BIOS 6312: Modern Biostatistics Methodology II

Andrew J. Spieker, Ph.D.

Associate Professor of Biostatistics
Vanderbilt University

Set 1: Review

Version: 04/26/2025

Table of Contents

Sampling distributions

Population parameters: Fixed and unknown

- Let's start with some of the underlying theory of estimation.
- The goal of many scientific studies is to estimate some population parameter (θ often denotes some general parameter).
  - Mean systolic blood pressure (SBP).
  - Risk of death from SARS-CoV-2 infection.
  - Hazard of cancer recurrence.
  - Incidence rate of flu.
  - Difference in mean HbA1c six months post-intervention.
- We presume these quantities fixed and unknown, but estimable nevertheless.

Sampling distributions

Estimation:

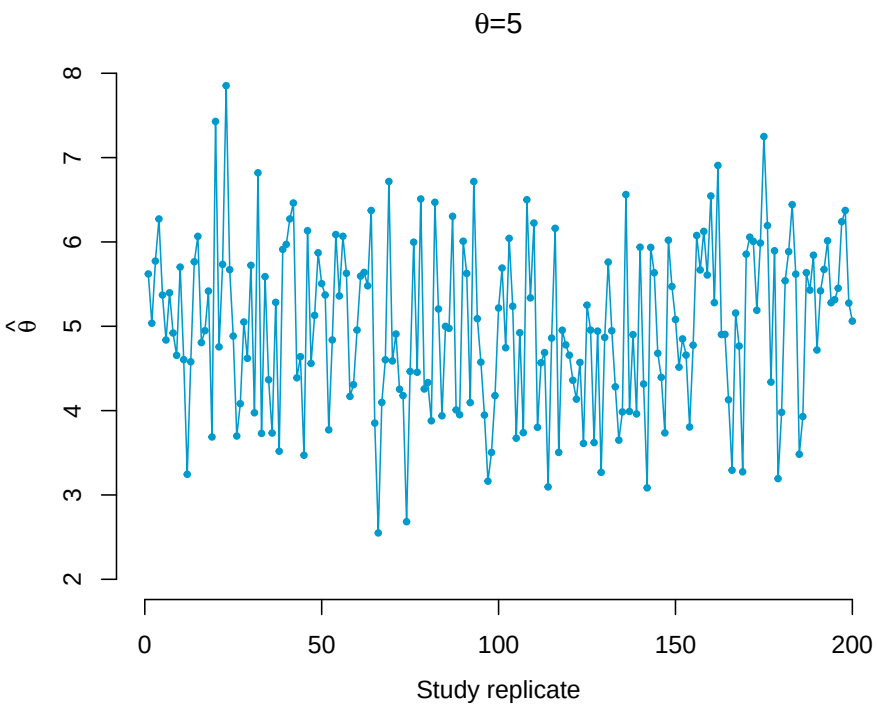
- An *estimator*, $\widehat{\theta}$, is loosely defined as a method of reducing data to a (set of) value(s) intended to stand as a reasonable representation of the unknown θ. There are many estimation approaches, including:
  - Maximum likelihood/maximum partial likelihood.
  - Least squares.
  - Weighted least squares.
  - Penalized least squares.
  - Targeted minimum loss estimation (TMLE).
  - Posterior mean/median/mode (Bayesian).
  - Minimax estimation.
  - Minimum risk equivariant estimation.
  - Minimum-risk linear unbiased estimation.
- Choice may depend upon parameter to be estimated, study design, assumptions, subjective "user-preferences."

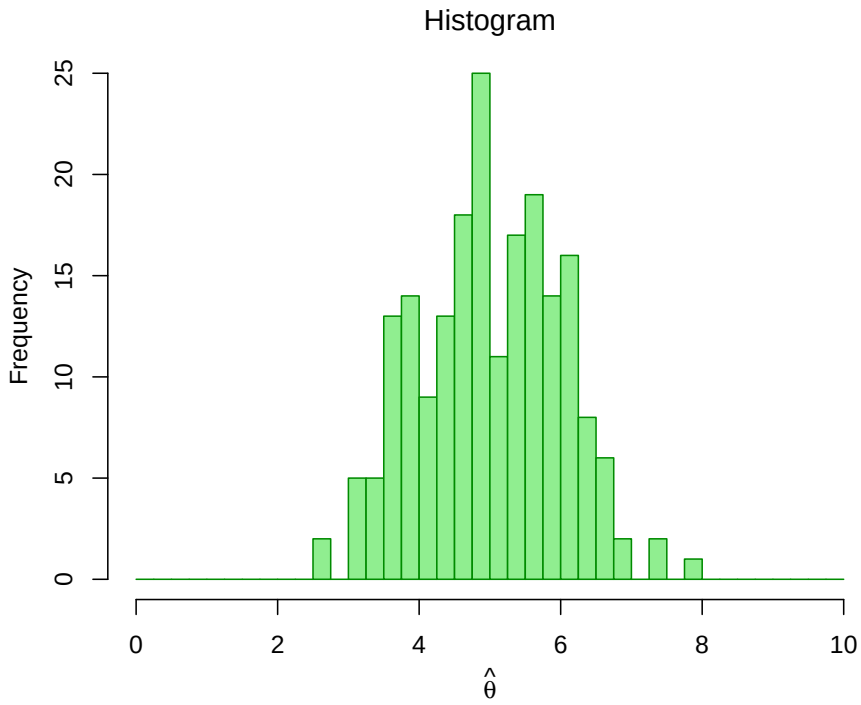
Sampling distributions

Estimation: Sampling variation

- Even if the quantity we are estimating is fixed, there is variation in the process of sampling the data to estimate that quantity.
- Variability in our data sample translates to variation in $\widehat{\theta}$.
- Were I to repeat a study again and again with random samples from the *same* population, under the *same* sample size, using the *same* method for estimation, the $\widehat{\theta}$'s would form a *distribution*.
- Frequentist framework often tries to understand (approximate) distribution of $\widehat{\theta}$ (at least in large samples) so that we may form confidence intervals (characterize precision of estimation) and compute p-values (characterize statistical strength of evidence).

Estimation: Sampling variation





Sampling distributions

Estimation: Bias

- $\mathrm{E}[\widehat{\theta}]$: Expected value of $\widehat{\theta}$.
  - Were I to repeat this study again and again with random samples from the same population, under the same sample size, and using the same method, what would be the "long-run average" of the resulting $\widehat{\theta}$'s?
- $\mathrm{Bias}[\widehat{\theta}] = \mathrm{E}[\widehat{\theta} - \theta] = \mathrm{E}[\widehat{\theta}] - \theta$.
  - To what degree does that "long-run average" miss the mark?

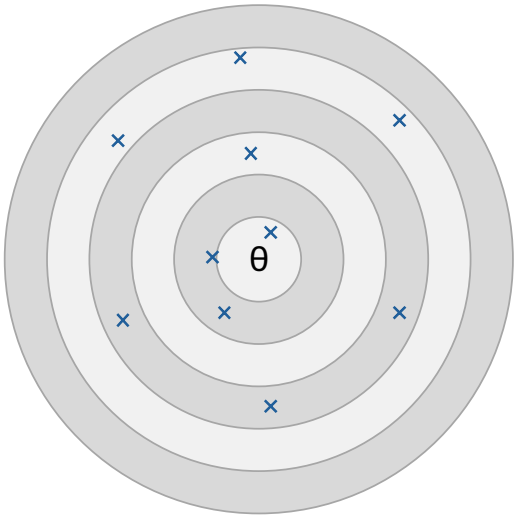
Sampling distributions

Estimation: Variability

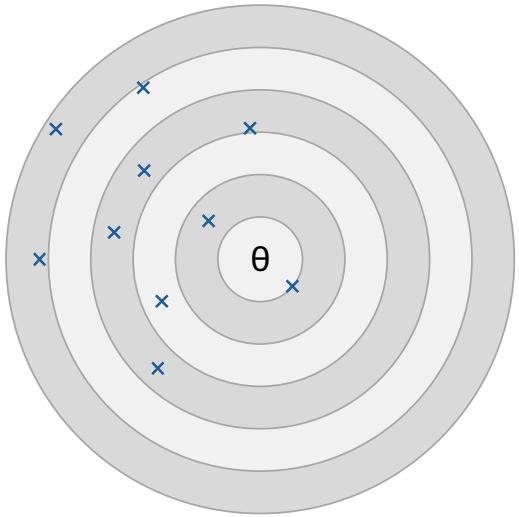
- $\text{Var}[\widehat{\theta}] = \text{E}[(\widehat{\theta} - \text{E}[\widehat{\theta}])^2]$: Variance of $\widehat{\theta}$.
  - Were I to repeat this study again and again with random samples from the same population, under the same sample size, and using the same method, what would be the "long-run variance" of the resulting $\widehat{\theta}$'s?
  - Or: what would be the "long-run average squared distance from $\text{E}[\widehat{\theta}]$"?
- Standard error: $\text{SE}[\widehat{\theta}] = \sqrt{\text{Var}[\widehat{\theta}]}$.
  - The standard error of $\widehat{\theta}$ is nothing more or less than the standard deviation of the sampling distribution.
- Precision refers to the *inverse* of the variance.
  - When variance is low, precision is high — and vice versa.

Sampling distributions

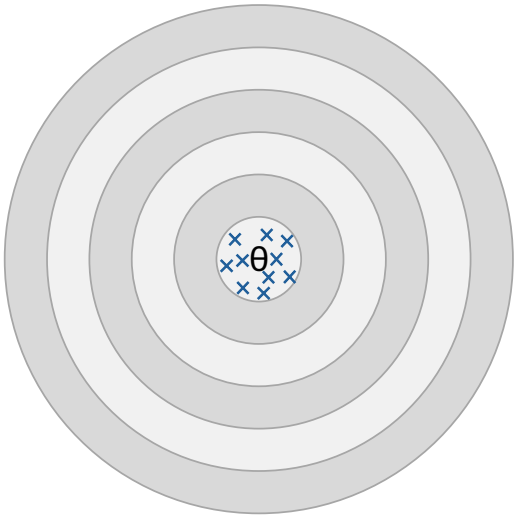
Low bias and low precision



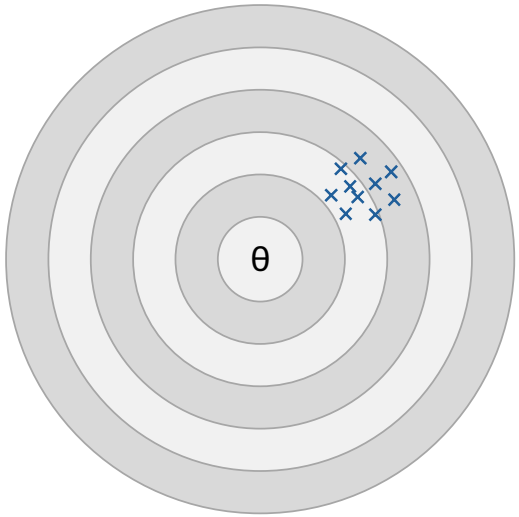
High bias and low precision



Low bias and high precision



High bias and high precision



Estimation: Mean squared error

- The bias-variance decomposition:

$$\text{MSE}[\widehat{\theta}] = \text{E}[(\widehat{\theta} - \theta)^2] = \overset{\text{(boring math)}}{\cdots} = \Big(\text{Bias}[\widehat{\theta}]\Big)^2 + \text{Var}[\widehat{\theta}].$$

- The mean squared error can be decomposed into the (squared) bias and variability, aggregating information on...
  1. ...how far $\widehat{\theta}$ is from θ on average, and
  2. ...how $\widehat{\theta}$ varies about its long-run average, $\text{E}[\widehat{\theta}]$

  ...in order to inform how much $\widehat{\theta}$ varies about θ.

Estimation: Consistency

- Discussion on bias and variability pertains to a *fixed* sample size.
- Of interest to know what happens to $\widehat{\theta}$ when the sample size grows.
- An estimator, $\widehat{\theta}$, is said to be consistent (or, $\widehat{\theta} \xrightarrow{p} \theta$) if for all $\epsilon > 0$,

$$\mathrm{P}(|\widehat{\theta} - \theta| > \epsilon) \to 0 \text{ as } n \to \infty.$$

- In plain language, as the sample size grows, the estimate should move closer to the truth (in a probabilistic sense).
- An estimator can be: unbiased and consistent, biased and consistent, unbiased and inconsistent, or biased and inconsistent.

Estimation: Consistency

- Let's talk about properties of the sample mean, $\overline{X}$, as an estimator of a population mean, μ. The weak law of large numbers (WLLN for short, or LLN for even shorter) tells us that if $X_1, \ldots, X_n$ are independent and identically distributed random variables, then:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu = \mathrm{E}[X]$$

- Notation:
  - $\overline{X}_n$ sample mean of the n observations.
  - $\mu = \mathrm{E}[X]$: population mean (expectation of X).
    - ★ May not always exist.
  - $\xrightarrow{p}$: convergence in probability (consistency).

Central limit theorem:

- The classic central limit theorem tells us that if $X_1, \ldots, X_n$ are independent and identically distributed random variables, then:

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

- Notation:
  - $\overline{X}_n$ sample mean of the n observations.
  - $\mu = \mathrm{E}[X]$: population mean (expectation of X).
  - σ^2: population variance of X.
  - $\xrightarrow{d}$: convergence in distribution.
  - $\mathcal{N}(0, \sigma^2)$: normal distribution with mean zero and variance σ^2.
- From this, we learn that if N is "large enough", then

$$\overline{X}_n \dot{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

- Here, $\dot{\sim}$ denotes an *approximate* distribution.

Central limit theorem:

- In this course, I will typically drop the subscript "n" on $\overline{X}_n$.
- From the central limit theorem: if n is "large enough", then

$$\overline{X} \dot{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

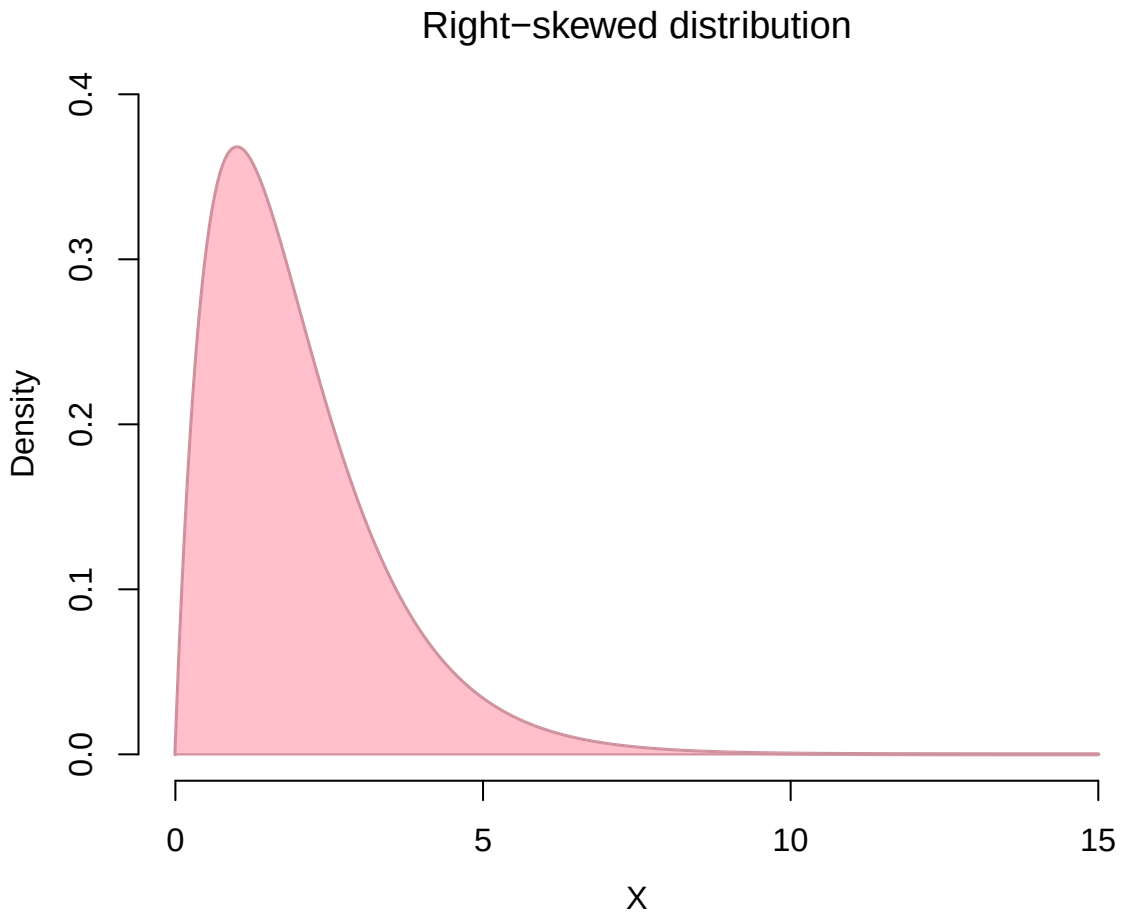
- Population variance (σ^2) unknown, but may be estimated:

$$\widehat{\sigma}^2 = \widehat{\text{Var}}[\overline{X}] = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

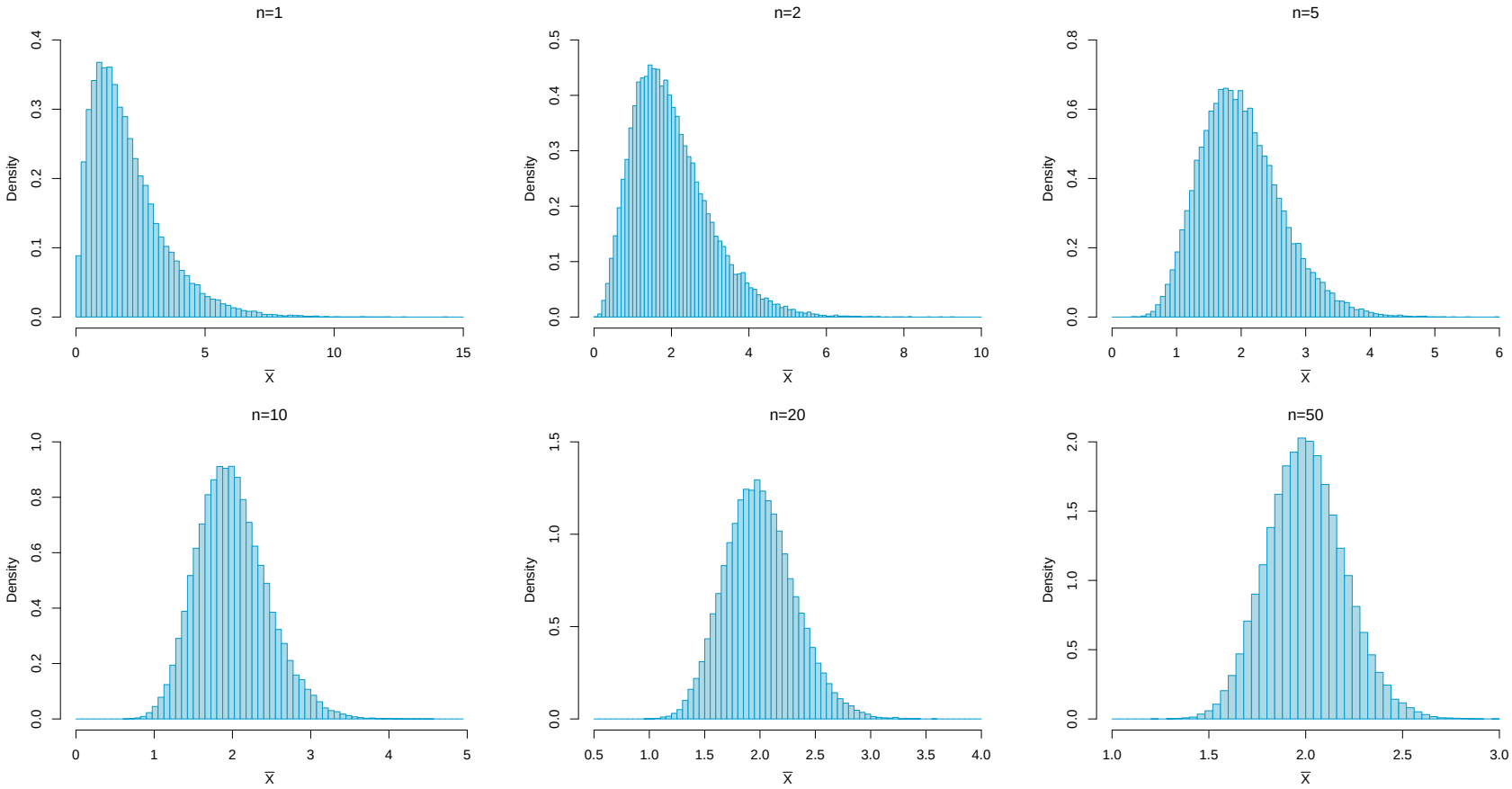
- In large samples, we can make statements about the entire sampling distribution of $\widehat{\mu} = \overline{X}$ that are approximately correct.
- The estimated standard error of $\widehat{\mu}$ is given by

$$\widehat{\text{SE}}[\widehat{\mu}] = \sqrt{\frac{\widehat{\sigma}^2}{n}} = \frac{\widehat{\sigma}}{\sqrt{n}}.$$

Central limit theorem: In action!



Central limit theorem: In action!



Note: The x-axis varies for visual clarity.

Table of Contents

Descriptive statistics

Motivating example: The REACH study

- Nelson et al. present results of a study in *Diabetes Care* (2021).
  - "Effects of a Tailored Text Messaging Intervention Among Diverse Adults With Type 2 Diabetes: Evidence From the 15-Month REACH Randomized Controlled Trial."
- Pseudo-data and documentation available on course website.

Descriptive statistics

Motivating example: The REACH study

- When glucose builds up in the blood, it binds to the hemoglobin in the erythrocytes (red blood cells). The HbA1c test measures the percent of glycated hemoglobin, which approximates average blood sugar level over prior three months (approximate lifespan of an erythrocyte).
- Blood sugar can be controlled in part with medication adherence. The REACH study sought to evaluate whether a daily text message intervention pertaining to self-efficacy could improve HbA1c in adult patients with type 2 diabetes.
  - A subset of the subjects receiving the REACH intervention received an additional intervention involving family coaching that we will not discuss at this particular time.
- HbA1c was measured several times over a fifteen-month period following randomization.

Descriptive statistics

Example: REACH and six-month HbA1c

- Suppose we seek to compare HbA1c between those receiving REACH and those not receiving REACH at the six-month mark.
- Two-group comparisons can be graphically represented using one of many methods:
  - Box plots.
  - Histograms.
- I will provide you with R code that I believe to be helpful. I am unlikely to explain every detail of the code in class; I expect you can figure a lot of it out from the documentation.
  - However, you are of course free to ask questions if something is still unclear or unanswered by the documentation!
- I tend to use base R for many tasks because it gives me control to make things look the way I want them to. However, I am in the minority. Feel free to use approaches that work for you and your needs.

Descriptive statistics

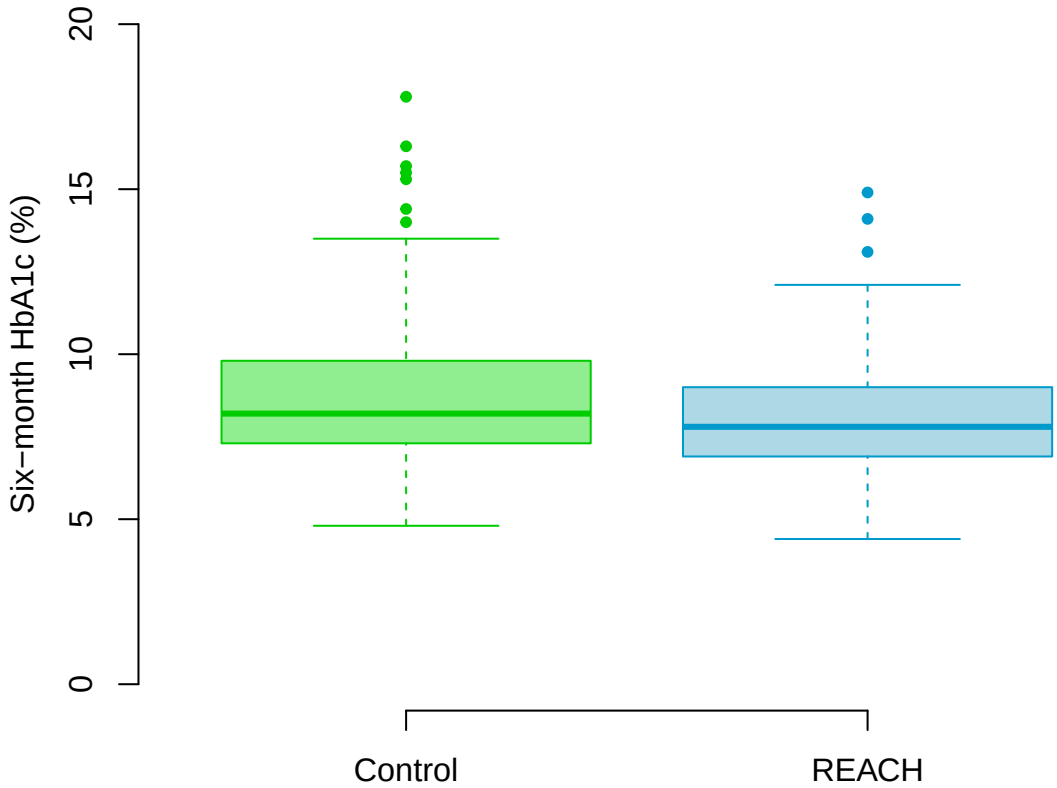
Example: REACH and six-month HbA1c (box plots)

```
## Read in data
dat <- read.csv("reach.csv")

## Create box plot
boxplot(a1c.6~reach, data = dat,
        frame.plot = FALSE,
        xlim = c(0.5,2.5),
        col = c("lightgreen", "lightblue"),
        pch = 20, ylim = c(0,20),
        xlab = "", ylab = "Six-month HbA1c (%)",
        names = c("Control", "REACH"),
        border = c("green3", "deepskyblue3"))
```

Note: This is based on my idiosyncratic preferences. So long as you're producing clear visuals and labeling everything correctly, you do you! :)

Example: REACH and six-month HbA1c (Box plots)



Descriptive statistics

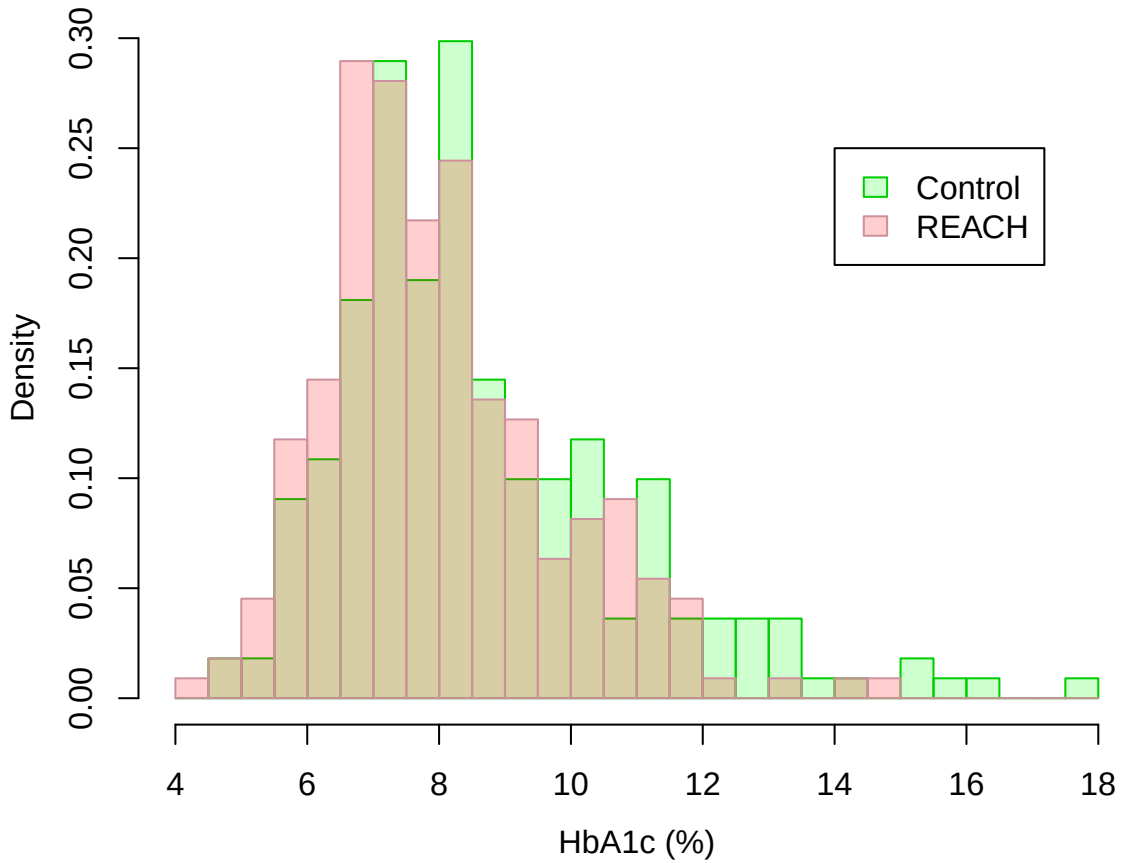
Example: REACH and six-month HbA1c (overlaid histograms)

```
## Colors of interest (transparency)
col1 <- rgb(0,255,0,alpha = 50, maxColorValue = 255)
col2 <- rgb(255,0,0,alpha = 50, maxColorValue = 255)

## Create histograms
hist(dat$a1c.6[dat$reach == 0],
     col = col1, border = "green3",
     freq = FALSE, breaks = seq(4,18,0.5),
     xlim = c(4, 18), ylim = c(0, 0.3),
     main = "", xlab = "HbA1c (%)")
hist(dat$a1c.6[dat$reach == 1],
     col = col2, border = "pink3",
     freq = FALSE, breaks = seq(4,18,0.5),
     add = TRUE)
legend(14,0.25, c("Control", "REACH"), col = NA,
       pch = 15, border = c("green3", "pink3"),
       fill = c(col1, col2))
```

Note: This is based on my idiosyncratic preferences.

Example: REACH and six-month HbA1c (Overlaid histograms)



Descriptive statistics

Example: REACH and six-month HbA1c

- The distribution of six-month HbA1c can be summarized numerically:
  - Count (and frequency of missing values).
  - Mean.
  - Median.
  - Variance (or its close relative, the standard deviation).
  - Quartiles (first and third).
  - Extreme values (minimum and maximum).
- Can be summarized in the whole sample or in each group separately.

Descriptive statistics

Example: REACH and six-month HbA1c

```
## Table of treatment assignments
> table(dat$reach)

  0   1
252 253

## Important library (contains descrip() function among others)
library("rigr")

## Descriptive statistics on six-month HbA1c
> descrip(dat$a1c.6, strata = dat$reach)

                          N    Msng   Mean       Std Dev   Min
 dat$a1c.6:  All          505    63   8.402      2.021     4.400
 dat$a1c.6:    Str  0     252    31   8.718      2.212     4.800
 dat$a1c.6:    Str  1     253    32   8.085      1.759     4.400
                          25%        Mdn        75%        Max
 dat$a1c.6:  All          7.000      8.000      9.400      17.80
 dat$a1c.6:    Str  0     7.300      8.200      9.800      17.80
 dat$a1c.6:    Str  1     6.900      7.800      9.000      14.90
```

R output: Putting it together

- Example of how to glean output in a format suitable for presentation to a collaborator:

HbA1c	*n* (msg*)	Mean (SD)	Median (IQR)	(Min, Max)
Control	221 (31)	8.72 (2.21)	8.2 (7.3, 9.8)	(4.8, 17.8)
REACH	221 (32)	8.09 (1.76)	7.8 (6.9, 9.0)	(4.4, 14.9)

* - Missing

- Why is there a discrepancy in significant digits in this example?
  - For instance, the sample mean HbA1c is reported as 8.72% in the controls, but then the sample median is reported as 8.2%.
- `NA` is recognized as missing. "`888`" means nothing!

Descriptive statistics

R output: Purpose of descriptive statistics

- Salient descriptive statistics on continuous variables:
  - *n* (missing) - *Sample size and missingness*
  - Mean (SD) - *Central tendency, spread*
  - Median (IQR) - *Central tendency, skewness*
  - Minimum, Maximum - *Range, skewness, outliers, errors*
- Skewness and kurtosis can also be computed but have less immediate clinical interpretability.

Descriptive statistics

Significant digits: In this course, please report **three** significant digits!!!

- Three significant digits on the arithmetic scale:
  - Correct: 142, 10.5, 2.72, 0.163, 0.0759.
  - Too **few** significant digits: 3.0, 0.006, 0.02.
  - Too **many** significant digits 36.65299.
- Three significant digits on the ratio scale (a/b):
  - Correct: 2.01 ("101% higher"), 1.523 ("52.3% higher"), 1.0492 ("4.92% higher"), 1.00231 ("0.231% higher").
  - Correct: 0.031 ("96.9% lower"), 0.342 ("65.8% lower"), 0.9234 ("7.66% lower"), and 0.99213 ("0.787% lower").
  - Principle: three significant digits should remain when the ratio is interpreted as a percentage:
    - $100 \times (a/b - 1)$ if $a/b > 1$, or $100 \times (1 - a/b)$ if $a/b < 1$.
    - Do not round the quantity a/b before converting to a percentage!
- The point is to be consistent and reproducible. It is unfortunate when people report risk ratios such as 1.1 in manuscripts (I want to know if my risk is 5.01% higher or 14.9% higher if exposed).

Table of Contents

Mean differences

Difference in means: Two groups

- Carrying on with our motivating example, suppose we wish to estimate and conduct inference on difference in mean six-month HbA1c between groups.
- Let n_0 and n_1 denote sample sizes in each group (control and REACH, respectively); let $n = n_0 + n_1$.
- Let $Y_{0,1}, \ldots, Y_{0,n_0}$ index the HbA1c values in control subjects, and let $Y_{1,1}, \ldots, Y_{1,n_1}$ index the HbA1c values in the REACH subjects (all subjects are independently sampled).
- Let μ_0 and μ_1 denote population means for HbA1c.
- Let σ_0^2 and σ_1^2 denote the population variances for HbA1c.
- We wish to estimate and conduct inference on $\delta = \mu_1 - \mu_0$.

Mean differences

Difference in means: Two groups

- Our *point estimate* is the difference in sample mean HbA1c:

$$\widehat{\delta} = \widehat{\mu}_1 - \widehat{\mu}_0 = \overline{Y}_1 - \overline{Y}_0 = \frac{1}{n_1}\sum_{j=1}^{n_1} Y_{1,j} - \frac{1}{n_0}\sum_{i=1}^{n_0} Y_{0,i}.$$

Mean differences (the equal-variance case)

Difference in means: Standard errors

- Assuming a common variance σ^2 between groups,

$$\text{Var}[\widehat{\delta}] = \text{Var}[\widehat{\mu}_1 - \widehat{\mu}_0] = \text{Var}[\widehat{\mu}_1] + \text{Var}[\widehat{\mu}_0] = \frac{\sigma^2}{n_0} + \frac{\sigma^2}{n_1} = \sigma^2\left(\frac{1}{n_0} + \frac{1}{n_1}\right).$$

  This may be estimated as:

$$\widehat{\text{Var}}[\widehat{\delta}] \quad = \quad s_p^2\left(\frac{1}{n_0} + \frac{1}{n_1}\right) = \frac{(n_0-1)s_0^2 + (n_1-1)s_1^2}{n_0+n_1-2}\left(\frac{1}{n_0} + \frac{1}{n_1}\right),$$

  where, s_0^2 and s_1^2 denote sample variances. For instance:

$$s_0^2 = \frac{1}{n_0 - 1}\sum_{i=1}^{n_0}(Y_{0i} - \overline{Y}_0)^2$$

- $\widehat{\text{SE}}[\widehat{\delta}] = \sqrt{\widehat{\text{Var}}[\widehat{\delta}]}$: standard deviation of the sample mean.

Mean differences

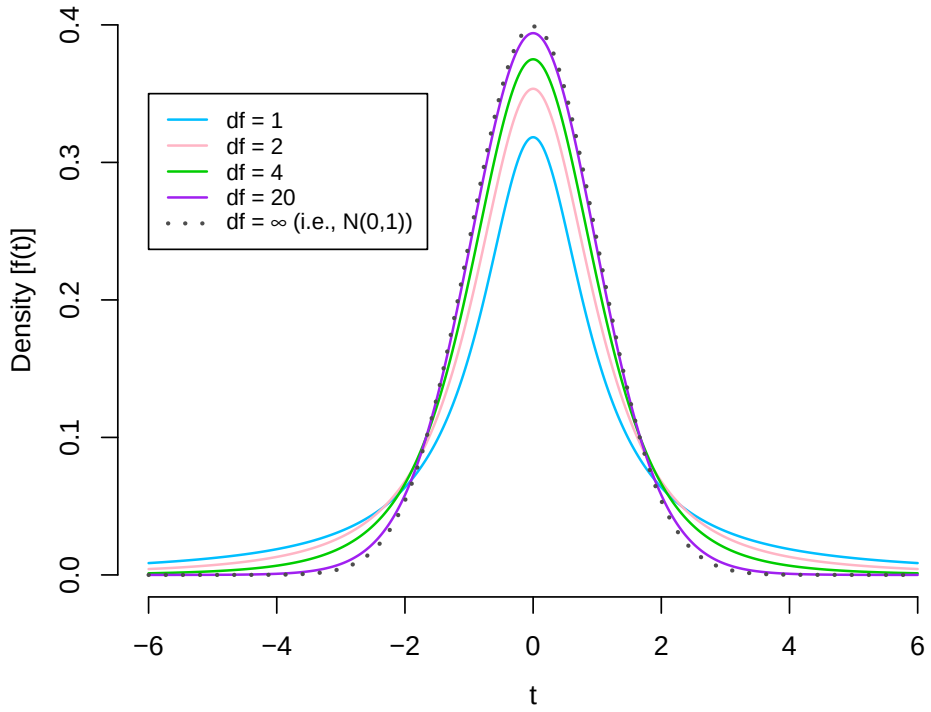
Difference in means: Asymptotic behavior

- Again under the assumption of equal variances:

$$\frac{\widehat{\delta} - \delta}{\widehat{\mathrm{SE}}[\widehat{\delta}]} \overset{\cdot}{\sim} t_{n-2};$$

  t_{n-2}: t-distribution with $n-2$ degrees of freedom (df).
- The t-distribution is exact if the outcomes in each group are each normally distributed, and is otherwise only *approximate* (for sufficiently large n).
- This key fact sets the stage for developing confidence intervals and performing hypothesis tests.

Reminder: The t-distribution family



- Symmetric with median, mode (and mean, if it exists) zero.
- Heavy tails (high kurtosis): high tendency for extremes.
- With $df \geq 50$, t_{df} nearly indistinguishable from $Z \sim \mathcal{N}(0, 1)$.

Mean differences: Interval estimation

- Two-sided $100(1-\alpha)\%$ confidence interval (CI):

$$\begin{aligned} \mathrm{P}\left(t_{\alpha/2,n-2} \leq \frac{\widehat{\delta}-\delta}{\widehat{\mathrm{SE}}[\widehat{\delta}]} \leq t_{1-\alpha/2,n-2} \right) &= 1-\alpha \\ \iff \mathrm{P}\left(t_{\alpha/2,n-2}\widehat{\mathrm{SE}}[\widehat{\delta}] \leq \widehat{\delta}-\delta \leq t_{1-\alpha/2,n-2}\widehat{\mathrm{SE}}[\widehat{\delta}] \right) &= 1-\alpha \\ \iff \mathrm{P}\left(\widehat{\delta} - t_{1-\alpha/2,n-2}\widehat{\mathrm{SE}}[\widehat{\delta}] \leq \delta \leq \widehat{\delta} + t_{1-\alpha/2,n-2}\widehat{\mathrm{SE}}[\widehat{\delta}] \right) &= 1-\alpha, \end{aligned}$$

  where $t_{q,df}$ denotes $(100 \times q)^{\text{th}}$ percentile of the t-distribution having df degrees of freedom.

- Very often (but not always), we choose $\alpha = 0.05$.
- Assuming equal variances, the CI can be formed as:

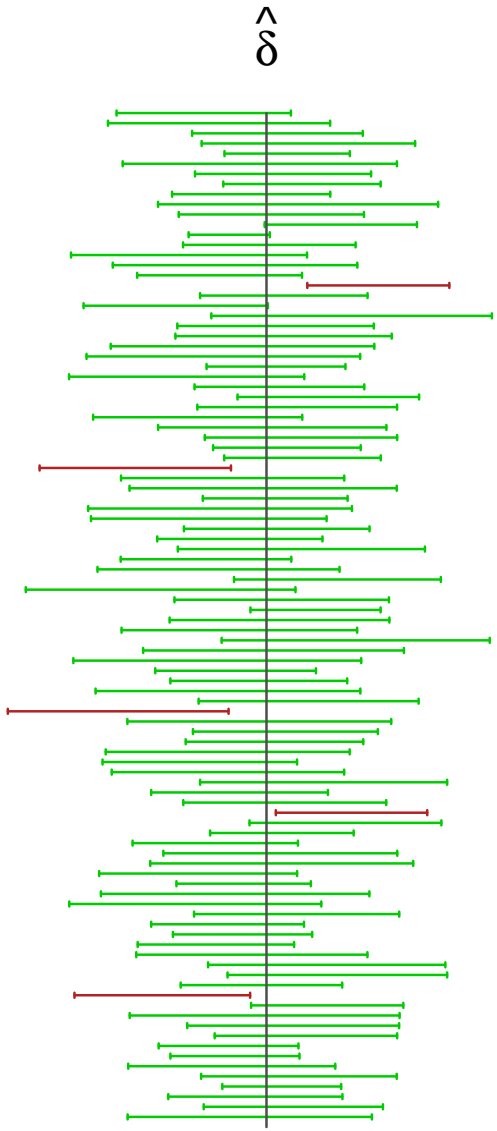
$$\widehat{\delta} \pm t_{1-\alpha/2,n-2}\left(s_p \sqrt{\frac{1}{n_0}+\frac{1}{n_1}} \right).$$

Mean differences: Interval estimate

- Interpreting a 95% confidence interval (CI) **incorrectly**:
  - There is a 95% chance that δ lies in my CI.
  - Were I to conduct this study repeatedly, 95% of the resulting $\widehat{\delta}$'s would lie between the endpoints of the CI that I obtained in my study.
- Interpreting a 95% confidence interval (CI) **correctly**:
  - Were I to conduct this study repeatedly, 95% of CIs derived in the manner described would contain the value of δ (assuming the necessary assumptions are met).

Mean differences

Recall: Confidence intervals (one correct interpretation)



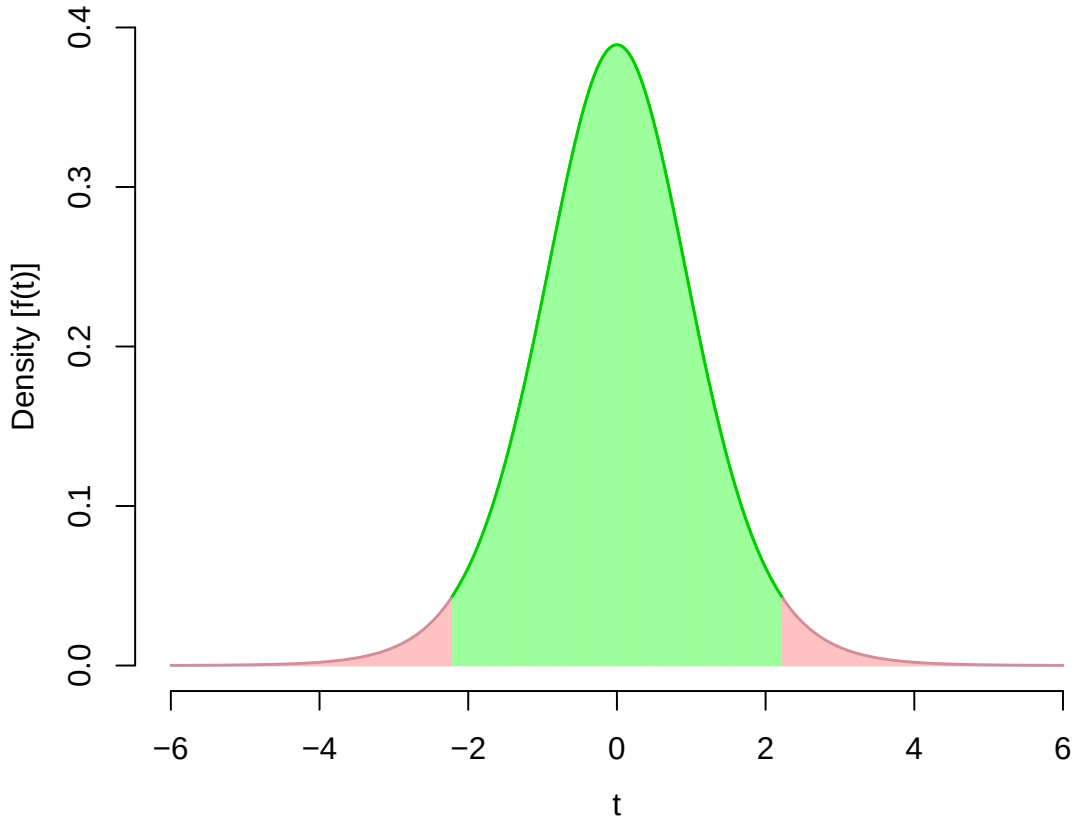
Mean differences: The t-test with equal variances

- Null vs. alternative hypotheses: $H_0 : \delta = 0$ vs. $H_1 : \delta \neq 0$.
- The t-statistic:

$$t = \frac{\widehat{\delta} - 0}{\widehat{\mathrm{SE}}[\widehat{\delta}]} = \frac{\widehat{\delta}}{s_p \cdot \sqrt{\frac{1}{n_0} + \frac{1}{n_1}}}.$$

- Under H_0, the t-statistic follows an approximate t-distribution with $df = (n_0 - 1) + (n_1 - 1) = n - 2$ degrees of freedom (*exact* if each group follows normal distribution).

Mean differences



- Red: Rejection region ($|T| > t_{1-\alpha/2,df}$); $(100 \times \alpha)\%$ of area.
  - We reject H_0 if t lies in this region.
- Green: We do not reject H_0 if t lies in this region.

Mean differences: The t-test with equal variances

- Also known as *Student's* t-test.
- Null vs. alternative hypotheses: $H_0 : \delta = 0$ vs. $H_1 : \delta \neq 0$.

$$t = \frac{\widehat{\delta} - 0}{\widehat{\mathrm{SE}}[\widehat{\delta}]} = \frac{\widehat{\delta}}{s_p \cdot \sqrt{\frac{1}{n_0} + \frac{1}{n_1}}}.$$

- Under $H_0 : \delta = 0$, t has *approximate* t_{n-2} distribution (exact if each group follows normal distribution).
- p-value given by $p = 2 \times \mathrm{P}(T > |t|)$; T is a random variable that follows a t-distribution with df degrees of freedom.
  - If $p < \alpha$, then we *reject* H_0.
  - Equivalently, if $|t| > t_{1-\alpha/2,n-2}$, then we *reject* H_0.
- Very often (but not always), we choose $\alpha = 0.05$.

Mean differences: Hypothesis testing

- Interpreting a p-value **incorrectly**:
  - There is a $(100 \times p)\%$ chance that H_0 is true.
  - (. . . *a p-value is obtained under the stipulation that the null is true, so it cannot possibly tell you the probability that it is*.)
- Interpreting a p-value **correctly**:
  - Null hypothesis, when true, produces results as extreme (or more so) as the one I got $(100 \times p)\%$ of the time.
  - (. . . *if this value is low, we reject the null hypothesis as a plausible mechanism by which our data were derived*.)

Mean differences: Inverting the test

- Note the connection between the CI and t-statistic.
- They both employ the estimated standard error.
  - Solved for CI endpoints by inverting the expression $(\widehat{\delta} - \delta)/\widehat{\text{SE}}[\widehat{\delta}]$.
- In this special case, we gain a more exciting interpretation for the confidence interval.

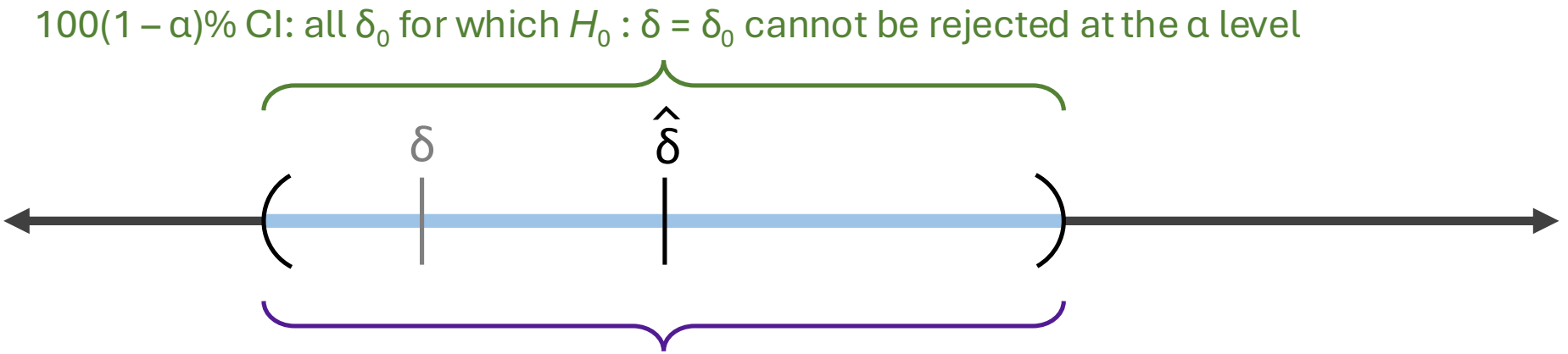
Mean differences: Interval estimate

- Interpreting a 95% confidence interval (CI) **correctly**:
  - Interpretation #1: Were I to conduct this study repeatedly, 95% of CIs derived in the manner described would contain the value of δ.
  - Correct, but more of a statement about the procedure than about the data. I never actually am able to conduct the study repeatedly, and there's nothing in my study that tells me which of the two cases I'm in (coverage or no coverage).

Mean differences: Interval estimate

- Interpreting a 95% confidence interval (CI) **correctly**:
  - Interpretation #2: If the CI *inverts the hypothesis test*, then it contains values of δ that cannot be ruled out by my data at my specified confidence level.
  - *More interesting!* It turns the CI into something like a range of plausible values based on my data. More precisely, it is the complement of the set containing all *implausible* values.

Recall: Confidence intervals (correct interpretation*)



Estimate, $\hat{\delta}$, would not be considered "atypical" if it were derived from a population with any true mean difference in this range.

All null hypotheses lying outside our CI would be rejected by our data.

* - *If the CI inverts the hypothesis test!*

Mean differences: Assumptions

- What are the assumptions we have used to formulate the confidence intervals and conduct a t-test?
  - Independent samples.
  - Finite variances (and finite means).
  - Equal variances between groups.
- A version of the t-test can be formed without having to make the third assumption.
- **My perspective**: I am typically in favor of methods that make as few assumptions as possible.

Mean differences: Standard error of the mean

- Let us allow possibility that the variances differ.
  - Say, σ_0^2 for the control group and σ_1^2 for the REACH group.
- In this case:

$$\text{Var}[\widehat{\delta}] = \text{Var}[\widehat{\mu}_1 - \widehat{\mu}_0] = \text{Var}[\widehat{\mu}_1] + \text{Var}[\widehat{\mu}_0] = \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}$$

  This may be estimated as:

$$\widehat{\text{Var}}[\widehat{\delta}] \quad = \quad \frac{s_0^2}{n_0} + \frac{s_1^2}{n_1}.$$

  where, s_0^2 and s_1^2 denote the sample variances as before.

- Again, $\widehat{\text{SE}}[\widehat{\delta}] = \sqrt{\widehat{\text{Var}}[\widehat{\delta}]}$ denotes standard deviation of the sample mean—the formulas differ based on whether you make the assumption of equal variances.

Mean differences: Asymptotic behavior

- Allowing unequal variances between groups:

$$\frac{\widehat{\delta} - \delta}{\widehat{\mathrm{SE}}[\widehat{\delta}]} \overset{\cdot}{\sim} t_{df}.$$

- The big question: What is df in this case?
- Welch-Satterthwaite approximation to df:

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}\right)^2}{\frac{s_1^4}{n_1^2 \nu_1} + \frac{s_0^4}{n_0^2 \nu_0}},$$

where $\nu_0 = n_0 - 1$ and $\nu_1 = n_1 - 1$.

Mean differences: Interval estimation

- Allowing unequal variances, the CI can be formed as:

$$\widehat{\delta} \pm t_{1-\alpha/2,\nu} \sqrt{\frac{s_0^2}{n_0} + \frac{s_1^2}{n_1}}.$$

Mean differences: The t-test allowing unequal variances

- Also known as *Welch's* t-test.
- Null vs. alternative hypotheses: $H_0 : \delta = 0$ vs. $H_1 : \delta \neq 0$.
- The t-statistic:

$$t = \frac{\widehat{\delta} - 0}{\widehat{\mathrm{SE}}[\widehat{\delta}]} = \frac{\widehat{\delta}}{\sqrt{\frac{s_0^2}{n_0} + \frac{s_1^2}{n_1}}}.$$

- Under H_0, follows approximate t_ν distribution.
- p-value given by $p = 2 \times \mathrm{P}(T > |t|)$; T is a random variable that follows a t-distribution with ν degrees of freedom.
  - If $p < \alpha$, then we *reject* H_0.
  - Equivalently, if $|t| > t_{1-\alpha/2,\nu}$, then we *reject* H_0.

Further points: Historical context

- Student's *t*-test was initially popular as there was no easy expression for the degrees of freedom absent the assumption of equal variances.
- Welch/Satterthwaite provide an approximation to the degrees of freedom when the assumption of equal variances is relaxed.
  - Previously: Examine critical values from tables using only whole numbers for degrees of freedom, so the Student *t*-test remained the approach of choice. I'll *never* make you do this.
- In the modern era, software accommodates continuous degrees of freedom, so this is no longer a justification to continue invoking this assumption.
- When the sample size is too small to support group-specific variance estimation (e.g., very small animal studies or phase 1 factorial designs), the equal-variance assumption *may* be more defensible.

Tables: Not in this class! Whew!!

***t*-table**: Critical values by degrees of freedom and quantiles

	0.60	0.667	0.75	0.80	0.875	0.90	0.95	0.975	0.99	0.995	0.999
df = 1	0.325	0.577	1.000	1.376	2.414	3.078	6.314	12.706	31.821	63.657	318.31
2	0.289	0.500	0.816	1.061	1.604	1.886	2.920	4.303	6.965	9.925	22.327
3	0.277	0.476	0.765	0.978	1.423	1.638	2.353	3.182	4.541	5.841	10.215
4	0.271	0.464	0.741	0.941	1.344	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.457	0.727	0.920	1.301	1.476	2.015	2.571	3.365	4.032	5.893
6	0.265	0.453	0.718	0.906	1.273	1.440	1.943	2.447	3.143	3.707	5.208
7	0.263	0.449	0.711	0.896	1.254	1.415	1.895	2.365	2.998	3.499	4.785
8	0.262	0.447	0.706	0.889	1.240	1.397	1.860	2.306	2.896	3.355	4.501
9	0.261	0.445	0.703	0.883	1.230	1.383	1.833	2.262	2.821	3.250	4.297
10	0.260	0.444	0.700	0.879	1.221	1.372	1.812	2.228	2.764	3.169	4.144
11	0.260	0.443	0.697	0.876	1.214	1.363	1.796	2.201	2.718	3.106	4.025
12	0.259	0.442	0.695	0.873	1.209	1.356	1.782	2.179	2.681	3.055	3.930
13	0.259	0.441	0.694	0.870	1.204	1.350	1.771	2.160	2.650	3.012	3.852
14	0.258	0.440	0.692	0.868	1.200	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.439	0.691	0.866	1.197	1.341	1.753	2.131	2.602	2.947	3.733
16	0.258	0.439	0.690	0.865	1.194	1.337	1.746	2.120	2.583	2.921	3.686
17	0.257	0.438	0.689	0.863	1.191	1.333	1.740	2.110	2.567	2.898	3.646
18	0.257	0.438	0.688	0.862	1.189	1.330	1.734	2.101	2.552	2.878	3.610
19	0.257	0.438	0.688	0.861	1.187	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.437	0.687	0.860	1.185	1.325	1.725	2.086	2.528	2.845	3.552
21	0.257	0.437	0.686	0.859	1.183	1.323	1.721	2.080	2.518	2.831	3.527
22	0.256	0.437	0.686	0.858	1.182	1.321	1.717	2.074	2.508	2.819	3.505
23	0.256	0.436	0.685	0.858	1.180	1.319	1.714	2.069	2.500	2.807	3.485
24	0.256	0.436	0.685	0.857	1.179	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.436	0.684	0.856	1.178	1.316	1.708	2.060	2.485	2.787	3.450
26	0.256	0.436	0.684	0.856	1.177	1.315	1.706	2.056	2.479	2.779	3.435
27	0.256	0.435	0.684	0.855	1.176	1.314	1.703	2.052	2.473	2.771	3.421
28	0.256	0.435	0.683	0.855	1.175	1.313	1.701	2.048	2.467	2.763	3.408
29	0.256	0.435	0.683	0.854	1.174	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.435	0.683	0.854	1.173	1.310	1.697	2.042	2.457	2.750	3.385
39	0.255	0.435	0.681	0.851	1.168	1.304	1.685	2.023	2.426	2.708	3.313
49	0.255	0.433	0.679	0.849	1.164	1.299	1.676	2.009	2.403	2.678	3.261
99	0.254	0.433	0.677	0.845	1.157	1.290	1.660	1.984	2.364	2.626	3.175
199	0.254	0.432	0.676	0.843	1.154	1.286	1.653	1.972	2.345	2.600	3.132
299	0.254	0.432	0.675	0.843	1.153	1.284	1.650	1.968	2.339	2.592	3.118
399	0.254	0.432	0.675	0.843	1.152	1.284	1.649	1.966	2.336	2.588	3.111
499	0.253	0.432	0.675	0.842	1.152	1.283	1.648	1.965	2.334	2.586	3.106
999	0.253	0.432	0.674	0.842	1.151	1.282	1.646	1.962	2.330	2.581	3.098
∞	0.253	0.431	0.674	0.842	1.150	1.282	1.645	1.960	2.326	2.576	3.090

Mean differences

Further points: Classic and modern approaches

- **Classic** approaches tend to be assumption-heavy: the attitude is that there is no reason to relax an assumption unless the data prove to me that my assumptions are wrong.
- **Modern** approaches tend to favor agnosticism and skepticism: failure to falsify an assumption does not prove it true, so we are often willing to accept a *slight* hit in precision to favor an approach that makes as few assumptions as possible. When possible, we also want to avoid using the data to make analytic choices.
- In this course, we're generally taking the modern approach.

Further points: Characterizing assumptions

- Inappropriate to say Welch's *t*-test *assumes* unequal variances. Rather, it *allows for the possibility of* unequal variances.
- Your data can provide descriptive/inferential evidence regarding assumptions, but remember that you cannot prove the null—only provide sufficient evidence against it.
  - Methods such as Jeffrey Blume's second generation p-value that can help circumvent some of these challenges in many practical settings.

R: *t*-test results

```
## Conduct t-test (defaults to unequal-variance approach)
## NOTE: This function is from the rigr library.

> ttest(dat$a1c.6, by = dat$reach)

Call:
ttest(var1 = dat$a1c.6, by = dat$reach)

Two-sample t-test allowing for unequal variances :

Summary:
            Group Obs Missing  Mean Std. Err. Std. Dev.         95% CI
   dat$reach = 0 252      31 8.718     0.149      2.21 [8.425, 9.01]
   dat$reach = 1 253      32 8.085     0.118      1.76 [7.852, 8.32]
      Difference 505      63 0.633      0.19      <NA> [0.259, 1.01]

 Ho: difference in  means = 0 ;
 Ha: difference in  means != 0
 t = 3.33 , df = 419
 Pr(|T| > t) =  0.000944275
```

Mean differences

Gleaning results: How do we describe our results?

- A write-up for an analysis should provide, where possible:
  - A description of what was to be tested and how we tested it.
  - The point estimate and its direction.
  - The confidence interval and its interpretation.
  - The p-value.
  - A statement summarizing our conclusions.

Mean differences

Gleaning results: How do we describe our results?

We applied a two-sample t-test allowing unequal variances to determine whether mean six-month HbA1c differed between the REACH and control groups. This study provides sufficient evidence to reject the null hypothesis that the means do not differ between the two groups ($p < 0.001$). We estimate the mean HbA1c to be 0.633% lower in the REACH group; based on a 95% CI, our estimate would not be deemed atypical if the true mean HbA1c were anywhere between 0.259% and 1.01% lower in the REACH group.

Mean differences

Gleaning results: Guiding principles on point estimation

- A point of frustration in reporting on HbA1c is that it is measured as a percent (we are not making a statement about the ratio of mean HbA1c values).
- Be certain that the direction of association is made apparent.
- In this example, it would be equally accurate to state that mean six-month HbA1c was estimated to be 0.633% **higher** in the **control** group; it's more standard to discuss the experimental group relative to the reference group.

Mean differences

Gleaning results: Guiding principles on interval estimation

- State in a manner consistent with the point estimate.
- Suppose R reported the 95% CI as $[-0.090287, 1.356351]$; the appropriate way to report it would be as follows:

  "*Based on a 95% CI, our estimate would not be deemed atypical if the true mean HbA1c were anywhere between 1.36% lower and 0.0903% higher in the REACH group.*"

- In the first few problem sets, I will make you go through some of these motions until you get the hang of it.

Gleaning results: Guiding principles on inference

- Report the p-value (even if it does not achieve statistical significance).
- When a p-value is smaller than 0.001, simply report ($p < 0.001$).
  - In fact, if the p-value is small enough, R may report $p = 0.0000$, or provide scientific notation (e.g., 2.63e-15).
  - This is either inaccurate or provides far too much granularity.
- The statement of inference could also appear at the end of a write-up rather than at the beginning. The statement of inference is longer than it strictly needs to be. It is acceptable to restate it in the affirmative: "This study provides sufficient evidence of a difference in means between groups."
- However, we cannot prove the null hypothesis true, so it is **never** okay to say that a study provides evidence that the means do not differ.

Table of Contents

Proportion differences

Difference in proportions: Another question

- Suppose we want to determine whether the proportion of individuals having a six-month HbA1c of at most 9.0% differs between the two groups.
- We can walk through the same process to obtain point/interval estimates and conduct inference.

Proportion differences

Proportion differences:

- Wish to estimate/conduct inference on difference in proportions of subjects in each group having six-month HbA1c $\leq 9\%$.
- Let p_0 and p_1 denote respective proportions in each group:
- We seek to estimate and conduct inference on $\delta = p_1 - p_0$.

Proportion differences

Proportion differences: Point estimate and standard error

- Our point estimate is the difference in sample proportions:

$$\widehat{\delta} = \widehat{p}_1 - \widehat{p}_0 = \frac{1}{n_1}\sum_{j=1}^{n_1} 1(Y_{1,j} \leq 9.0) - \frac{1}{n_0}\sum_{i=1}^{n_0} 1(Y_{0,j} \leq 9.0).$$

- Variance of $\widehat{\delta}$ is given by:

$$\text{Var}[\widehat{\delta}] = \text{Var}[\widehat{p}_1] + \text{Var}[\widehat{p}_0] = \frac{p_0(1-p_0)}{n_0} + \frac{p_1(1-p_1)}{n_1}.$$

- Standard error: $\text{SE}[\widehat{\delta}] = \sqrt{\text{Var}[\widehat{\delta}]}$. May be estimated by plugging in the respective sample proportions.

Proportion differences

Proportion differences: Asymptotic behavior

- If p_0 and p_1 are strictly between zero and one:

$$\frac{\widehat{\delta} - \delta}{\text{SE}[\widehat{\delta}]} \dot{\sim} \mathcal{N}(0, 1)$$

- With laser-focused attention, you'll notice that denominator refers to the theoretical standard error instead of the estimated standard error.
- The purpose of this formulation will be made apparent momentarily.

Proportion differences: Interval estimate

- Two-sided $100(1-\alpha)\%$ CI:

$$\widehat{\delta} \pm z_{1-\alpha/2}\widehat{\mathrm{SE}}[\widehat{\delta}],$$

or,

$$\widehat{p}_1 - \widehat{p}_0 \pm z_{1-\alpha/2}\sqrt{\frac{\widehat{p}_0(1-\widehat{p}_0)}{n_0} + \frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1}}.$$

Proportion differences: Hypothesis testing

- Null and alternative hypothesis: $H_0 : \delta = 0$ vs. $H_1 : \delta \neq 0$.
- Test statistic:

$$z = \frac{\widehat{p}_1 - \widehat{p}_0}{\sqrt{\widehat{p}_t(1-\widehat{p}_t)\left(\frac{1}{n_1} + \frac{1}{n_0}\right)}},$$

  where $\widehat{p}_t$ is the pooled proportion, $\frac{1}{n}\sum_{i=1}^{n} 1(Y_i \leq 9.0)$.
- Under H_0, z has approximate standard normal distribution.
- The p-value is given by $p = 2 \times \mathrm{P}(Z > |z|)$.
  - If $p < \alpha$, then we *reject* H_0.
  - Equivalently, if $|z| > z_{1-\alpha/2}$, then we *reject* H_0.
- Importantly, the CI does *not* invert the test.
  - Idea: test statistic takes into account the mean-variance relationship of a binary variable for theoretical reasons. The denominator is the standard error *under the null*.

Proportion differences

R: Two-sample proportion test

```
## Conduct proportion test
## NOTE: This function is from the rigr library.
> proptest(dat$a1c.6 <= 9, by = dat$reach)

Call:
proptest(var1 = dat$a1c.6 <= 9, by = dat$reach)

Two-sample proportion test (approximate) :

            Group Obs Missing       Mean Std. Err.             95% CI
   dat$reach = 0 252      31  0.6696833    0.0316  [0.608, 0.73169]
   dat$reach = 1 253      32  0.7511312    0.0291  [0.694, 0.80813]
      Difference 505      63 -0.0814479     0.043 [-0.166, 0.00278]
Summary:

 Ho: Difference in proportions = 0
 Ha: Difference in proportions != 0
 Z = -1.89
 p.value = 0.0591
```

Proportion differences

Gleaning: How do we describe our results?

Based on a two-sample test of proportions, there is not sufficient evidence to suggest that the proportion of individuals with six-month HbA1c of at most 9.0% differed between those assigned to REACH and those assigned to the control condition ($p = 0.059$). We estimate the difference in proportions to be 0.0814, with the REACH group having the higher proportion. The 95% CI for the difference in proportions is given by: [-0.00278, 0.166]; were we to repeat this study over and over under the same sample size, approximately 95% of the resulting CIs derived in this fashion would contain the true proportion.

Proportion differences

Further points:

- Take care in analyzing differences in proportions. It is one of the few examples in this course in which the confidence interval does *not* invert the test.
- We construct *Wald*-based confidence intervals of the form:

$$\text{Estimate} \pm \text{Tolerance} \times \text{Estimated standard error},$$

but to conduct testing, we use a *score* statistic of the form:

$$\frac{\text{Estimate}}{\text{Standard error under null}}.$$

- We must interpret the confidence interval cautiously.

Proportion differences

Further points:

- This example is generally just for illustration. Ordinarily, it's not wise to dichotomize a continuous variable.
- As was illustrated in this example, doing so often reduces *power* (the probability of rejecting the null when it is not true).

Summary

This unit:

- Sampling distributions.
- Point and interval estimation.
- Inference.
- Two-sample t-test.
- Two-sample proportion test.

Summary

So far:

- Review.

Summary

Coming up:

- Simple linear regression.
- Multiple linear regression (foundations).
- Multiple linear regression (interactions and strata).
- Transformations and basis expansions.
- Regression with binary outcomes.
- Regression with nominal, ordinal, and count outcomes.
- Introduction to clustered data.
- Methods for time-to-event outcomes.
- Predictive capacity of regression models.