015 - Midterm Review

EPIB 607 - FALL 2020

Sahir Rai Bhatnagar
Department of Epidemiology, Biostatistics, and Occupational Health
McGill University

`sahir.bhatnagar@mcgill.ca`

slides compiled on October 21, 2020

Exam Details

Data visualization

Tidy Data

Descriptive statistics

Sampling Distributions, CLT, Confidence Intervals and p-values

Bootstrap

One sample mean

One sample proportion

p-values

Power and sample size

Exam Details

- **When:** Friday October 23, 3:35pm - 5:35pm.
- This is a 2 hour, open book exam. Any material on myCourses (EPIB607/613) and personal notes are permitted.
- You are not permitted to use the internet and you must work alone. Using the internet or obtaining help from anyone else is considered Cheating as per Article 17 of the Code of Student Conduct and Disciplinary Procedures
- Provide units and state your assumptions when applicable. Label axes and write answers in complete sentences when appropriate.
- The format of the exam will follow the assignments. That is, you will be required to complete a series of questions in an RMarkdown document and knit to pdf. Your solutions for each question must then be uploaded to Crowdmark. A template will be provided.
- You will be given 1 extra hour to upload your solutions.

Topics to be covered

1. Data visualization (histograms, boxplots, scatterplots, line plots), Tidy Data, Color Palettes
2. Descriptive statistics (mean, median, range, IQR, sd, correlation)
3. Normal Curve Calculations
4. Sampling Distributions, CLT, Bootstrap
5. Confidence intervals and p-values for one sample mean and one sample proportion
6. Hypothesis Testing
7. Power and Sample size calculations

Exam Details

Data visualization

Tidy Data

Descriptive statistics

Sampling Distributions, CLT, Confidence Intervals and p-values

Bootstrap

One sample mean

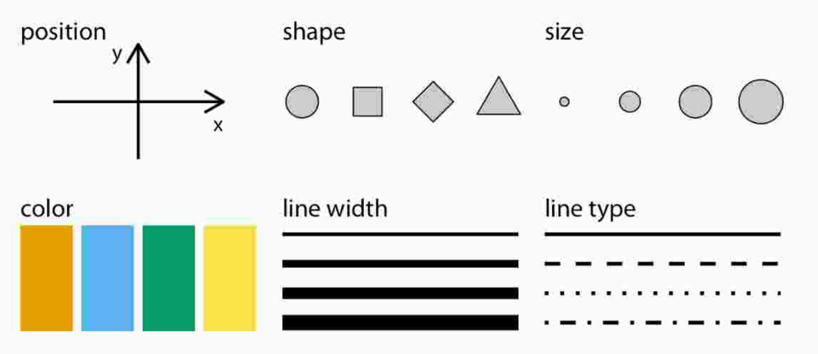
One sample proportion

p-values

Power and sample size

Aesthetics

- Aesthetics



- Commonly used aesthetics in data visualization: position, shape, size, color, line width, line type. Some of these aesthetics can represent both continuous and discrete data (position, size, line width, color) while others can only represent discrete data (shape, line type)

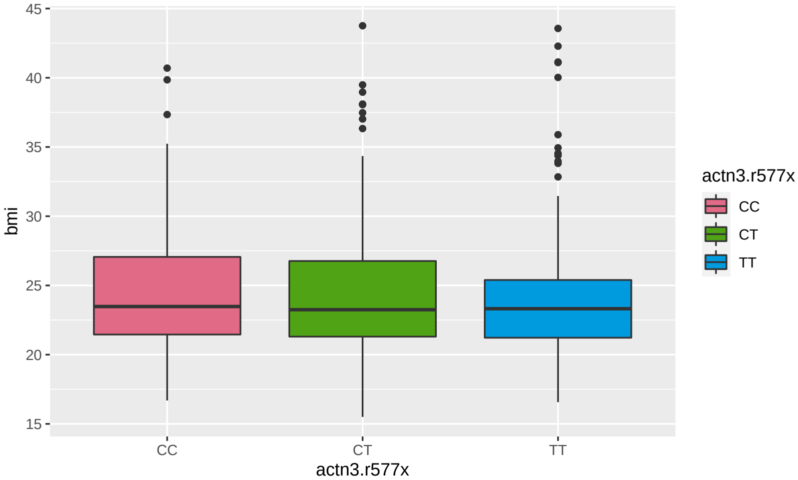
Types of Graphs

- Review the types of graphs created in the assignments.
- You should be able to critique a graph and propose appropriate graphics for a given dataset. Be mindful of the research question. The graphic should try to answer the research question.
- `https://serialmentor.com/dataviz/directory-of-visualizations.html`
- `https://www.data-to-viz.com/`

Boxplots with qualitative palette

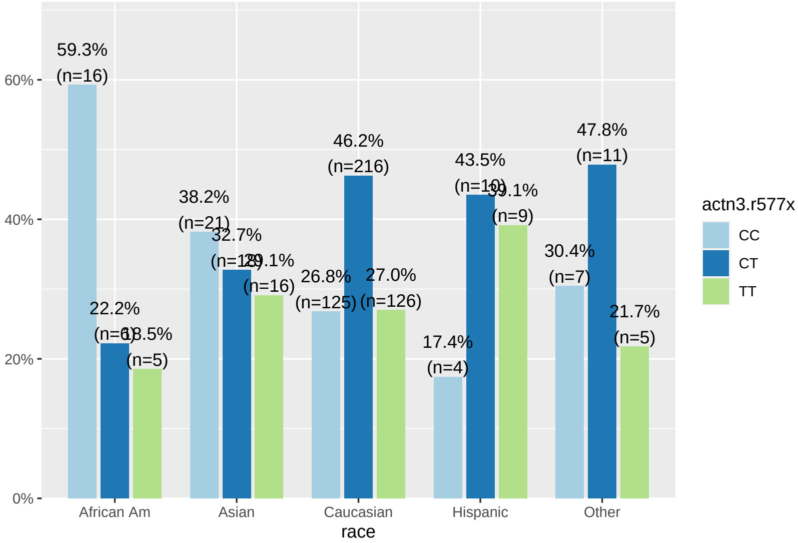
```r
library(oibiostat); data("famuss")
library(ggplot2)
library(colorspace)

ggplot(famuss, aes(x = actn3.r577x, y = bmi, fill = actn3.r577x)) +
geom_boxplot() +
colorspace::scale_fill_discrete_qualitative()
```



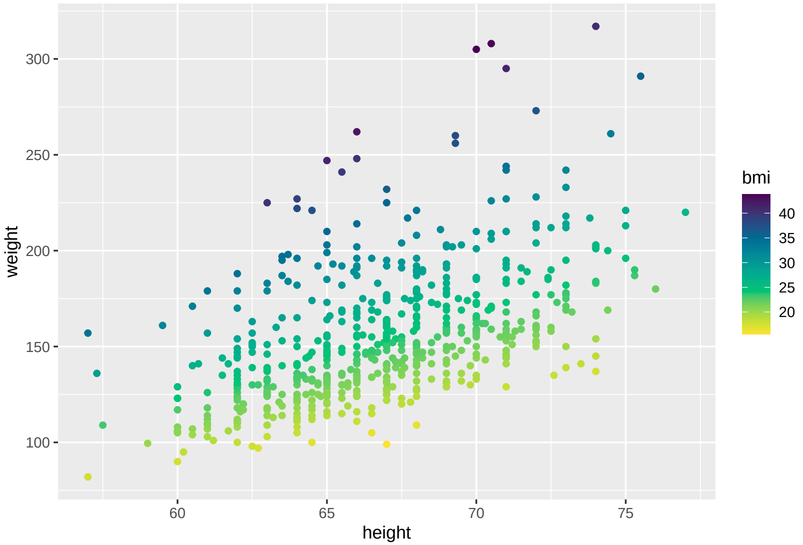
Conditional distribution of genotype *given* race

```
sjPlot::plot_xtab(famuss$race, famuss$actn3.r577x, margin = "row")
```



Scatter plots with sequential palette

```r
ggplot(famuss, aes(x = height, y = weight, color = bmi)) +
geom_point() +
colorspace::scale_color_continuous_sequential(palette = "Viridis")
```

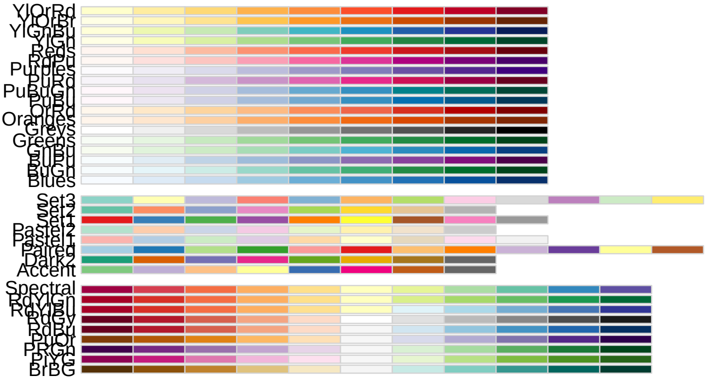


Variable Types

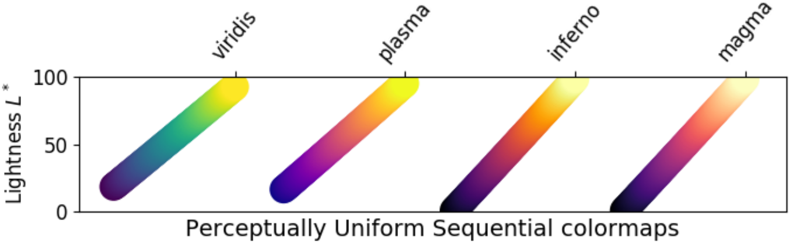
- quantitative/numerical continuous (1.3, 5.7, 83, 1.5×10^{-2})
- quantitative/numerical discrete (1,2,3,4)
- qualitative/categorical unordered (dog, cat, fish)
- qualitative/categorical ordered (good, fair, poor)

Color Palettes: Cynthia Brewer

```r
pacman::p_load(RColorBrewer)
RColorBrewer::display.brewer.all()
```



Color Palettes: viridis



Exam Details

Data visualization

Tidy Data

Descriptive statistics

Sampling Distributions, CLT, Confidence Intervals and p-values

Bootstrap

One sample mean

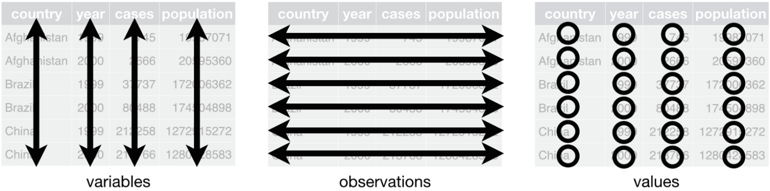
One sample proportion

p-values

Power and sample size

Tidy data

- Each variable forms a column.
- Each observation forms a row.
- Each type of observational units forms a table
- Tidy data is ready for regression routines and plotting



Example: Is it tidy?

Mode of Delivery	Covariate			No. of Mother–Child Pairs	No. of HIV-1–Infected Children
	NO. OF PERIODS OF ANTIRETROVIRAL THERAPY	ADVANCED MATERNAL DISEASE	LOW BIRTH WEIGHT OF INFANT (<2500 g)		
Elective cesarean	0	No	No	372	30
Other	0	No	No	3850	652
Elective cesarean	0	Yes	No	28	5
Other	0	Yes	No	303	74
Elective cesarean	0	No	Yes	110	17
Other	0	No	Yes	767	196
Elective cesarean	0	Yes	Yes	27	4
Other	0	Yes	Yes	114	40
Elective cesarean	1 or 2	No	No	41	0
Other	1 or 2	No	No	441	49
Elective cesarean	1 or 2	Yes	No	23	3
Other	1 or 2	Yes	No	186	33
Elective cesarean	1 or 2	No	Yes	7	0
Other	1 or 2	No	Yes	83	22
Elective cesarean	1 or 2	Yes	Yes	10	3
Other	1 or 2	Yes	Yes	54	19
Elective cesarean	3	No	No	124	2
Other	3	No	No	878	49
Elective cesarean	3	Yes	No	34	1
Other	3	Yes	No	208	24
Elective cesarean	3	No	Yes	25	0
Other	3	No	Yes	109	11
Elective cesarean	3	Yes	Yes	8	1
Other	3	Yes	Yes	38	6

Exam Details

Data visualization

Tidy Data

Descriptive statistics

Sampling Distributions, CLT, Confidence Intervals and p-values

Bootstrap

One sample mean

One sample proportion

p-values

Power and sample size

Descriptive statistics

- Boxplots, histograms, density plot
- IQR, median, mode, mean, min, max, range
- Q1, Q3
- Skewness (long left/right tail)
- Correlation

Descriptive stats by group

```
library(oibiostat); data("famuss")
library(dplyr)

famuss %>%
  dplyr::group_by(actn3.r577x) %>%
  dplyr::summarise(mean_bmi = mean(bmi),
                   sd_bmi = sd(bmi))

## # A tibble: 3 x 3
##   actn3.r577x mean_bmi sd_bmi
##   <fct>          <dbl>  <dbl>
## 1 CC              24.5   4.41
## 2 CT              24.5   4.55
## 3 TT              24.2   4.81
```

Subsetting data

```r
library(oibiostat); data("famuss")
library(dplyr)

f.male <- famuss %>%
             dplyr::filter(sex == "Male")


f.male.cauc <- famuss %>%
                  dplyr::filter(sex == "Male" & race == "Caucasian")

    f.bmi.low <- famuss %>%
                     dplyr::filter(bmi <= 23)
```

Standard error (SE) of a sample statistic

- Recall: When we are talking about the variability of a **statistic**, we use the term **standard error** (not standard deviation). The standard error of the sample mean is $\sigma/\sqrt{n}$.

Remark (SE vs. SD)

In quantifying the instability of the sample mean ($\bar{y}$) statistic, we talk of SE of the mean (SEM)

SE($\bar{y}$) describes how far $\bar{y}$ could (typically) deviate from μ;

SD(y) describes how far an individual y (typically) deviates from μ (or from $\bar{y}$).

Exam Details

Data visualization

Tidy Data

Descriptive statistics

Sampling Distributions, CLT, Confidence Intervals and p-values

Bootstrap

One sample mean

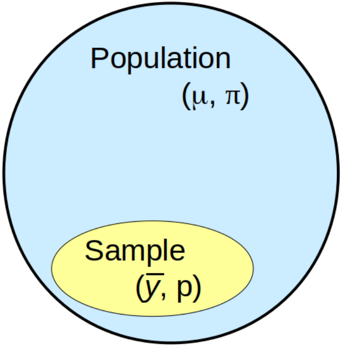
One sample proportion

p-values

Power and sample size

Parameters, Samples, and Statistics

- **Paramter**: An unknown numerical constant pertaining to a population/universe, or in a statistical model.
  - μ: population mean π: population proportion
- **Statistic**: A numerical quantity calculated from a sample. The empirical counterpart of the parameter, used to *estimate* it.
  - $\bar{y}$: sample mean p: sample proportion



Samples must be random

- The validity of inference will depend on the way that the sample was collected. If a sample was collected badly, no amount of statistical sophistication can rescue the study.

- Samples should be **random**. That is, there should be no systematic set of characteristics that is related to the scientific question of interest that causes some people to be more likely to be sampled than others. The simplest type of randomization selects members from the population with equal probability (a uniform distribution).

- **Do not cheat by**
    - Taking 5 people from the same household to estimate
        - proportion of Québécois who don't have a family doctor
        - who saw a medical doctor last year
        - average rent
    - Sampling the depth of the ocean only around Montreal to estimate
        - proportion of Earth's surface covered by water

Sampling Distributions

Definition (Sampling Distribution)

- *The sampling distribution of a statistic is the distribution of values taken by the statistic in* ***all possible samples of the same size*** *from the same population.*

- *The standard deviation of a sampling distribution is called a* ***standard error***

Sampling Distributions

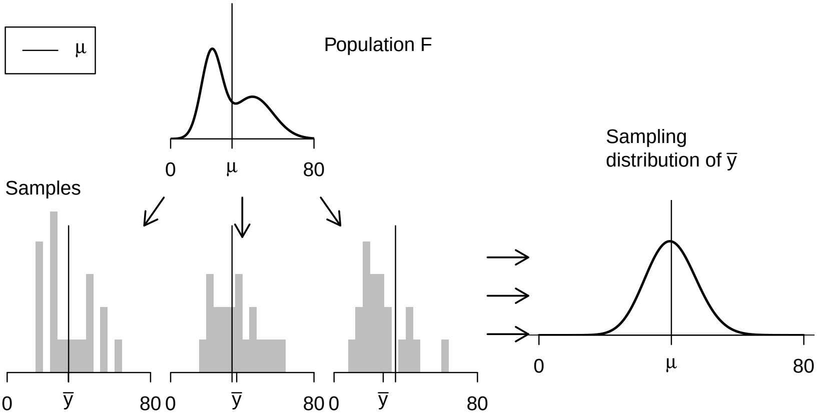


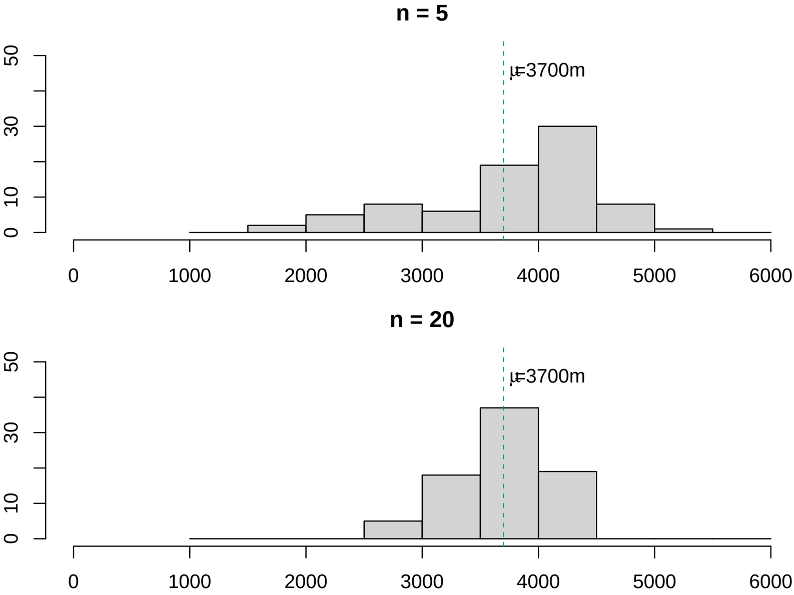
Figure: Ideal world. Sampling distributions are obtained by drawing repeated samples from the population, computing the statistic of interest for each, and collecting (an infinite number of) those statistics as the sampling distribution

Why are sampling distributions important?

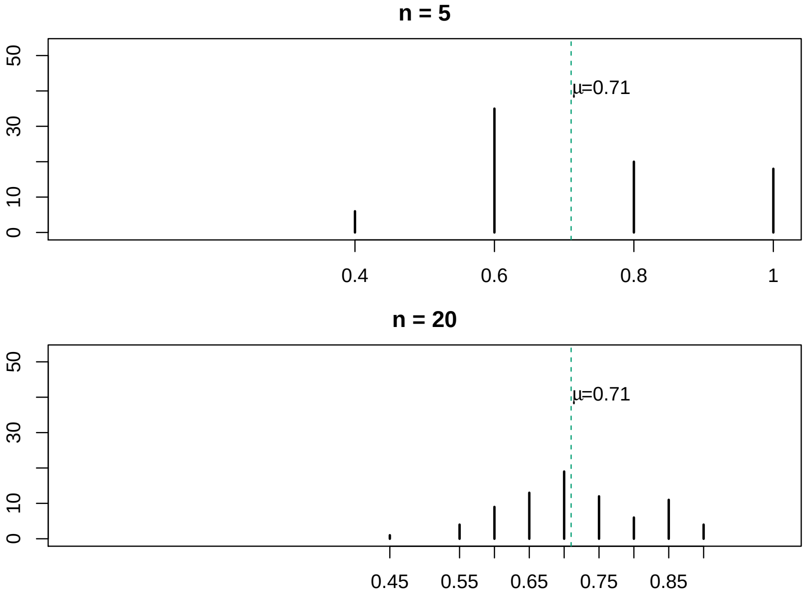
- They tell us how far from the target (true value of the parameter) our statistical shot at it (i.e. the statistic calculated form a sample) is likely to be, or, to have been.

- Thus, they are used in confidence intervals for parameters. Specific sampling distributions (based on a null value for the parameter) are also used in statistical tests of hypotheses.

Sampling distribution: mean depth of the ocean



Sampling distribution: proportion covered by water

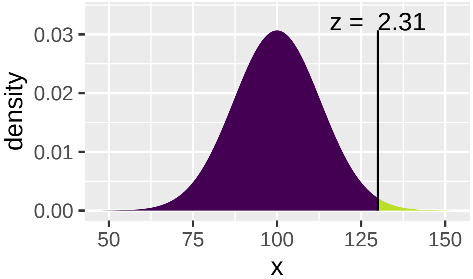


Normal Distribution: For probabilities we use *pnorm*

```
stats::pnorm(q = 130, mean = 100, sd = 13)

## [1] 0.99
```

```
mosaic::xpnorm(q = 130, mean = 100, sd = 13)
```



```
## [1] 0.99
```

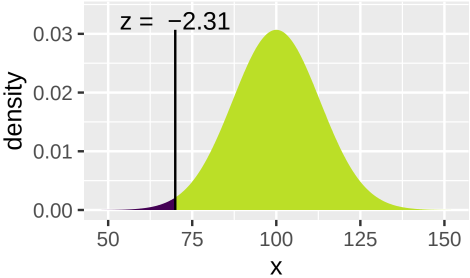
- `pnorm` returns the integral from $-\infty$ to q for a $\mathcal{N}(\mu, \sigma)$
- `pnorm` goes from *quantiles* (think Z scores) to probabilities

Normal Distribution: For quantiles we use *qnorm*

```
stats::qnorm(p = 0.0104, mean = 100, sd = 13)

## [1] 70
```

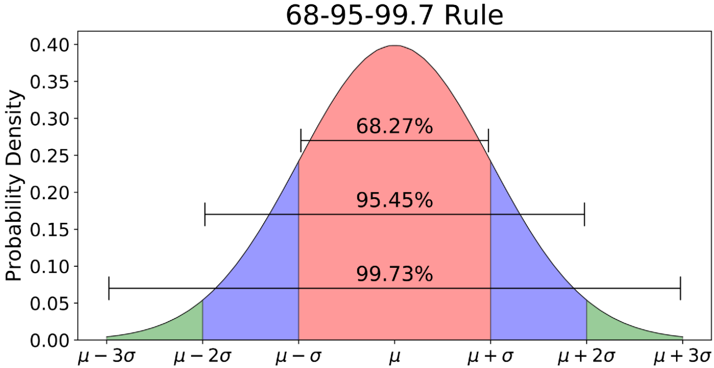
```
mosaic::xqnorm(p = 0.0104, mean = 100, sd = 13)
```



```
## [1] 70
```

- `qnorm` answers the question: What is the Z-score of the *p*th percentile of the normal distribution?
- `qnorm` goes from *probabilities* to quantiles

Empirical Rule or 68-95-99.7% Rule



Quadruple the work, half the benefit

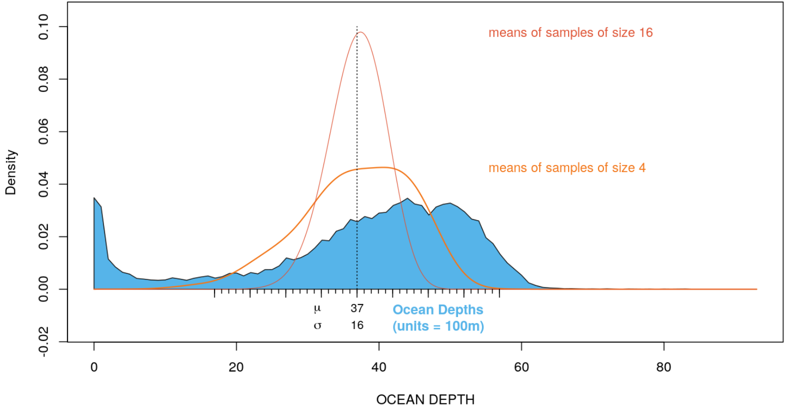


Figure: When the sample size increases from 4 to 16, the spread of the sampling distribution for the mean is reduced by a half, i.e., the range is cut in half. This is known as the curse of the $\sqrt{n}$

The Central Limit Theorem (CLT)

- The sampling distribution of $\bar{y}$ is, for a large enough n, close to Gaussian in shape no matter what the shape of the distribution of individual Y values.
- This phenomenon is referred to as the CENTRAL LIMIT THEOREM
- The CLT applied also to a $\underline{\text{sample proportion}}$, $\underline{\text{slope}}$, $\underline{\text{correlation}}$, or any other statistic created by $\underline{\text{aggregation of individual observations}}$

Theorem (Central Limit Theorem)

$$if\ Y \sim ???(\mu_Y, \sigma_Y),\ then$$

$$\bar{y} \sim \mathcal{N}(\mu_Y, \sigma_Y / \sqrt{n})$$

Confidence Interval

Definition (Confidence Interval)

A level C confidence interval for a parameter has two parts:

1. *An interval calculated from the data, usually of the form*

   $$\textit{estimate} \pm \textit{margin of error}$$

   where the estimate is a sample statistic and the margin of error represents the accuracy of our guess for the parameter.

2. *A confidence level C, which gives the probability that the interval will capture the true parameter value in different possible samples. That is, the confidence level is the success rate for the method*

Confidence Interval: A simulation study

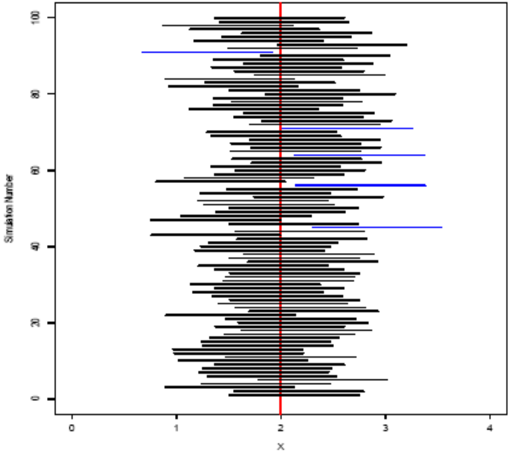


Figure: True parameter value is 2 (red line). Each horizontal black line represents a 95% CI from a sample and contains the true parameter value. The blue CIs do not contain the true parameter value. 95% of all samples give an interval that contains the population parameter.

Interpreting a frequentist confidence interval

- The confidence level is the success rate of the method that produces the interval.

- We don't know whether the 95% confidence interval from a particular sample is one of the 95% that capture θ (the unknown population parameter), or one of the unlucky 5% that miss.

- To say that we are 95% confident that the unknown value of θ lies between U and L is shorthand for "We got these numbers using a method that gives correct results 95% of the time."

68% Confidence interval using `qnorm`

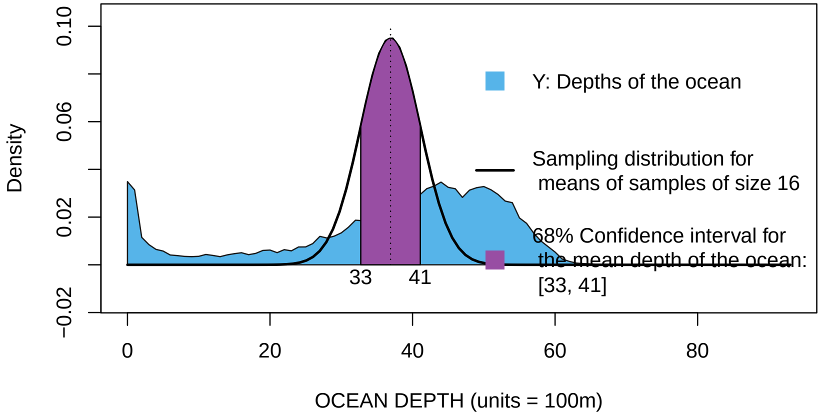


Figure: 68% Confidence interval calculated using `qnorm(p = c(0.16,0.84), mean = 37, sd = 4.2)`

95% Confidence interval using `qnorm`

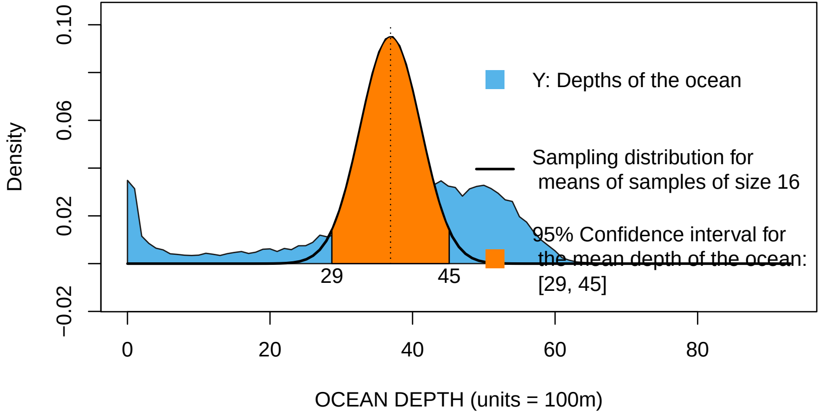


Figure: 95% Confidence interval calculated using
`qnorm(p = c(0.025,0.975), mean = 37, sd = 4.2)`

Example: Inference for a single population mean

So what does the CI allow us to learn about μ??

- It tells us that if we repeated this procedure again and again (collecting a sample mean, and constructing a 95% CI), 95% of the time, the CI would *cover* μ.

- That is, with 95% probability, the *procedure* will include the true value of μ. Note that we are making a probability statement about the CI, not about the parameter.

- Unfortunately, we do not know whether the true value of μ is contained in the CI in the particular experiment that we have performed.

Exam Details

Data visualization

Tidy Data

Descriptive statistics

Sampling Distributions, CLT, Confidence Intervals and p-values

Bootstrap

One sample mean

One sample proportion

p-values

Power and sample size

Motivation for the Bootstrap

- The ± and `qnorm` methods to calculate a CI both require the CLT

Q: What happens if the CLT hasn't 'kicked in'? Or you don't believe the CLT?

Q: What happens if there is no formula available to calculate a CI?

A: Bootstrap

Ideal world: known sampling distribution

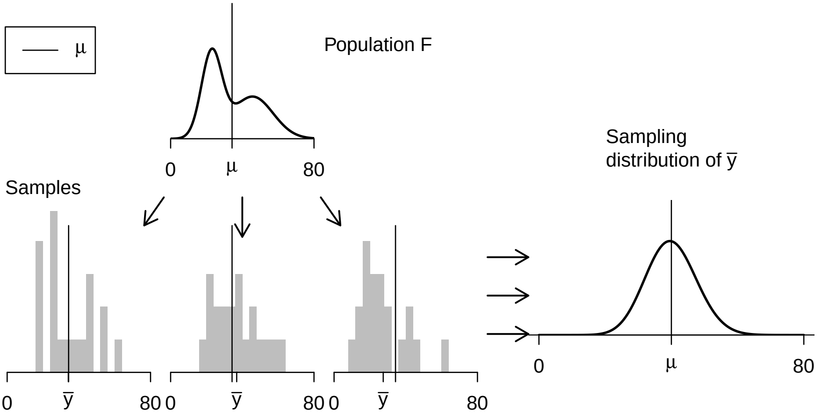


Figure: Ideal world. Sampling distributions are obtained by drawing repeated samples from the population, computing the statistic of interest for each, and collecting (an infinite number of) those statistics as the sampling distribution

Reality: use the bootstrap distribution instead

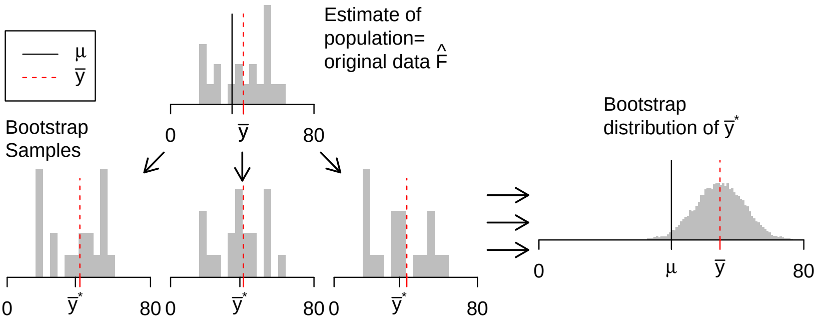
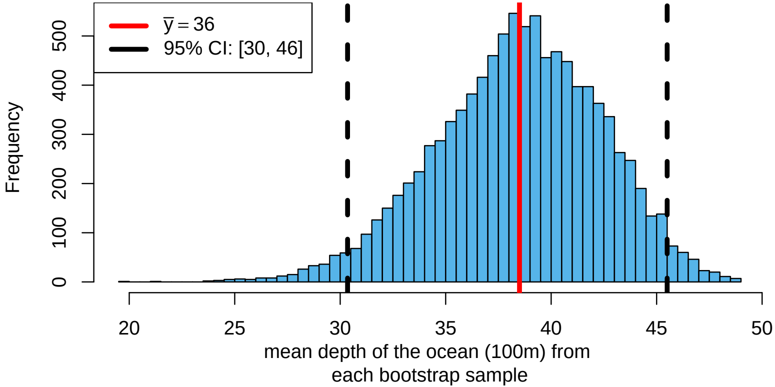


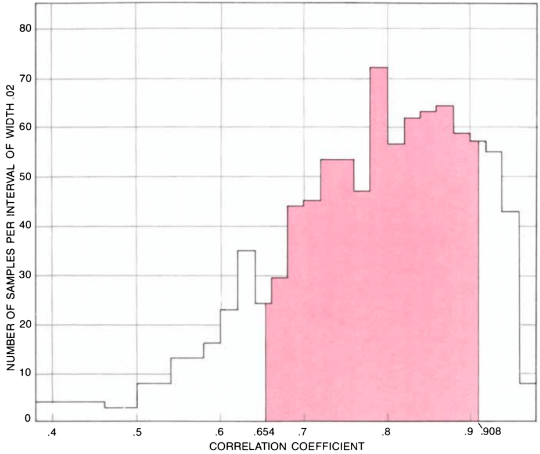
Figure: Bootstrap world. The bootstrap distribution is obtained by drawing repeated samples from an estimate of the population, computing the statistic of interest for each, and collecting those statistics. The distribution is centered at the observed statistic ($\bar{y}$), not the parameter (μ).

Main idea: simulate your own sampling distribution

```
R <- replicate(B, {
dplyr::sample_n(depths.n.20, size = N, replace = TRUE) %>%
dplyr::summarize(r = mean(alt)) %>%
dplyr::pull(r)
})
CI_95 <- quantile(R, probs = c(0.025, 0.975))
```



Bootstrap can be used for other statistics (e.g. R^2)



source: Bootstrap article in Scientific American

Exam Details

Data visualization

Tidy Data

Descriptive statistics

Sampling Distributions, CLT, Confidence Intervals and p-values

Bootstrap

One sample mean

One sample proportion

p-values

Power and sample size

σ known vs. unknown

σ	known	unknown
Data	$\{y_1, y_2, \ldots, y_n\}$	$\{y_1, y_2, \ldots, y_n\}$
Pop'n param	μ	μ
Estimator	$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$	$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$
SD	σ	$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$
SEM	$\sigma/\sqrt{n}$	$s/\sqrt{n}$
$(1-\alpha)100\%$ CI	$\bar{y} \pm z^{\star}_{1-\alpha/2}(\text{SEM})$	$\bar{y} \pm t^{\star}_{1-\alpha/2,(n-1)}(\text{SEM})$
test statistic	$\frac{\bar{y}-\mu_0}{\text{SEM}} \sim \mathcal{N}(0,1)$	$\frac{\bar{y}-\mu_0}{\text{SEM}} \sim t_{(n-1)}$

Assumptions

	z	t	Bootstrap
SRS	✓	✓	✓
Normal population	✓*	✓*	✗
needs CLT	✓*	✓*	✗
σ known	✓	✗	✗
Sampling dist. center at	μ	μ	$\bar{y}$
SD	σ	s	s
SEM	$\sigma/\sqrt{n}$	$s/\sqrt{n}$	SD(bootstrap statistics)

[a] *If population is Normal then CLT is not needed. If population is not Normal then CLT is needed.

Exam Details

Data visualization

Tidy Data

Descriptive statistics

Sampling Distributions, CLT, Confidence Intervals and p-values

Bootstrap

One sample mean

One sample proportion

p-values

Power and sample size

- Binomial calculations
- Nomogram, Clopper-Pearson CI
- Normal approximation

Exam Details

Data visualization

Tidy Data

Descriptive statistics

Sampling Distributions, CLT, Confidence Intervals and p-values

Bootstrap

One sample mean

One sample proportion

p-values

Power and sample size

p-values and statistical tests

Definition (*p*-value)

*A **probability concerning the observed data**, calculated under a **Null Hypothesis** assumption, i.e., assuming that the only factor operating is sampling or measurement variation.*

Use To assess the evidence provided by the sample data in relation to a pre-specified claim or 'hypothesis' concerning some parameter(s) or data-generating process.

Basis As with a confidence interval, it makes use of the concept of a *distribution*.

Caution A *p*-value is NOT the probability that the null 'hypothesis' is true

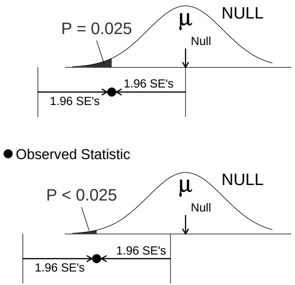
More about the *p*-value

- The *p*-value is a **probability concerning data**, **conditional on the Null Hypothesis being true**.
- **It is not the probability that Null Hypothesis is true**, *conditional on the data.*

$$\begin{aligned} p_{value} &= P(\text{this or more extreme data}|H_0) \\ &\neq P(H_0|\text{this or more extreme data}). \end{aligned}$$

- Statistical tests are often coded as statistically significant or not according to whether results are extreme or not with respect to a reference (null) distribution. But a test result is just one piece of data, and needs to be considered *along with rest of evidence* before coming to a ‘conclusion.’
- **Likewise with statistical ‘tests’: the *p*-value is just one more piece of *evidence*, hardly enough to ‘conclude’ anything**.

Close relationship between *p*-value and CI



- (Upper graph) If upper limit of 95% CI *just touches* null value, then the 2 sided *p*-value is 0.05 (or 1 sided *p*-value is 0.025).
- (Lower graph) If upper limit *excludes* null value, then the 2 sided *p*-value is less than 0.05 (or 1 sided *p*-value is less than 0.025).
- (Graph not shown) If CI *includes* null value, then the 2-sided *p*-value is greater than (the conventional) 0.05, and thus observed statistic is “not statistically significantly different” from hypothesized null value.

Exam Details

Data visualization

Tidy Data

Descriptive statistics

Sampling Distributions, CLT, Confidence Intervals and p-values

Bootstrap

One sample mean

One sample proportion

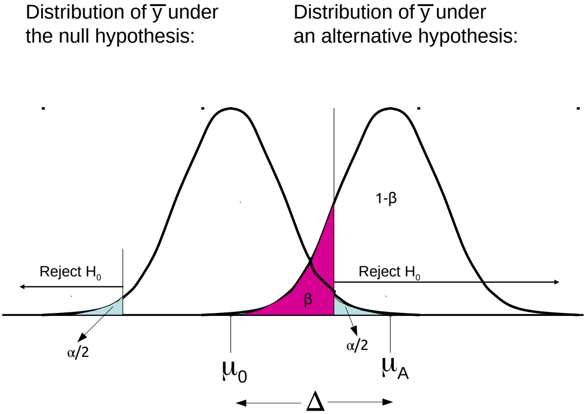
p-values

Power and sample size

Power = $1 - \beta$

Definition (Power = $1 - \beta$)

*The probability that a fixed level α significance test will reject H_0 when a particular alternative value of the parameter is true is called the **power** of the test to detect the alternative.*


Distribution of $\overline{y}$ under
the null hypothesis:
Distribution of $\overline{y}$ under
an alternative hypothesis:
1-β
Reject H_0
Reject H_0
β
α/2
α/2
μ_0
μ_A
Δ

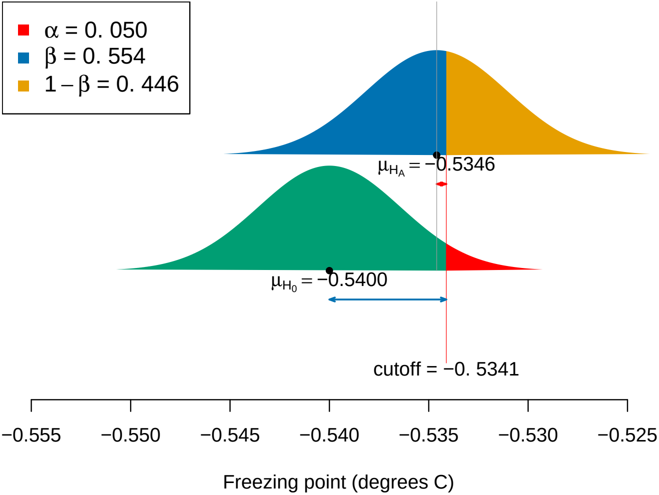
Power and Sample Size: 3 questions

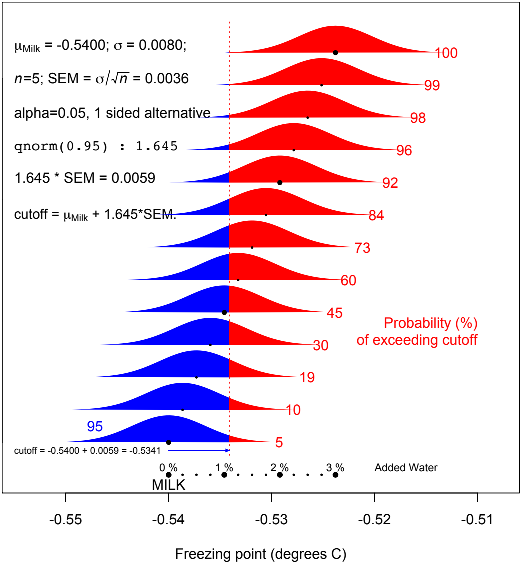
1. How much water a supplier could add to the milk before they have a 10% , 50%, 80% chance of getting caught, i.e., of the buyer detecting the cheating ?

2. Assume a 99:1 mix of milk and water. What are the chances of detecting cheating if the buyer uses samples n=10, 15 or 20 rather than just 5 measurements?

3. At what n does the chance of detecting cheating reach 80%? (*a commonly used, but arbitrary, criterion used in sample-size planning by investigators seeking funding for their proposed research*)

If the supplier added 1% water to the milk

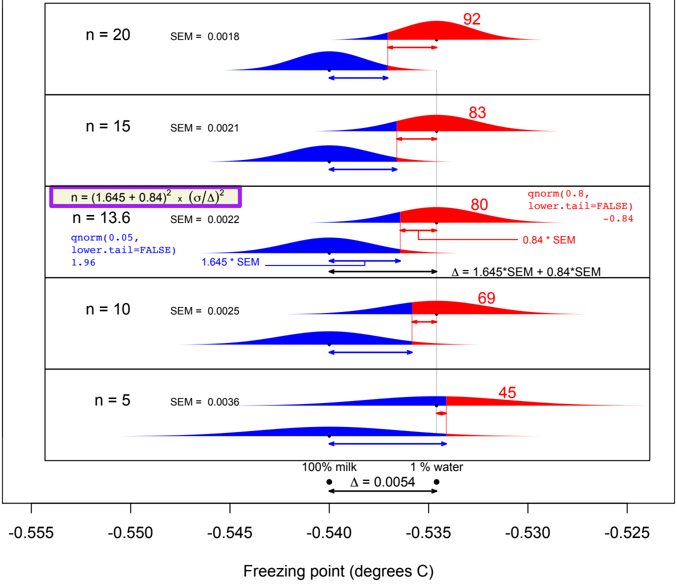




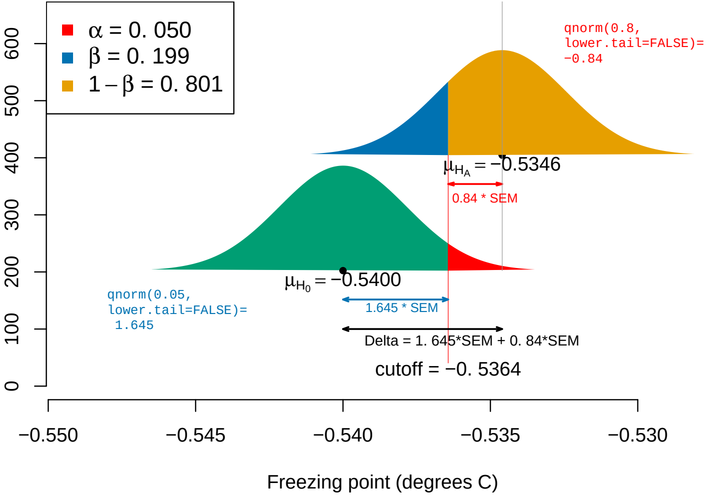
The probabilities in red were calculated using the formula: `stats::pnorm(cutoff, mean = mu.mixture, sd = SEM, lower.tail=FALSE)`

σ = 0.0080; SEM = $\sigma/\sqrt{n}$

cutoff = -0.54 + 1.645*SEM (alpha=0.05, 1 sided alternative)



The balancing formula



What sample size needed?

- The ‘balancing formula’, in SEM terms, is simply the *n* where

  $$1.645 \times \mathit{SEM} + 0.84 \times \mathit{SEM} = \Delta.$$

  Replacing each of the SEMs (assumed equal, because we assumed the variability is approx. the same under both scenarios) by $\sigma/\sqrt{n}$, i.e.,

  $$1.645 \times \sigma/\sqrt{n} + 0.84 \times \sigma/\sqrt{n} = \Delta.$$

  and solving for *n*, one gets

  $$n = (1.645 + 0.84)^2 \times \left\{ \frac{\sigma}{\Delta} \right\}^2 = (1.645 + 0.84)^2 \times \left\{ \frac{\mathit{Noise}}{\mathit{Signal}} \right\}^2.$$

What sample size needed? General Formula

- Two sided alternative:

$$\Delta = z_{1-\alpha/2} \times SEM + z_{1-\beta} \times SEM$$

- One sided alternative:

$$\Delta = z_{1-\alpha} \times SEM + z_{1-\beta} \times SEM$$

Session Info

```
R version 4.0.2 (2020-06-22)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Pop!_OS 20.04 LTS

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/liblapack.so.3

attached base packages:
[1] tools     stats     graphics  grDevices utils     datasets  methods  
[8] base     

other attached packages:
 [1] latex2exp_0.4.0    RColorBrewer_1.1-2 colorspace_1.4-1   oibiostat_0.2.0   
 [5] NCStats_0.4.7      FSA_0.8.30         forcats_0.5.0      stringr_1.4.0     
 [9] dplyr_1.0.2        purrr_0.3.4        readr_1.3.1        tidyr_1.1.2       
[13] tibble_3.0.3       ggplot2_3.3.2      tidyverse_1.3.0    knitr_1.29        

loaded via a namespace (and not attached):
  [1] minqa_1.2.4        TH.data_1.0-10     ellipsis_0.3.1    
  [4] rio_0.5.16         leaflet_2.0.3      sjlabelled_1.1.7  
  [7] snakecase_0.11.0   estimability_1.3   ggstance_0.3.4    
 [10] parameters_0.8.6   ggdendro_0.1.22    fs_1.5.0          
 [13] rstudioapi_0.11    farver_2.0.3       ggrepel_0.8.2     
 [16] fansi_0.4.1        mvtnorm_1.1-1      lubridate_1.7.9   
 [19] xml2_1.3.2         mosaic_1.7.0       codetools_0.2-16  
 [22] splines_4.0.2      sjmisc_2.8.5       polyclip_1.10-0   
 [25] jsonlite_1.7.1     nloptr_1.2.2.2     ggeffects_0.16.0  
 [28] broom_0.7.0        dbplyr_1.4.4       ggforce_0.3.2     
 [31] effectsize_0.3.3   compiler_4.0.2     httr_1.4.2        
 [34] sjstats_0.18.0     emmeans_1.5.1      backports_1.1.9   
 [37] assertthat_0.2.1   Matrix_1.2-18      cli_2.0.2         
 [40] tweenr_1.0.1       htmltools_0.5.0    coda_0.19-4       
 [43] gtable_0.3.0       glue_1.4.2         Rcpp_1.0.5        
 [46] carData_3.0-4      cellranger_1.1.0   vctrs_0.3.4       
 [49] sjPlot_2.8.5       nlme_3.1-149       crosstalk_1.1.0.1 
 [52] insight_0.9.6      xfun_0.17          lme4_1.1-23       
 [55] openxlsx_4.1.5     rvest_0.3.6        lifecycle_0.2.0   
 [58] mosaicCore_0.8.0   pacman_0.5.1       statmod_1.4.34    
 [61] MASS_7.3-53        zoo_1.8-8          scales_1.1.1      
 [64] hms_0.5.3          sandwich_2.5-1     curl_4.3          
```