Language Technologies Institute

Carnegie Mellon University

Multimodal Affective Computing

Lecture 7: Statistical Foundations

Jeffrey Girard
Louis-Philippe Morency

Outline of this week's lecture

1. Exploratory data analysis
2. Statistical hypothesis testing
3. Point estimation and effect sizes
4. Interval estimation and confidence intervals

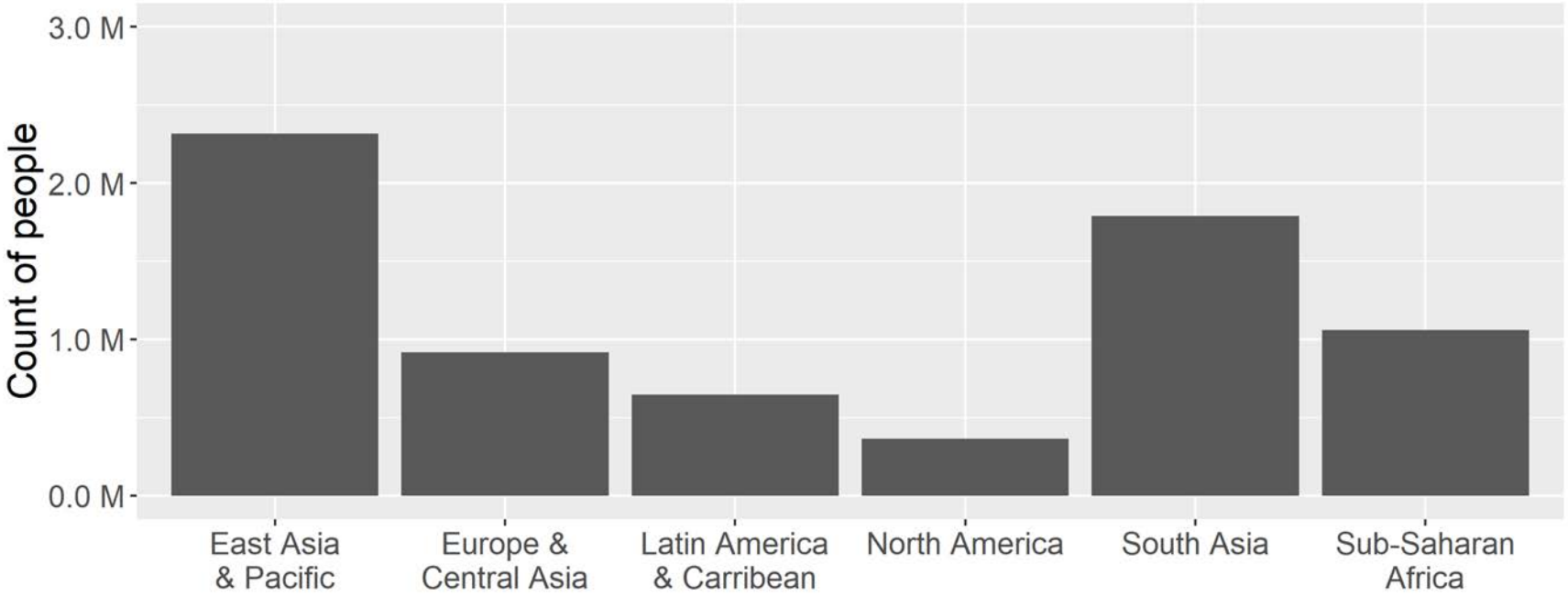
Exploratory data analysis

Exploratory data analysis

- The goal is to better understand your data
  - How are the variables distributed? Any surprises?
  - Are there missing, impossible, or outlier values?
  - Do the variables seem to be associated?
- May help detect and correct errors
- May suggest analytical approaches
- May inspire new research hypotheses

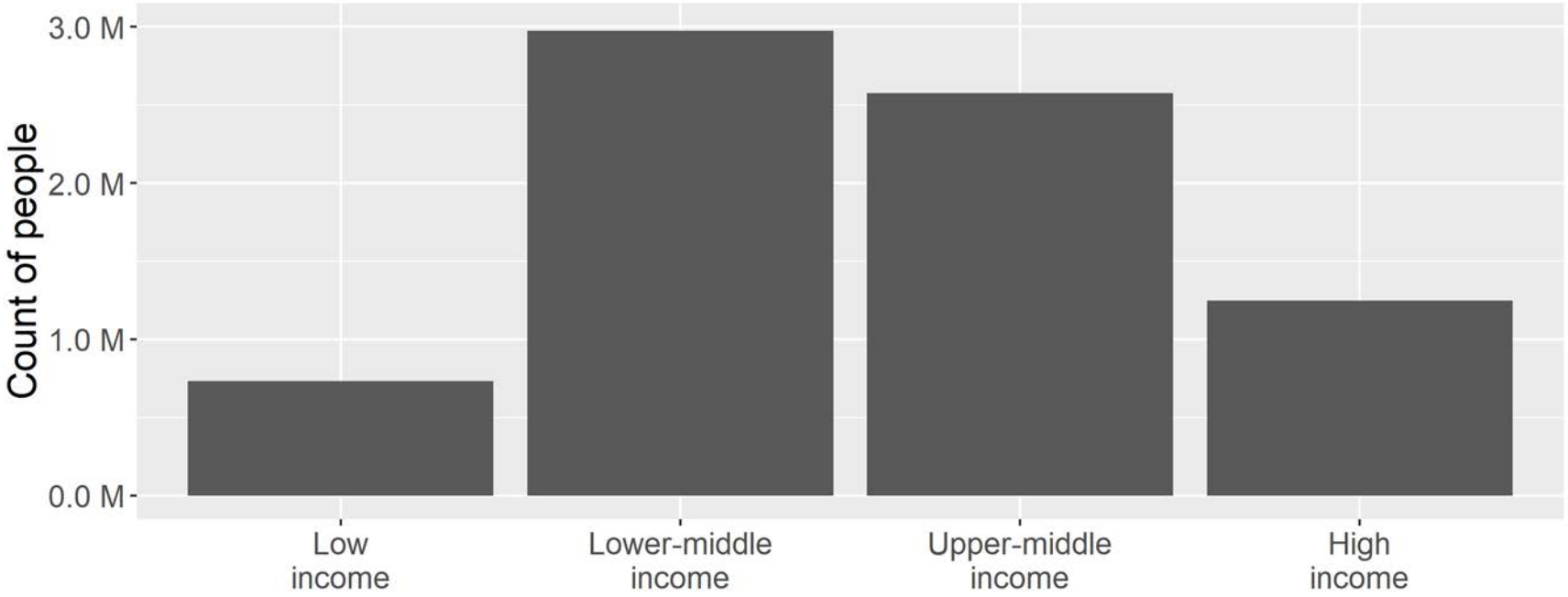
Visualizing distributions

- A categorical variable's distribution represents the size (frequency, count, probability) of each unique category



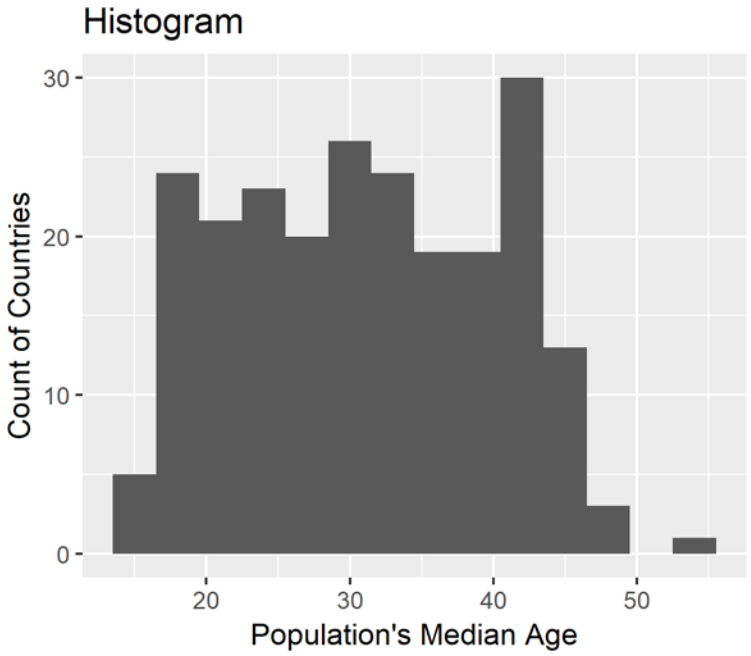
Visualizing distributions

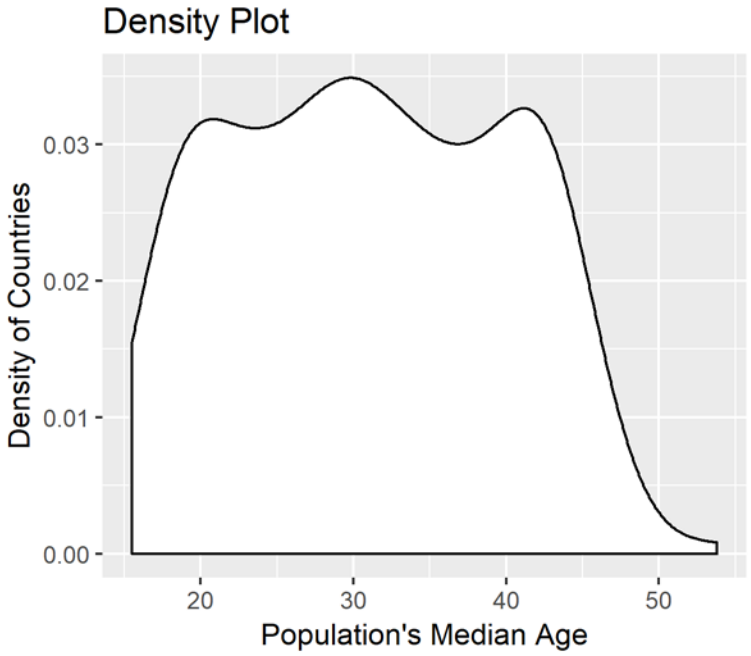
- A categorical variable's distribution represents the size (frequency, count, probability) of each unique category



Visualizing distributions

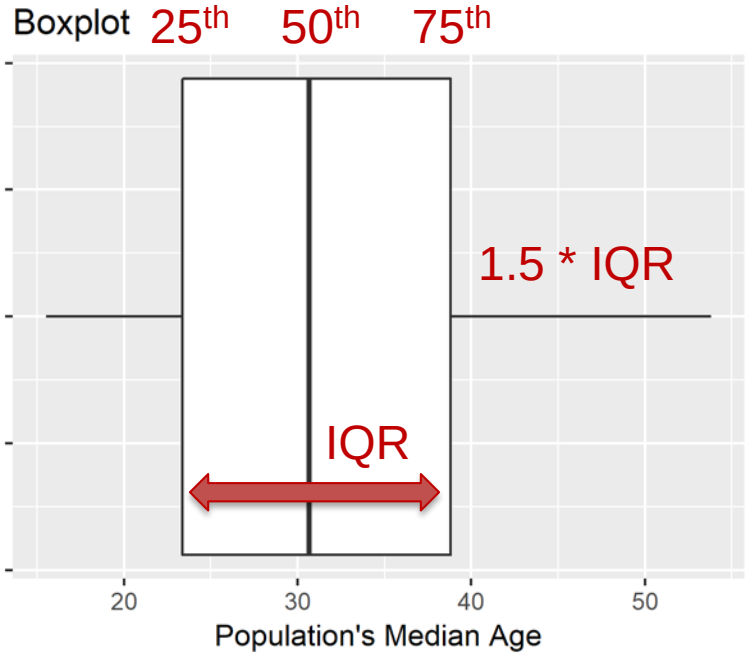
- A continuous variable's distribution represents the size (count, frequency, probability) of different value ranges



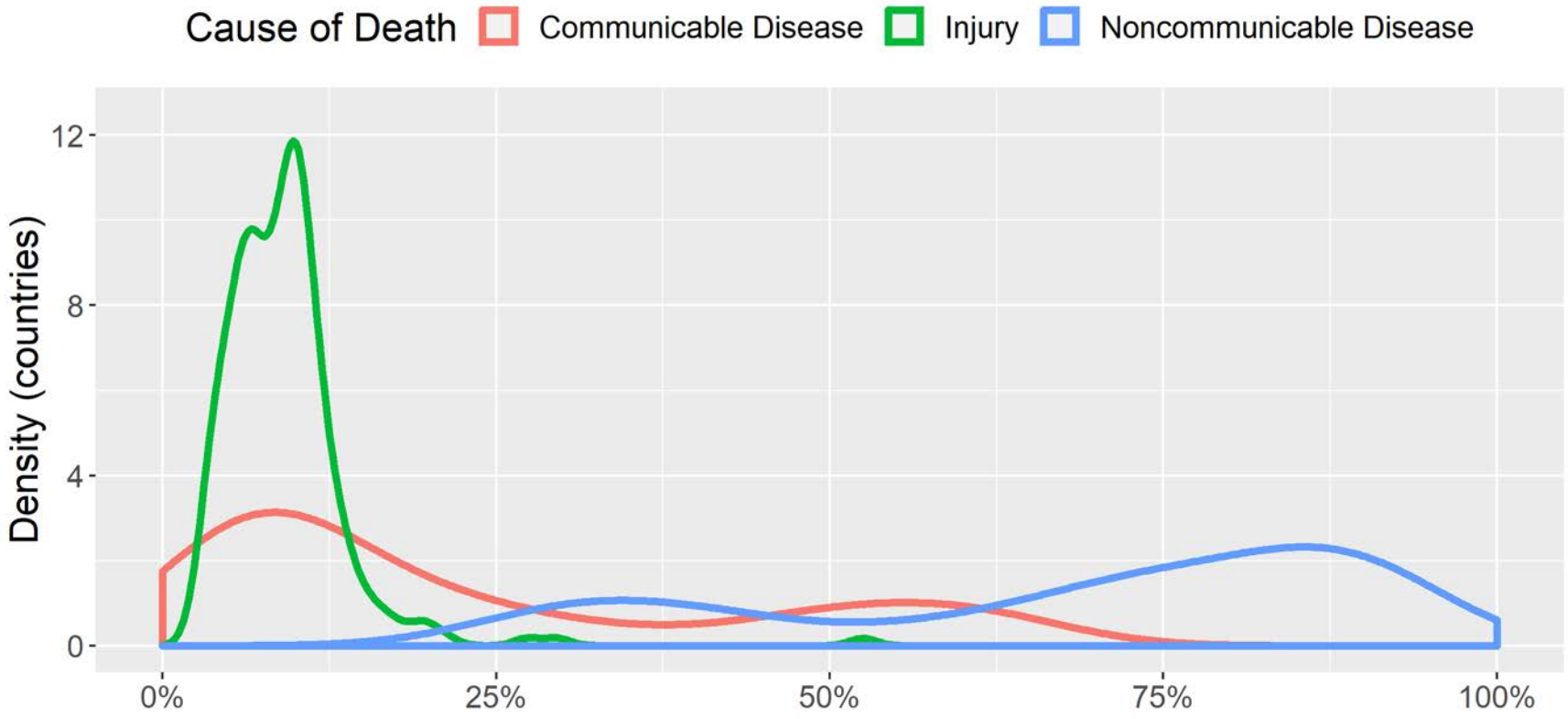


Visualizing distributions

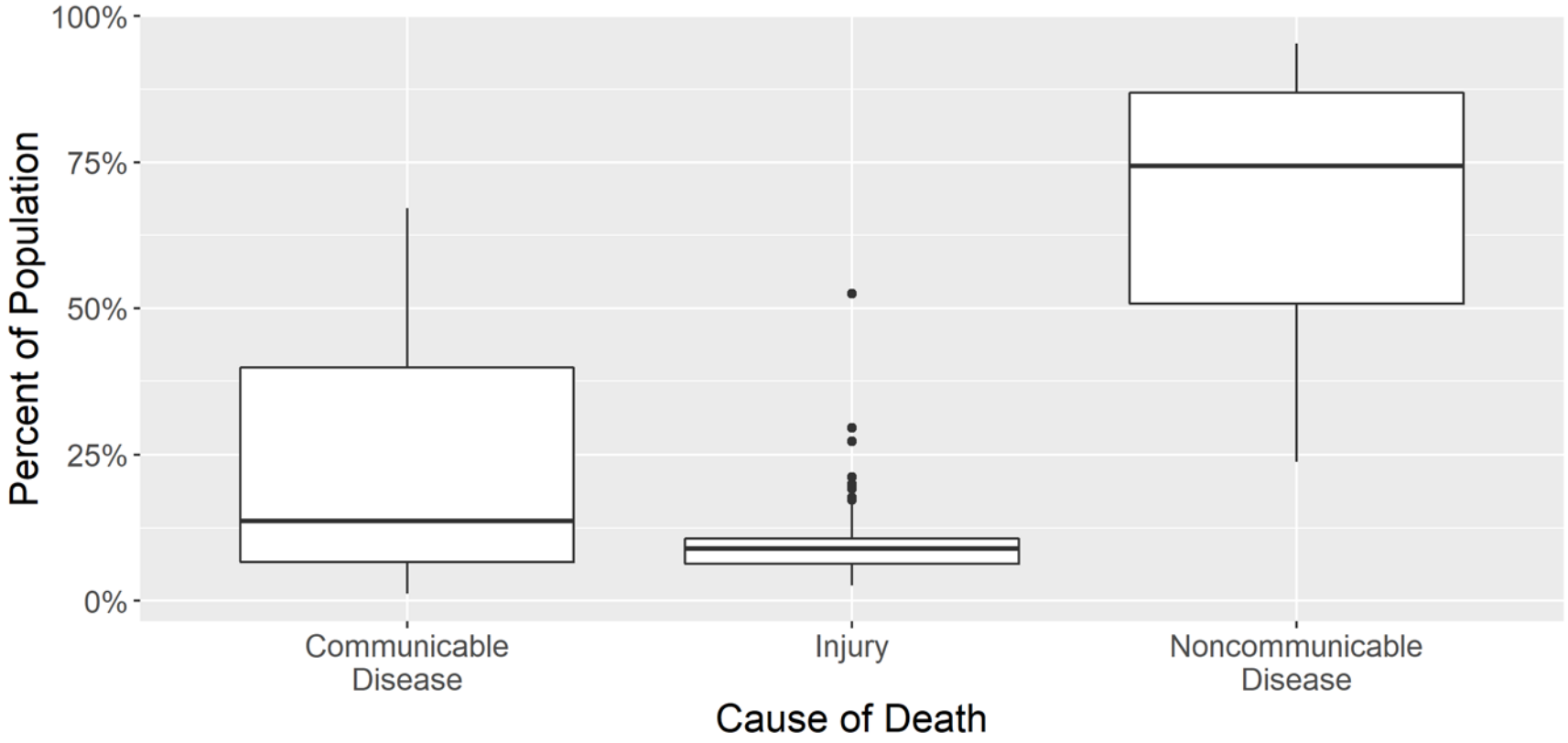
- A continuous variable's distribution describes the count, frequency, or probability of different value ranges



Comparing distributions



Comparing distributions



Summary statistics

- A summary statistic describes some aspect of a variable's distribution and can be viewed quickly
- For categorical variables, the most important thing is to represent the size of each unique category

```
# A tibble: 6 x 3
  Area                       Count Percent
  <fct>                      <dbl>   <dbl>
1 East Asia & Pacific      2314365   0.327
2 Europe & Central Asia     915546   0.129
3 Latin America & Carribean 644138   0.0909
4 North America             362493   0.0512
5 South Asia               1788389   0.252
6 Sub-Saharan Africa       1061108   0.150
```

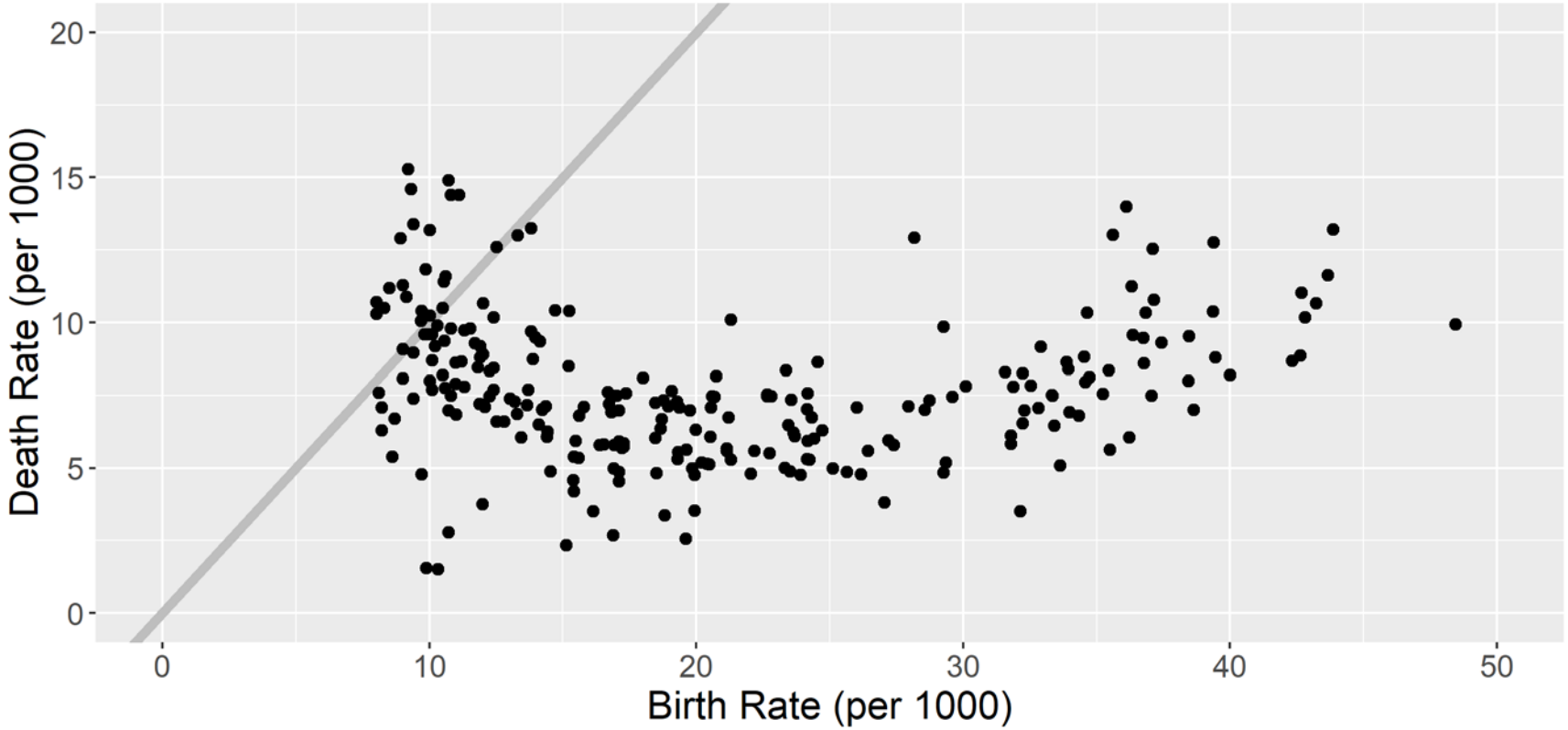
```
# A tibble: 4 x 3
  Income                 Count Percent
  <ord>                  <dbl>   <dbl>
1 Low income            732449  0.0973
2 Lower-middle income  2972643  0.395
3 Upper-middle income  2576203  0.342
4 High income          1249066  0.166
```

Summary statistics

- A summary statistic describes some aspect of a variable's distribution and can be viewed quickly
- For continuous variables, you want to represent the number, range, central tendency, spread, and shape

```
# A tibble: 4 x 11
  var              n_c   n_m   min   max median   iqr  mean    sd  skew  kurt
  <chr>          <int> <int> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 PopMedianAge     228     0  15.5  53.8   30.6  15.4  31.0  8.96  0.06  1.89
2 DeathByInjury    229    35   2.6  52.5    8.98  4.4   9.11 4.82  3.92 32.1
3 DeathByCommD     229    35   1.2  67.1   13.7  33.2  22.9  20.3   0.87  2.24
4 DeathByNcommD    229    35  23.8  95.3   74.3  36.1  68    21.8  -0.65  2.01
```

Exploring relationships



Statistical hypothesis testing

Hypotheses and effect sizes

- A hypothesis is a testable prediction about a phenomenon
  - Our smile detection algorithm will perform better than chance
  - Women tend to be more facially expressive than men
  - Ratings of rapport will be associated with fewer interruptions
- An effect size is a measure of a phenomenon's magnitude
  - The average smile performance for our algorithm
  - The difference in expressivity between men and women
  - The association of rapport and interruptions during interactions
  - (Note that there are many other types of effect sizes as well)

Populations and population parameters

- Hypotheses are usually about specific population parameters
- A <u>population</u> is the group we want to know the effect size in
  - The group of all possible face images
  - The group of all men and women
  - The group of all social interactions
- A <u>population parameter</u> is the effect size in that population
  - The average smile performance for all possible face images
  - The difference in expressivity between all men and women
  - The association of rapport and interruptions for all social interactions
- Population parameters are represented as Greek letters (μ, σ)

Samples and sample statistics

- We can't measure an *entire* population so we use samples
- A sample is a subset that we draw from a population
  - A subset of 30,000 face images from Flickr
  - A subset of 50 male and female students
  - A subset of 100 web calls from Skype
- A sample statistic is the effect size in the sample
  - The average smile performance for 30,000 face images from Flickr
  - The difference in expressivity between 50 male and female students
  - The association of rapport and interruptions for the 100 Skype web calls
- Sample statistics are represented as Latin letters ($\bar{x}$, s)

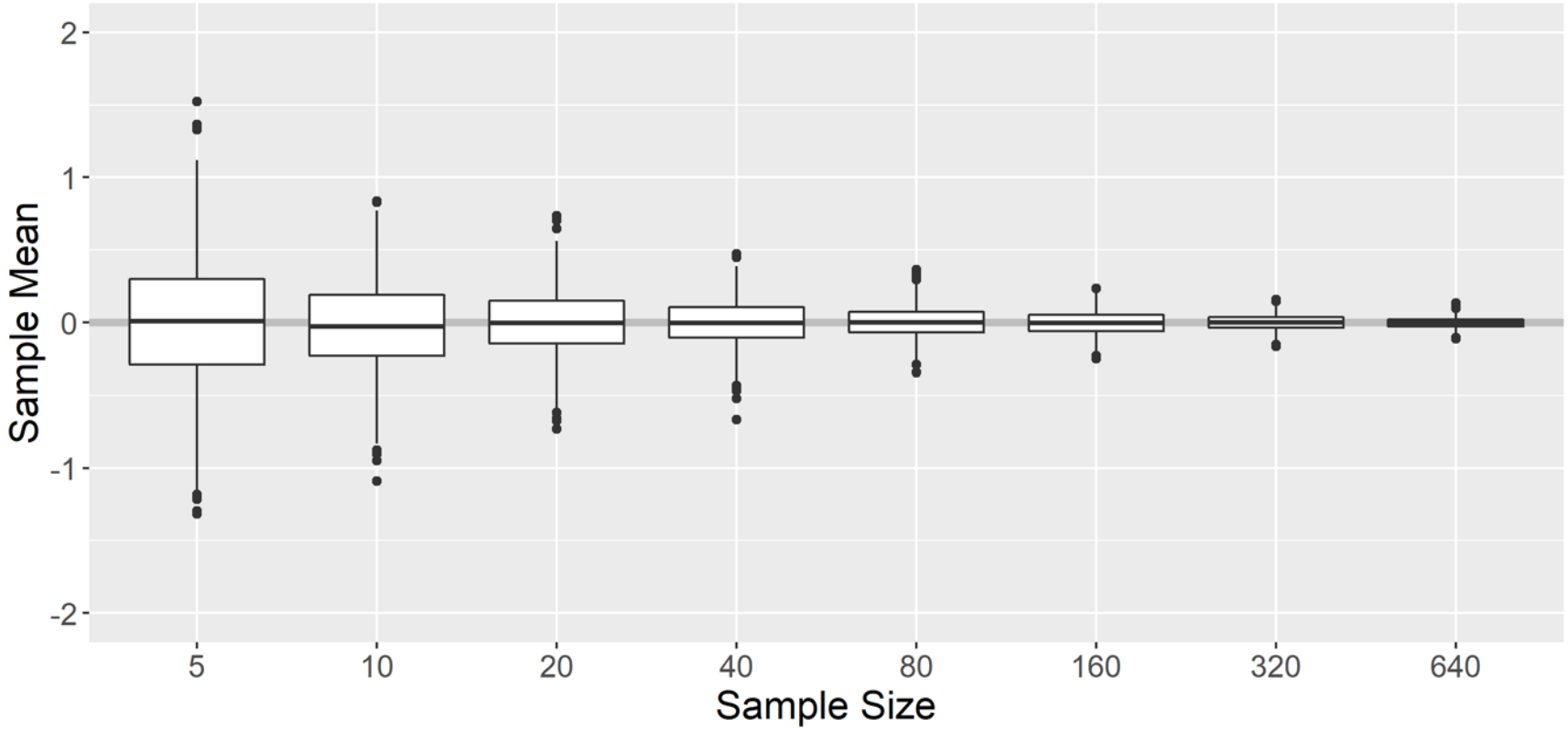
Estimation and representativeness

- The sample statistic *estimates* the population parameter
- The population parameter is what we truly care about
  - The average smile performance for all possible face images
  - The difference in expressivity between all men and women
  - The association of rapport and interruptions for all social interactions
- The sample statistic is what we actually have
  - The average smile performance for the 30,000 face images from Flickr
  - The difference in expressivity between 50 male and female students
  - The association of rapport and interruptions for the 100 Skype web calls
- Generalizing from sample to population depends on this match

Sampling error and bias

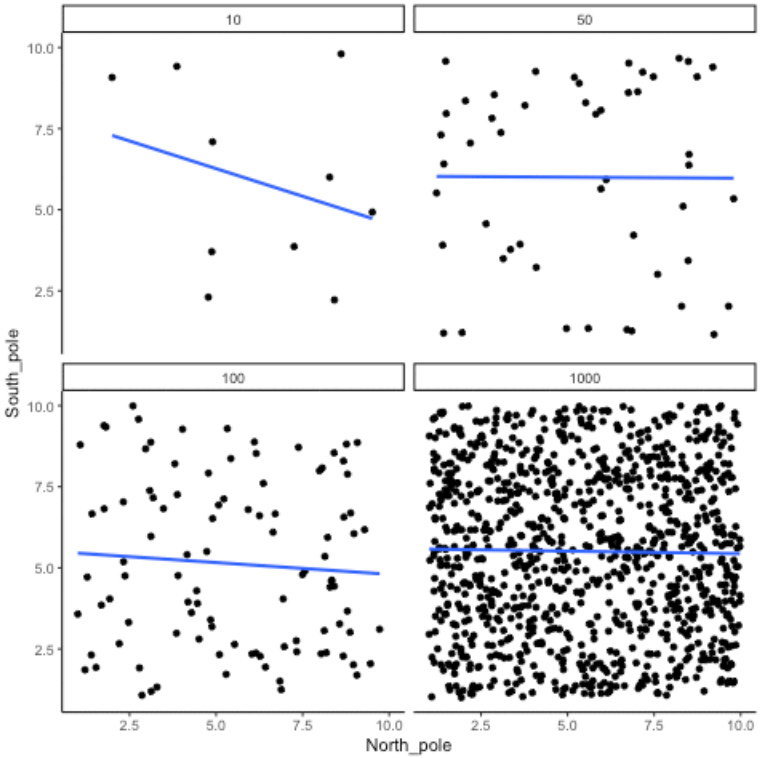
- Sampling error comes from imperfect estimation
  - Sample Statistic ≠ Population Parameter (e.g., $\bar{x} \neq \mu$)
- Sampling error can be stochastic or systematic
  - Stochastic error comes from random chance in sampling
  - Systematic error comes from bias in the sampling process
- To improve estimation, we want to minimize sampling error
  - Stochastic error can be reduced with larger samples
  - Systematic error can be reduced with better samples
  - We want samples to be "representative" of populations

Sampling error for means

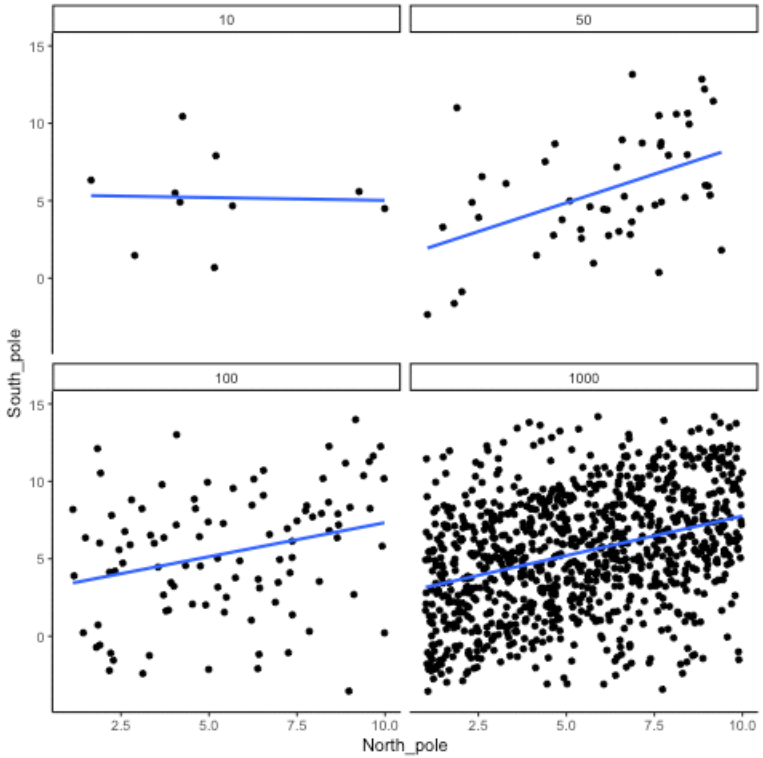


Samples drawn randomly from a standard normal distribution

Sampling error for associations



Sample associations between two variables with 0 relation in population



Sample associations between two variables with + relation in population

Statistical hypothesis testing

- To test hypotheses, we compare effect sizes to predictions
  - Is the average smile performance actually better than chance?
  - Are women actually more facially expressive than men?
  - Is rapport actually associated with the number of interruptions?
- But effect size estimates are influenced by sampling error
- How do we decide if the predictions were supported or not?
- Statistical hypothesis testing tries to answer this question

Statistical hypothesis testing

- To evaluate a hypothesis, we need two pieces of information

1. What is the magnitude of the effect?
   - *How large do we think the effect is?*
   - What was the sample mean, difference, association, etc.?
2. What is the precision of the estimate?
   - *How confident are we in our estimate?*
   - How much sampling error is there likely to be?
   - How much variability was present in the data?

Paradigms for hypothesis testing

- A common approach to hypothesis testing uses p-values
  - A p-value combines magnitude and precision into one number
  - If p-value is less than α (e.g., 0.05) it is statistically significant
  - Having one number is easy but hides some of the information
  - It would be nice to know magnitude and precision separately
- We will instead use an approach that keeps them separate
  - Magnitude can be represented by a point estimate
  - Precision can be represented by an interval estimate
  - This is sometimes called the "New Statistics" approach
  - It still evaluates statistical significance but also much more

Point estimation and effect sizes

Common measures as effect sizes

- An effect size measures the magnitude of an effect
- A point estimate is a single best-guess of the effect size
- There are many different types of effect size measures
  - Some use original units and others use standardized units
  - Some are signed (+/−) and others are unsigned
  - Some are bounded within a range and others are unbounded
- Some common measures you already know are effect sizes
  - Statistical moments: *mean, variance, skewness, kurtosis, etc.*
  - Performance metrics: *accuracy, F_1, AUC, MAE, RMSE, etc.*

The difference family of effect size measures

- Difference measures (d_*) compare the mean of two groups
  - Comparisons may be between different groups (e.g., men vs. women)
  - Comparisons may be within the same group (e.g., before vs. after)
  - Difference measures are signed, standardized, and unbounded
- There are many variants with different denominators

Parameter $\delta = \frac{\mu_1 - \mu_2}{\sigma^*}$ Statistic $d = \frac{\bar{x}_1 - \bar{x}_2}{s^*}$

d_{pool}	d_{s1}	d_{s2}	d_{Δ}
$s^* = s_{\text{pool}}$ Weighted mean of SDs	$s^* = s_1$ SD of group 1	$s^* = s_2$ SD of group 2	$s^* = s_{\Delta}$ SD of paired differences

The difference family of effect size measures

- Worked example for two *independent* samples:
  - x_1 are the patient group's scores, $\mathrm{n}_1 = 30,\ \bar{x}_1 = 27.0,\ s_1 = 9.1$
  - x_2 are the control group's scores, $\mathrm{n}_2 = 40,\ \bar{x}_2 = 23.7,\ s_2 = 12.8$

$$s_{\text{pool}} = \sqrt{\frac{(30-1)9.1^2 + (40-1)12.8^2}{(30-1)+(40-1)}} = \sqrt{\frac{8791.3}{68}} = 11.4$$

$$d_{\text{pool}} = \frac{27.0 - 23.7}{11.4} = 0.29 \qquad d_{s1} = \frac{27.0 - 23.7}{9.1} = 0.36 \qquad d_{s2} = \frac{27.0 - 23.7}{12.8} = 0.26$$

"The mean of the patient group was 0.29 pooled standard deviations greater than the mean of the control group ($d_{\text{pool}} = 0.29$)."

The difference family of effect size measures

- Worked example for two *dependent* samples:
  - x_1 are the before-treatment scores, $n = 25,\ \bar{x}_1 = 96.9,\ s_1 = 8.5$
  - x_2 are the after-treatment scores, $n = 25,\ \bar{x}_2 = 85.6,\ s_2 = 9.4$
  - Paired differences ($x_\Delta = x_1 - x_2$), $n = 25,\ \bar{x}_\Delta = 11.4,\ s_\Delta = 3.8$

$$d_\Delta = \frac{96.9 - 85.6}{3.8} = 3.0$$

"The standardized mean change was a reduction of 3.0 from before treatment to after treatment ($d_\Delta = 3.0$)."

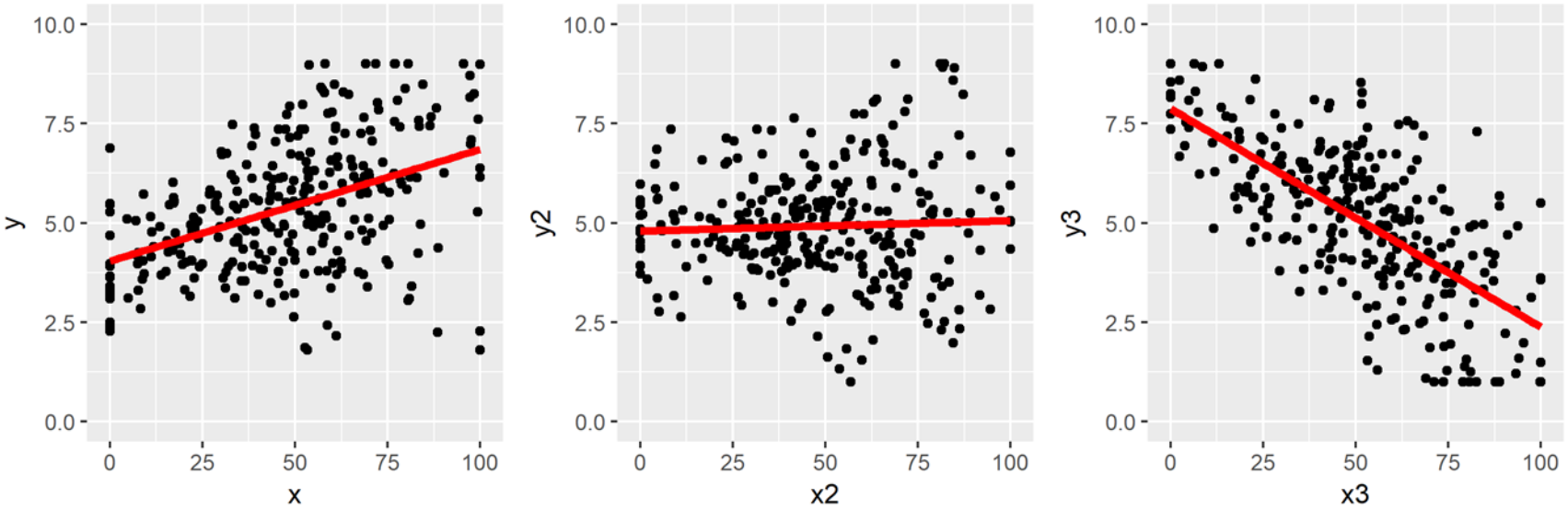
$$d_{s1} = \frac{96.9 - 85.6}{8.5} = 1.3$$

"The pre-treatment mean was 1.3 standard deviations higher than the post-treatment mean ($d_{s1} = 1.3$)."

The association family of effect size measures

- Association measures (r_*) evaluate the relation of variables
  - Covariances are signed, unstandardized, and unbounded
  - Correlations are signed, standardized, and bounded -1.0 to $+1.0$
  - Variance explained is unsigned, standardized, and bounded 0.0 to 1.0
- There are many variants of the association measures
  - r is the correlation between two continuous variables
  - r_{pb} is the correlation between a binary and a continuous variable
  - r^2 is the amount of variance explained by a correlation coefficient
  - R^2 is the amount of variance explained by a multiple regression model
  - η^2, ε^2, and ω^2 measure the variance explained by an ANOVA model

The association family of effect size measures



$r = +0.44$
$r^2 = 0.19$

$r = +0.05$
$r^2 = 0.003$

$r = -0.70$
$r^2 = 0.49$

"The x and y variables were positively correlated ($r = 0.44$); this association explained 19% of their variance ($r^2 = 0.19$)."

Practical advice for using effect sizes

- Selecting effect size measures
  - Pick a measure that matches your research questions/design
  - Unstandardized measures are sometimes easier to interpret
  - Standardized measures are usually easier to compare
- Interpreting effect sizes
  - Do not blindly use "rules of thumb" for effect sizes
  - Consider the research context and look at previous work
  - What effect sizes are common and uncommon in this field?
  - What effect sizes would have practical implications in this field?

Interval estimation and confidence intervals

Confidence intervals

- An interval estimate is a range of plausible effect sizes
- It takes the point estimate and accounts for its precision
- More precise estimates have narrower interval estimates
- A confidence interval is a popular type of interval estimate

$$r = 0.36, 95\% \text{ CI: } [0.34, 0.40]$$

Sample Effect Size

Confidence Level $(1 - \alpha)$

Lower and Upper Bounds

Interpreting confidence intervals

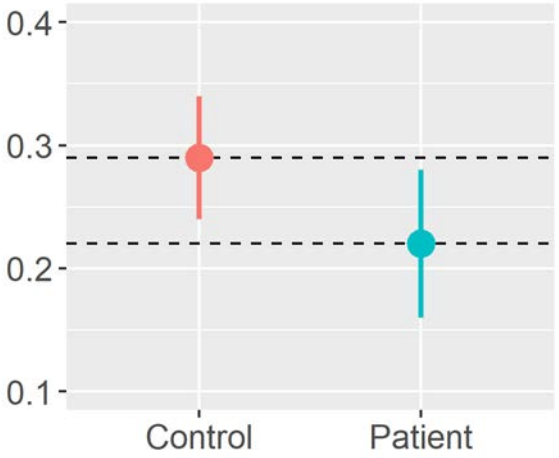
- Correct interpretations of a confidence interval
  - "If the CI is calculated for indefinitely many replications, in the long run, 95% of these intervals will include the population parameter value."
  - "The CI is a set of values that are plausible for the parameter value."
  - "We can be 95% confident that the CI contains the parameter value."
- Incorrect interpretation of a confidence interval
  - "The CI has 95% chance of including the population parameter value."

Presenting confidence intervals

- In text, present the CIs in full and then short format
- In tables, present the CIs in short format in a column
- In figures, present the CIs as error bars (define in caption)

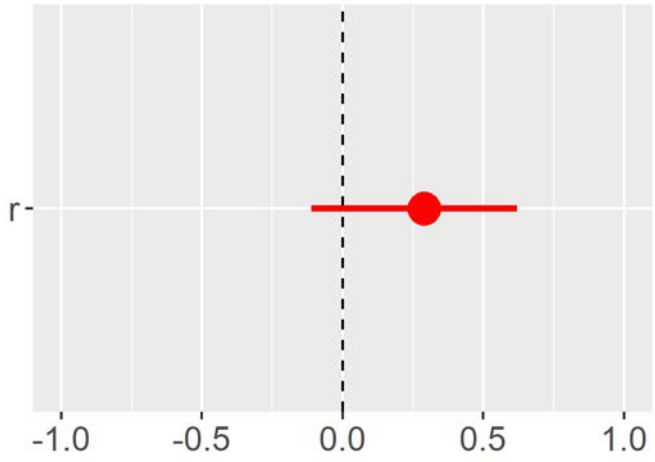
"The mean of the control group was 0.29, 95% CI: [0.24, 0.34] and the mean of the patient group was 0.22 [0.16, 0.28]."

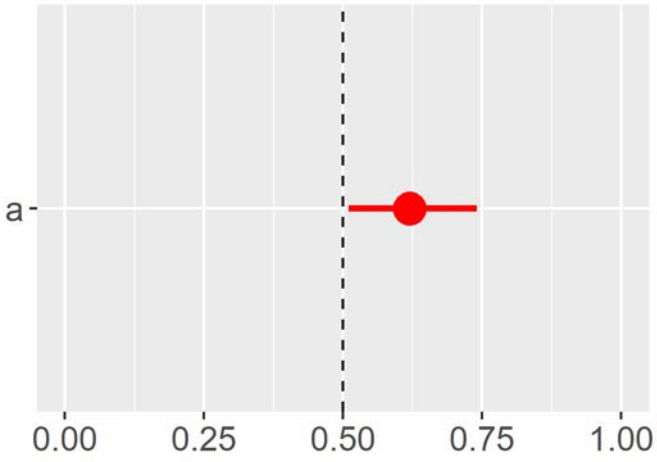
Group	Mean	95% CI
Control	0.29	[0.24, 0.34]
Patient	0.22	[0.16, 0.28]



Significance testing with confidence intervals

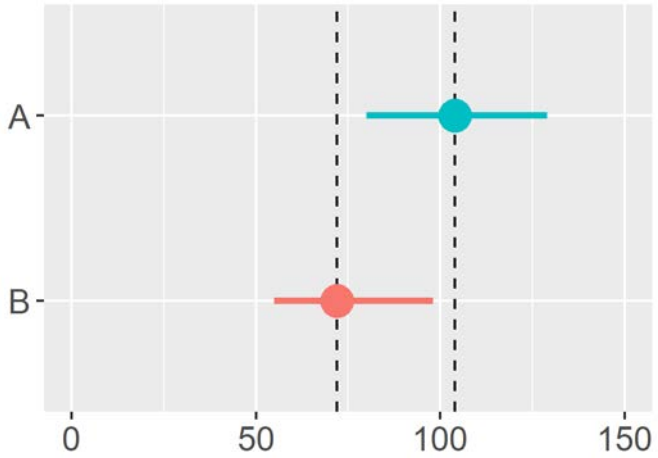
- Comparing a confidence interval to a "reference value"
  - If a reference value is outside of a CI, then it is significantly different
  - When using a 95% CI, significant differences imply that $p < .05$
  - If $r = 0.29, 95\%\ \text{CI}: [-0.11, 0.62]$, is r significantly different from 0.00?
  - If $a = 0.62, 95\%\ \text{CI}: [0.51, 0.74]$, is a significantly different from 0.50?

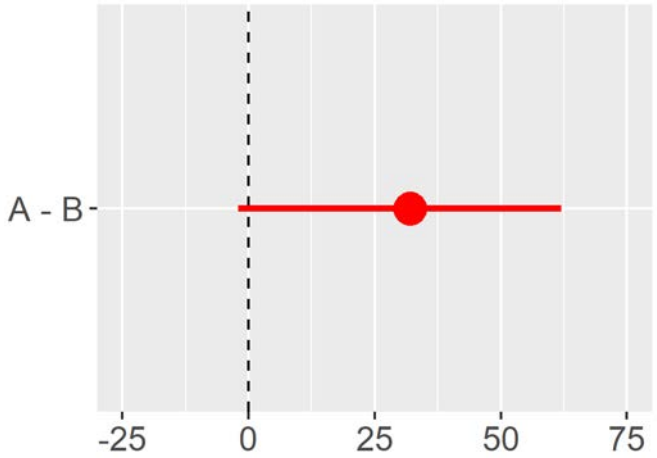




Significance testing with confidence intervals

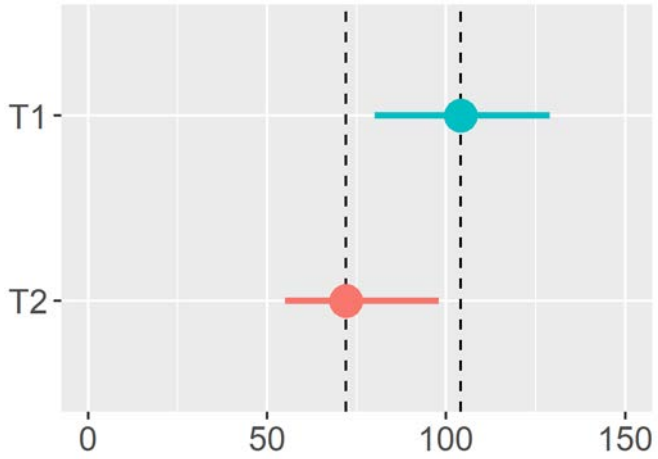
- Comparing two independent confidence intervals
  - If two independent CIs do not overlap, they are significantly different
  - If two independent CIs overlap, they may or may not be significant
  - So it is helpful to calculate a confidence interval for their difference
  - Then you can compare this difference CI to the reference value of 0.0

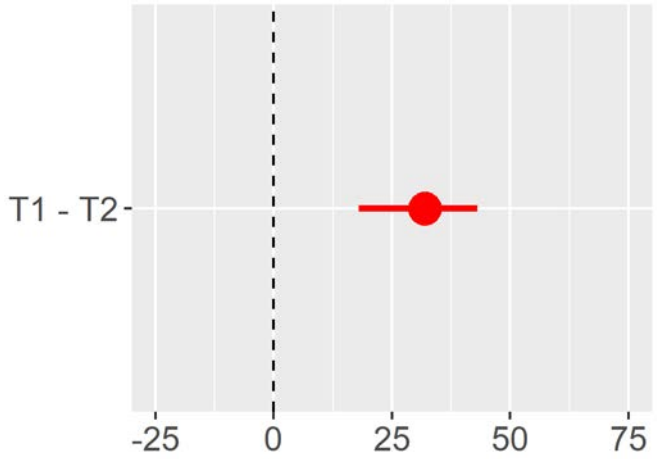




Significance testing with confidence intervals

- Comparing two dependent confidence intervals
  - The overlap of two dependent CIs does not tell you about significance
  - You must calculate a confidence interval for the *paired* differences
  - Then you can compare this difference CI to the reference value of 0.0
  - The standard error is often smaller with dependent samples





Estimating confidence intervals

- To construct a CI, we need to estimate sampling error
  - The standard error is the amount of sampling error we want to estimate
  - What if we sampled n observations from the population many times?
  - The sampling distribution would be the distribution of sample statistics
  - The standard error is the standard deviation of the sampling distribution
- Parametric approaches
  - Assume the sampling distribution is normally distributed
  - Estimate the standard error as a function of sample size and variability

$$\bar{x} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \qquad SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Estimating confidence intervals

- Nonparametric bootstrapping
  - Treat the sample as the population and repeatedly re-sample from it
  - The bootstrap approach resamples n observations with replacement
  - This process is repeated a large number of times (e.g., $b = 2000$)
  - The resampling distribution is the distribution of statistics in resamples
  - This allows us to approximate a sampling distribution with *any* shape
  - Once we have a resampling distribution, we can calculate any CI
  - The percentile approach is one fast and easy way to calculate CIs
  - More complex approaches (BCa and ABC) tend to be more accurate

Pseudo-code for the bootstrap approach

Matrix **X** has n rows (observations)

> Pre-allocate vector **R** with length b

> Repeat b times:

    > Resample n observations from **X** with replacement

    > Calculate the statistic of interest using the resampled data

    > Store the calculated resample statistic in vector **R**

To save time, you can bootstrap multiple statistics at the same time by calculating them all using each resample and saving them all to a matrix **R**

Pseudo-code for the percentile approach

Matrix **X** has n rows (observations)

Vector **R** has b resamples of the statistic of interest

Calculate the observed estimate and 95% CI:

> The **estimate** is the observed sample statistic in **X**

> The **lower bound** of the CI is the 2.5th percentile of **R**

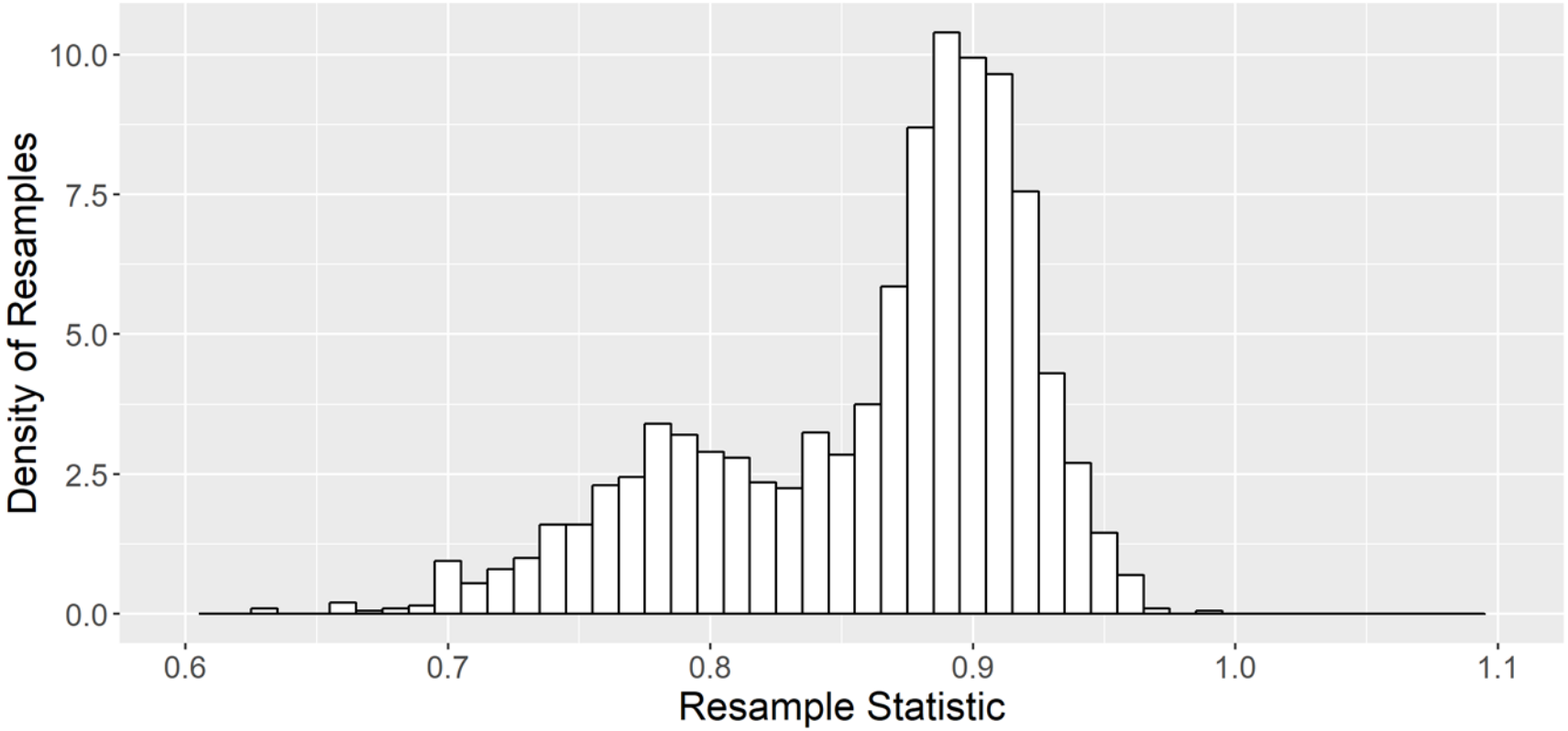
> The **upper bound** of the CI is the 97.5th percentile of **R**

To calculate a confidence level other than 95%

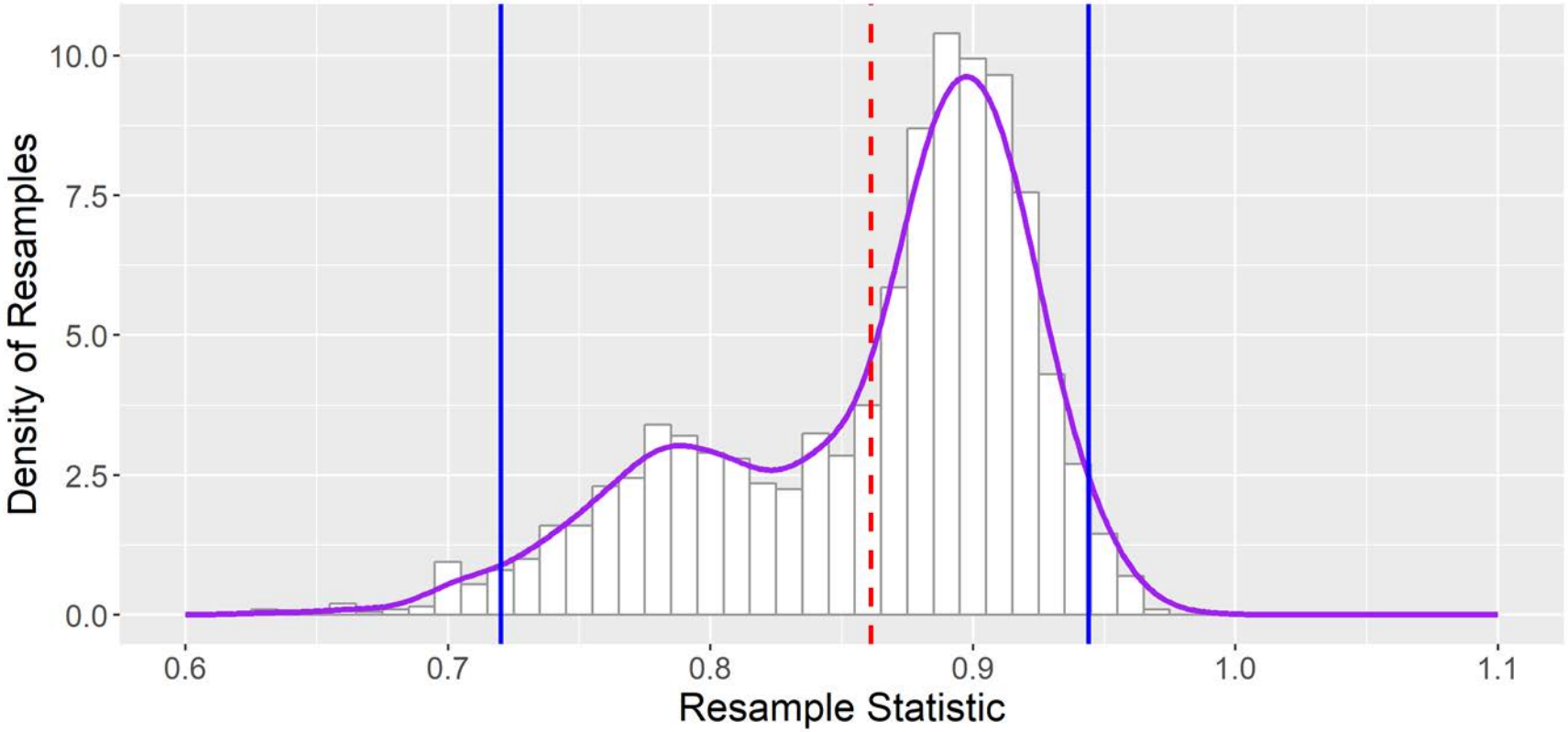
Lower bound = $[(100 - \text{level})/2]^{th}$ percentile

Upper bound = $[(100 + \text{level})/2]^{th}$ percentile

Visualizing the resampling distribution

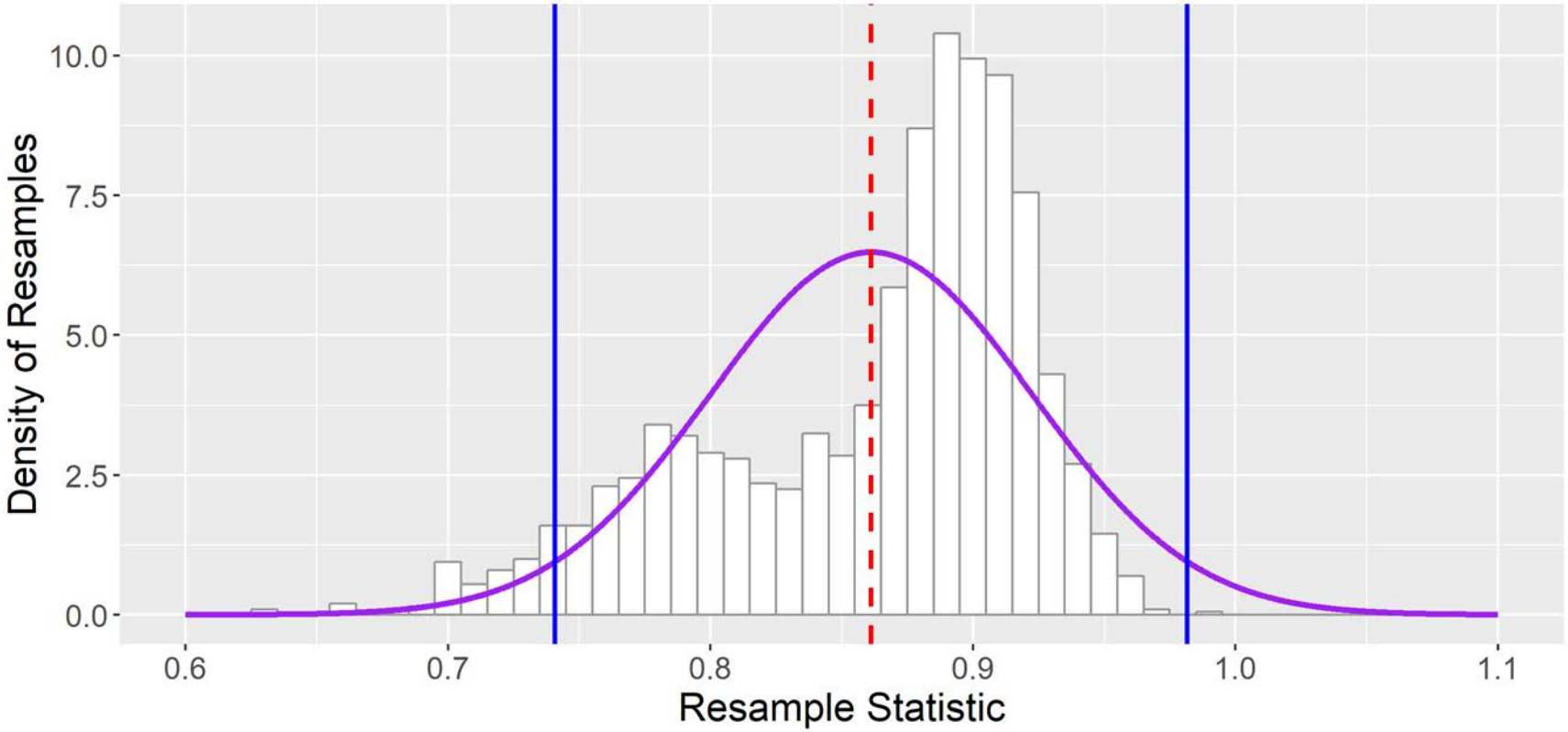


Visualizing the resampling distribution



Percentile Approach: $\text{ES} = 0.86, 95\%\ \text{CI}: [0.72, 0.94]$

Visualizing the resampling distribution



Parametric Approach: $\text{ES} = 0.86, 95\%\ \text{CI}: [0.74, 0.98]$

Further Reading

Optional

- Kline, R. B. (2013). Sampling and estimation. In *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed., pp. 29–65).
- Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *The American Psychologist*, *60*(2), 170–180.
- Cumming, G., & Fidler, F. (2010). Effect sizes and confidence intervals. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 79–92).

Advanced

- Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.).
- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*.
- McElreath, R. (2016). *Statistical rethinking: A Bayesian course in R and Stan.*