Day 9: Confidence intervals (4.2)

BSTA 511/611

Meike Niederhausen, PhD
OHSU-PSU School of Public Health

2023-10-30

Last time -> Goals for today

Day 8: Section 4.1

- Sampling from a population
  - population **parameters** vs. **point estimates**
  - sampling variation

- **Sampling distribution** of a mean
- **Central Limit Theorem**

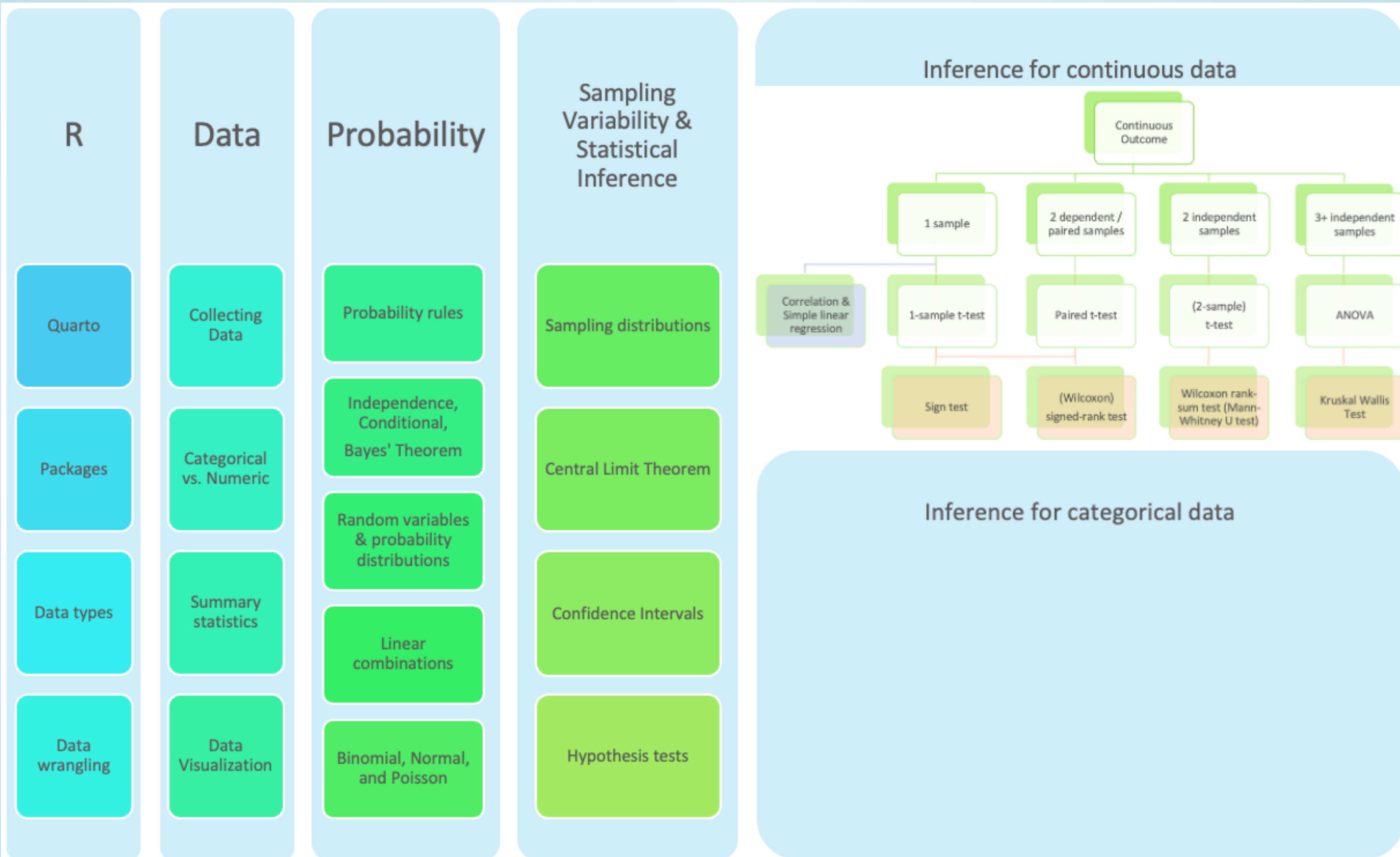
Day 9: Section 4.2

What are **Confidence Intervals**?

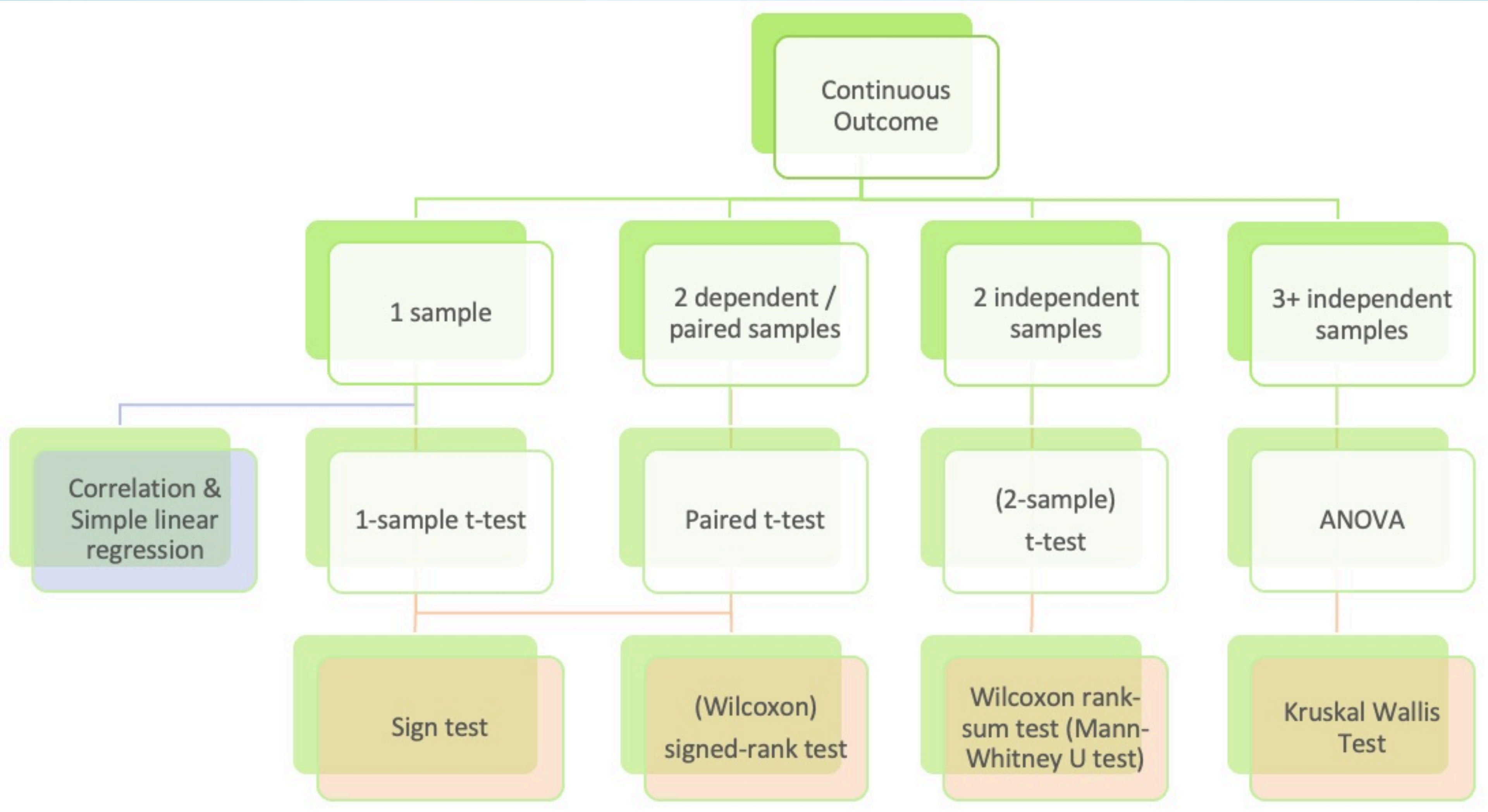
- How to **calculate** CI's?
- How to **interpret** & **NOT** interpret CI's?

- What if we don't know σ?
- Student's **t-distribution**

Where are we?


R
Quarto
Packages
Data types
Data wrangling
Data
Collecting Data
Categorical vs. Numeric
Summary statistics
Data Visualization
Probability
Probability rules
Independence, Conditional, Bayes' Theorem
Random variables & probability distributions
Linear combinations
Binomial, Normal, and Poisson
Sampling Variability & Statistical Inference
Sampling distributions
Central Limit Theorem
Confidence Intervals
Hypothesis tests
Inference for continuous data
Continuous Outcome
1 sample
2 dependent / paired samples
2 independent samples
3+ independent samples
Correlation & Simple linear regression
1-sample t-test
Paired t-test
(2-sample) t-test
ANOVA
Sign test
(Wilcoxon) signed-rank test
Wilcoxon rank-sum test (Mann-Whitney U test)
Kruskal Wallis Test
Inference for categorical data

Where are we? Continuous outcome zoomed in



Our hypothetical population: YRBSS

Youth Risk Behavior Surveillance System **(YRBSS)**

- Yearly survey conducted by the US Centers for Disease Control (CDC)
- "A set of surveys that track behaviors that can lead to poor health in students grades 9 through 12."[1]
- Dataset `yrbss` from `oibiostat` pacakge contains responses from *n* = 13,583 participants in 2013 for a subset of the variables included in the complete survey data

```r
library(oibiostat)
data("yrbss")  #load the data
# ?yrbss
```

```r
dim(yrbss)
```

```
[1] 13583    13
```

```r
names(yrbss)
```

```
 [1] "age"                        "gender"
 [3] "grade"                      "hispanic"
 [5] "race"                       "height"
 [7] "weight"                     "helmet.12m"
 [9] "text.while.driving.30d"     "physically.active.7d"
[11] "hours.tv.per.school.day"    "strength.training.7d"
[13] "school.night.hours.sleep"
```


1. Youth Risk Behavior Surveillance System https://www.cdc.gov/healthyyouth/data/yrbs/index.htm (YRBSS)

Transform height & weight from metric to to standard

Also, drop missing values and add a column of id values

```
yrbss2 <- yrbss %>%                      # save new dataset with new name
  mutate(                                # add variables for
    height.ft = 3.28084*height,          #     height in feet
    weight.lb = 2.20462*weight           #     weight in pounds
  ) %>%
  drop_na(height.ft, weight.lb) %>% # drop rows w/ missing height/weight values
  mutate(id = 1:nrow(.)) %>%             # add id column
  select(id, height.ft, weight.lb)  # restrict dataset to columns of interest

head(yrbss2)
```

```
  id height.ft weight.lb
1  1  5.675853  186.0038
2  2  5.249344  122.9957
3  3  4.921260  102.9998
4  4  5.150919  147.9961
5  5  5.413386  289.9957
6  6  6.167979  157.0130
```

```
dim(yrbss2)
```

```
[1] 12579     3
```

```
# number of rows deleted that had missing values for height and/or weight:
nrow(yrbss) - nrow(yrbss2)
```

```
[1] 1004
```

yrbss2: stats for height in feet

```r
summary(yrbss2)
```

```
       id          height.ft       weight.lb     
 Min.   :    1   Min.   :4.167   Min.   : 66.01  
 1st Qu.: 3146   1st Qu.:5.249   1st Qu.:124.01  
 Median : 6290   Median :5.512   Median :142.00  
 Mean   : 6290   Mean   :5.549   Mean   :149.71  
 3rd Qu.: 9434   3rd Qu.:5.840   3rd Qu.:167.99  
 Max.   :12579   Max.   :6.923   Max.   :399.01  
```

```r
(mean_height.ft <- mean(yrbss2$height.ft))
```

```
[1] 5.548691
```

```r
(sd_height.ft <- sd(yrbss2$height.ft))
```

```
[1] 0.3434949
```

10,000 samples of size n = 30 from `yrbss2`

Take 10,000 random samples of size n = 30 from `yrbss2`:

```
samp_n30_rep10000 <- yrbss2 %>%
  rep_sample_n(size = 30,
               reps = 10000,
               replace = FALSE)
samp_n30_rep10000
```

```
# A tibble: 300,000 × 4
# Groups:   replicate [10,000]
   replicate    id height.ft weight.lb
       <int> <int>     <dbl>     <dbl>
 1         1  5869      5.15      145.
 2         1  6694      5.41      127.
 3         1  2517      5.74      130.
 4         1  5372      6.07      180.
 5         1  5403      6.07      163.
 6         1  2329      6.07      182.
 7         1  8863      5.25      125.
 8         1  8058      5.84      135.
 9         1   335      6.17      235.
10         1  4698      5.58      124.
# ℹ 299,990 more rows
```

Calculate the mean for each of the 10,000 random samples:

```
means_hght_samp_n30_rep10000 <-
  samp_n30_rep10000 %>%
  group_by(replicate) %>%
  summarise(mean_height =
              mean(height.ft))

means_hght_samp_n30_rep10000
```

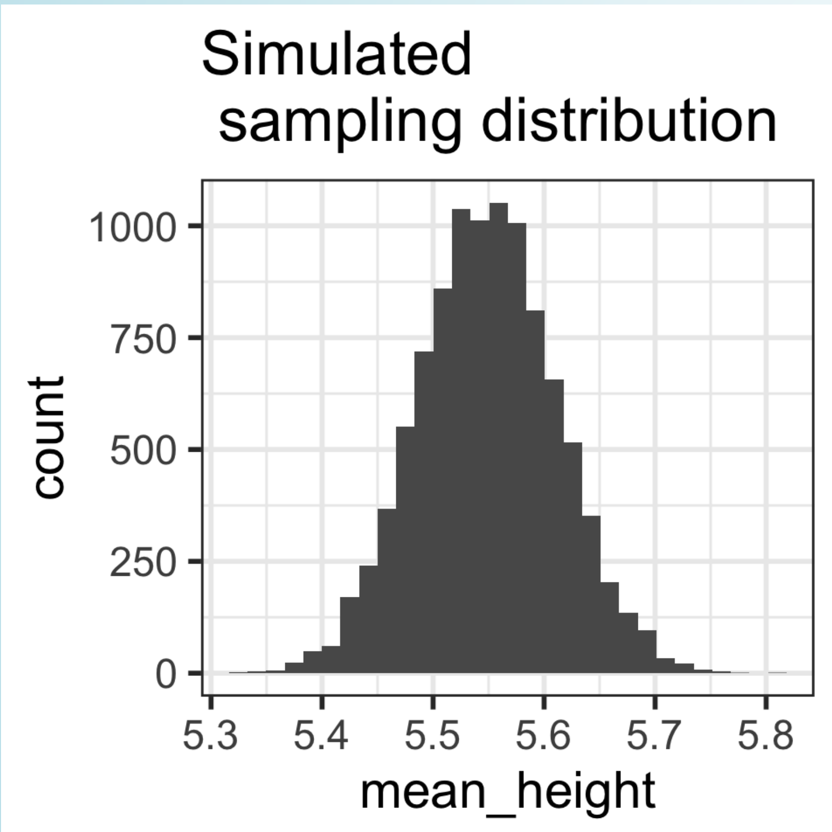
```
# A tibble: 10,000 × 2
   replicate mean_height
       <int>       <dbl>
 1         1        5.59
 2         2        5.59
 3         3        5.51
 4         4        5.65
 5         5        5.64
 6         6        5.57
 7         7        5.61
 8         8        5.60
 9         9        5.52
10        10        5.64
# ℹ 9,990 more rows
```

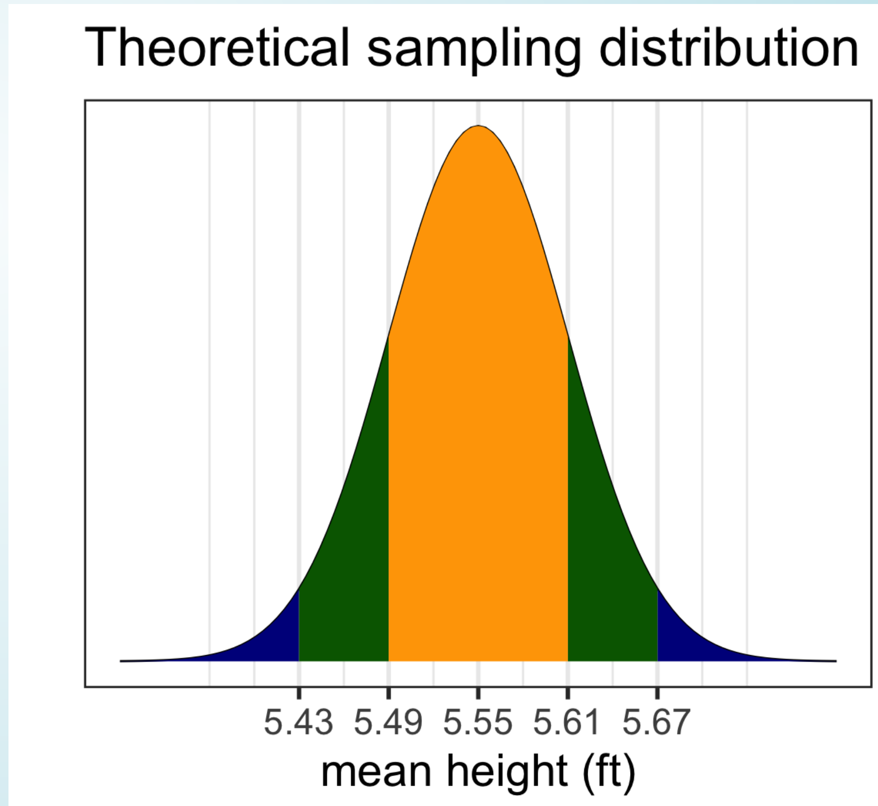
How close are the mean heights for each of the 10,000 random samples?

Simulated sampling distribution for n = 30 using 10,000 sample mean heights

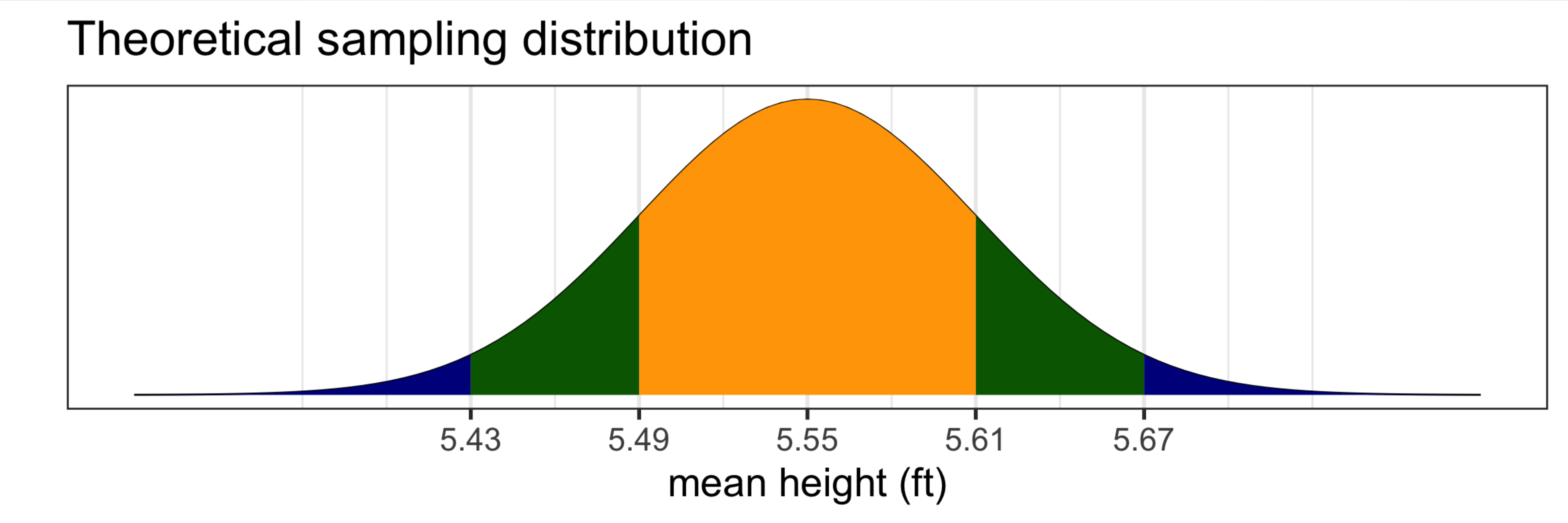
```r
ggplot(
  means_hght_samp_n30_rep10000,
  aes(x = mean_height)) +
  geom_histogram() +
  labs(title = "Simulated \n sampling
```

CLT tells us that we can model the sampling distribution of mean heights using a normal distribution.





Given $\bar{x}$, what are plausible values of μ?

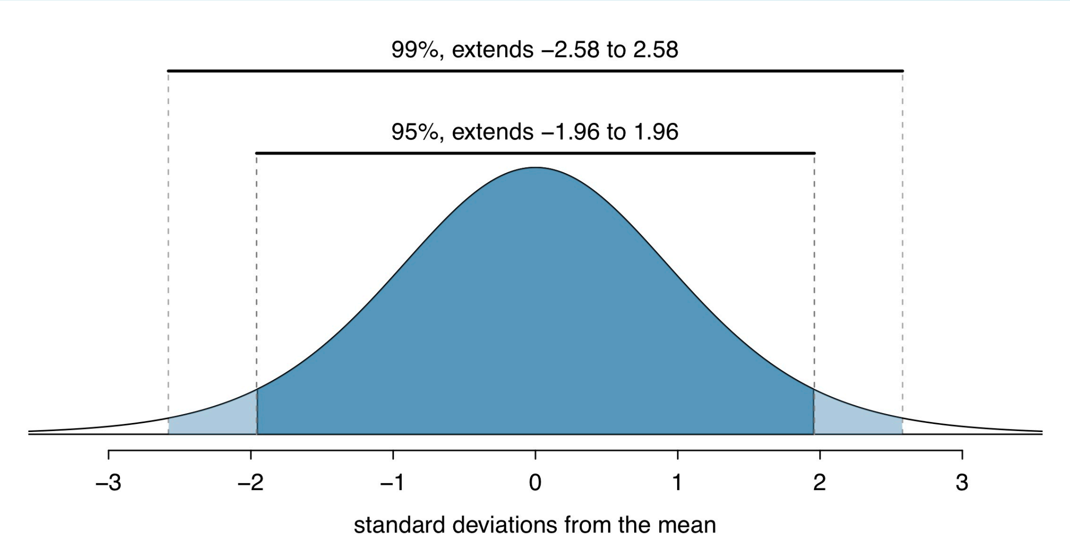


Confidence interval (C I) for the mean μ

$$\bar{x} \pm z^* \times \mathrm{SE}$$

where

- $SE = \frac{\sigma}{\sqrt{n}}$



- z^* depends on the confidence level
  - For a 95% CI, z^* is chosen such that 95% of the standard normal curve is between $-z^*$ and z^*

```r
qnorm(.975)
```

```
[1] 1.959964
```

```r
qnorm(.995)
```

```
[1] 2.575829
```

When can this be applied?

Example: C I for mean height

- A random sample of 30 high schoolers has mean height 5.6 ft.
- Find the 95% confidence interval for the population mean, assuming that the population standard deviation is 0.34 ft.

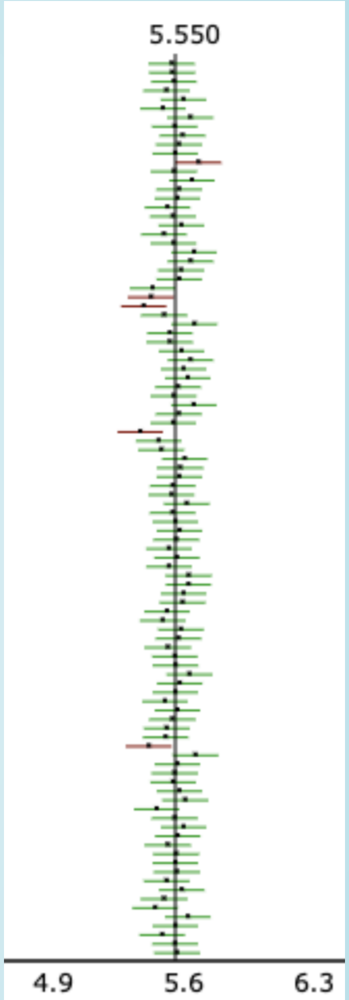
How to interpret a C I? (1/2)

Simulating Confidence Intervals:

http://www.rossmanchance.com/applets/ConfSim.html

The figure shows CI's from 100 simulations.

- The true value of μ = 5.55 is the vertical black line.
- The horizontal lines are 95% CI's from 100 samples.
  - **Green**: the CI "captured" the true value of μ
  - **Red**: the CI *did not* "capture" the true value of μ



Question:

What percent of CI's captured the true value of μ ?

How to interpret a C I? (2/2)

Actual interpretation:

- If we were to
  - **repeatedly take random samples** from a population and
  - calculate a 95% CI for each random sample,
- then we would **expect 95% of our CI's to contain the true population parameter** μ.

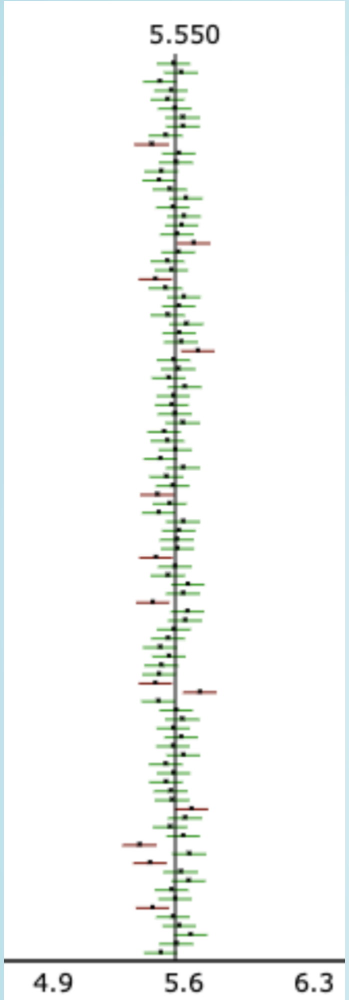
What we typically write as "shorthand":

- We are 95% *confident* that (the 95% confidence interval) captures the value of the population parameter.

WRONG interpretation:

- There is a 95% *chance* that (the 95% confidence interval) captures the value of the population parameter.
  - For one CI on its own, it either does or doesn't contain the population parameter with probability 0 or 1. We just don't know which!

What percent CI was being simulated in this figure?



100 CI's are shown in the figure.

Interpretation of the mean heights C I

Correct interpretation:

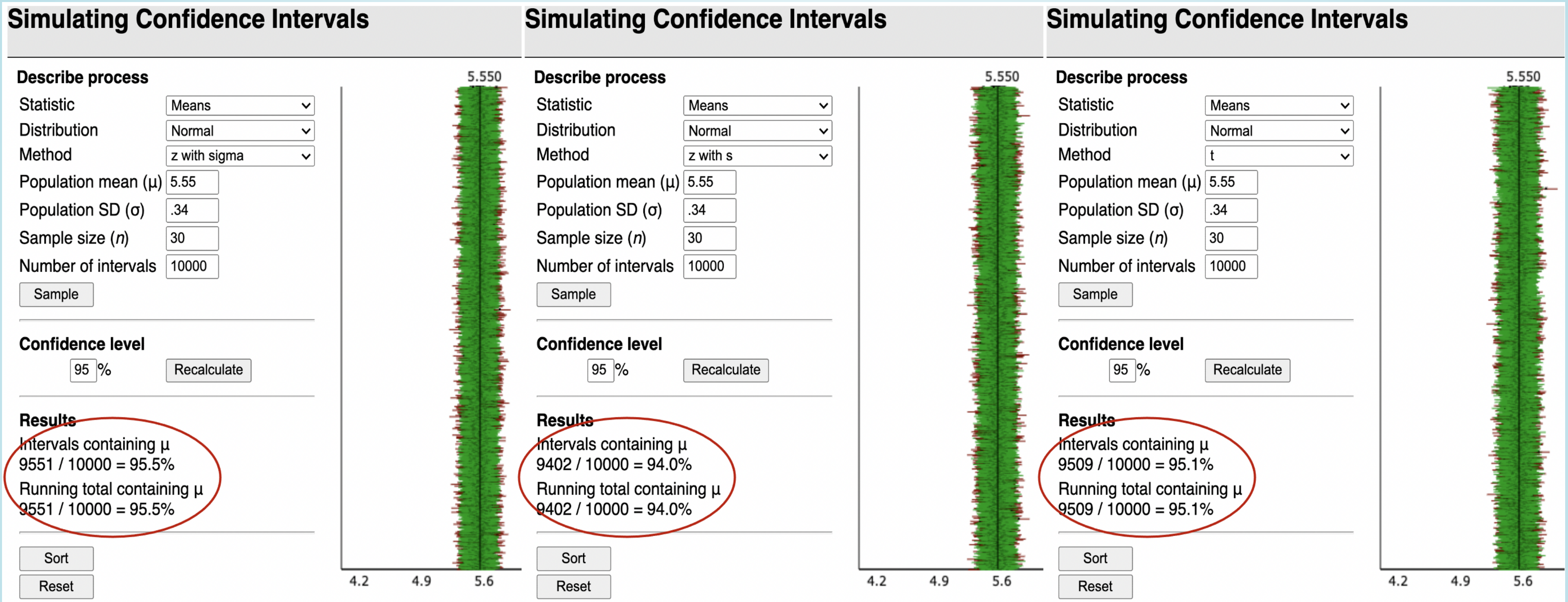
- We are 95% *confident* that the mean height for high schoolers is between 5.43 and 5.67 feet.

WRONG:

- There is a 95% *chance* that the mean height for high schoolers is between 5.43 and 5.67 feet.

What if we don't know σ ? (1/3)

Simulating Confidence Intervals:
http://www.rossmanchance.com/applets/ConfSim.html



The normal distribution doesn't have a 95% "coverage rate" when using s instead of σ

What if we don't know σ ? (2/3)

- In real life, we don't know what the population sd is (σ)
- If we replace σ with s in the SE formula, we add in additional variability to the SE!

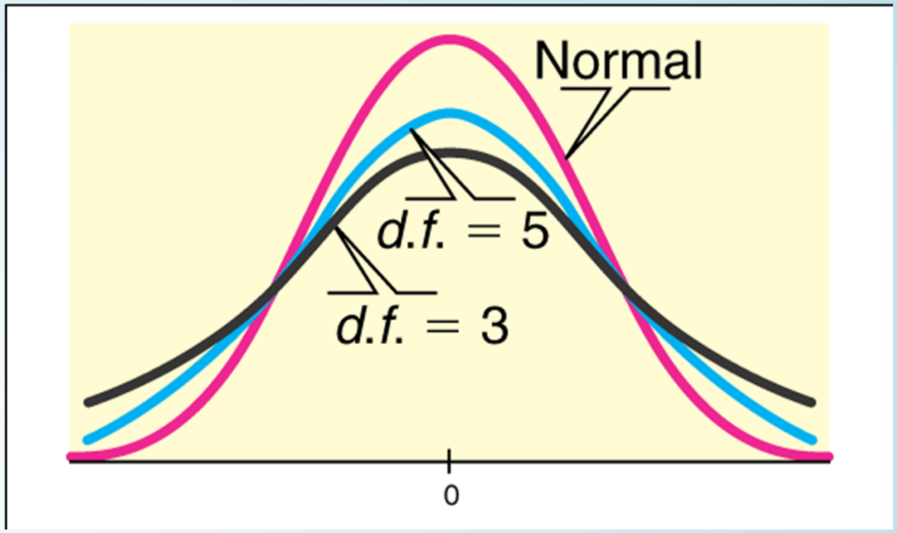
$$\frac{\sigma}{\sqrt{n}} \quad \text{vs.} \quad \frac{s}{\sqrt{n}}$$

- Thus when using s instead of σ when calculating the SE, we **need a different probability distribution** with thicker tails than the normal distribution.
  - In practice this will mean using a different value than 1.96 when calculating the CI.

What if we don't know σ ? (3/3)

The **Student's t-distribution**:

- Is bell shaped and symmetric with mean = 0.
- Its tails are a thicker than that of a normal distribution
  - The "thickness" depends on its **degrees of freedom**: $df = n – 1$, where n = sample size.
- As the degrees of freedom (sample size) increase,
  - the tails are less thick, and
  - the t-distribution is more like a normal distribution
  - in theory, with an infinite sample size the *t*-distribution is a normal distribution.



Calculating the C I for the population mean using s

CI for μ:

$$\bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$

where t^* is determined by the t-distribution and dependent on the **df =** $n-1$ and the confidence level

- `qt` gives the quartiles for a t-distribution. Need to specify
  - the percent under the curve to the left of the quartile
  - the degrees of freedom = n-1
- Note in the R output to the right that t^* gets closer to 1.96 as the sample size increases.

```r
qt(.975, df=9)   # df = n-1
```

```
[1] 2.262157
```

```r
qt(.975, df=49)
```

```
[1] 2.009575
```

```r
qt(.975, df=99)
```

```
[1] 1.984217
```

```r
qt(.975, df=999)
```

```
[1] 1.962341
```

Using a t-table to get t^*

B.2 t-Probability Table

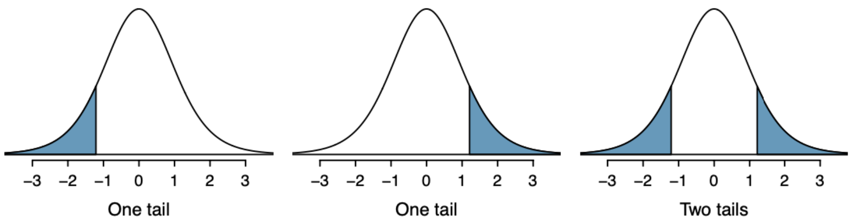


Figure B.1: Tails for the t-distribution.

one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05
13	1.35	1.77	2.16	2.65	3.01
14	1.35	1.76	2.14	2.62	2.98
15	1.34	1.75	2.13	2.60	2.95
16	1.34	1.75	2.12	2.58	2.92
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
21	1.32	1.72	2.08	2.52	2.83
22	1.32	1.72	2.07	2.51	2.82
23	1.32	1.71	2.07	2.50	2.81
24	1.32	1.71	2.06	2.49	2.80
25	1.32	1.71	2.06	2.49	2.79
26	1.31	1.71	2.06	2.48	2.78
27	1.31	1.70	2.05	2.47	2.77
28	1.31	1.70	2.05	2.47	2.76
29	1.31	1.70	2.05	2.46	2.76
30	1.31	1.70	2.04	2.46	2.75

one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 31	1.31	1.70	2.04	2.45	2.74
32	1.31	1.69	2.04	2.45	2.74
33	1.31	1.69	2.03	2.44	2.73
34	1.31	1.69	2.03	2.44	2.73
35	1.31	1.69	2.03	2.44	2.72
36	1.31	1.69	2.03	2.43	2.72
37	1.30	1.69	2.03	2.43	2.72
38	1.30	1.69	2.02	2.43	2.71
39	1.30	1.68	2.02	2.43	2.71
40	1.30	1.68	2.02	2.42	2.70
41	1.30	1.68	2.02	2.42	2.70
42	1.30	1.68	2.02	2.42	2.70
43	1.30	1.68	2.02	2.42	2.70
44	1.30	1.68	2.02	2.41	2.69
45	1.30	1.68	2.01	2.41	2.69
46	1.30	1.68	2.01	2.41	2.69
47	1.30	1.68	2.01	2.41	2.68
48	1.30	1.68	2.01	2.41	2.68
49	1.30	1.68	2.01	2.40	2.68
50	1.30	1.68	2.01	2.40	2.68
60	1.30	1.67	2.00	2.39	2.66
70	1.29	1.67	1.99	2.38	2.65
80	1.29	1.66	1.99	2.37	2.64
90	1.29	1.66	1.99	2.37	2.63
100	1.29	1.66	1.98	2.36	2.63
150	1.29	1.66	1.98	2.35	2.61
200	1.29	1.65	1.97	2.35	2.60
300	1.28	1.65	1.97	2.34	2.59
400	1.28	1.65	1.97	2.34	2.59
500	1.28	1.65	1.96	2.33	2.59
∞	1.28	1.65	1.96	2.33	2.58

Example: C I for mean height (revisited)

- A random sample of 30 high schoolers has mean height 5.6 ft and standard deviation 0.34 ft.
- Find the 95% confidence interval for the population mean.

z vs t??

(& important comment about Chapter 4 of textbook)

Textbook's rule of thumb

- (Ch 4) If $n \geq 30$ and population distribution not strongly skewed:
  - **Use normal distribution**
  - **No matter if using σ or s for the SE**
  - If there is skew or some large outliers, then need $n \geq 50$
- (Ch 5) If $n < 30$ and data approximately symmetric with no large outliers:
  - Use Student's t-distribution

BSTA 511 rule of thumb

- Use normal distribution ONLY if know σ
  - If using s for the SE, then use the Student's t-distribution

For either case, can apply if either

- $n \geq 30$ and population distribution not strongly skewed
  - If there is skew or some large outliers, then $n \geq 50$ gives better estimates
- $n < 30$ and data approximately symmetric with no large outliers

If do not know population distribution, then check the distribution of the data.