:
  - Transforming quantitative variables (linearly or nonlinearly)
  - Bivariate measures of association and hypothesis tests
    - Correlations for quantitative variables
    - Contingency table associations of categorical variables
  - Decision errors in hypothesis testing
    - Type I and Type II errors
    - Power analysis and sample size planning

Review: Univariate Statistics

- What kind of **univariate summary statistics** are relevant to report depends on the type of variable to be described:

  - **Categorical variables (numbers are just labels):**
    - **Binary** (0 or 1): Mean (= **proportion** of 1 values); variance and skewness are then determined by the mean (i.e., they are redundant)
    - **Ordinal** or **Nominal** with **3+ categories**: **percentage** of each category; a single mean (or variance or skewness) makes no kind of sense
    - You may see ordinal variables treated as quantitative, but keep in mind this assumes real distances between the numbers used as labels
    - Bar graphs of the percentage in each category make a good visual

  - **Quantitative variables (numbers are numbers):**
    - If "symmetric enough": Min, Max, Mean, SD (or SD^2 = variance)
    - If not, add median (for central tendency) and IQR (for dispersion) that are "robust" to outliers (extreme values) or general skewness
    - Binned-value histograms or boxplots (or violin plots) make good visuals

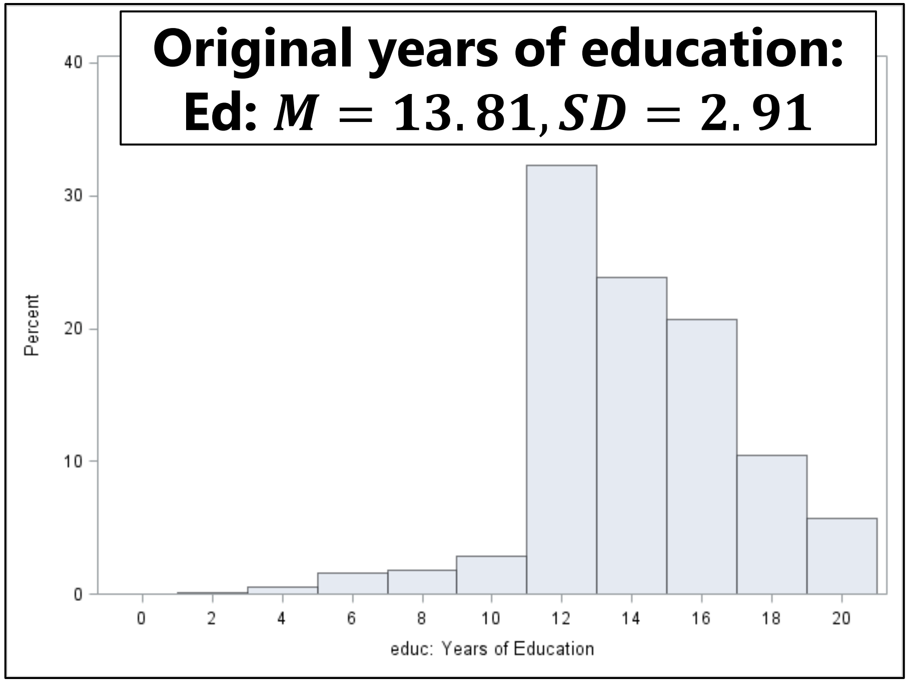
Transforming Quantitative Variables

- **Metric of quantitative variables** can vary greatly across contexts
  - May be familiar scales of "real" units: e.g., income in $1000s, height in inches/centimeters, weight in pounds/kilograms
  - May be frequencies: e.g., packs of cigarettes smoked weekly, length of hospital stay, number of hurricanes this year
  - May be induced by the number and format of contributing items: e.g., a score on a depression screener of 31; a score on a vocabulary test of 47

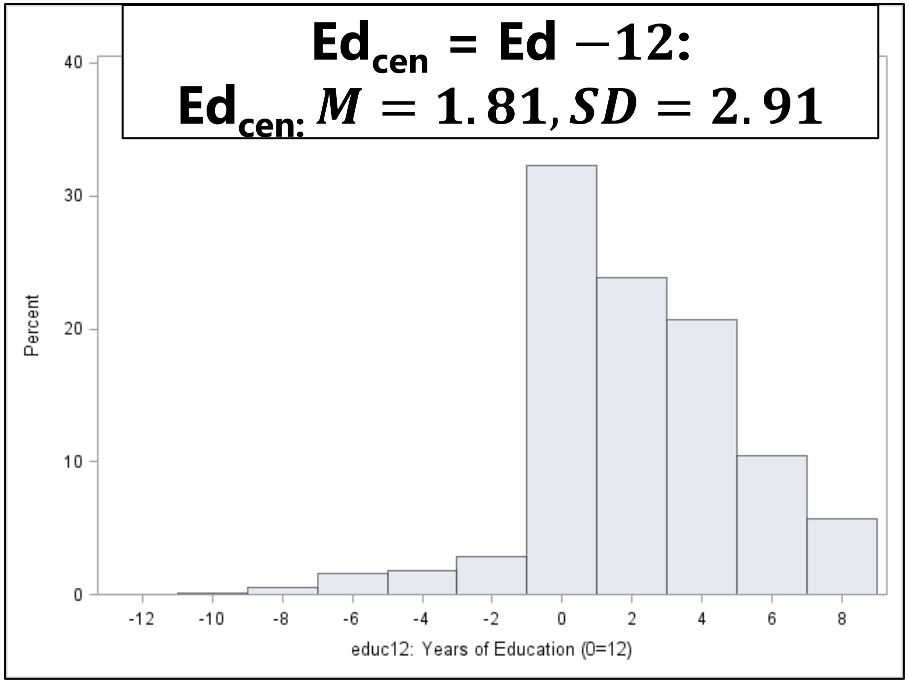
- **Arbitrary metrics are often transformed for interpretability**
  - e.g., number correct → percent correct (to range from 0-100%)
  - e.g., for 10 items, each with choices of 1-5, a sum score of 31 → item mean of 3.1 (i.e., near whatever "3" means on average)
  - e.g., test scores get converted to common "standardized" scale, e.g., M=100, SD=15 (see also GRE scores with M~150, SD~10)
  - These are all examples of **linear transformations**—transformations to the mean and/or variance of a variable that **changes all of its values evenly**

Linear Transformation: Centering

- Another example is **centering → adding or subtracting a constant** so that 0 is then a meaningful value for the new (centered) variable
  - If the sample mean $\bar{y}$ is chosen as the centering constant, this is known as "**mean-centering**" (or "grand-mean-centering")
  - Predictors will be centered when we build models (lecture 3)…



Original years of education:
Ed: $M = 13.81, SD = 2.91$



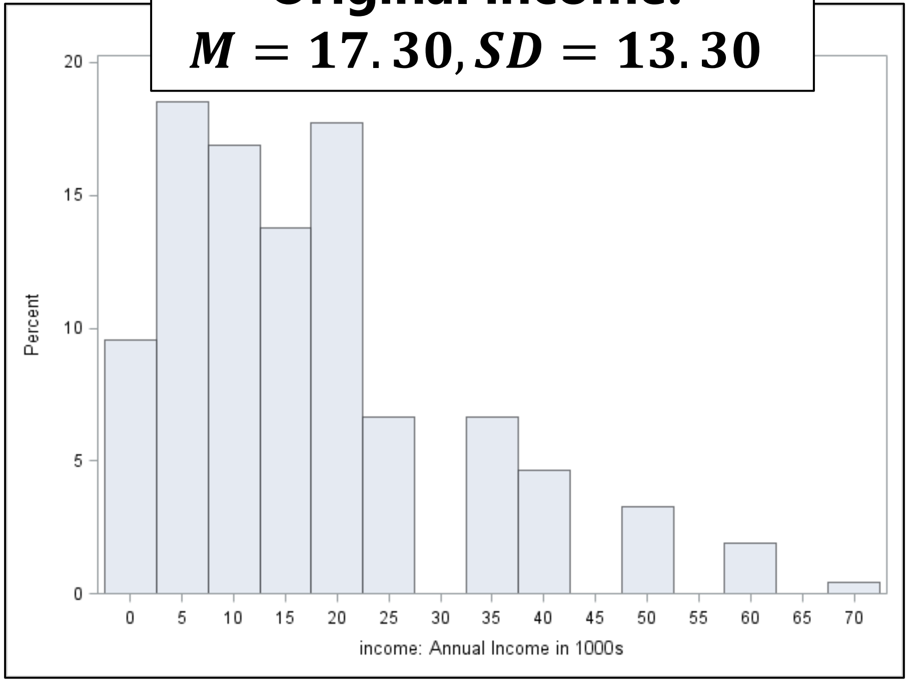
Ed$_{\text{cen}}$ **= Ed** -12**:**
Ed$_{\text{cen:}}$ $M = 1.81, SD = 2.91$

Linear Transformation: z-scoring

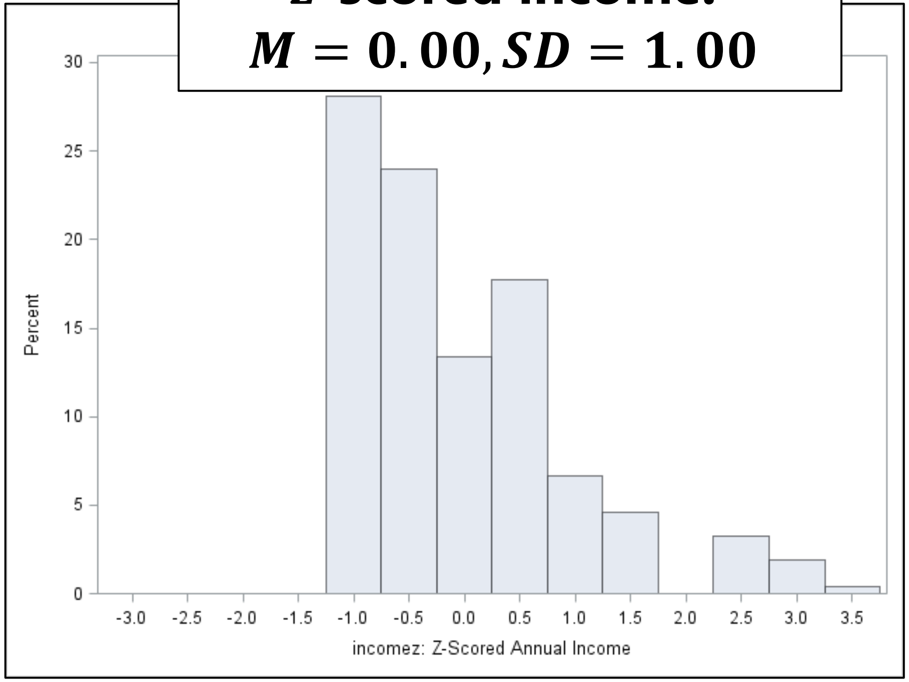
- Prevalent in statistics is the use of **"z-scoring"** = **standardize to scale of $M = 0$, $SD = 1$** using: $\boldsymbol{z_i = \frac{y_i - \bar{y}}{s}}$
- Despite the name, z-scoring does NOT make a variable normally distributed!

To unstandardize back from z_i to y_i:
$y_i = \bar{y} + (z_i * s)$

Original income:
$\boldsymbol{M = 17.30, SD = 13.30}$



z-scored income:
$\boldsymbol{M = 0.00, SD = 1.00}$



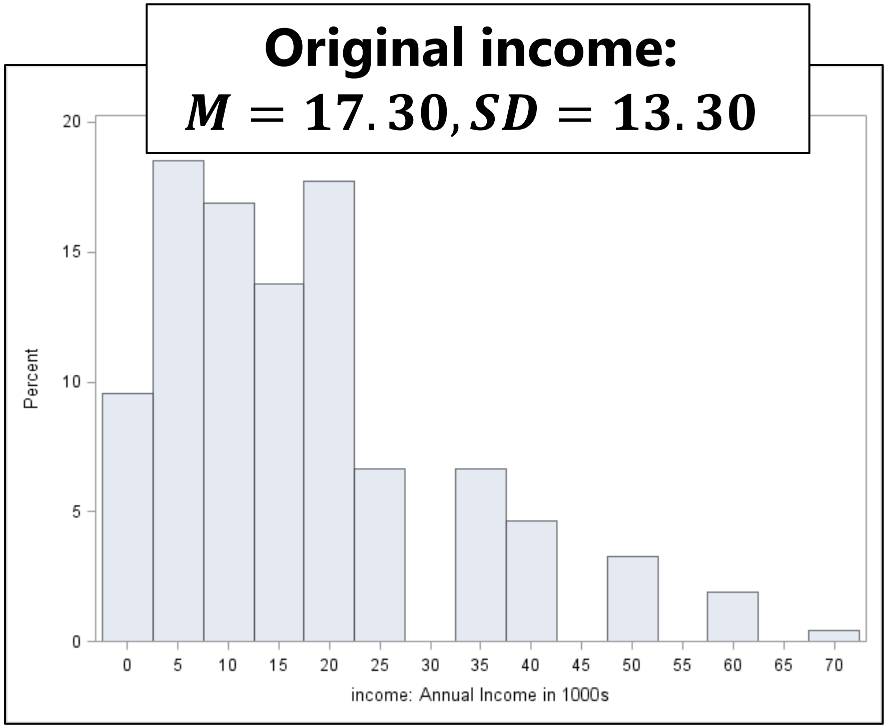
Linear vs. Nonlinear Transformations

- Primary uses of **linear transformations**:
  - To make the variable's values more **interpretable** (0 especially)
  - To put **different variables onto the same scale** so the strength of their associations with other variables can be compared more easily

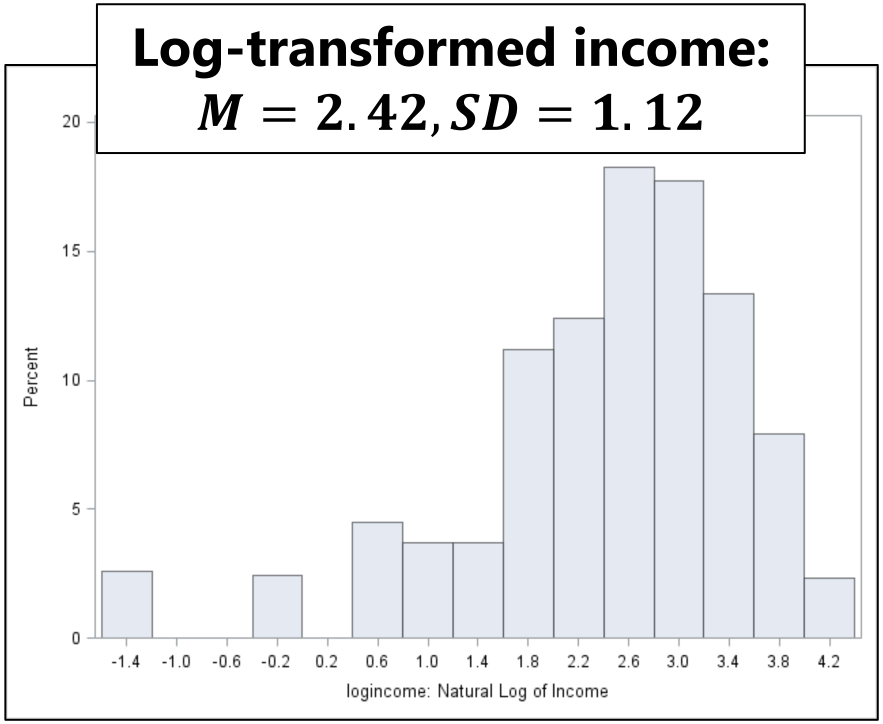
- In contrast, **nonlinear transformations change a variable's values unevenly**, often done for one of these reasons:
  - To create an **unbounded version of a bounded variable** (to be used when predicting variables with boundaries)
    - We will see an example of this in creating confidence intervals (stay tuned)
  - To **reduce the impact** of extreme (positive) values—two examples:
    - Replace values with **rank-order** (also used for associations of ordinal variables)
    - Reshape values with **natural-log transformation**… let's see an example of this

Nonlinear Transformation: Natural Log

- One example of a nonlinear transformation uses the "**logarithm**" → the exponent to which the base must be raised to produce a number x: so $Log_{base}(x) = y$ exactly if $base^{y} = x$
- The only one you will likely see in statistics is the "**Natural log**" (Log_e) that uses ***e*** (~2.718281828459) as its base: $Log_e(x) = e^{x} = exp(x)$
- Log_e spreads out lower values, and reels in upper values



Original income:
$\boldsymbol{M = 17.30, SD = 13.30}$



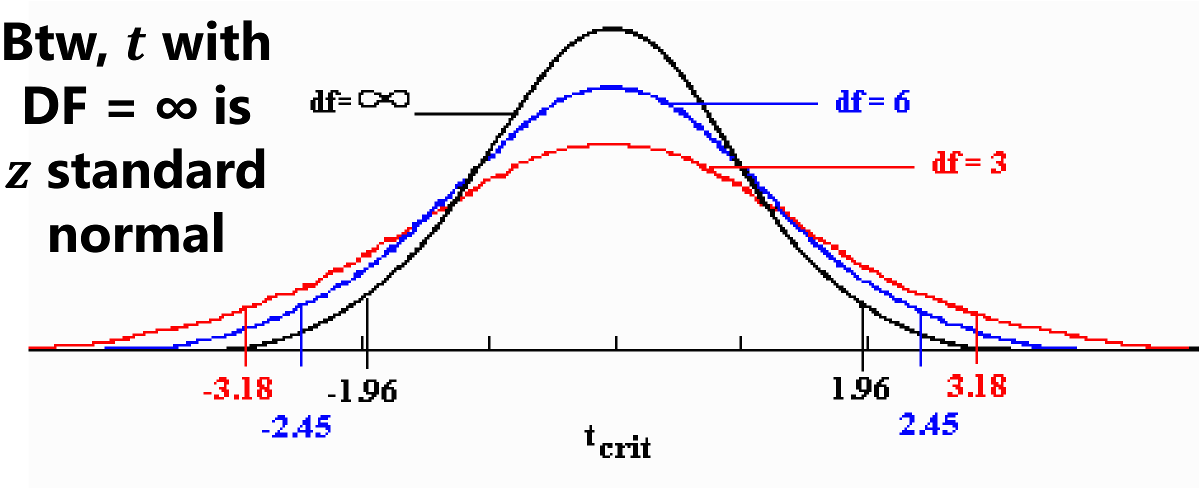
Log-transformed income:
$\boldsymbol{M = 2.42, SD = 1.12}$

For details, see https://en.wikipedia.org/wiki/Natural_logarithm

Review: From Sample to Population

- In lecture 1, we explored how to make inferences about a population mean (μ) from a sample mean ($\bar{y}$):
  - Relies on the **standard error (SE) of the mean ($SE = s/\sqrt{N}$)**, which is the average deviation of any sample mean from the population mean
  - Use SE to form a **confidence interval** (CI) around the sample mean estimate
    - $Estimate \pm t_{critical} * SE$, where % confidence and DF_{den} $(N-1)$ → $t_{critical}$
  - Use SE to form a **significance test**: How often would we see a sample mean $\bar{y}$ so discrepant from the population mean μ if μ really was true?
    - **p-value** = probability of more extreme result (from t-distribution given alpha)

Btw, t with DF = ∞ is z standard normal



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

From Univariate to Bivariate

- So far we’ve seen how to address **univariate research questions** involving a comparison of a sample statistic to a known population value (e.g., mean)
- But to answer questions about **relationships** between two variables, we need measures of **bivariate association → bi = “two”** variables
- Which measure of bivariate association should be used depends on the **kind of variables being paired** (binary, nominal, ordinal, or quantitative)

- For each measure of association, we need a **point estimate** and a test of its “**statistical significance**”: the probability of observing the association we found in the sample *if the association in the population were truly 0*
  - More formally, the process of testing an association between variables against a population value (e.g., 0) is known as “**Null Hypothesis Significance Testing**”
  - Let’s see how NHST works with a common measure of association between pairs of quantitative variables: **Pearson’s correlation**…
    - Pearson correlations are available in SAS PROC CORR or STATA PWCORR

Introducing Pearson's Correlation r

- Let's say we have **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**"; one specific type for *quantitative* variables is **Pearson's**
  - A **population** correlation is denoted as $\boldsymbol{\rho}$ ("rho"), and a **sample** correlation is $\boldsymbol{r}$
  - Correlations range continuously from $-\mathbf{1}$ **to** $\mathbf{1}$ (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 ($r = -1$):

$r = 1$ | $1 > r > 0$ | $r = 0$ | $0 > r > -1$ | $r = -1$

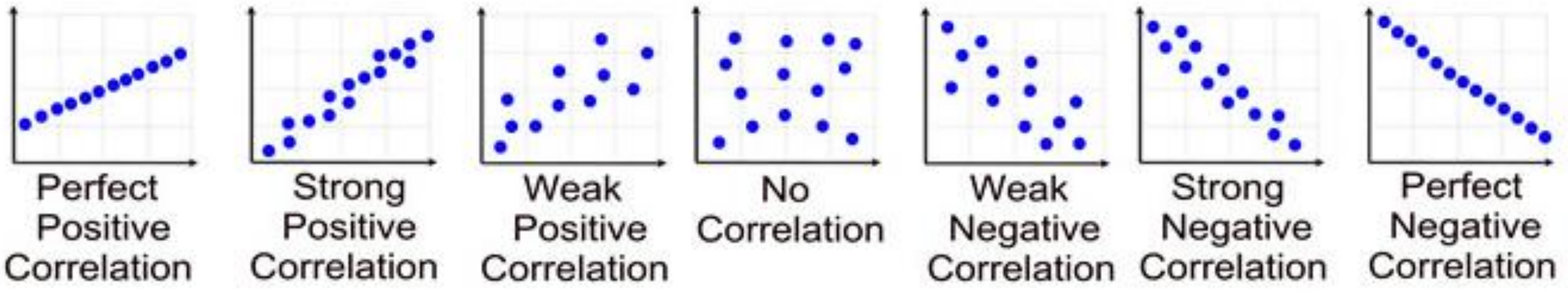


Image borrowed from: https://mathbitsnotebook.com/Algebra1/StatisticsReg/ST2CorrelationCoefficients.html

Computing Pearson's Correlation r

- To compute Pearson's r for quantitative variables x and y, we first need their univariate statistics of mean and variance:
  - Means: $\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}$, $\overline{y} = \frac{\sum_{i=1}^{N} y_i}{N}$
  - Variances: $s_x^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$, $s_y^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1}$

Note the change in notation: we identify to which variable the s^2 variance refers using a subscript

- Second, we compute their **covariance**: an unbounded measure of **association** in the **original metric** of the two variables
  - Covariance of x and y: $Cov(x, y) = \frac{\sum_{i=1}^{N}[(x_i - \bar{x})(y_i - \bar{y})]}{N-1}$
  - **Positive** covariance → same-direction match
    - **High** x values go with **High** y values; **Low** x values go with **Low** y values
  - **Negative** covariance → opposite-direction match
    - **High** x values go with **Low** y values; **Low** x values go with **High** y values
  - **Zero** covariance → no correspondence of any kind
  - Btw, the covariance of a variable with itself is its variance

Within each variable, we have only spent 1 DF_{den} → so still $N - 1$

Computing Pearson's Correlation r

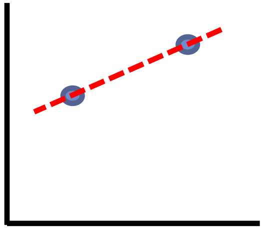
- Covariance of x and y: $Cov(x, y) = \frac{\sum_{i=1}^{N}[(x_i - \bar{x})(\mathrm{y}_i - \bar{\mathrm{y}})]}{N-1}$
  - Although a **covariance's direction is informative, its value is not directly informative** because it is **specific to the x and y units**
  - Example: the association between **height and weight** in $N = 10$ men:
    - Height in inches: $\bar{x} = 72.20$, $s_x = 6.51$, $s_x^2 = 42.40$, $range = 62 - 82$
    - Weight in pounds: $\bar{y} = 235.90$, $s_y = 20.89$, $s_y^2 = 436.54$, $range = 201 - 269$
    - Covariance: $Cov(x, y) = 135.24$ "inch–pounds" indicates ?????
    - It's a **positive covariance**, which tells us that **taller men tend to be heavier**, but it does not give the size of this relationship in a standardized way… we need $\boldsymbol{r}$

- Third: we rescale the covariance by adjusting it for the SD of each variable, which leads to **Pearson's $\boldsymbol{r}$, a standardized association**
  - $r = \frac{Cov(x,y)}{s_x s_y} = \frac{135.24}{6.51 * 20.89} = .99408$
  - Positive association is almost perfect!

If both variables have SD=1 (e.g., they have each been z-scored so Mean=0, SD=1), then Covariance = Correlation

Adjusting* Pearson's r for Sample Size

- Note what is **not included** in the formula for Pearson's r:
  - $r = \frac{Cov(x,y)}{s_x s_y}$ → There is no reference to DF_{den} to reflect **sample size**!
  - To illustrate why this is a problem, think about what would happen if we picked two points randomly and fit a line through them… perfect ($r = 1$)!



- To solve this problem in small samples (like our example of $N = 10$), one could instead choose to report an "**adjusted correlation**"***:
  - $r_{adj} = \sqrt{1 - \frac{(1-r^2)(N-1)}{N-2}} = \sqrt{1 - \frac{(1-.99^2)(10-1)}{10-2}}$ = .99339 (instead of .99408)
  - r and r_{adj} will be more similar the stronger the correlation is, and the bigger the sample is

*** I have never actually reported r_{adj}, but I include it here for completeness just in case Reviewer 3 asks for it someday…

Testing Pearson's r for "Significance"

- More generally, we are doing a "**Null Hypothesis Significance Test**"; in this example, we are asking "what is the probability of observing the sample r we found if the population $\rho = 0$"?
  - A "**hypothesis**" is a statement about a population parameter

- A "**null hypothesis**" ($\boldsymbol{H_0}$) is a statement about the population parameter being equal to some specific (expected) value
  - In Lecture 1 testing the sample mean $\bar{y}$, $H_0\colon \mu = 10$
  - In current example testing the sample correlation r, $H_0\colon \rho = 0$

- An "**alternative hypothesis**" ($\boldsymbol{H_A}$) is a statement that contradicts the null hypothesis and **conveys allowed directionality of deviations** from value given by H_0
  - In Lecture 1 with the sample mean $\bar{y}$, $H_A\colon \mu \neq 10$
  - In current example with the sample correlation r, $H_A\colon \rho \neq 0$
  - These are both "two-tailed" hypotheses (allow either direction)

Steps in Significance Testing

- **Choose critical region: % alpha ("unexpected") and possible direction**
  - Two sides or just one side?
  - Alpha (α) ($1 - \%$ confidence)?
  - Distribution for test-statistic will be dictated as follows:

Uses Denominator Degrees of Freedom?	**Test 1 thing***	**Test >1 thing***
No: implies infinite N	z	$\chi^2 (= z^2$ if 1)
Yes: adjusts based on N	$\boldsymbol{t}$	$F (= t^2$ if 1)

- If the **test-statistic exceeds** the distribution's critical value(s), then the obtained $\boldsymbol{p}$**-value is less than the chosen alpha** level:
  - You "**reject the null hypothesis**"—it is sufficiently **unexpected** to get a test-statistic that extreme *if the null hypothesis is true;* result is "**significant**"

- If the **test-statistic does NOT exceed** the distribution's critical value(s), then the $\boldsymbol{p}$**-value is greater than or equal to the chosen alpha** level:
  - You "**DO NOT reject the null hypothesis**"—it is sufficiently **expected** to get a test-statistic that extreme *if the null hypothesis is true;* result is "**not significant**"

* *Thing = numerator DF for association (stay tuned)*

Testing Pearson's r for "Significance"

- Sample correlation $\boldsymbol{r}$ is tested against population correlation $\boldsymbol{\rho}$ using a $\boldsymbol{t}$**-distribution** (with denominator degrees of freedom, DF_{den})
  - For $H_0\!: \rho = 0$, test-statistic $t = r\sqrt{\frac{N-2}{1-r^2}}$, $DF_{den} = N - 2$
- Choose a **two-tailed test** (because either a negative or positive correlation would be meaningful), and **typical alpha (**$\boldsymbol{\alpha}$**) = .05**
  - For $\boldsymbol{\alpha = .05}$ (95% confidence) and $\boldsymbol{DF_{den} = 8}$, then $\boldsymbol{t_{critical} = \pm 2.31}$
- For our example, testing $H_0\!: \rho = 0$

Either way, **we reject** $\boldsymbol{H_0}$**:**
$\boldsymbol{r}$ is **"significantly"** positive

  - **Pearson's** $\boldsymbol{r}$: $t = \boldsymbol{.99408}\sqrt{\frac{10-2}{1-(.99408)^2}} = 25.88$, $p = .00000000534$ (5.34E-09)
  - **Adjusted** $\boldsymbol{r}$: $t = \boldsymbol{.99334}\sqrt{\frac{10-2}{1-(.99334)^2}} = 24.38$, $p = .00000000855$ (8.55E-09)
  - It's **REALLY UNLIKELY** to observe $r = .99$ with $N = 10$ if the true $\rho = 0$

Testing Pearson's r for "Significance"

- Another example using $N = 10$ and two random variables simulated to have no relationship in the population ($\rho = 0$)
  - $t = r\sqrt{\frac{N-2}{1-r^2}}$, for $DF_{den} = N - 2 = 8$ and $\alpha = .05$, $t_{critical} = \pm 2.31$

Either way, we **do not reject $\boldsymbol{H_0}$**: $\boldsymbol{r}$ is "**nonsignificantly**" negative

- New example, testing $H_0\!: \rho = 0$
  - **Pearson's $\boldsymbol{r}$**: $t = -\mathbf{.250}\sqrt{\frac{10-2}{1-(-.250)^2}} = -0.732\ p = .485$
  - **Adjusted $\boldsymbol{r}$**: $t = -\mathbf{.237}\sqrt{\frac{10-2}{1-(-.237)^2}} = -0.691,\ p = .498$
  - It's **sufficiently expected** to obtain $r = \pm .25$ with $N = 10$ if the true $\rho = 0$; a more extreme t test-statistic would have been found about 49% of the time

- When reporting results, 2 or 3 decimal places is sufficient
- Quantities that cannot go past 1 (like r and p) do not need leading zeros, but you should use them for everything else

Pearson correlation r: From estimate of relationship directly to significance

- To estimate the Pearson correlation r between two variables in a sample, we need their means, variances (→SD), and covariance:
  - $Cov(x, y) = \frac{\sum_{i=1}^{N}[(x_i - \bar{x})(\mathrm{y}_i - \bar{\mathrm{y}})]}{N-1}$ → Pearson $r = \frac{Cov(x,y)}{s_x s_y}$

- We then **directly compute a t test-statistic** for sample correlation r against population correlation $\rho = 0$ **using sample size N**:
  - $\boldsymbol{t} = r\sqrt{\frac{N-2}{1-r^2}}$, $DF_{den} = N - 2$ and chosen alpha → $t_{critical}$
  - Note: the same r will result in a greater t test-statistic (i.e., t-value) with greater N → **more people, easier to say** obtained correlation r is "**unexpected**" if population correlation is really $\rho = 0$
  - In software, the t-value is generally omitted and given instead is the **exact p-value** → **probability of sample r if population $\rho = 0$**
    - If p-value < alpha, reject H_0: $\rho = 0$ → r is "significantly" different than 0

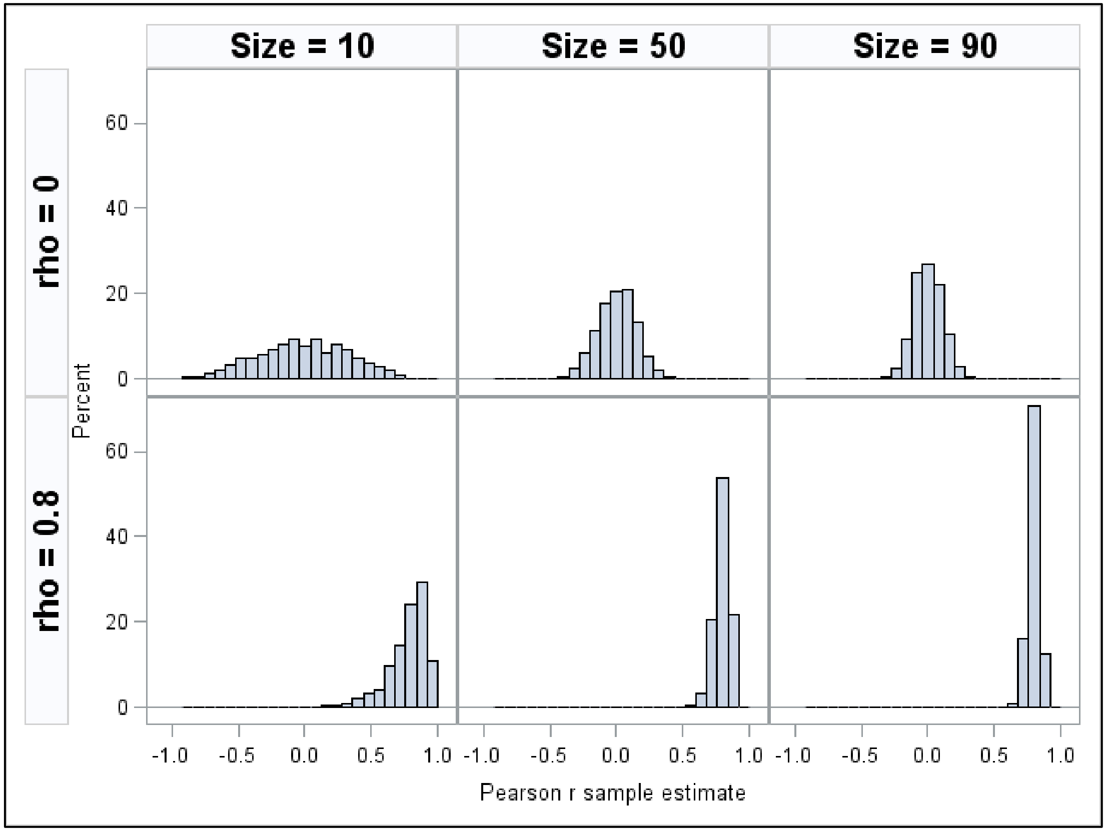
What about a CI for correlation r?

- Knowing a correlation r is “significant” doesn’t speak to its **expected inconsistency** across samples…
  - Remember **confidence intervals**? CI = range that should include the population value in chosen % of samples
    - A **symmetric interval** around any sample statistic (like correlation r here) is given by: $CI = estimate\ \pm (critcal * SE)$
    - $\boldsymbol{critical}$ refers to threshold value on PDF capturing the statistic’s sampling distribution given chosen alpha + directionality (one side or both) and degrees of freedom (numerator and/or denominator)
    - $\boldsymbol{SE}$ refers to standard error of the correlation estimate r: the average deviation of a sample correlation from the population correlation

- Relative to the SE and CI for a sample mean previously, finding the SE and CI for a sample correlation is more complicated because r only ranges from -1 to 1

Sampling Distribution of correlation r

- Demo: I simulated two bivariate normal distributions ($\rho = 0$ or $\rho = .8$) of 100,000 fake persons for variables x_i and y_i, each in a z-score metric (so $M = 0, SD = 1$)
- Drew 1000 random samples each of $N = 10$, 50, or 80



Pop ρ	N per sample	Mean r_s	SD r_s
0	10	-.02	.34
	50	.00	.14
	90	.00	.11
.8	**10**	**.77**	**.15**
	50	.79	.06
	90	.80	.04

What would happen to $CI = r \pm (2ish * SE)$???

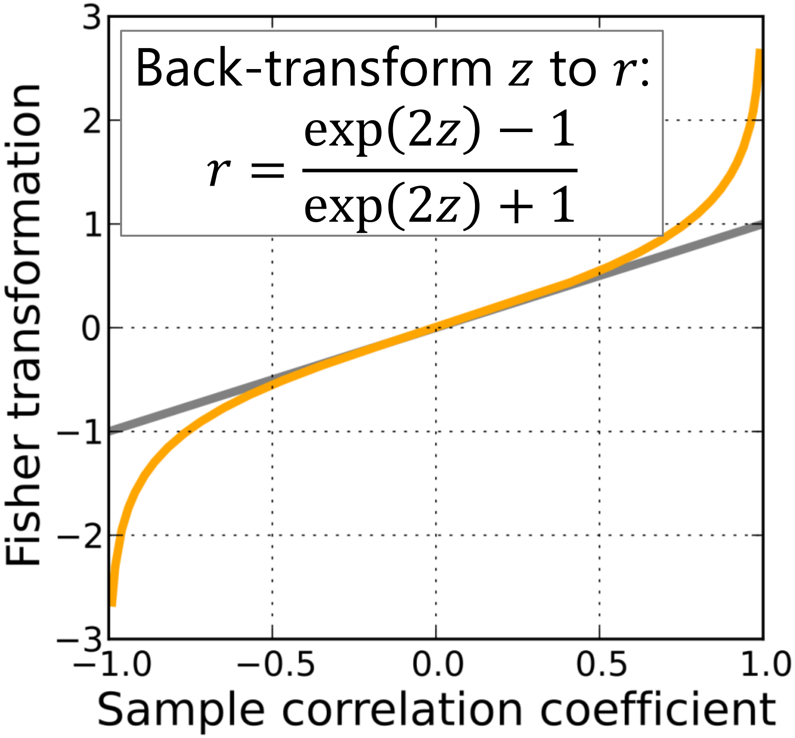
SE and CI for Pearson’s r

- Finding an SE and CI for r is more complicated because **r is bounded between ±1**
  - This means that a symmetric CI (i.e., from $r \pm critical * SE$) will not work for extreme r values
- One solution is a **nonlinear** “**Fisher transformation**” →
  - It’s called “**Fisher’s z**”, but it’s not the same z as in z-score (sorry)
- A **more general solution** is to form a symmetric CI around the **unbounded slope (implied by bounded r) in a model**
  - Stay tuned…

Fisher $z_r = 0.5\left[Log_e\left(\frac{1+r}{1-r}\right)\right]$,

$$SE\ z_r = \frac{1}{\sqrt{N-3}}\ ,\ CI = z_r \pm z_{crit} * SE$$

Steps: convert r to z_r, compute lower and upper bounds in z-scale, back-transform bounds to r-scale



Back-transform z to r:

$$r = \frac{\exp(2z) - 1}{\exp(2z) + 1}$$

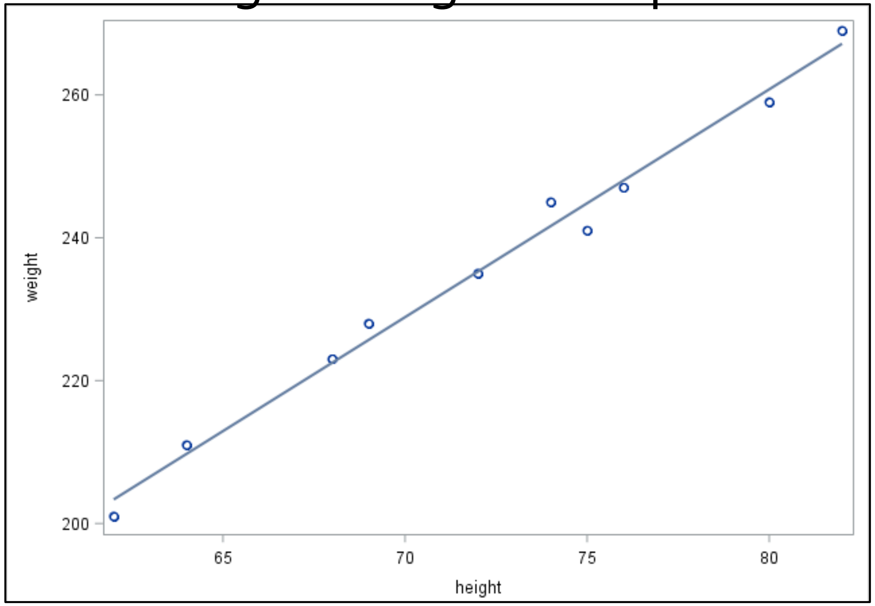
Image borrowed from: https://en.wikipedia.org/wiki/Fisher_transformation

Pearson's Correlation and Linearity

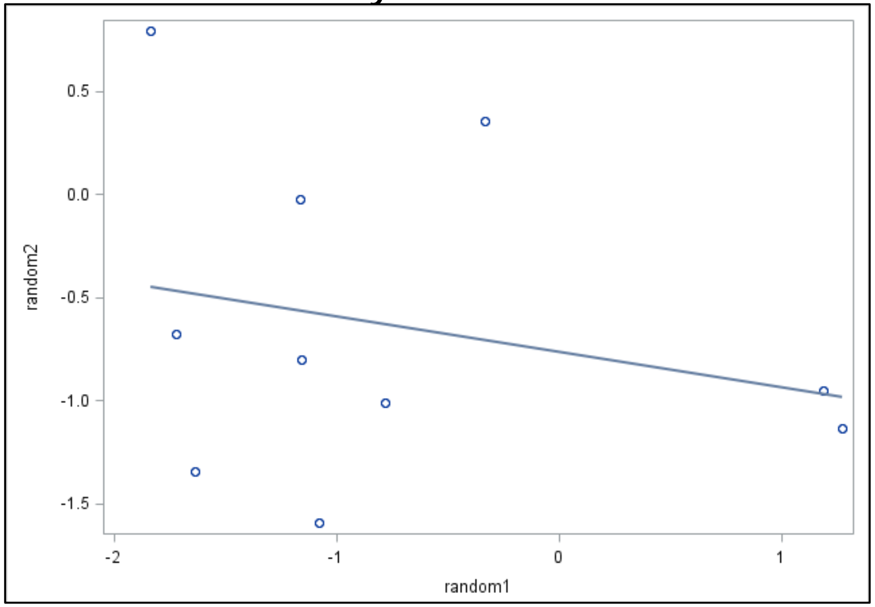
- The bivariate association between quantitative variables provided by Pearson's correlation r has a specific assumed form: **linear relationship**
- The $\boldsymbol{r}$ value is indicated by the **slope of the prediction ("regression") line**

How did the regression line get determined? Stay tuned…

$r = .994$ for
Height–Weight example

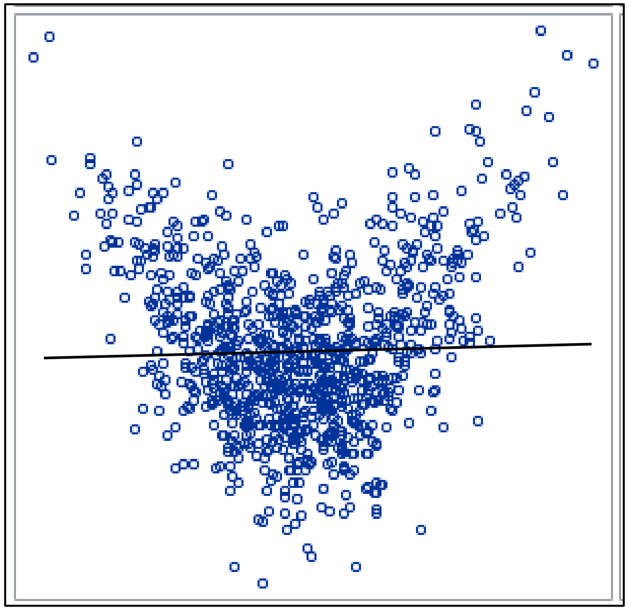


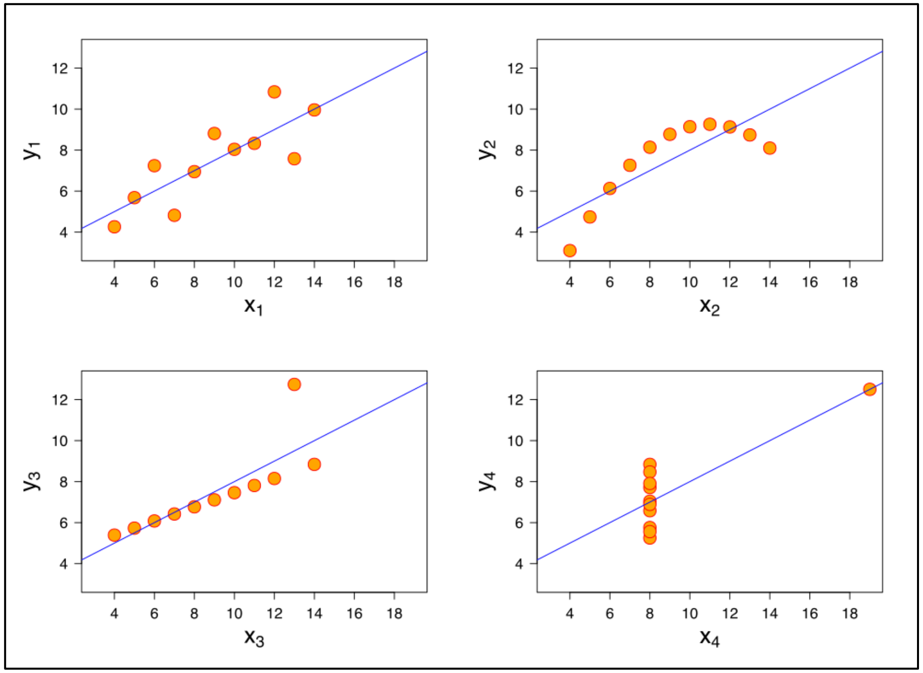
$r = -.250$ for
two randomly created variables



Pearson Correlation and Linearity

- **Pearson's r will not capture any nonlinear relationships**
- Right: line reflects $r = .05$, but it misses the real story—a U-shaped relationship
  - X and Y are negatively related up to some point, after which they are positively related





Left: Anscombe's quartet, in which $r = .82$ in each of 4 datasets with nearly identical statistics (but which show very different types of association)

Right image borrowed from Ryan Walters IDC 625 (Creighton University)
Left image borrowed from: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Pearson's r vs. Spearman's rho (ρ)

- Computational shortcuts for Pearson's r with special names:
  - Pearson's r for two binary variables = "**phi**" $\boldsymbol{r}$
  - Pearson's r for a binary and a quantitative variable = "**point-biserial**" $\boldsymbol{r}$

- To reduce influence of "outliers" (extreme values), choose another kind of correlation: **Spearman's rank correlation coefficient (or $\boldsymbol{\rho}$, rho)**
  - Sort variables by value, then do **Pearson's $\boldsymbol{r}$ on the rank order** of values (using same process to find SE, CIs, and t test-statistics for significance)
  - Available in SAS PROC CORR or STATA SPEARMAN

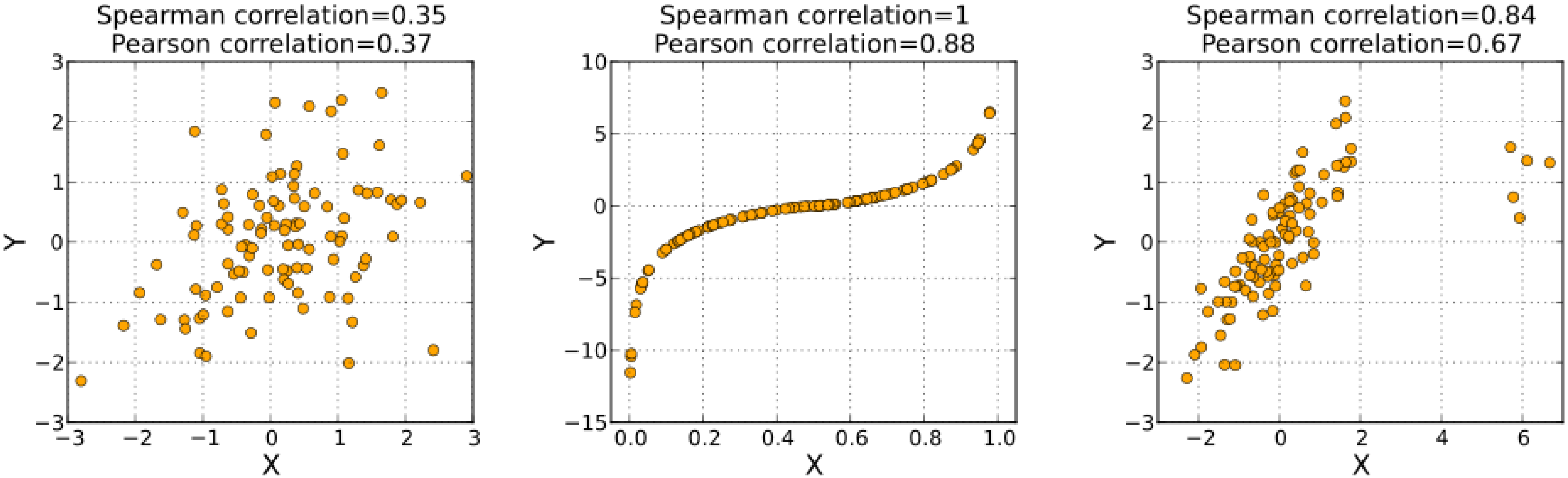


Image borrowed from: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

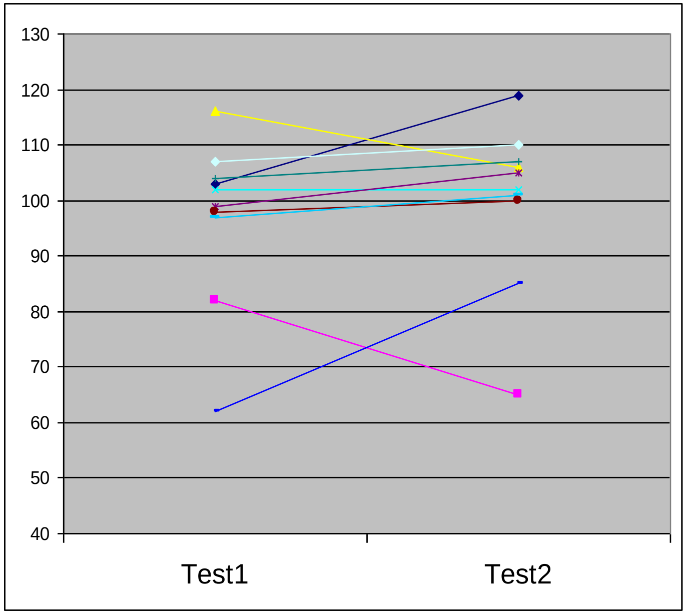
Pearson vs. Intraclass Correlation

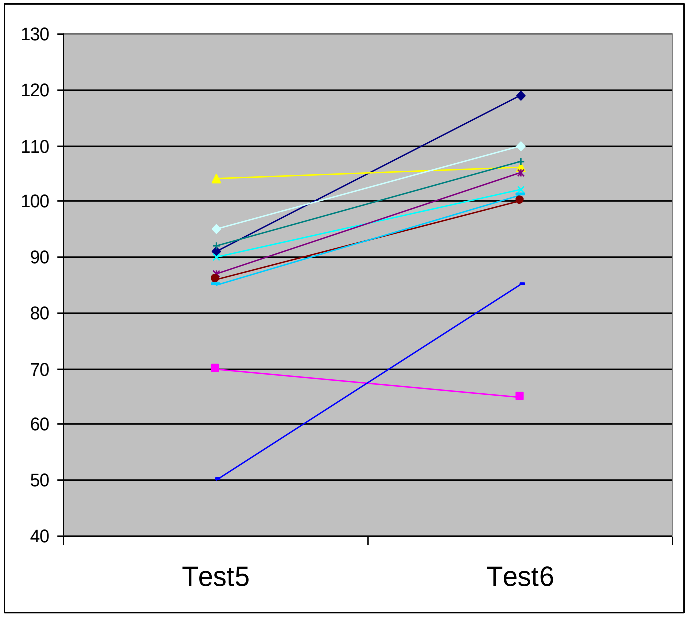
- Correlations are sometimes computed to measure **reliability**: the extent of agreement between two or more sources (variables)
  - e.g., **multiple raters** (y_1, y_2) each provide scores for the same set of targets
- **Pearson's r is problematic for reliability**, because it ignores differences in mean and variance across raters by standardizing each variable separately
- Solution: use an "**Intraclass Correlation**" (ICC) instead, which standardizes across all raters using a **common mean and variance** instead
  - For example, for two raters: $\mathrm{ICC}(y_1, y_2) = \frac{\sum_{i=1}^{N}[(y_{1i}-\bar{\mathrm{y}})(\mathrm{y}_{2i}-\bar{\mathrm{y}})]}{(N-1)*s^2}$

    where $\bar{y} = \frac{\sum_{i=1}^{N}[(y_{1i}+\mathrm{y}_{2i})]}{2N}$ and $s_y^2 = \frac{\sum_{i=1}^{N}(y_{1i}-\bar{y})^2+\sum_{i=1}^{N}(y_{2i}-\bar{y})^2}{2N-1}$
  - ICC is also a ratio of variances: $ICC = \frac{s^2_{Between-Targets}}{s^2_{Between-Targets}+s^2_{Between-Raters}+s^2_{within-both}}$
- **ICCs can readily be extended** to more than two raters, as well as to quantify the effect of multiple distinct sources of systematic variance
  - e.g., multiple raters of multiple targets across days—how much variance for each?
  - This is the basis of "Generalizability Theory" (or G-Theory) in measurement

For more info, see: https://en.wikipedia.org/wiki/Intraclass_correlation

Intraclass Correlation Example





M: 97 100

SD: 15 15

$$Pearson\ r = .670$$

$$Intraclass\ r = .679$$

M: 85 100

SD: 15 15

$$Pearson\ r = .670$$

$$Intraclass\ r = .457$$

$$ICC = \frac{s^2_{Between-Targets}}{s^2_{Between-Targets} + \boldsymbol{s^2_{Between-Raters}} + s^2_{within-both}}$$

Correlations for Binary Variables?

- The possible **Pearson's r for binary variables will be limited** when they are not evenly split into 0/1 because their variance depends on their mean
  - Remember: Mean = p, Variance = $p * (1 - p)$

- If two variables (x and y) differ in p, such that $p_y > p_x$
  - Maximum covariance: $Cov(x, y) \;=\; p_x(1 - p_y)$
  - This problem is known as **"range restriction"**
  - **Here this means the maximum Pearson's r will be smaller than ± 1 it should be:**

$$r_{x,y} = \sqrt{\frac{p_x(1-p_y)}{p_y(1-p_x)}}$$

  - Some examples using this formula to predict maximum Pearson r values →

px	py		max r
0.1	0.2		0.67
0.1	0.5		0.33
0.1	0.8		0.17
0.5	0.6		0.82
0.5	0.7		0.65
0.5	0.9		0.33
0.6	0.7		0.80
0.6	0.8		0.61
0.6	0.9		0.41
0.7	0.8		0.76
0.7	0.9		0.51
0.8	0.9		0.67

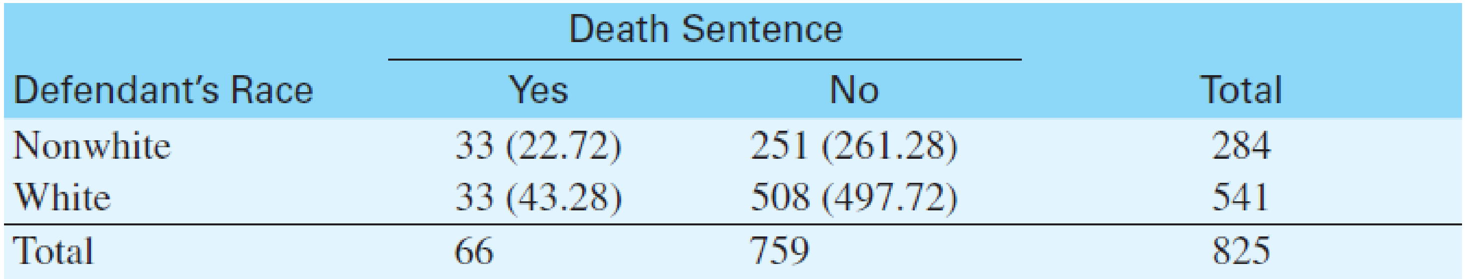
Correlations for Binary or Ordinal Variables

- To solve this range restriction, you may want to report a different type of correlation based on the idea of a "continuous underlying variable" for the binary or ordinal variables (≠ Pearson's r)
- Here are four you will hear of in **advanced** quant classes…
  - **Tetrachoric correlation**: between 'underlying continuous' distributions of two actually binary variables (not = Pearson or Spearman);
  - **Biserial correlation**: between 'underlying continuous' (but really binary) variable and observed quantitative variable (not = Pearson or Spearman)
  - **Polychoric correlation**: between 'underlying continuous' distributions of two ordinal variables (not = Pearson or Spearman)
  - **Polyserial correlation**: between 'underlying continuous' distributions of one ordinal variable and observed quantitative variable (not = Pearson or Spearman)
- Tetrachoric and polychoric correlations are used in latent variable measurement models for categorical outcomes (Item Response Theory)

Bivariate Association for Categorical Variables

- **Associations among categorical variables** are more often described using test statistics from **cross-tabulations** (aka, contingency tables)
  - Frequencies of each possible observed combinations across variables
  - Each combination is a "**cell**"; total across a row or column is a "**margin**"
  - All cells must be **independent** (or else you need a different approach)
  - Available in SAS PROC FREQ or STATA TABULATE, TAB2, and CS (for effect sizes)

- For example: relationship of defendant race to death sentence



	Death Sentence		
Defendant's Race	Yes	No	Total
Nonwhite	33 (22.72)	251 (261.28)	284
White	33 (43.28)	508 (497.72)	541
Total	66	759	825

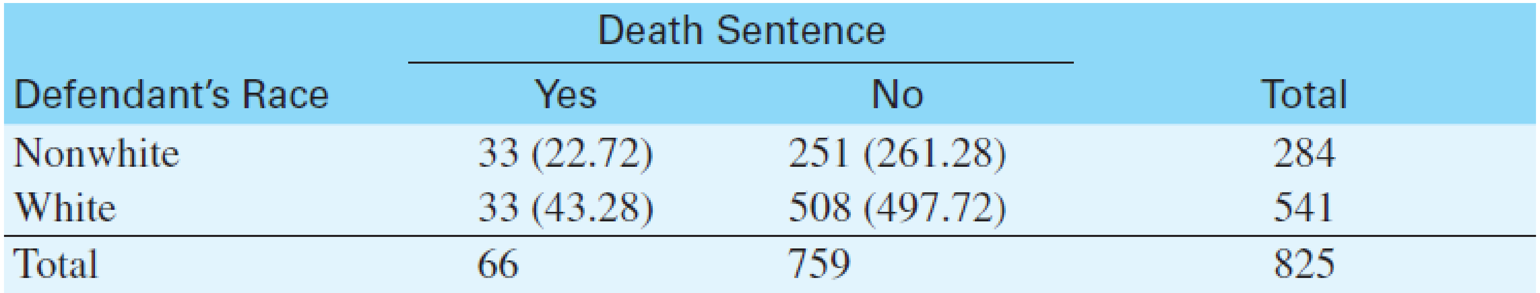
  - (Numbers) are expected cell counts for row r and column c: $E_{rc} = \frac{N_r N_c}{N}$

N_r = row total
N_c = column total

  - For $r = 1$ and $c = 1$ → Nonwhite Yes: $E_{11} = \frac{284 * 66}{825} = 22.72$
  - For $r = 1$ and $c = 2$ → Nonwhite No: $E_{12} = \frac{284 * 759}{825} = 261.28$

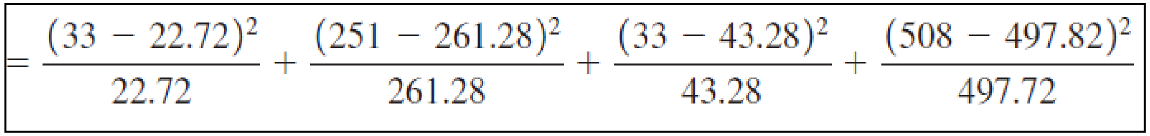
Example borrowed from: Howell, D. C. (2010). *Statistical methods for psychology* (7th ed). Belmont, CA: Cengage Wadsworth.

Bivariate Association for Categorical Variables



	Death Sentence		
Defendant's Race	Yes	No	Total
Nonwhite	33 (22.72)	251 (261.28)	284
White	33 (43.28)	508 (497.72)	541
Total	66	759	825

- **Pearson's χ^2 test-statistic →** how far off the expected (E_{rc}) from observed (O_{rc}) frequencies are for cell $t = rc$, summed over T cells:
- $\chi^2 = \sum_{t=1}^{T} \frac{(O_{rc} - E_{rc})^2}{E_{rc}} = 7.71$



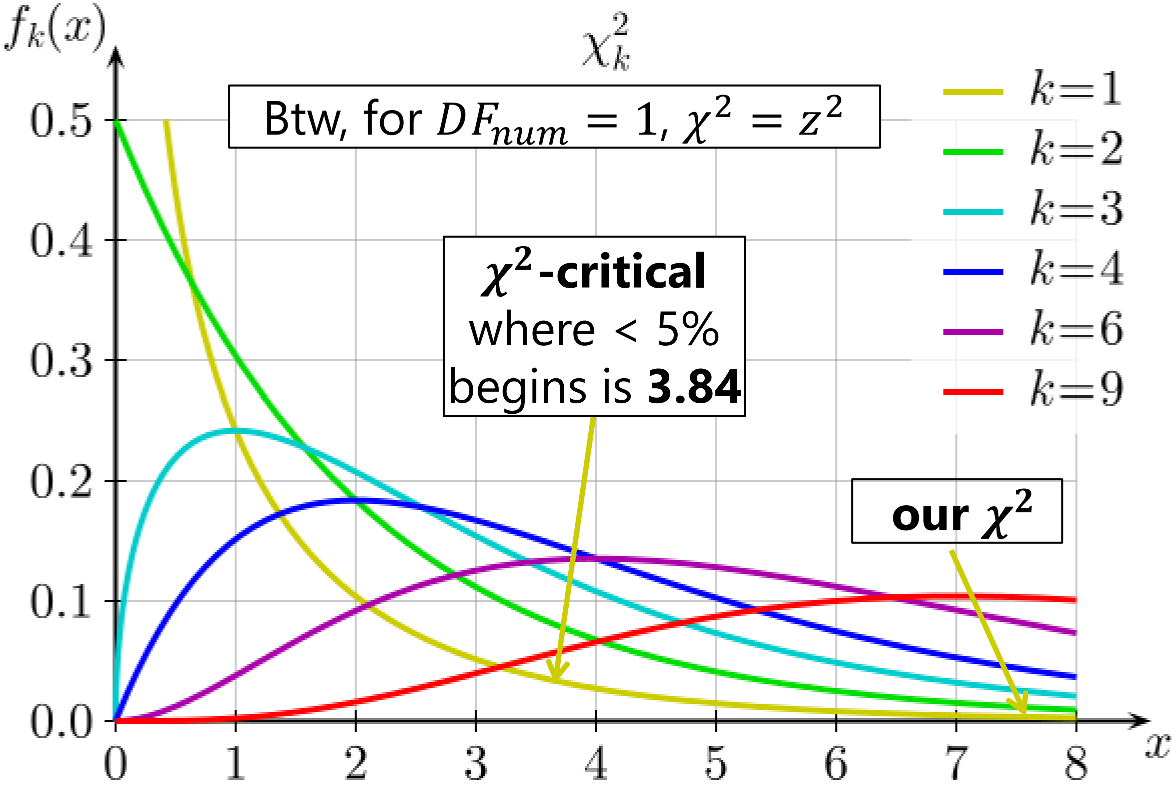
$$= \frac{(33 - 22.72)^2}{22.72} + \frac{(251 - 261.28)^2}{261.28} + \frac{(33 - 43.28)^2}{43.28} + \frac{(508 - 497.82)^2}{497.72}$$

- To get the χ^2 test-statistic's critical value ($\chi^2_{critical}$), you need to know degrees of freedom—but in this case, it is ***numerator* degrees of freedom** (DF_{num}) instead
  - Based on R = # of rows and C = # of columns: $DF_{num} = (R-1)(C-1) = 1$
  - Because χ^2 doesn't use denominator DF , the label "DF" is sufficient, but I want to distinguish each kind of DF (numerator = relationship parameters tested, denominator = "points" left over from sample size minus parameters tested)
  - $DF_{num} = 1$ is written as $\chi^2(1) = 7.71$ or $\chi^2_1 = 7.71$; $\chi^2_{critical} = 3.84$ for $\alpha = .05$

Example borrowed from: Howell, D. C. (2010). Statistical methods for psychology (7th ed). Belmont, CA: Cengage Wadsworth.

The Chi-square (χ^2) Distribution

- The expected value of the χ^2 for H_0 = "no association" is its (numerator) degrees of freedom (DF_{num}, labeled "k" below)
  - χ^2 has only positive values → only right tail for "unexpected" area



Btw, for $DF_{num} = 1$, $\chi^2 = z^2$

χ^2-critical where < 5% begins is **3.84**

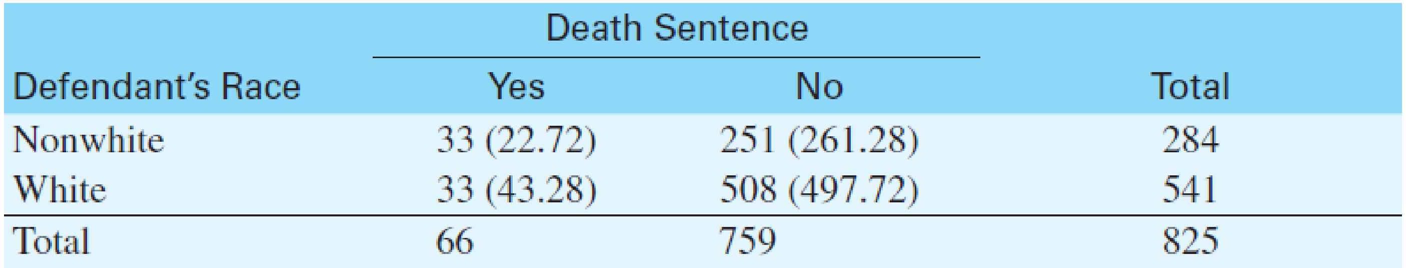
our χ^2

For current example with $DF_{num} = 1$ (yellow line), our obtained $\boldsymbol{\chi^2 = 7.71}$ is above the critical value of **3.84 at $\boldsymbol{\alpha} = .05$**, so we reject H_0 = no association, **exact p-value = .00549**.

The χ^2distribution is used more generally to test ≥ 1 effects simultaneously, but without denominator DF (i.e., no adjustment for N).

Image borrowed from: https://en.wikipedia.org/wiki/File:Chi-square_pdf.svg

Bivariate Association for Categorical Variables



	Death Sentence		
Defendant’s Race	Yes	No	Total
Nonwhite	33 (22.72)	251 (261.28)	284
White	33 (43.28)	508 (497.72)	541
Total	66	759	825

- **Conclusion?** Obtained $\chi^2_1 = 7.71 \ > \chi^2_{critical} = 3.84$, **so reject $\boldsymbol{H_0}$**
  - From CHIDIST in excel, p-value = .00549 → gives the percentage of time we’d find $\chi^2_1 \geq 7.71$ if there were no association in the population (which is $\chi^2 = DF_{num}$)
  - **Conclusion in English?** We need to **determine the pattern** that created this significant result—in this case, this is straightforward to do because there is only one distinction to make across columns or rows ($DF_{num} = 1$)
  - **Across columns:** Among nonwhite defendants, there is a greater proportion given the death sentence than would be expected (where “expected” → based on the proportion of nonwhite defendants and the proportion of any persons given death sentences); Among white defendants, there is a smaller proportion given the death sentence than would be expected (based on the proportion of white defendants and the proportion of any persons given death sentences)
  - **Across rows:** Among persons receiving the death penalty, more of them are nonwhite (and fewer or them are white); Among persons not receiving the death penalty, more of them are white (and fewer of them are nonwhite)

Example borrowed from: Howell, D. C. (2010). Statistical methods for psychology (7th ed). Belmont, CA: Cengage Wadsworth.

Bivariate Association for Categorical Variables

- Pearson's χ^2 can be used for variables with > 2 categories, but determining the reason for a significant result is then more challenging—for example:

Number of Child Abuse Categories Checked	Abused as Adult No	 Yes	Total
0	512 (494.49)	54 (71.51)	566
1	227 (230.65)	37 (33.35)	264
2	59 (64.65)	15 (9.35)	74
3-4	18 (26.21)	12 (3.79)	30
Total	816	118	934

- $\chi^2 = \sum_{t=1}^{T} \frac{(O_{rc}-E_{rc})^2}{E_{rc}} = 29.63,\ DF_{numerator} = (R-1)(C-1) = 3$
  - Obtained $\chi^2_3 = 29.63 > \chi^2_{critical} = 7.82$; reject H_0 (exact $p = 0.0000017$)
  - There are 3 unique $2x2$ ("2 by 2") combinations to consider ("unique" implies that others can be found once you know those 3)
  - You can break the analysis into $2x2$ tables to see what the patterns are, but this situation is better handled in a general*ized* linear model…
    - Come back in a few semesters for "Applied Generalized Linear Models"! ☺

Example borrowed from: Howell, D. C. (2010). Statistical methods for psychology (7th ed). Belmont, CA: Cengage Wadsworth.

Other Measures of Bivariate Association You May See for Categorical Variables

- When $DF_{num} = 1$ (testing 1 thing), $z^2 = \chi^2$, and both ignore N!
- **χ^2 p-values may not be accurate when any expected cell count < 5**, and so various (non-t)"fixes" have been developed:
  - "**Exact**" tests: use simulation (not assumed distributions) to get p-values
  - Likelihood ratio test: $\chi^2 = 2\sum_{t=1}^{T}\left[O_{rc} * Log_e \frac{O_{rc}}{E_{rc}}\right]$
    - Equivalent to Pearson's χ^2 in "big enough" samples; shows up in models for categorical outcomes (like "log-linear"; "generalized")
- **What if some categories are a lot more frequent?**
  - **Kappa** (κ): χ^2 used for measuring agreement (e.g., reliability) that corrects for chance levels of agreement
  - Other ways of correcting for disproportionate numbers of people in certain categories (e.g., McNemar's test for consistency in responses)

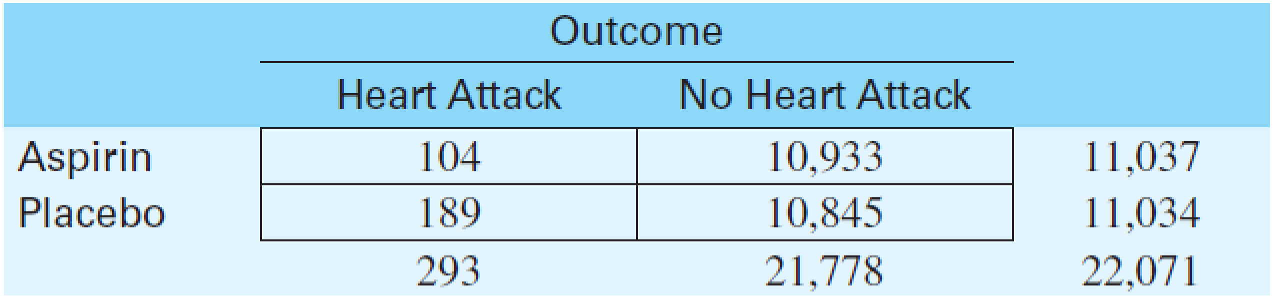
Effect Sizes for Measures of Association

- The **correlation metric $\boldsymbol{r}$** is more generally known as an index of "**effect size**"—a **standardized metric** that conveys the size of an effect, irrespective of statistical significance (and N)
  - Another effect size is $\boldsymbol{d}$: standardized mean difference (stay tuned)

- **Test-statistics** (that use both effect size and sample size N in significance testing) can be **converted back into effect sizes:**
  - e.g., Pearson's χ^2 between two **binary variables** is called a "phi" correlation that is exactly the same as Pearson's r: $r = \sqrt{\chi_1^2/N}$
    - However: p-values may not match! This is because Pearson r is tested using a t-distribution with DF_{den}, but χ^2 (like standard normal z) does not account for DF_{den}
  - e.g., convert any t test-statistic to an r effect size: $r = \dfrac{t}{\sqrt{(t^2 + DF_{den})}}$

- **Pearson's $\boldsymbol{\chi^2}$ has other special types of effect sizes, too…**

Effect Size via Risk Ratios (Relative Risk)



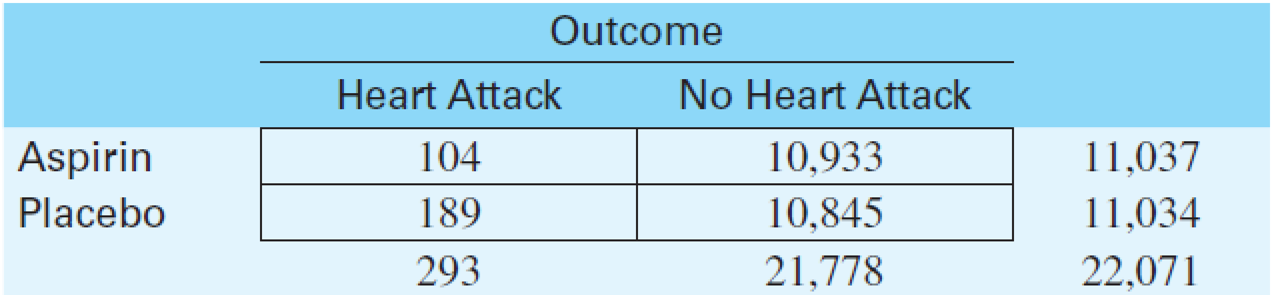
	Outcome		
	Heart Attack	No Heart Attack	
Aspirin	104	10,933	11,037
Placebo	189	10,845	11,034
	293	21,778	22,071

$\chi^2_1 = 25.014 >$
$\chi^2_{critical} = 3.84;$
$p < .0001$ (5.69E-07)

- **Risk** = single cell proportion within a row or a column
  - e.g., aspirin: heart attack **risk** $= \frac{104}{11{,}037} = 0.94\%$
  - e.g., placebo: heart attack **risk** $= \frac{189}{11{,}034} = 1.71\%$
  - *Note that total number of each row is used as the denominator*
  - Difference ($= 0.77\%$) doesn't seem like much, but it's a bigger deal when you consider how small the base rates of heart attacks are
- **Risk ratio** (= **relative risk**) $= \frac{1.71\%}{0.94\%} = 1.819$
  - Without aspirin, your risk of a heart attack is 1.819 times greater

Example borrowed from: Howell, D. C. (2010). Statistical methods for psychology (7th ed). Belmont, CA: Cengage Wadsworth.

Effect Size via Odds Ratios



	Outcome		
	Heart Attack	No Heart Attack	
Aspirin	104	10,933	11,037
Placebo	189	10,845	11,034
	293	21,778	22,071

$\chi^2_1 = 25.014 >$
$\chi^2_{critical} = 3.84;$
$p < .0001$ (5.69E-07)

- **Odds** = ratio of cell frequencies across a row or a column
  - e.g., aspirin: heart attack **odds** $= \frac{104}{10{,}933} = 0.95\%$
  - e.g., placebo: heart attack **odds** $= \frac{189}{10{,}845} = 1.74\%$
  - *Note that frequency of other condition in the row is used as the denominator*
- **Odds ratio** (OR) = $\frac{1.74\%}{0.95\%} = 1.832$
  - Without aspirin, your risk of a heart attack is 1.832 times greater
  - With aspirin, your risk of a heart attack is 0.546 times smaller
    - Thus, odds are not symmetric, and that drives me crazy…
  - Odds ratios are common measures of effect size in health-related research

Example borrowed from: Howell, D. C. (2010). Statistical methods for psychology (7th ed). Belmont, CA: Cengage Wadsworth.

Intermediate Summary

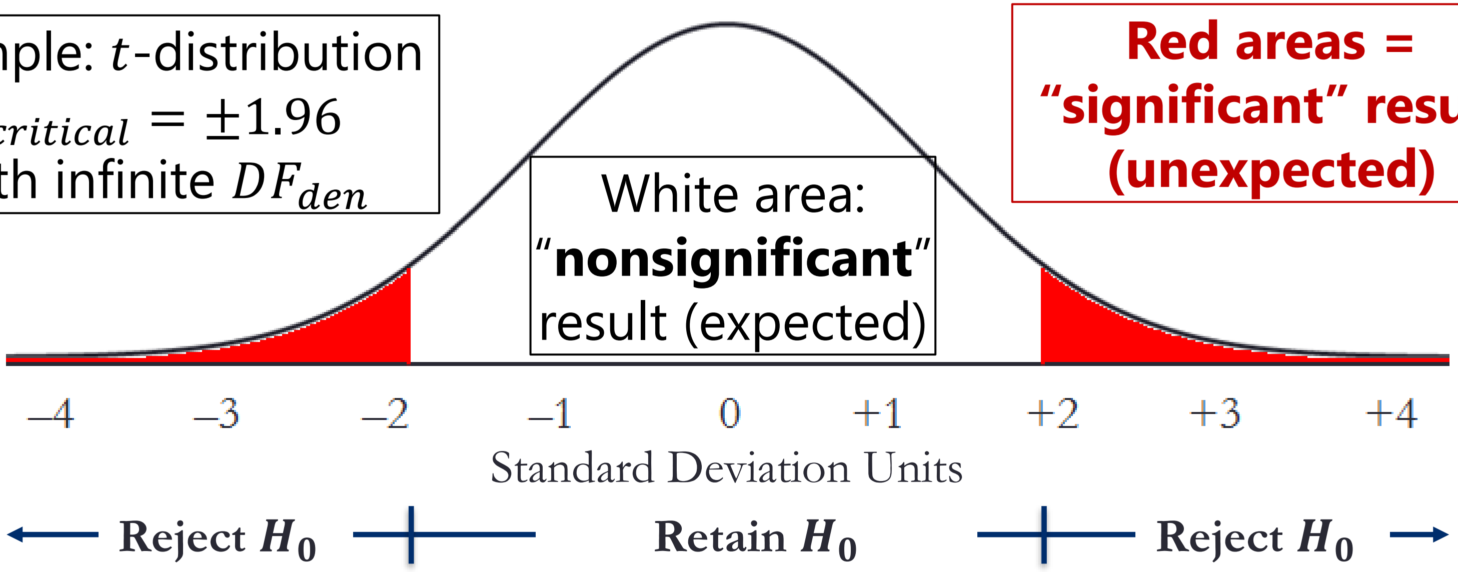
- Measures of **bivariate association** come in many flavors:
  - Two **quantitative** or binary variables: **Pearson's $\boldsymbol{r}$** (which measures linear relationships only, has special names of "phi" and "point-biserial")
  - Two **ordinal** variables (or quant with extreme values): **Spearman's $\boldsymbol{r}$**
    - Both kinds of r can be tested for statistical significance against a null hypothesis of no correlation ($H_0\!: \rho = 0$) using a t test-statistic with $DF_{den} = N - 2$
  - Two **categorical** variables: **Pearson's $\boldsymbol{\chi^2}$** (which assumes nominal variables; has many related variants to correct small sample issues)
    - Tested for statistical significance against a null hypothesis of no association using a χ^2 test-statistic with numerator degrees of freedom, such that ($H_0\!: \chi^2 = DF_{num}$)
  - I skipped the combination of quantitative with nominal variables that have 3+ categories, as that is best handled with a model

- In deciding whether or not to claim a result is significant (i.e., to reject H_0), we can screw this up in 2 distinct and important ways…

Significance Tests Require:

- A **distribution** (e.g., t, z, F, or χ^2) that goes with the test-statistic
- A **rejection region** = alpha (α) → how extreme the test-statistic value must be to declare it "significant" and thus "unexpected"
  - e.g., $\alpha = .05$ (95% confidence) implies that a result that extreme must only happen less than 5% of the time if the null hypothesis (H_0) is true
  - You also have to decide if you want the rejection region at both ends (a **two-tailed test**; usually) or only at one end (one-tailed test; rarely)

Example: t-distribution
$t_{critical} = \pm 1.96$
with infinite DF_{den}



White area: "**nonsignificant**" result (expected)

Red areas = "significant" result (unexpected)

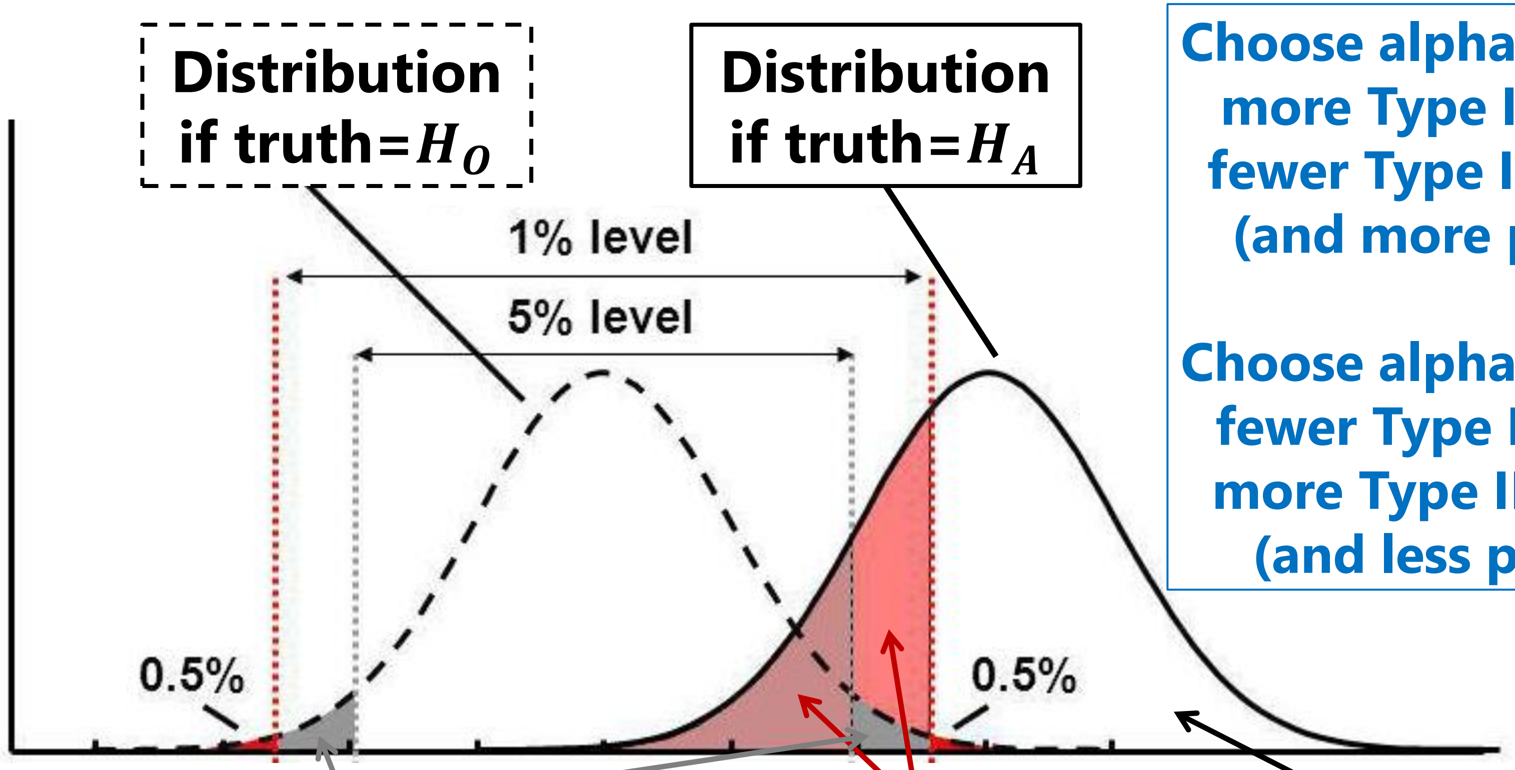
Image borrowed from Ryan Walters IDC 625 (Creighton University)

Decision Errors in Hypothesis Testing

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

	Truth: H_0	**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



Distribution if truth=H_0

Distribution if truth=H_A

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)

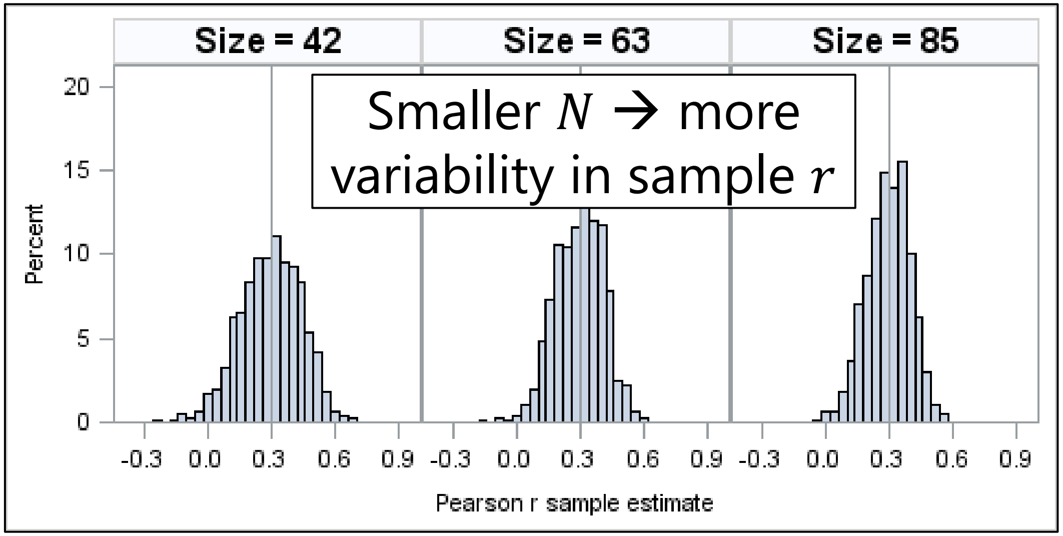
H_0 gray areas = α = % Type I errors (false alarms)

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

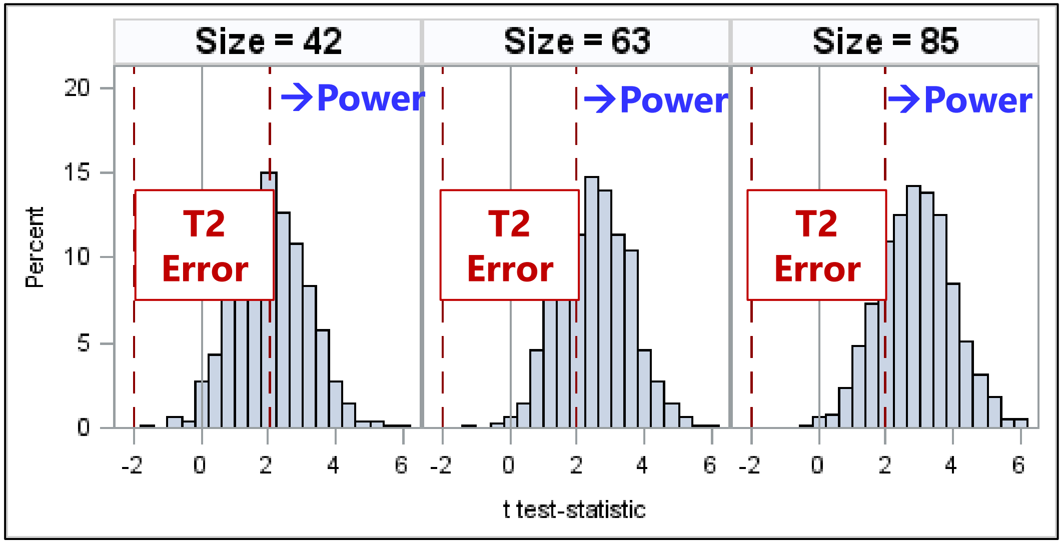
H_A white area = $1-\beta$ % power

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

Anticipating Statistical Power



Smaller N → more variability in sample r



→Power

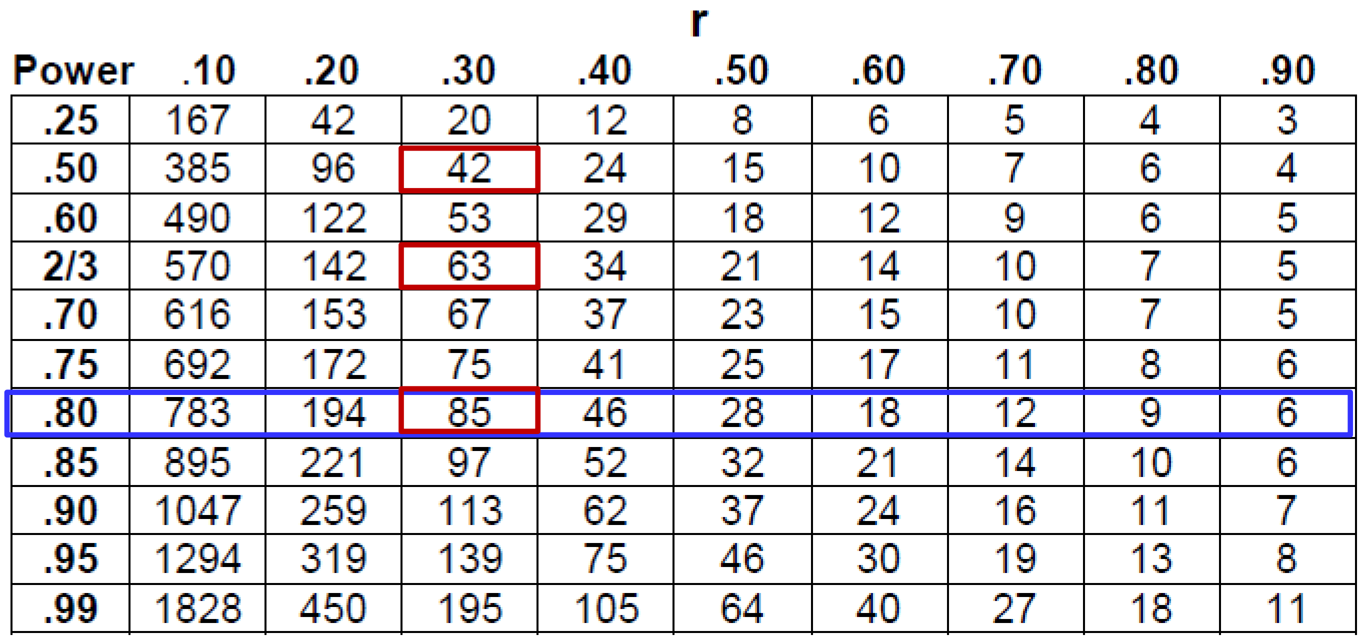
T2 Error

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

N	**Statistical Power: % significant**	**Type II Error: % not significant**
42	**50%**	**50%**
63	**66%**	**37%**
85	**79%**	**21%**

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"
  - 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

For every hypothesis test, the following will be reported in a known format:

- **Estimate** of parameter (from a model); value of obtained **test-statistic** (t, z, F, or χ^2)
- **Numerator degrees of freedom** (DF_{num}) when testing more than one relationship parameter simultaneously (used with F or χ^2; $DF_{num} = 1$ for t or z)
- **Denominator degrees of freedom** ($DF_{denominator}$) when not assuming infinite sample size (used with t or F; not used with z or χ^2)
- $\boldsymbol{p}$**-value**: probability of obtained test-statistic if null hypothesis H_0 is true
- **Effect size** (e.g., r, d, or odds ratio)—you have an r effect size already if your association is a type of correlation (or else compute it); effect size CIs are nice to include, too

Conditional on your decision about significance, what can happen?

- If you **reject $\boldsymbol{H_0}$** and claim your result as "**significant**" given your chosen alpha (α):
  - **DO** have to worry about probability of **Type I error** (given by your p-value): **a false alarm**
  - DO NOT have to worry about the probability of a Type II error: a miss
  - Power is related to replicability—a significant result with low power is less likely to replicate!
- If you **retain $\boldsymbol{H_0}$** and claim your result as "**nonsignificant**" given your chosen alpha (α):
  - DO NOT have to worry about probability of Type I error (given by your p-value): a false alarm
  - **DO** have to worry about the probability of a **Type II error: a miss** (power = 1 – Type II error)
  - In planning studies, the conventional level of power to aim for is 80% (harder to do with smaller effects)