p–values Had a Good Run: A Primer on the 'New Statistics'

Rob Cribbie
Quantitative Methods Program
Department of Psychology
York University

Day Two

- Part 4: Confidence Intervals, Effect Sizes and Confidence Intervals for Effect Sizes
- Part 5: Replication
- Part 6: Bayesian Analysis

Part 4: Confidence Intervals, Effect Sizes and Confidence Intervals for Effect Sizes

- When we use a sample statistic (e.g., M) to estimate a population parameter (e.g., μ), an important question relates to how precisely we have measured the parameter
- A confidence interval can give us an estimate of that precision
  - X% Confidence Interval (CI)
    - Interpretation: If we sample repeatedly from a population, X% of the confidence intervals are expected to contain the population parameter

Confidence Intervals

- Example: 95% CI
  - If we sample repeatedly from a population (i.e., we extract thousands of samples from a given population), 95% of the confidence intervals computed from the samples are expected to contain the population parameter

- Note: We CANNOT say “there is a 95% chance that the true mean lies within our calculated CI”

Confidence Intervals

- (1–α)% CI for the Mean (σ known)
  - To calculate the CI, we need to know the *z* value that cuts off the highest α/2 of the cases from the rest
  - For a 95% CI:
    - (1–(.05/2) = .975 (area in lower tail)
  - In R:
    - qnorm(.975, lower.tail=TRUE) = 1.96

Confidence Interval for the Mean

- $M \pm SEM * z_{(1-(\alpha/2)}$
- $M \pm {}^{\sigma}/_{\sqrt{N}} * z_{(1-(\alpha/2)}$

- Example – PTSD
  - We sampled N = 25 students from a population with σ = 5 and gave them PTSD training
    - After the training we obtain M = 28
  - We are interested in how precisely we are estimating the true mean
  - Calculate and interpret the 95% confidence interval

Confidence Intervals

- $95\%\ CI = M \pm SEM * z_{1-(.05/2)}$
- $95\%\ CI = M \pm {}^{\sigma}/_{\sqrt{N}} * z_{.975}$

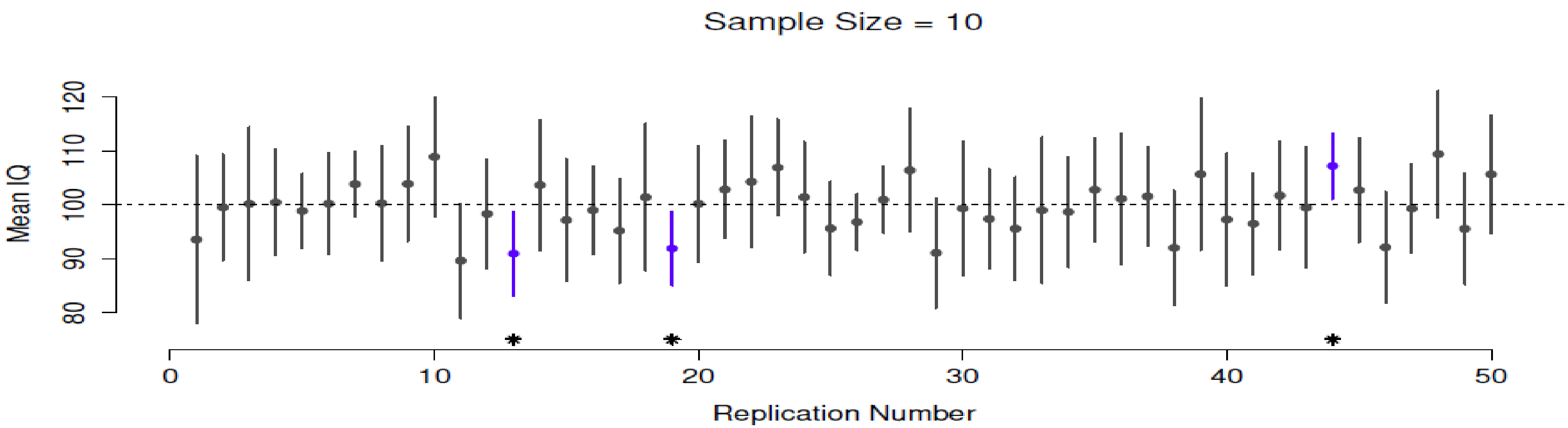
- $95\%\ CI = 28 \pm {}^{5}/_{\sqrt{25}} * 1.96$
  - $28 - (1)(1.96) = 26.04$
  - $28 + (1)(1.96) = 29.96$
- Thus, the 95% CI = (26.04, 29.96)
- Interpretation
  - If we conduct the study over and over again, 95% of the CIs are expected to include the population mean
  - Researcher's job: Determine if the precision (i.e., width) of the CI is acceptable
    - Researchers often don't report CIs because they are much wider than what they would like

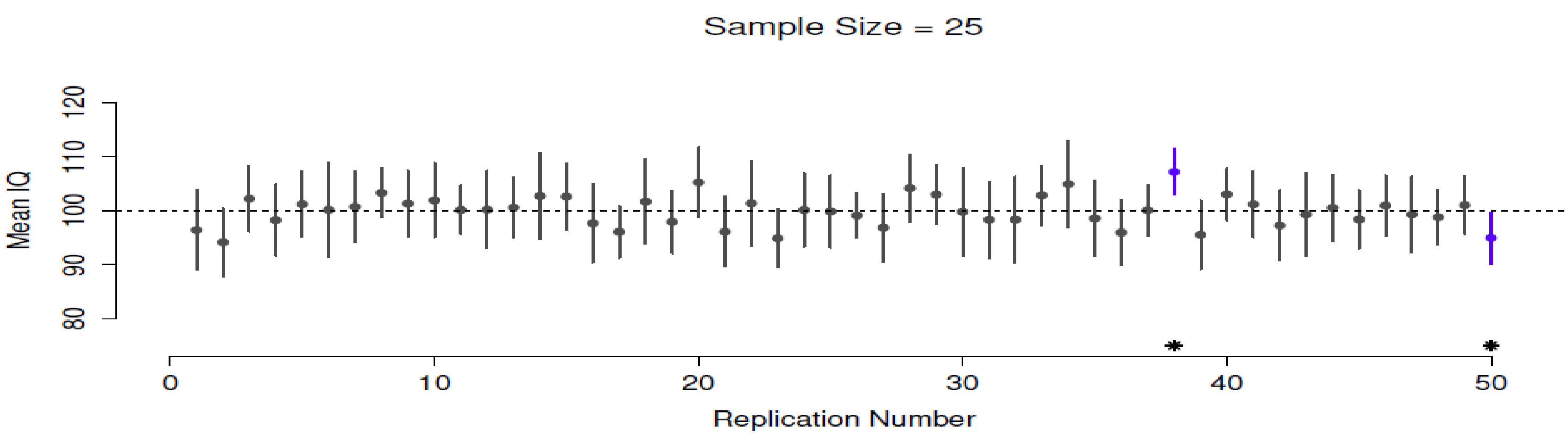
Confidence Intervals

- What determines the width of a CI?
  - Recall: M $\pm$ ${}^{\sigma}/_{\sqrt{N}}$ * $z_{(1-(\alpha/2))}$

  - Size of the standard deviation
    - Less variability, narrower CI
  - Sample Size
    - Larger N, narrower CI
  - Level of Confidence (e.g., 95%)
    - Lower the confidence, the narrower the CI

Simulated CIs: N = 10, 25





Issues with Confidence Intervals

- 1) They are simply used to determine statistical significance
  - If the CI does not include the null value then the effect is statistically significant
    - E.g., If testing H_0: $\mu = 10$, then if the CI does not include 10 then it is concluded that the mean is significantly different from 10
      - Logically then, if the CI does include 10 then we cannot reject the null hypothesis
  - But …. that is not the primary purpose of CIs

Issues with Confidence Intervals

- 2) Misinterpretations of CIs
  - Recall the correct interpretation:
    - If we conducted the study over and over again, X% of the CIs are expected to include the population parameter
  - However, many researchers wish/hope that they could say that there is an X% chance that the population parameter falls within their *single computed CI*

Issues with Confidence Intervals

- 2) Misinterpretations of CIs
  - A recent study by Hoekstra, Morey, Rouder and Wagenmakers (2014) looked at misinterpretations of CIs
  - They presented participants with the following statement:
    - “A researcher reports a 95% CI for the mean that ranges from 0.1 to 0.4.”
  - They then asked the participants a series of T/F questions

Issues with Confidence Intervals

- Hoekstra et al. T/F questions
  - 1. The probability that the true mean is greater than 0 is at least 95%
  - 2. The probability that the true mean equals 0 is smaller than 5%
  - 3. The “null hypothesis” that the true mean equals 0 is likely to be incorrect
  - 4. There is a 95% probability that the true mean lies between 0.1 and 0.4
  - 5. We can be 95% confident that the true mean lies between 0.1 and 0.4
  - 6. If we were to repeat the experiment over and over, then 95% of the time the true mean falls between 0.1 and 0.4

Issues with Confidence Intervals

- Hoekstra et al. T/F questions
  - 1. The probability that the true mean is greater than 0 is at least 95%
  - 2. The probability that the true mean equals 0 is smaller than 5%
  - 3. The “null hypothesis” that the true mean equals 0 is likely to be incorrect
- These discuss probabilities associated with a hypothesis, which is not allowed in a frequentist framework

Issues with Confidence Intervals

- Hoekstra et al. T/F questions
  - 4. There is a 95% probability that the true mean lies between 0.1 and 0.4
  - 5. We can be 95% confident that the true mean lies between 0.1 and 0.4
  - 6. If we were to repeat the experiment over and over, then 95% of the time the true mean falls between 0.1 and 0.4
- These make reference to the specific interval, which is not how we interpret CIs
  - We reference hypothetical future intervals, but not the single current interval

Issue with CIs: Hoekstra et al.

Table 1 Percentages of students and teachers endorsing an item

Statement	First Years ($n = 442$)	Master Students ($n = 34$)	Researchers ($n = 118$)
The probability that the true mean is greater than 0 is at least 95 %	51 %	32 %	38 %
The probability that the true mean equals 0 is smaller than 5 %	55 %	44 %	47 %
The "null hypothesis" that the true mean equals 0 is likely to be incorrect	73 %	68 %	86 %
There is a 95 % probability that the true mean lies between 0.1 and 0.4	58 %	50 %	59 %
We can be 95 % confident that the true mean lies between 0.1 and 0.4	49 %	50 %	55 %
If we were to repeat the experiment over and over, then 95 % of the time the true mean falls between 0.1 and 0.4	66 %	79 %	58 %

Effect Size

- Nakagawa and Cuthill (2007) discuss how *effect size* encompasses:
  - (a) a statistic which estimates the magnitude of an effect (e.g., r)
  - (b) the actual values calculated from certain effect statistics (e.g., $r = .3$)
  - (c) a relevant interpretation of an estimated magnitude of an effect from the effect statistics (e.g., “medium”)

Definitions of Effect Size

- Olejnik and Algina (2003) define an effect size measure as:
  - A *standardized* index that estimates a parameter that is independent of sample size and quantifies the magnitude of the difference between populations or the relationship between explanatory and response variables
- Grissom and Kim (2012)
  - Whereas a test of statistical significance is traditionally used to provide evidence (attained *p* level) that a null hypothesis is wrong, an effect size (ES) measures the degree to which such a null hypothesis is wrong (if it is wrong)

Definitions of Effect Size

- Cohen (1988)
  - The degree to which the null hypothesis is false
- Thompson (2004)
  - Effect sizes quantify by how much sample results diverge from the null hypothesis
- Kelley & Preacher (2012)
  - A quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest

- Summary: Lots of different ways of defining "effect size"

Definitions of Effect Size

- What about the Sample Size issue?
  - It is interesting that some definitions of effect sizes don't mention that they should be immune to any effects of sample size
  - One of the primary reasons to focus on effect sizes is that we want a statistic that is not highly related to N
    - Recall that many statistics (e.g., t, F) are highly related to sample size (larger N -> more extreme statistic)
- It is important that any measure of effect size be independent of the size of the sample

Characteristics of Effect Size

- Following Nakagawa and Cuthill, we can outline the following characteristics of effect size
  - Dimension
    - Abstract conceptualization regarding the effect of interest
      - E.g., “Difference in Central Tendencies” is a dimension (that could be measured by mean difference, median difference, trimmed mean difference, etc.)
  - Measure/Index
    - The operational definition of the dimension
      - E.g., “standardized mean difference” could be the measure of differences in central tendency
  - Value
    - The raw value calculated from the measure
      - E.g., the standardized mean difference is .5

Are Effect Sizes Descriptive or Inferential?

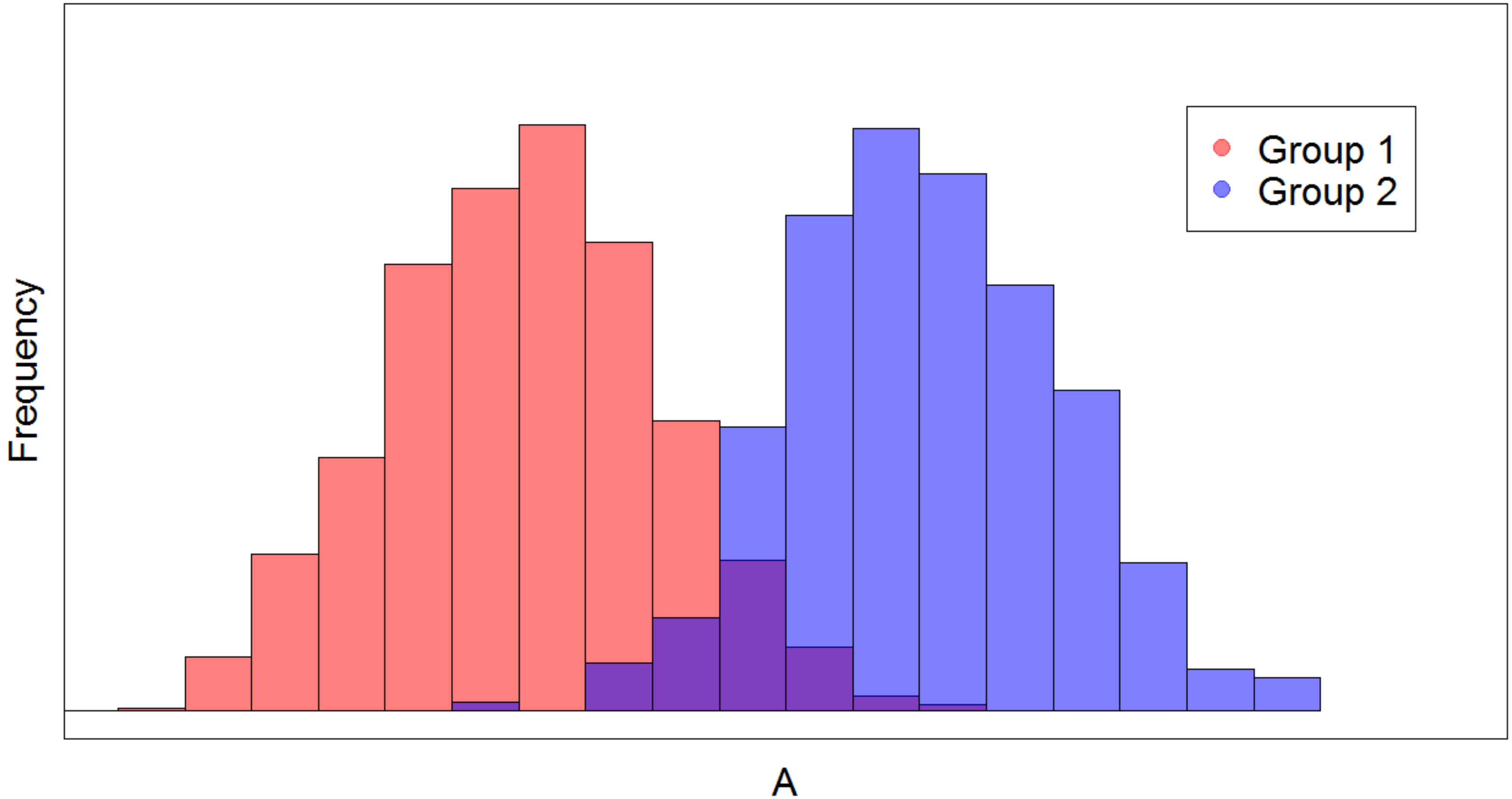
- Effect sizes quantify sample information, so the clearest answer is that an effect size value is descriptive
- However, since the sample effect size is an estimate of the population parameter, can we think of effect sizes as inferential?
  - A *t*-test is calculated on a sample, however we try to make inferences regarding mean differences in the population, so this could also apply to effect sizes

Should Effect Sizes be Standardized or Unstandardized

- One of the most interesting debates regarding effect sizes is whether unstandardized or standardized effect sizes are most useful
  - When the units of measurement are meaningful, many researchers recommend unstandardized effect sizes
    - E.g., Canadians spend 3 hours less a week watching TV relative to Americans
    - E.g., In academia, males earn $5000 more than females for the same work
  - However, don't we need to know something about the variability?

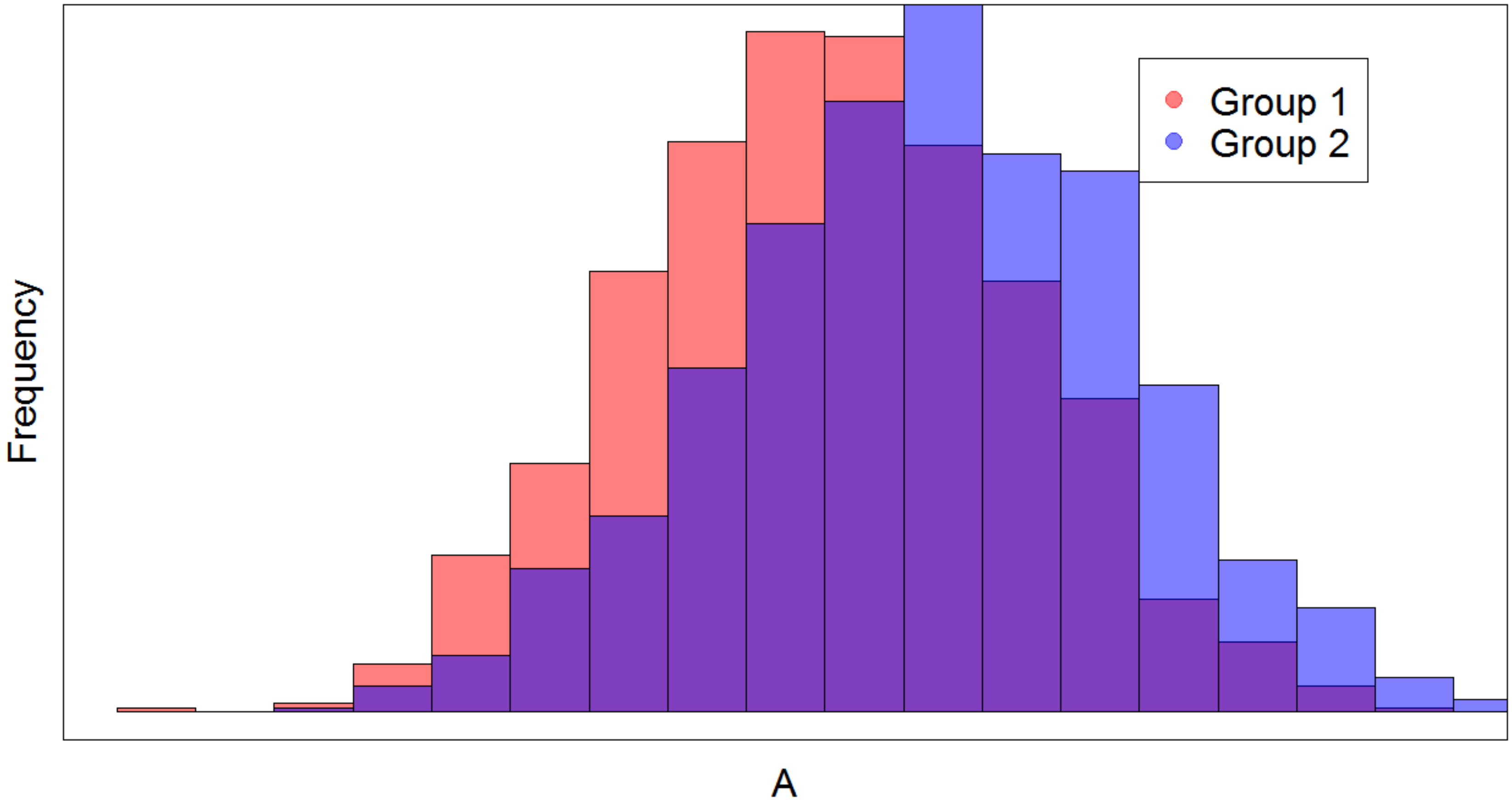
Mean Difference = 3, SDs = 1

- Seems clear here that the groups differ meaningfully



Mean Difference = 3, SDs = 5

- But what about here? Same raw mean difference



Example: Unstandardized vs Standardized Effect Sizes

- Multiple Regression
  - Effect of Years of Education (SD = 3) and Years on the Job (SD = 10) on Income
- Unstandardized Coefficients
  - Income' = 1995 + 201*YrsEd + 100*YrsJob
    - One more year of education increases income by about $200, where one more year on the job increases income by about $100
- Standardized Coefficients
  - YrsEd = .45, YrsJob = .95
    - One SD increase in years of education (~ 3 years) increases income by about a half a SD, where one SD increase in years on the job (about 10 years) increases income by about a SD
- In the first case YrsEd has the larger effect size, whereas in the second case YrsJob has the larger effect size
  - Which is more appropriate to use in this situation?

Are Effect Sizes More Useful for Omnibus or for Specific Effects?

- We often observe that researchers provide an effect size for an omnibus test (e.g., effect size for a one-way ANOVA with 4 groups), but do not provide effect sizes for the specific comparisons of the means of each of the groups (e.g., Grp 1 vs Grp 2)
  - Another example would be providing an effect size for a multiple regression model (e.g., R^2) instead of for each predictor
- It is more important to provide effect sizes for targeted effects than for omnibus effects

Types of Effect Size Measures

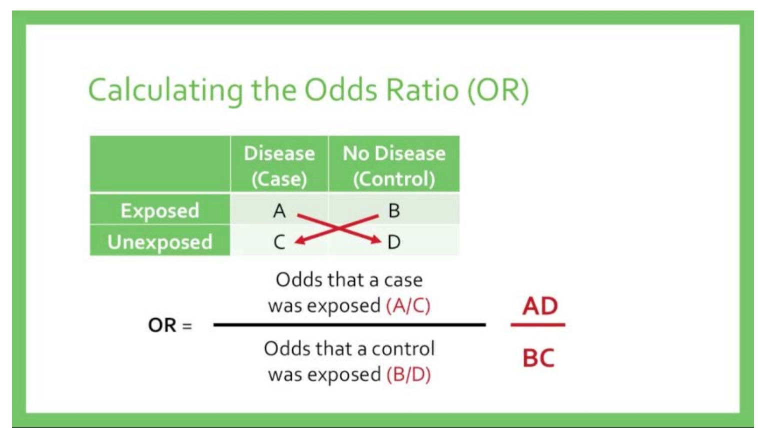
- Correlation/Percent of Variance Explained
  - r/r^2
    - Also encompasses partial/semi-partial r, which are popular in multiple regression
    - Pratt Indices – relative importance of predictors in linear models, in a correlation metric
  - η^2/ω^2
    - Biased/less biased estimates of the proportion of variability in the outcome that is explained by a predictor
      - Partial versions of η^2 and ω^2 are also available for multiple predictor models, however much caution should be used in interpreting these statistics
  - f^2
    - Generally used for omnibus F tests

Types of Effect Size Measures

- Mean Difference
  - $M_1 - M_2$ (unstandardized)
  - $d = \frac{M_1 - M_2}{s}$ (standardized)
    - There are also variations, such as Hedges *g* (e.g., *g* is better with very small N)
    - s can vary depending on what measure of variability you feel is most appropriate for standardization

- Regression Coefficients
  - b (unstandardized) – change in DV for 1 unit change in predictor
  - β (standardized) – SD change in DV for 1 SD change in predictor

- Categorical Relations
  - Odds Ratio
  - Relative Risk (probability metric)
  - Cramer's V $= \sqrt{\frac{\chi^2}{N[\min(r,c)-1]}}$
    - can be interpreted like a correlation



Types of Effect Size Measures

- Common Language Effect Size Estimators
  - Differences Among Two Groups
    - The probability that a randomly selected score from the one population will be greater than a randomly sampled score from the other population
    - CLES $= \Phi(z) = \Phi\left(\frac{d}{\sqrt{2}}\right) = \Phi\left(\frac{M_1 - M_2}{s_p\sqrt{2}}\right)$ $\quad s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$
      - Φ = lower tail probability under the standard normal distribution
    - E.g., Heights of men (M = 69.7, SD = 2.8) and women (M = 64.3, SD =2.6)
      - $\Phi\left(\frac{M_1 - M_2}{s_p\sqrt{2}}\right) = \Phi\left(\frac{69.7-64.3}{2.7\sqrt{2}}\right) = \Phi(1.41) = .92$
      - Thus, there is a 92% chance that the male will be taller than the female if each are randomly drawn from their population

Types of Effect Size Measures

- Common Language Effect Size Estimators
  - Correlation
    - Assume that we have randomly sampled two individuals' scores on X and Y
    - If individual one is defined as the individual with the larger score on X, then the CL statistic is the probability that individual one also has the larger score on Y
    - CLES $= \frac{\sin^{-1}(r)}{\pi} + .5$
    - E.g., father's and son's heights have r = .4
      - CLES $= \frac{\sin^{-1}(.4)}{\pi} + .5 = .63$
      - If father A is taller than father B, there is a 63% chance that son A will be taller than son B

Confidence Intervals for Effect Sizes

- Although effect sizes are extremely meaningful on their own, without a CI we have no information regarding the precision of the effect size
  - Without a measure of precision, of what value is an effect size?
    - E.g., say we measured depression for two males and two females and calculated Cohen's d to be .4
    - Would you be confident in reporting that effect size to others? Could we make any inferences to the population of males and females?
      - The CI could be {.39, .41} or {-2,4}

Confidence Intervals for Effect Sizes

- CI for Cohen's *d*
  - Noncentral *t* distribution for one sample
  - $t = \frac{M-\mu_0}{s_M} + \frac{\mu-\mu_0}{\sigma_M} = \frac{M-\mu_0}{s/\sqrt{N}} + \frac{\mu-\mu_0}{\sigma/\sqrt{N}} = \frac{M-\mu_0}{s_M} + ncp$
    - *ncp* = noncentrality parameter
  - The left part of the equation is the usual *t* (central *t*) statistic, whereas the right part is the *ncp*
  - The *ncp* "shifts" the distribution right or left depending on the sign of $\mu - \mu_0$
  - We know that the population value of d (let's call it d*) is $\frac{\mu-\mu_0}{\sigma}$, so $d^* = ncp\sqrt{N}$
    - If can get the CI for the *ncp*, we can easily find the CI for d

Confidence Intervals for Effect Sizes

- CI for *ncp*
  - To find the CI for *ncp*, we are looking for the values of *ncp* that cutoff the lower α/2 and upper α/2 from the noncentral *t* distribution
    - This is messy, so let's cheat and use R
  - pt(t,df,ncp) can be used to find the value of *ncp* that cutoff the upper and lower tails
  - E.g., let say we want to know if the depression scores for a group of prison inmates differ from 10 (a previously published value). We sample 20 inmates, and M = 9.2, s = 1.4
    - $t = \frac{9.2-10}{1.4/\sqrt{20}}$ =-2.55, d $= \frac{9.2-10}{1.4}$ =-.57

Confidence Intervals for Effect Sizes

- CI for *ncp* … by trial and error
  - Lower tail
    - > pt(-2.56,df=19,ncp=-1)
      - [1] 0.07886299
    - > pt(-2.56,df=19,ncp=-.45)
      - [1] 0.0271863
    - > pt(-2.56,df=19,ncp=-.41)
      - [1] 0.02493609
  - Upper Tail
    - > pt(-2.56,df=19,ncp=-2)
      - [1] 0.3136319
    - > pt(-2.56,df=19,ncp=-4.75)
      - [1] 0.9798424
    - > pt(-2.56,df=19,ncp=-4.65)
    - [1] 0.9749218

Confidence Intervals for Effect Sizes

- Thus, the 95% CI for *ncp* is:
  - −4.65,−.41

- CI for d
  - $d_{lower} = ncp_{lower} / \sqrt{N} = -4.65/\sqrt{20} = -1.04$
  - $d_{upper} = ncp_{upper} / \sqrt{N} = -.41/\sqrt{20} = -.09$
  - 95% CI for d = {−1.04, −.09}

Confidence Intervals for Effect Sizes

- CI for d
  - An easier way to compute the CI for *d* is to use a built-in function from R

```
>library(psych)
> d.ci(-.57,n1=20)
     lower          effect         upper
[1,] -1.037391 -0.57 -0.08981953
```

Confidence Intervals for Effect Sizes

- CI for d
  - A more general way, that works for most statistics, is to bootstrap
    - E.g., we can do this via a *for* loop in R

  - Depression Example:

```
> dep_bs_d<-numeric(1000)
> for (i in 1:1000) {
>    samp<-sample(dep, replace=TRUE)
>    dep_bs[i]<-(mean(samp)-10)/sd(samp)
> }
> quantile(dep_bs,c(.025,.975))
          2.5%          97.5%
          -1.0181445 -0.1685463
```

Summary

- Confidence Intervals should be included for each effect of interest
- Effect size values should be included for each effect of interest
  - Effect sizes must be scaled appropriately, given the measurement and the question of interest
  - The point estimate of the population effect size value should be independent of sample size
  - Effect size values should be accompanied with confidence intervals
  - Estimates of effect sizes values should have desirable estimation properties; namely, they should be:
    - unbiased (their expected values should equal the corresponding population values)
    - consistent (they should converge to the corresponding population value as sample size increases)
    - efficient (they should have minimal variance among competing measures)