018 - Midterm Review

EPIB 607

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

sahir.bhatnagar@mcgill.ca

slides compiled on October 19, 2021

Exam Details

Data visualization

Tidy Data

Descriptive statistics

Grammar of graphics

Parameters, probability, random variables

Sampling Distributions, CLT, Confidence Intervals and p-values

Bootstrap

One sample mean

p-values

Exam Details

- **When:** Thursday October 21. Exam will be released at 9am and will be open for 24 hours. A 5% per hour late penalty will be applied. No submissions will be accepted after the 24 hour window has passed.
- This is a 2 hour, timed, open book exam. You are given a total of 5 hours from the moment you start the exam to complete it. This extra time is to account for uploading your solutions to Crowdmark.
- Any material on myCourses (EPIB607/613), the course website, 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.
- You must upload your signed academic integrity statement to Crowdmark.

Exam Details

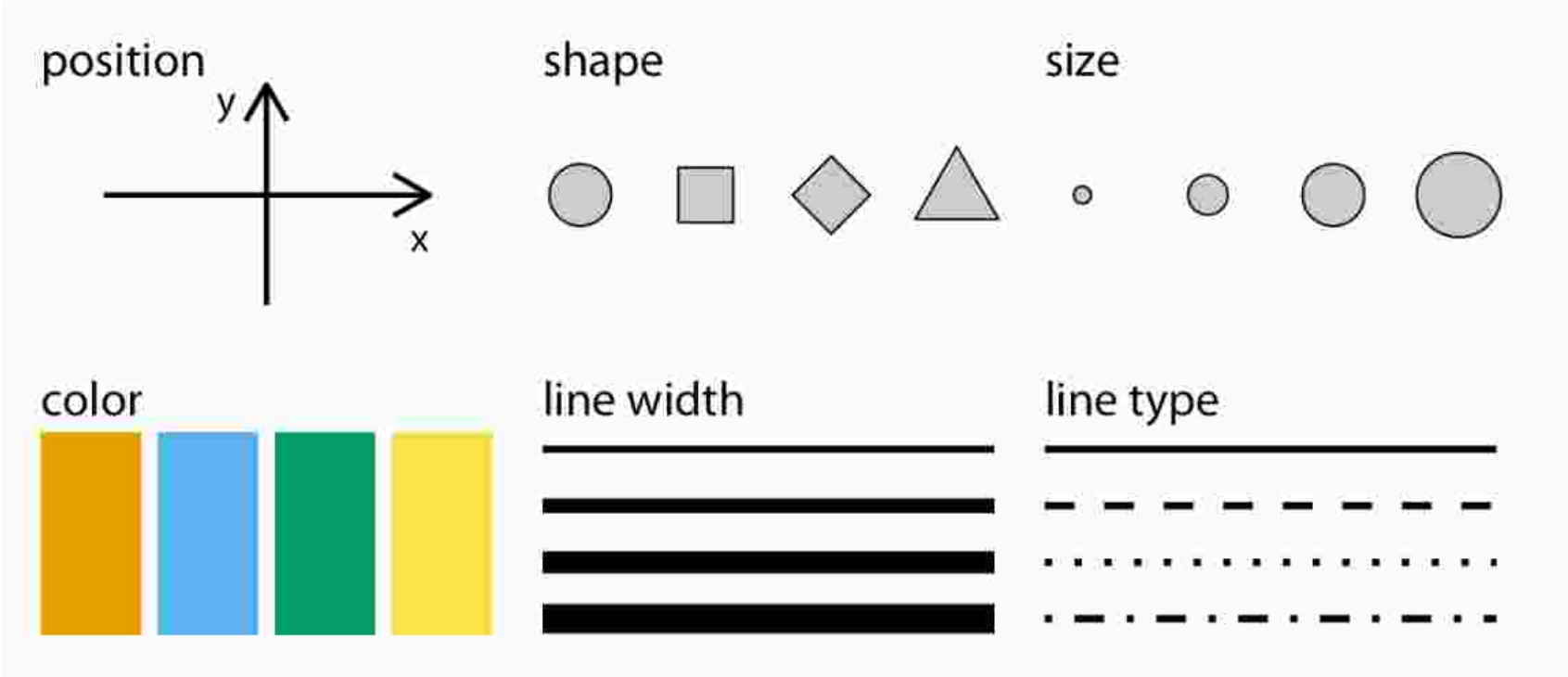
- Provide units and state your assumptions when applicable. Label axes and write answers in complete sentences when appropriate.
- The format of the exam will follow the assignments. That is, you will be required to complete a series of questions in an RMarkdown document and knit to pdf. Your solutions for each question must then be uploaded to Crowdmark. A template will be provided.
- .Rmd files will not be accepted. Any files emailed directly to the instructor will not be accepted. Your solutions must be uploaded to Crowdmark as a pdf.

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. Grammar of graphics
4. Statistical parameters, probability, random variables
5. Normal Curve Calculations
6. Sampling Distributions, CLT, Bootstrap
7. Confidence intervals and p-values
8. Hypothesis Testing

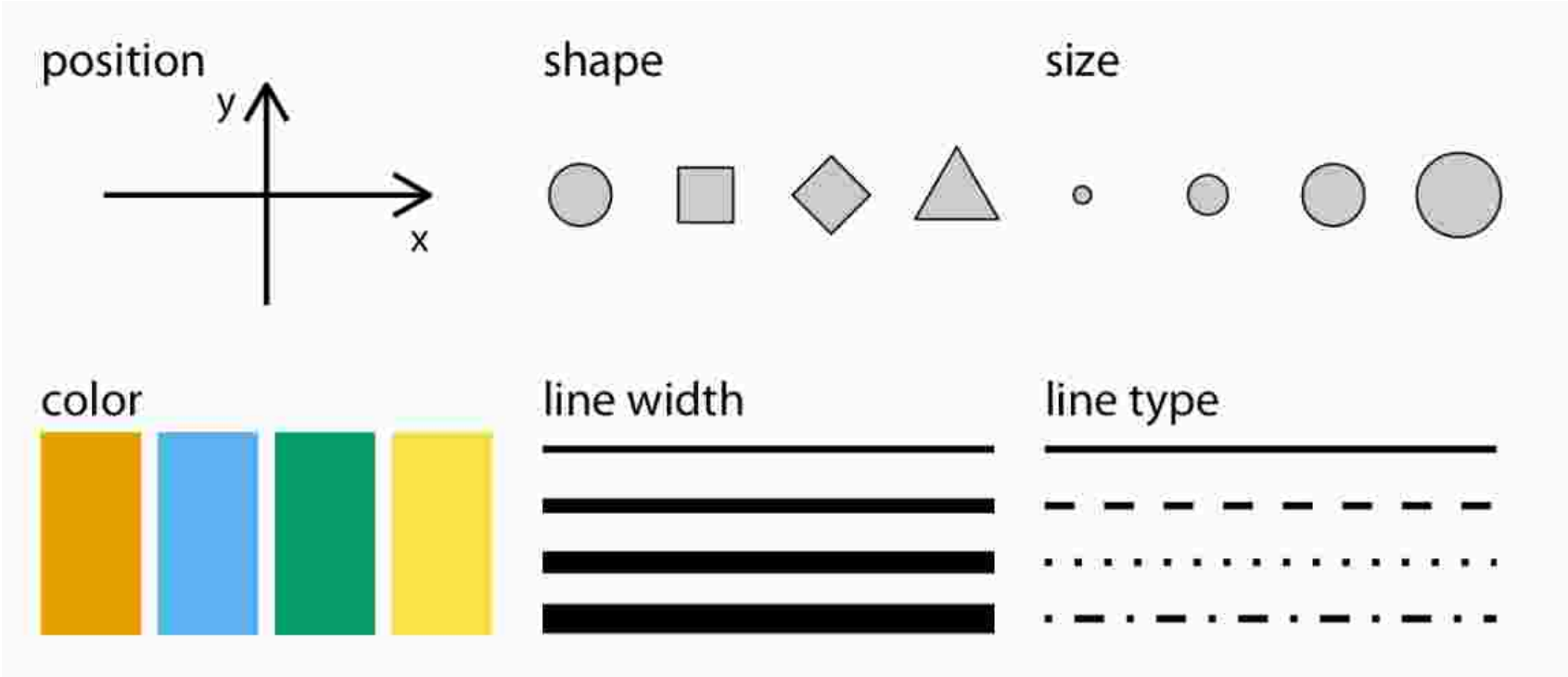
Aesthetics

- Aesthetics



Aesthetics

- Aesthetics

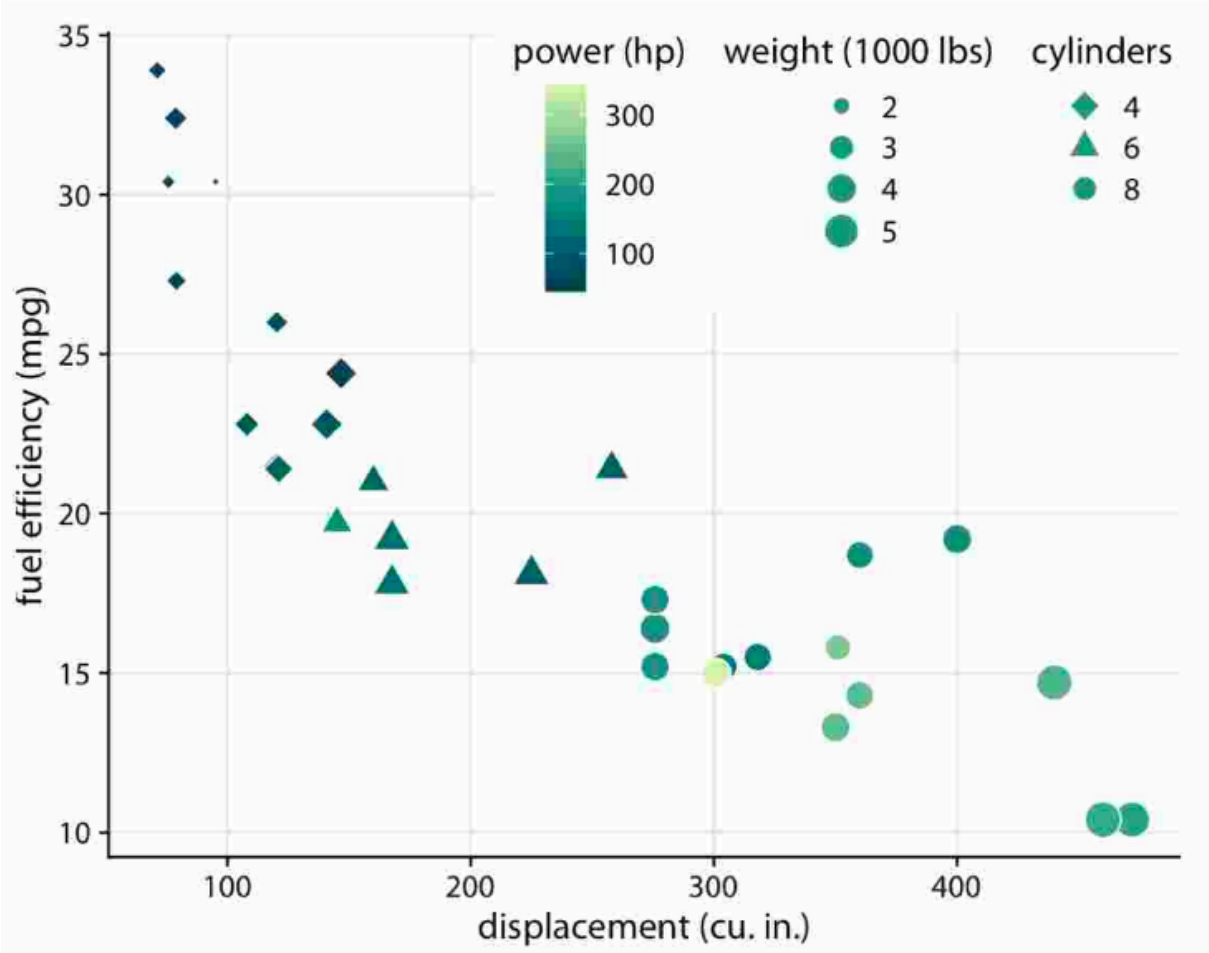


- Commonly used aesthetics in data visualization: position, shape, size, color, line width, line type. Some of these aesthetics can represent both continuous and discrete data (position, size, line width, color) while others can only represent discrete data (shape, line type)

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/

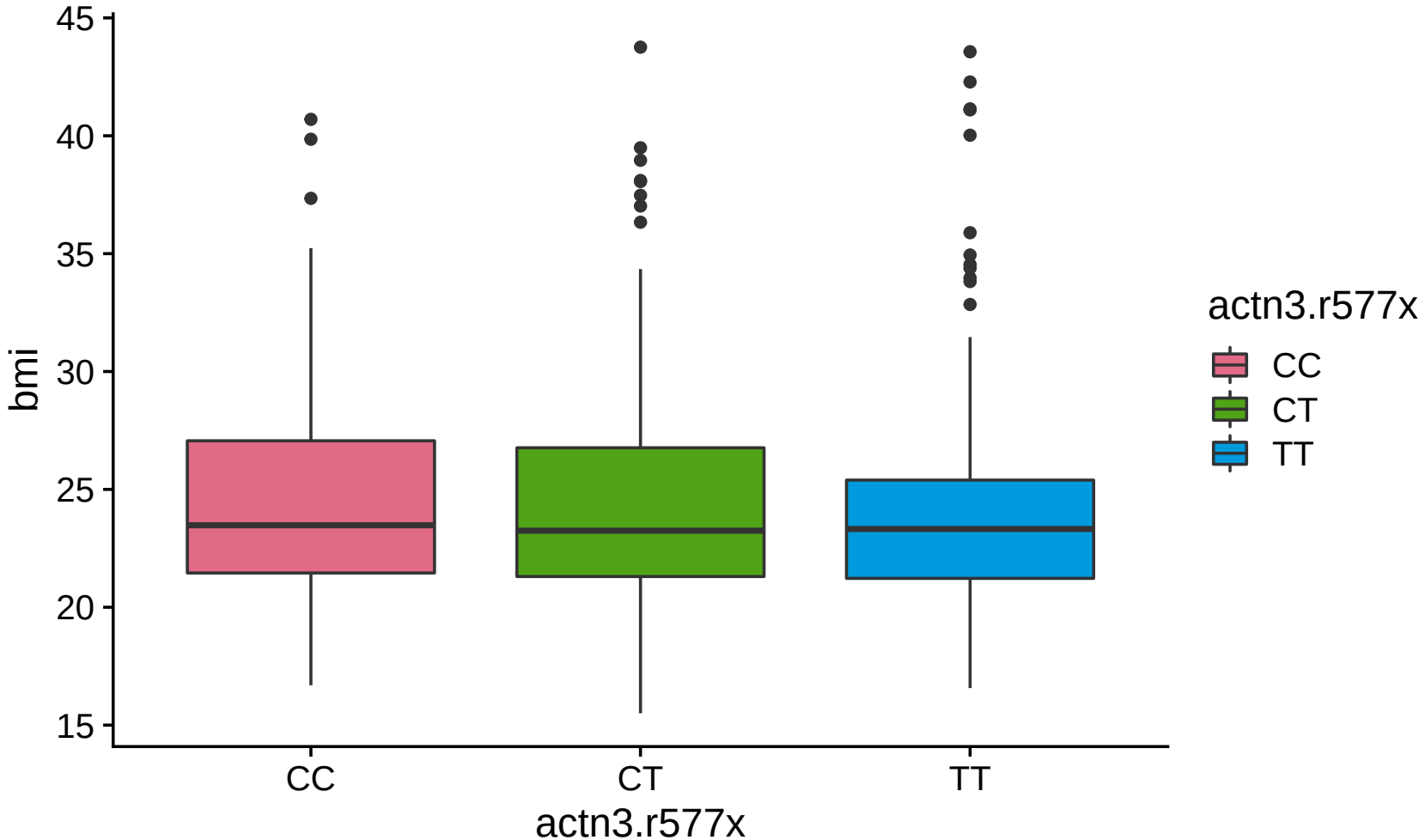
How many scales are being used?



Boxplots with qualitative palette

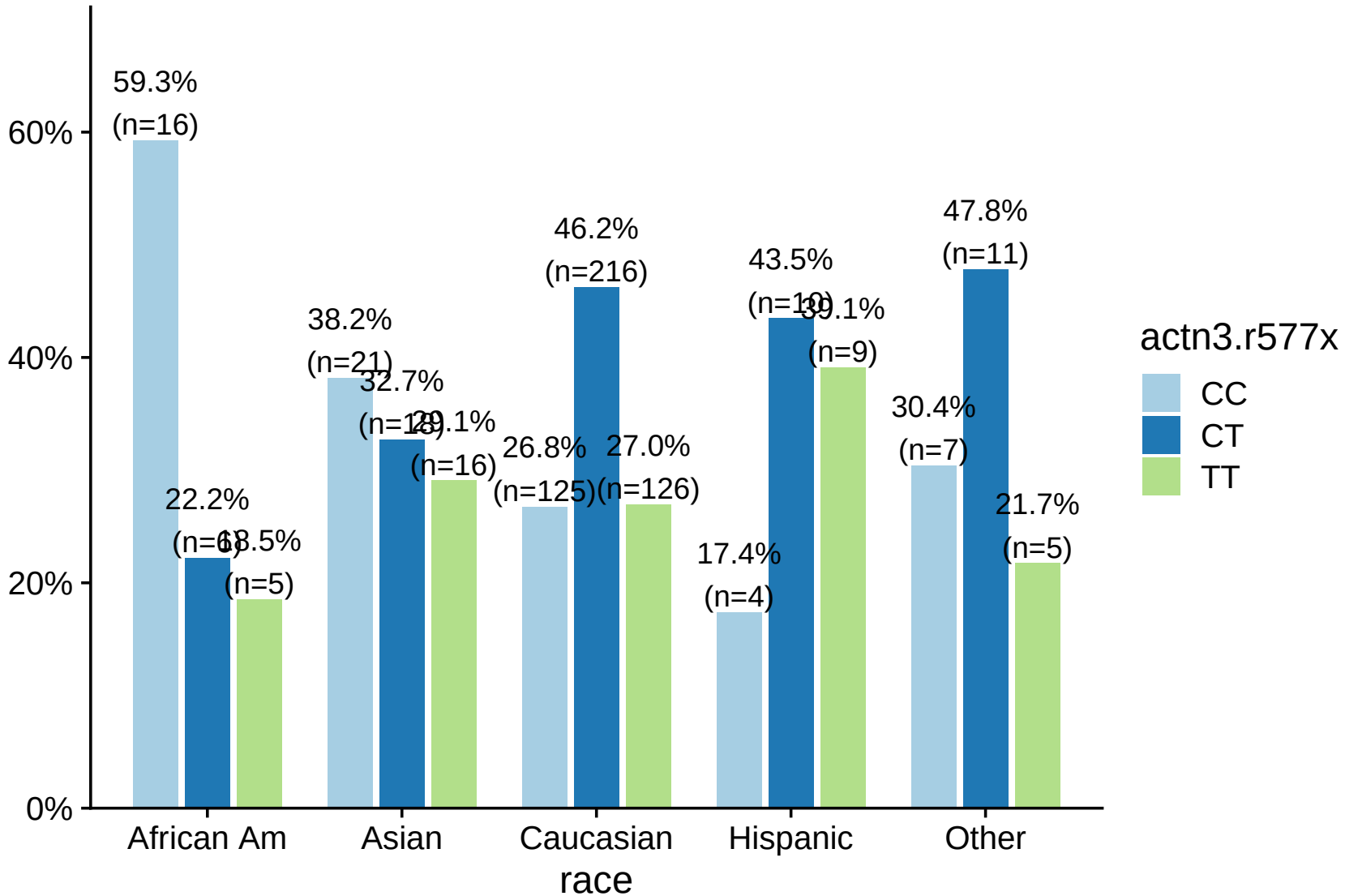
```
library(oibiostat); data("famuss")
library(ggplot2)
library(colorspace)

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



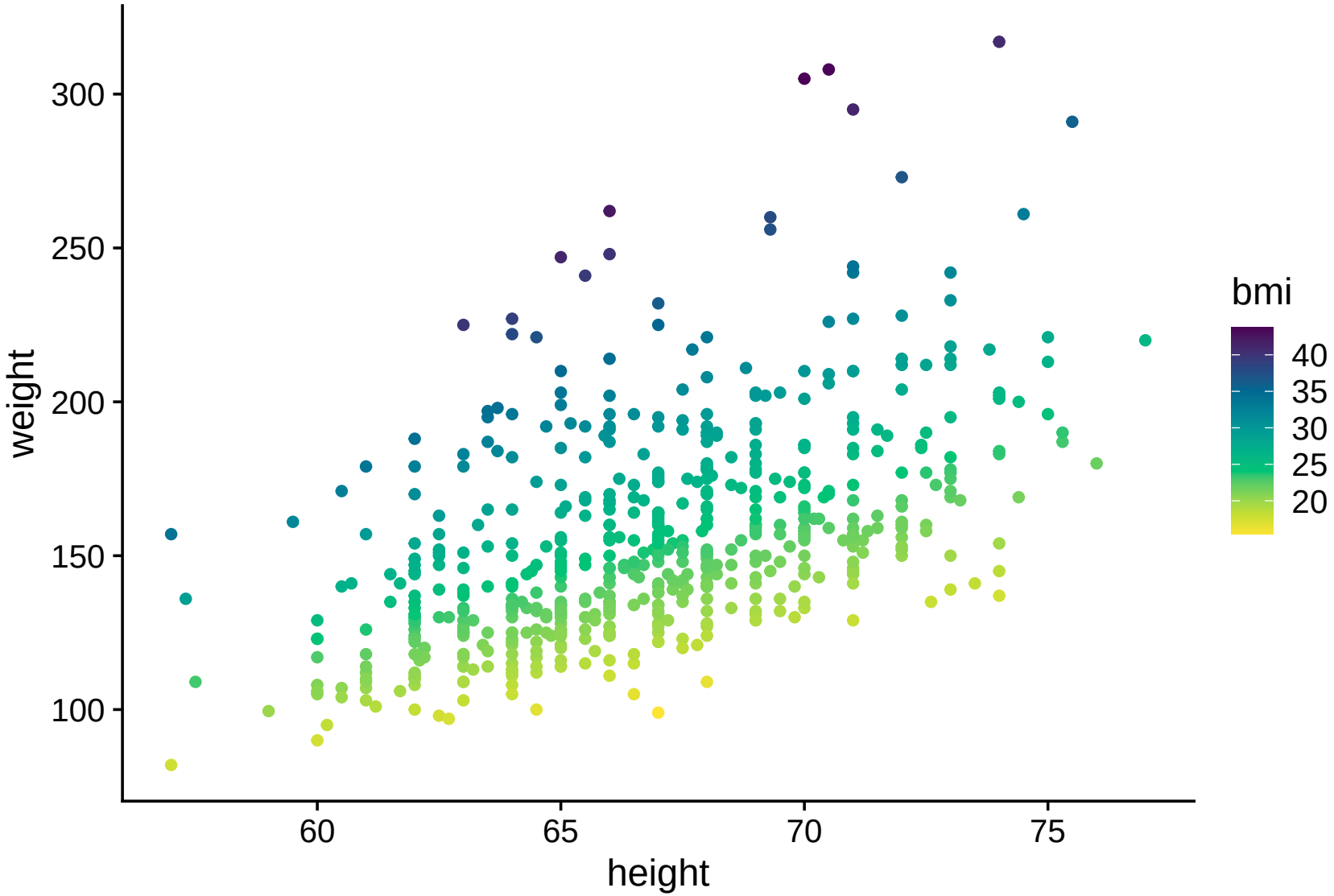
Conditional distribution of genotype *given* race

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



Scatter plots with sequential palette

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

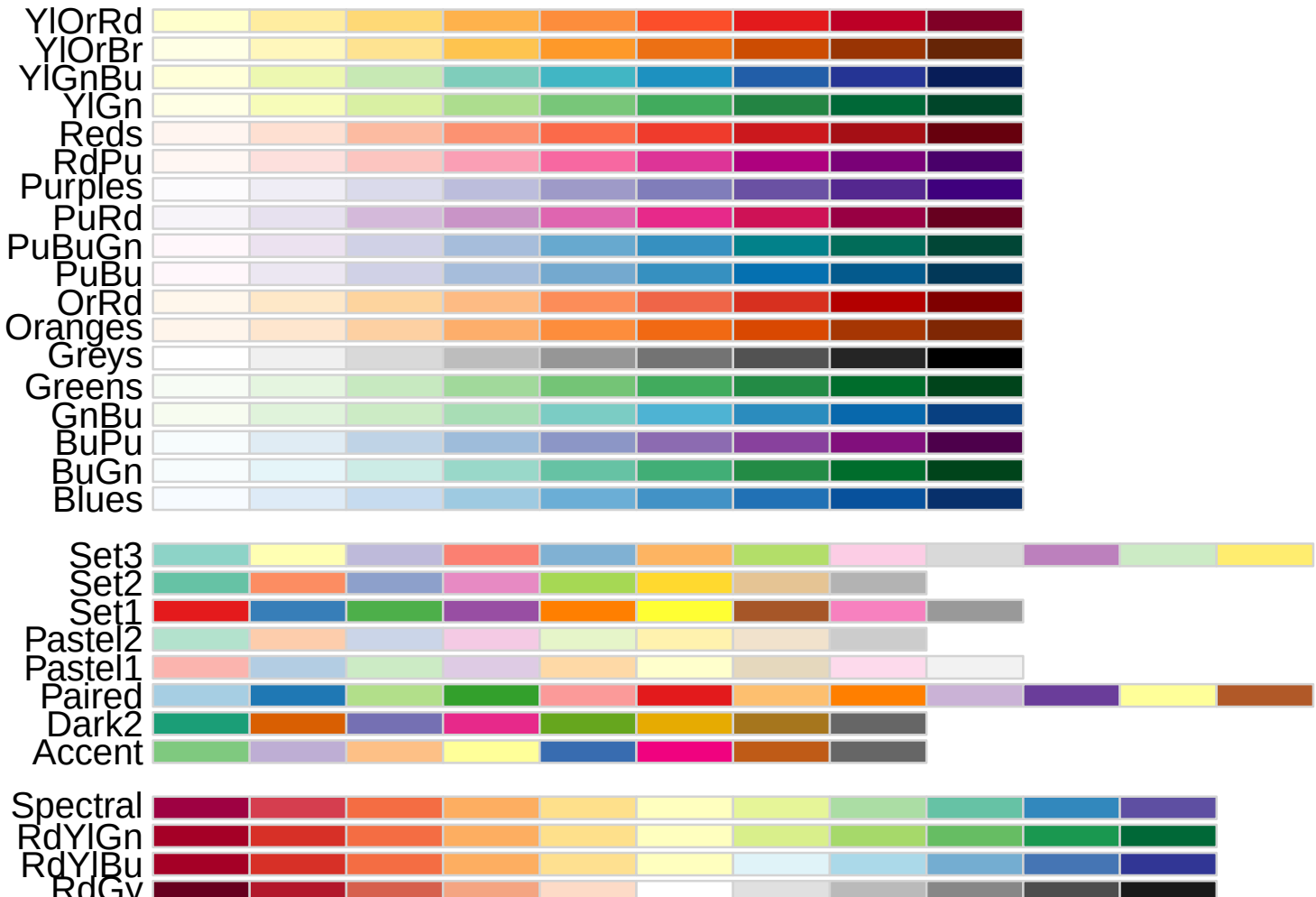


Variable Types

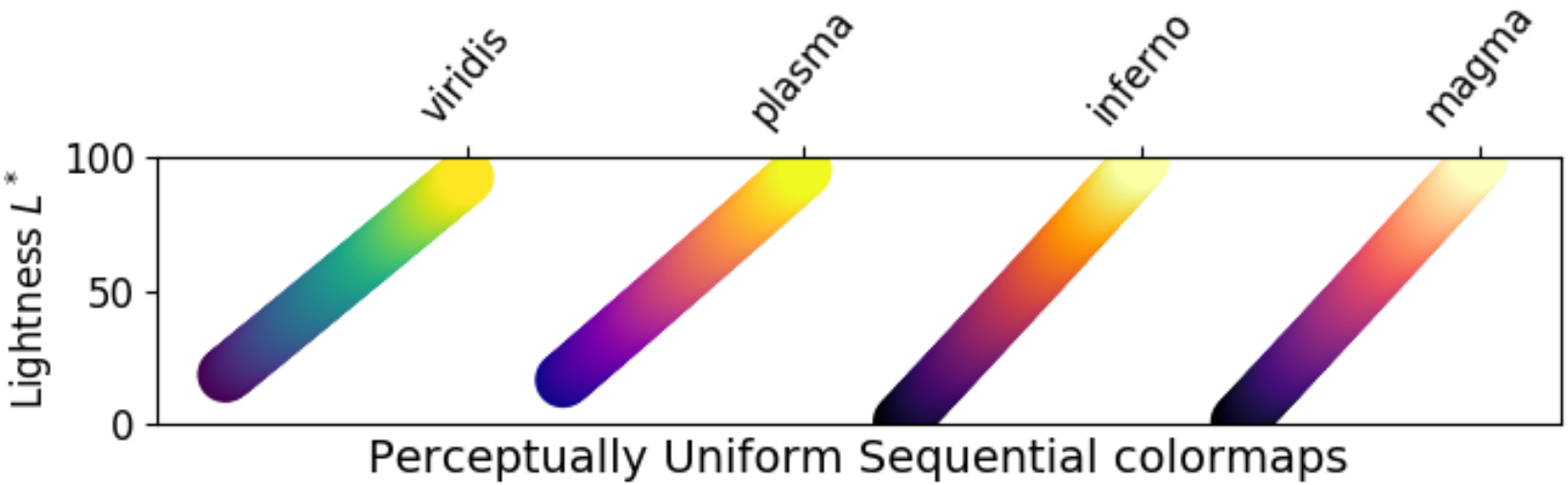
- quantitative/numerical continuous (1.3, 5.7, 83, 1.5×10^{-2})
- quantitative/numerical discrete (1,2,3,4)
- qualitative/categorical unordered (dog, cat, fish)
- qualitative/categorical ordered (good, fair, poor)

Color Palettes: Cynthia Brewer

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



Color Palettes: viridis



Exam Details

Data visualization

Tidy Data

Descriptive statistics

Grammar of graphics

Parameters, probability, random variables

Sampling Distributions, CLT, Confidence Intervals and p-values

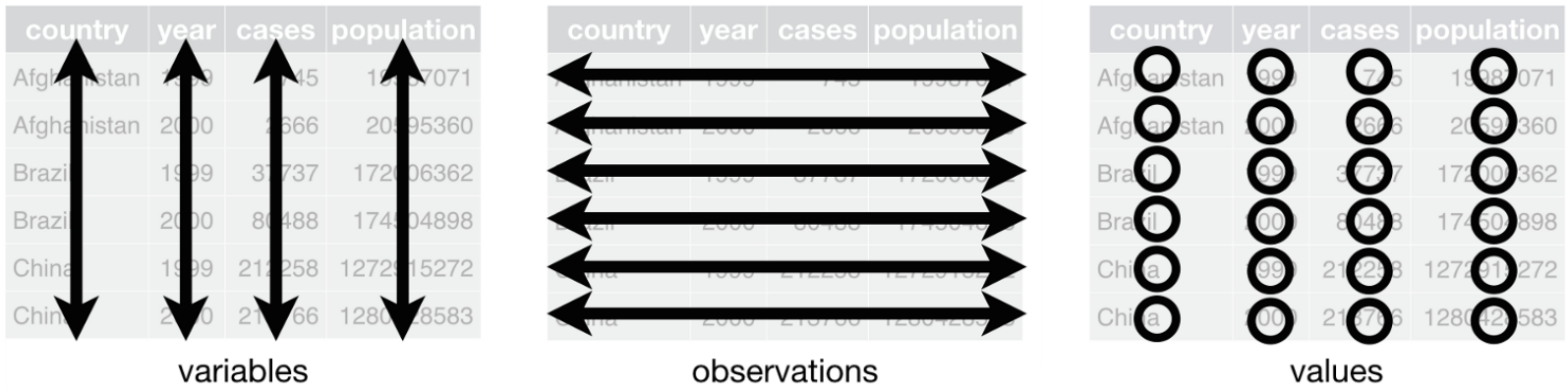
Bootstrap

One sample mean

p-values

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: Does a full moon affect behaviour?

- Many people believe that the moon influences the actions of some individuals.
- A study of dementia patients in nursing homes recorded various types of disruptive behaviors every day for 12 weeks.
- Days were classified as moon days if they were in a 3-day period centered at the day of the full moon.
- For each patient, the average number of disruptive behaviors was computed for moon days and for all otherdays.

patient	moon_days	other_days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32
4	3.33	0.19
5	3.33	1.26
6	3.67	0.11
7	4.67	0.30

Not tidy vs. tidy data

patient	moon_days	other_days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32
4	3.33	0.19
5	3.33	1.26

patient	day_type	mean_behavior
1	moon_days	3.33
1	other_days	0.27
2	moon_days	3.67
2	other_days	0.59
3	moon_days	2.67
3	other_days	0.32
4	moon_days	3.33
4	other_days	0.19
5	moon_days	3.33
5	other_days	1.26

Not tidy vs. tidy data

patient	moon_days	other_days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32
4	3.33	0.19
5	3.33	1.26

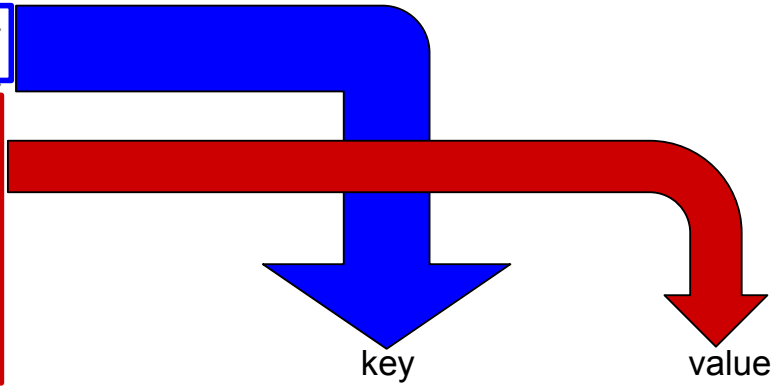
Not tidy

tidy

patient	day_type	mean_behavior
1	moon_days	3.33
1	other_days	0.27
2	moon_days	3.67
2	other_days	0.59
3	moon_days	2.67
3	other_days	0.32
4	moon_days	3.33
4	other_days	0.19
5	moon_days	3.33
5	other_days	1.26

tidyr::pivot_longer()

patient	moon_days	other_days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32
4	3.33	0.19
5	3.33	1.26

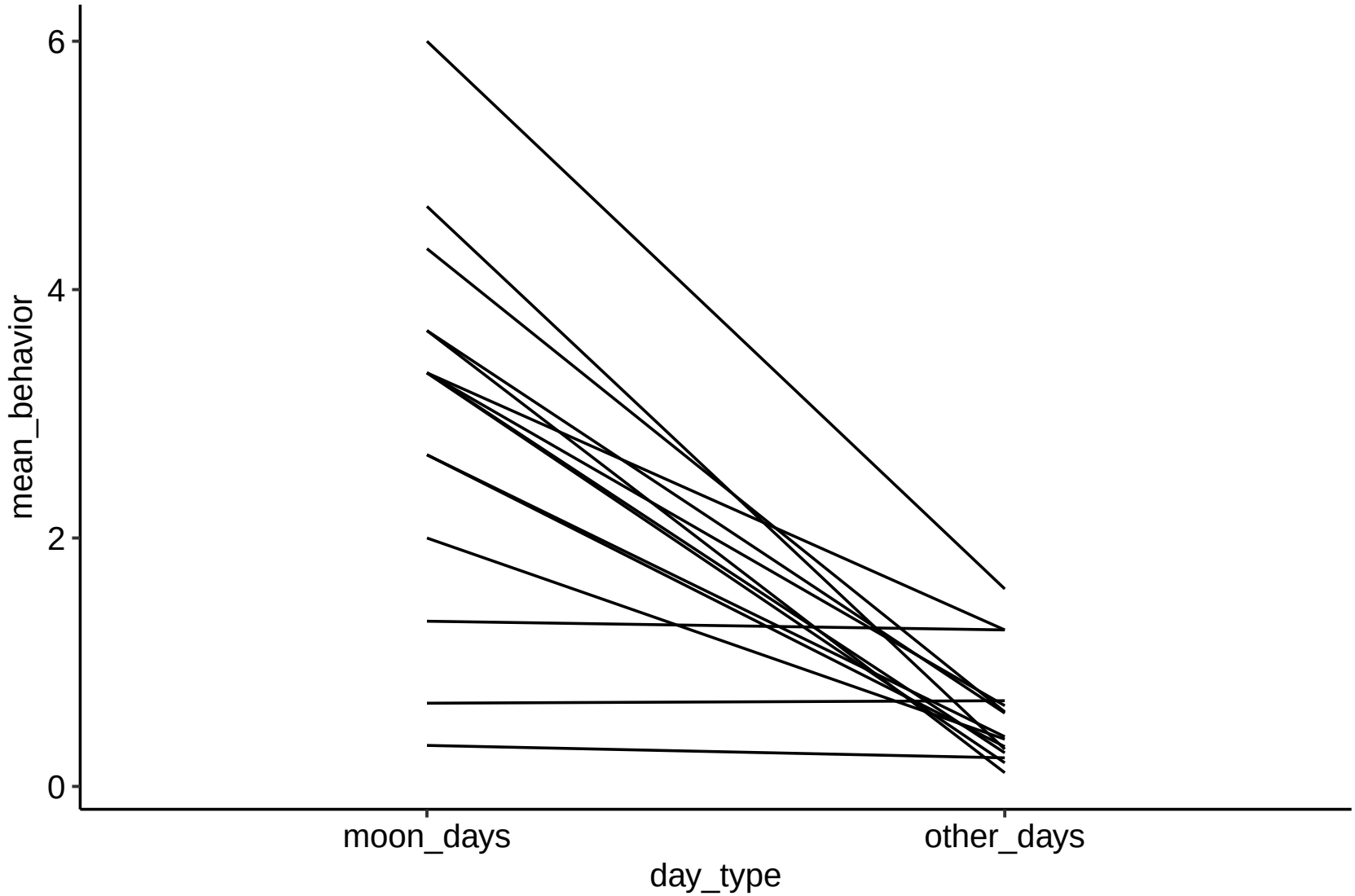


patient	day_type	mean_behavior
1	moon_days	3.33
1	other_days	0.27
2	moon_days	3.67
2	other_days	0.59
3	moon_days	2.67
3	other_days	0.32
4	moon_days	3.33
4	other_days	0.19
5	moon_days	3.33
5	other_days	1.26

```
tidyr::pivot_longer(data = df, cols = -patient, names_to = "day_type", values_to = "mean_behavior")
```

Plotting with tidy data

```
ggplot(data = df_tidy, mapping = aes(x = day_type, y = mean_behavior, group = patient)) + geom_line()
```



Regression with tidy data

```
fit <- lme4::lmer(mean_behavior ~ day_type + (1|patient), data = df_tidy)
summary(fit)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: mean_behavior ~ day_type + (1 | patient)
##    Data: df_tidy
##
## REML criterion at convergence: 90.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.2728 -0.3014 -0.0408  0.4860  2.4482
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  patient  (Intercept) 0.1559   0.3948
##  Residual             1.0663   1.0326
## Number of obs: 30, groups:  patient, 15
##
## Fixed effects:
##                   Estimate Std. Error t value
## (Intercept)          3.0220     0.2854  10.587
## day_typeother_days  -2.4327     0.3771  -6.452
##
## Correlation of Fixed Effects:
##             (Intr)
## dy_typthr_d -0.660
```

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

Exam Details

Data visualization

Tidy Data

Descriptive statistics

Grammar of graphics

Parameters, probability, random variables

Sampling Distributions, CLT, Confidence Intervals and p-values

Bootstrap

One sample mean

p-values

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

```
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 1 (SE vs. SD).

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

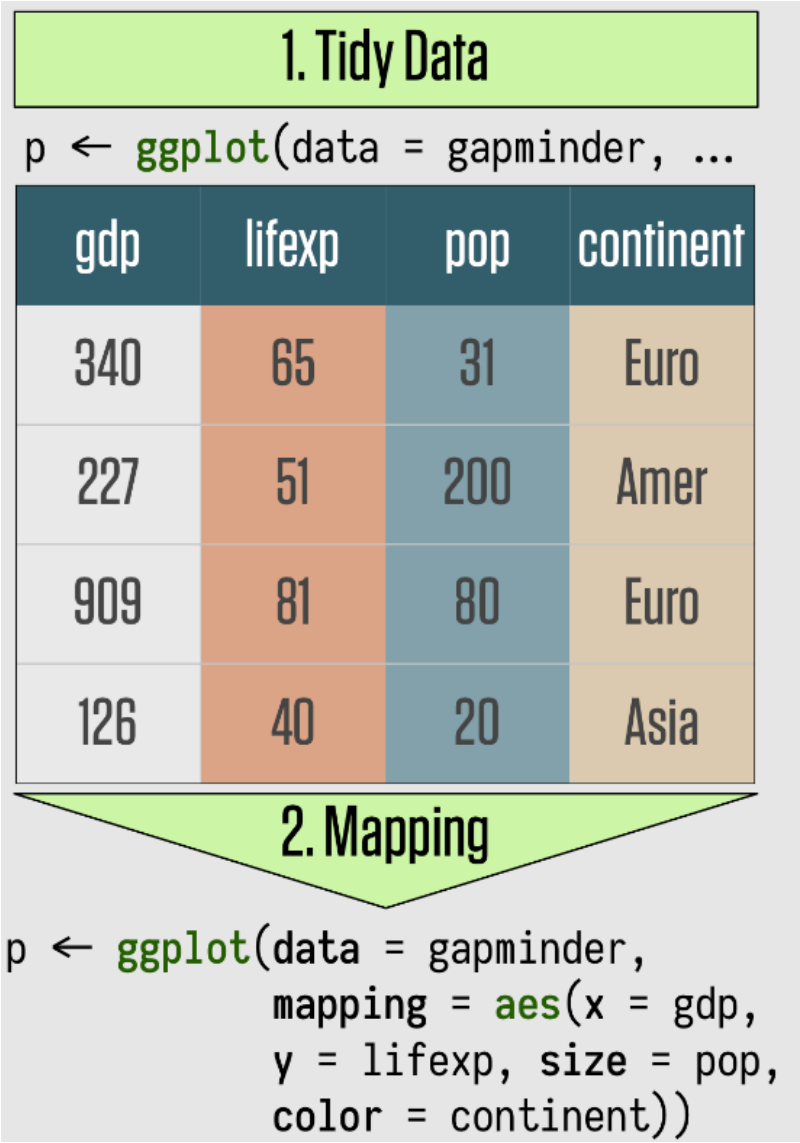
SE($\bar{y}$) describes how far $\bar{y}$ could (typically) deviate from μ;

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

ggplot2 to make plots

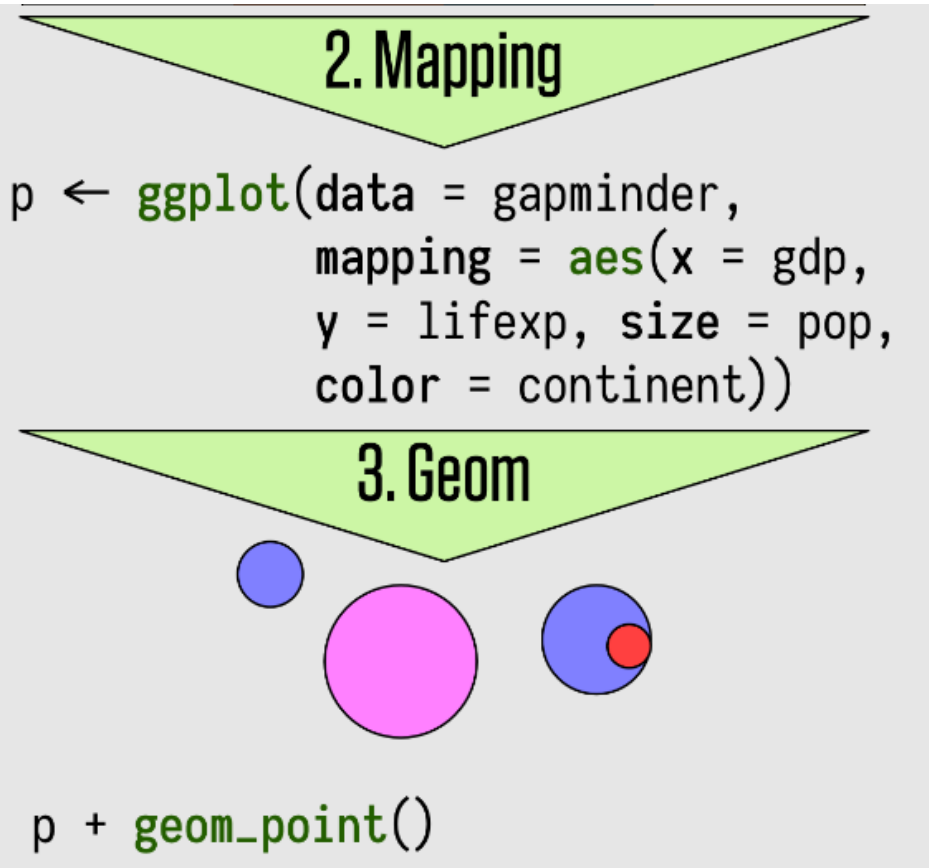
- ggplot provides you with a set of tools to map data
  1. to visual elements on your plot
  2. to specify the kind of plot you want, and
  3. then subsquently to control the fine details of how it will be displayed.

Aesthetic mappings



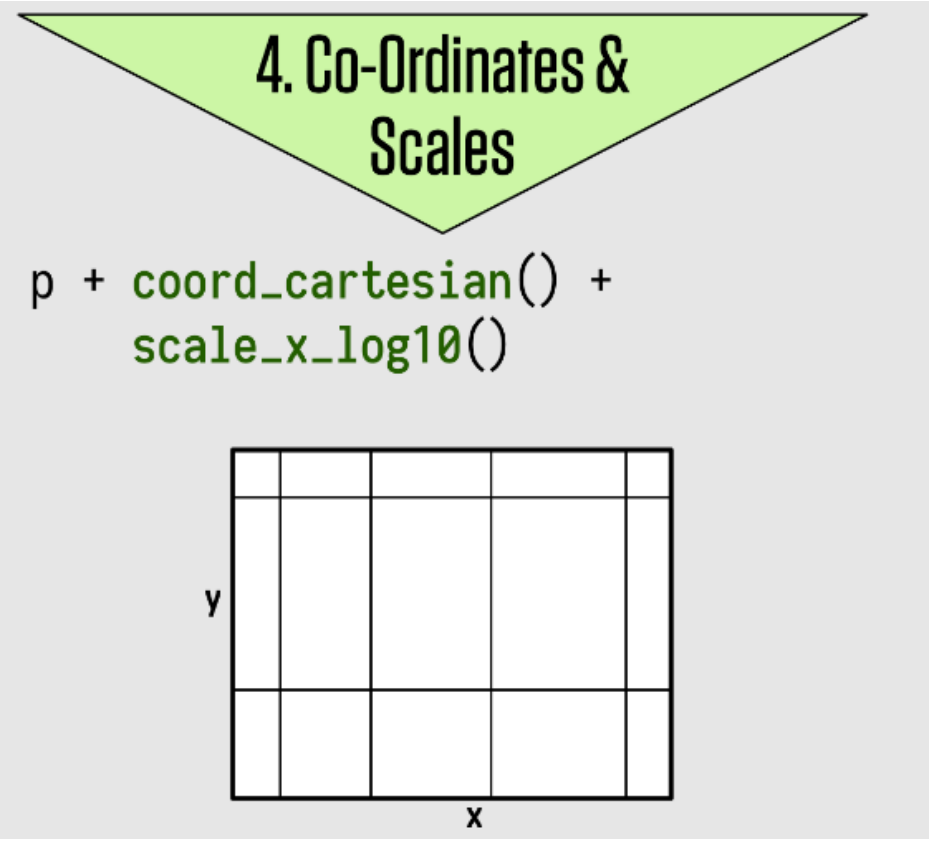
- The code you write specifies the connections between the variables in your data, and the colors, points, and shapes you see on the screen.
- In ggplot, these logical connections between your data and the plot elements are called *aesthetic mappings* or just *aesthetics*.
- You begin every plot by telling the `ggplot()` function what your data is, and then how the variables in this data logically map onto the plot's aesthetics.

Geometry



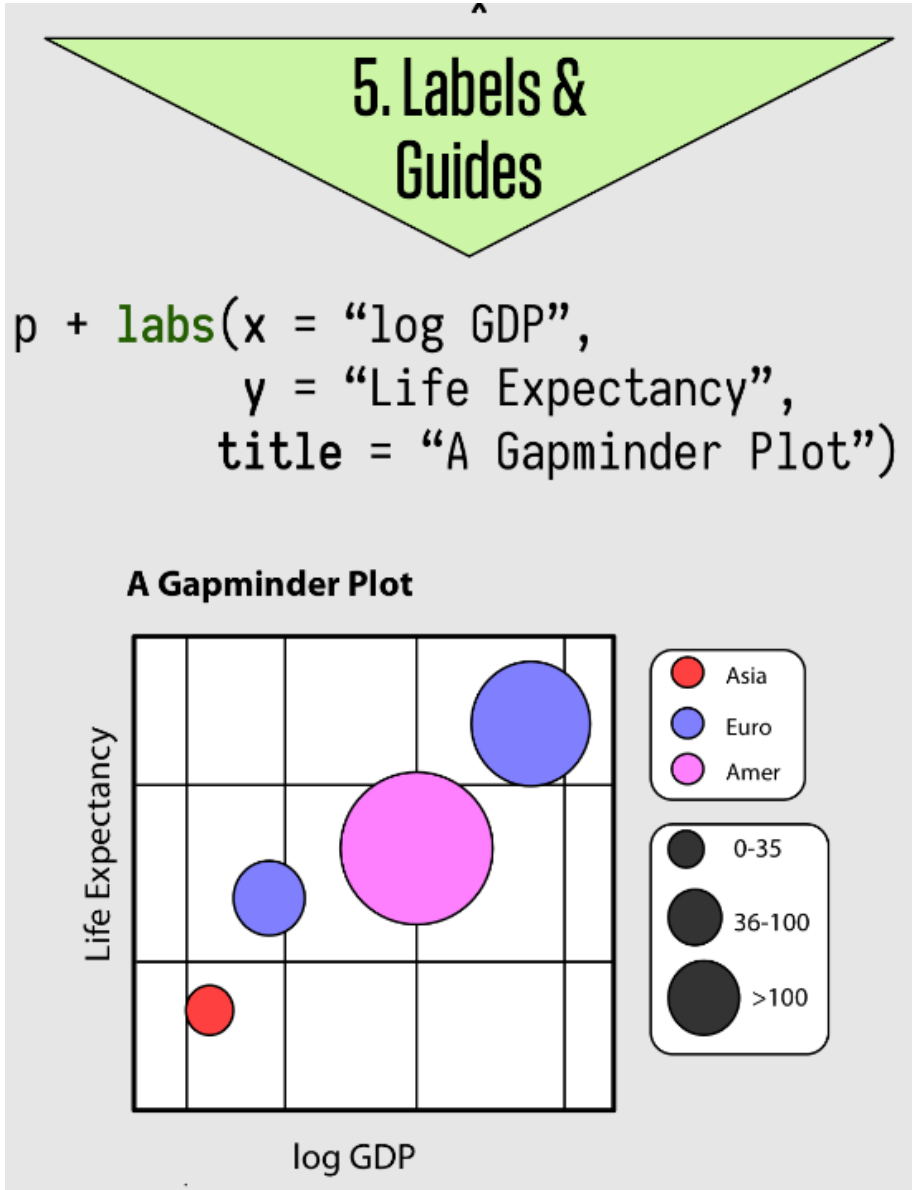
- Then you take the result and say what general sort of plot you want, such as a scatterplot, a boxplot, or a bar chart. In `ggplot`, the overall type of plot is called a `geom`.
- Each geom has a function that creates it. For example, `geom_point()` makes scatterplots, `geom_bar()` makes barplots, `geom_boxplot()` makes boxplots, and so on.
- You combine these two pieces, the `ggplot()` object and the `geom`, by literally adding them together in an expression, using the "+" symbol.

Customization



- At this point, ggplot will have enough information to be able to draw a plot for you. ggplot will use a set of defaults that try to be sensible about what gets drawn.
- But more often, you will want to specify exactly what you want, including information about the scales, the labels of legends and axes, and other guides that help people to read the plot.
- Each component has it own function, you provide arguments to it specifying what to do, and you literally add it to the sequence of instructions.
- In this way you systematically build your plot piece by piece.

Customization



Exam Details

Data visualization

Tidy Data

Descriptive statistics

Grammar of graphics

Parameters, probability, random variables

Sampling Distributions, CLT, Confidence Intervals and p-values

Bootstrap

One sample mean

p-values

Probability function

- The probability function of a random variable is defined for any value that the random variable may obtain and produces the **distribution** of the random variable. The probability function may emerge as a relative frequency as in the given example or it may be a result of theoretical modeling.
- Consider the following probability distribution:

Value	Probability	Cum.Prob
0	0.50	0.50
1	0.25	0.75
2	0.15	0.90
3	0.10	1.00

- What is $\mathrm{P}(Y = 0)$, the probability that Y is equal to 0?:
- What is the probability of Y falling in the interval $[0.5, 2.3]$?

Expectation

- The average of the data can be computed as the weighted average of the values that are present in the data, with weights given by the relative frequency. Specifically, for the data

$$1,\ 1,\ 1,\ 2,\ 2,\ 3,\ 4,\ 4,\ 4,\ 4,\ 4,$$

the mean can be calculated via

$$\begin{aligned}\bar{y} &= \frac{1+1+1+2+2+3+4+4+4+4+4}{11} \\ &= 1 \times \frac{3}{11} + 2 \times \frac{2}{11} + 3 \times \frac{1}{11} + 4 \times \frac{5}{11}\end{aligned}$$

producing the value of $\bar{y} = 2.727$ in both representations.

- Using a formula, the equality between the two ways of computing the mean is given in terms of the equation:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \sum_{y} \left(y \times (f_y/n) \right),$$

where f_y represents the frequency of y in the data.

Expectation

- Using a formula, the equality between the two ways of computing the mean is given in terms of the equation:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \sum_{y} \left(y \times (f_y / n) \right) ,$$

  where f_y represents the frequency of y in the data.

- The expectation of a random variable is computed in the spirit of the second formulation, and is define via the equation:

$$\mathrm{E}(Y) = \sum_{y} \left(y \times \mathrm{P}(y) \right) .$$

Variance

- The sample variance (s^2) is obtained as the sum of the squared deviations from the average, divided by the sample size (n) minus 1:

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1} \ .$$

- A second formulation for the computation of the same quantity is via the use of relative frequencies. The formula for the sample variance takes the form

$$s^2 = \frac{n}{n-1} \sum_{y} \left((y - \bar{y})^2 \times (f_y / n) \right) \ .$$

- In a similar way, the variance of a random variable may be defined via the deviation from the expectation. This deviation is then squared and multiplied by the probability of the value. The multiplications are summed up in order to produce the variance:

$$\mathrm{Var}(Y) = \sum_{y} \left((y - \mathrm{E}(Y))^2 \times \mathrm{P}(y) \right) \ .$$

Expected value for a discrete RV

Definition 1.

Let Y be a discrete random variable with set of possible values $D = \{y_1, y_2, \ldots, y_k\}$ and corresponding probabilities for each value, e.g., y_1 with probability $\mathrm{P}(y_1)$, y_2 with probability $\mathrm{P}(y_2)$, y_3 with probability $\mathrm{P}(y_3)$, ..., y_k with probability $\mathrm{P}(y_k)$. Furthermore, let $g(Y)$ be some real-valued function of Y. Then the expected value of $g(Y)$ is:

$$\mathrm{E}(g(Y)) = \sum_{y \in D} g(y) \times \mathrm{P}(y) \ .$$

i.e. it is a weighted mean of the $g(y)$'s, with $\mathrm{P}(y)$'s as weights.

Expected value for a discrete RV

Definition 1.

Let Y be a discrete random variable with set of possible values $D = \{y_1, y_2, \ldots, y_k\}$ and corresponding probabilities for each value, e.g., y_1 with probability $\mathrm{P}(y_1)$, y_2 with probability $\mathrm{P}(y_2)$, y_3 with probability $\mathrm{P}(y_3)$, ..., y_k with probability $\mathrm{P}(y_k)$. Furthermore, let $g(Y)$ be some real-valued function of Y. Then the expected value of $g(Y)$ is:

$$\mathrm{E}(g(Y)) = \sum_{y \in D} g(y) \times \mathrm{P}(y) \ .$$

i.e. it is a weighted mean of the $g(y)$'s, with $\mathrm{P}(y)$'s as weights.

Let c be a constant and Z another random variable

- $g(Y) = Y + c \rightarrow$
- $g(Y) = cY \rightarrow$
- $g(Y, Z) = Y + Z \rightarrow$

Exercise: $\mathrm{E}(g(Y)) = g(E(Y))$?

- Y = Noon Temperature (C) in Montreal on a randomly selected day of the year; $g(Y)$ = Temperature (F) = 32 + (9/5) Y
- Y_1 and Y_2 are two random variables that might or might not be related; $g(Y_1, Y_2) = Y_1 + Y_2$
- $g(Y_1, Y_2) = \frac{Y_1 + Y_2}{2}$
- $g(Y_i) = \frac{1}{n} \sum_{i=1}^{n} Y_i$

Example: $g(Y)$ = Volume of sphere = $\frac{\pi}{6} Y^3$

Example: Checking via simulation

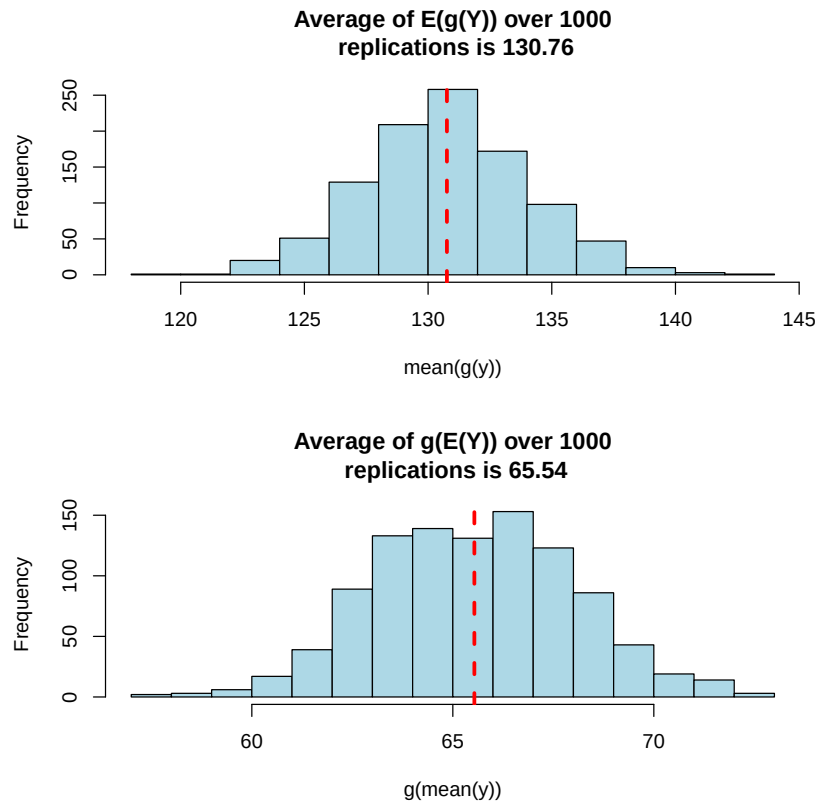
```r
g.y <- function(y) {
(pi / 6) * y^3
}

set.seed(12)
B <- 1000; N <- 2000
E_g.y <- replicate(B, {
diameter <- runif(N, min = 0, max = 10)
mean(g.y(diameter)) # E(g(y))
})

g_E.y <- replicate(B, {
diameter <- runif(N, min = 0, max = 10)
g.y(mean(diameter)) # g(E(y))
})

par(mfrow = c(1,2))
hist(E_g.y, col = "lightblue", xlab = "mean(g(y))",
main = sprintf("Average of E(g(Y)) over 1000\n replications
is %0.2f",mean(E_g.y)))
abline(v = mean(E_g.y), col = "red", lty = 2, lwd = 3)

hist(g_E.y, col = "lightblue", xlab = "g(mean(y))",
main = sprintf("Average of g(E(Y)) over 1000\n replications
is %0.2f",mean(g_E.y)))
abline(v = mean(g_E.y), col = "red", lty = 2, lwd = 3)
```



A sum of n random variables

- Up to now, to keep things general, we used n non-identical – but independent – random variables. If we consider the Variance and the sum of n **identical** – and independent – random variables, so the n Variances (each abbreviated to Var) are all equal, the laws simplify:

- First, since the variances add, we have that

  $$\text{Var}(RV_1 + RV_2 + \cdots + RV_n) = \text{Var}_1 + \text{Var}_2 + \cdots + \text{Var}_n = n \times \textit{each Var}\,.$$

- Taking square roots,

  $$SD(\ RV_1 + RV_2 + \cdots + RV_n\ ) = \sqrt{\ n \times \text{ each Var}} = \sqrt{n}\ \times \text{ each } SD$$

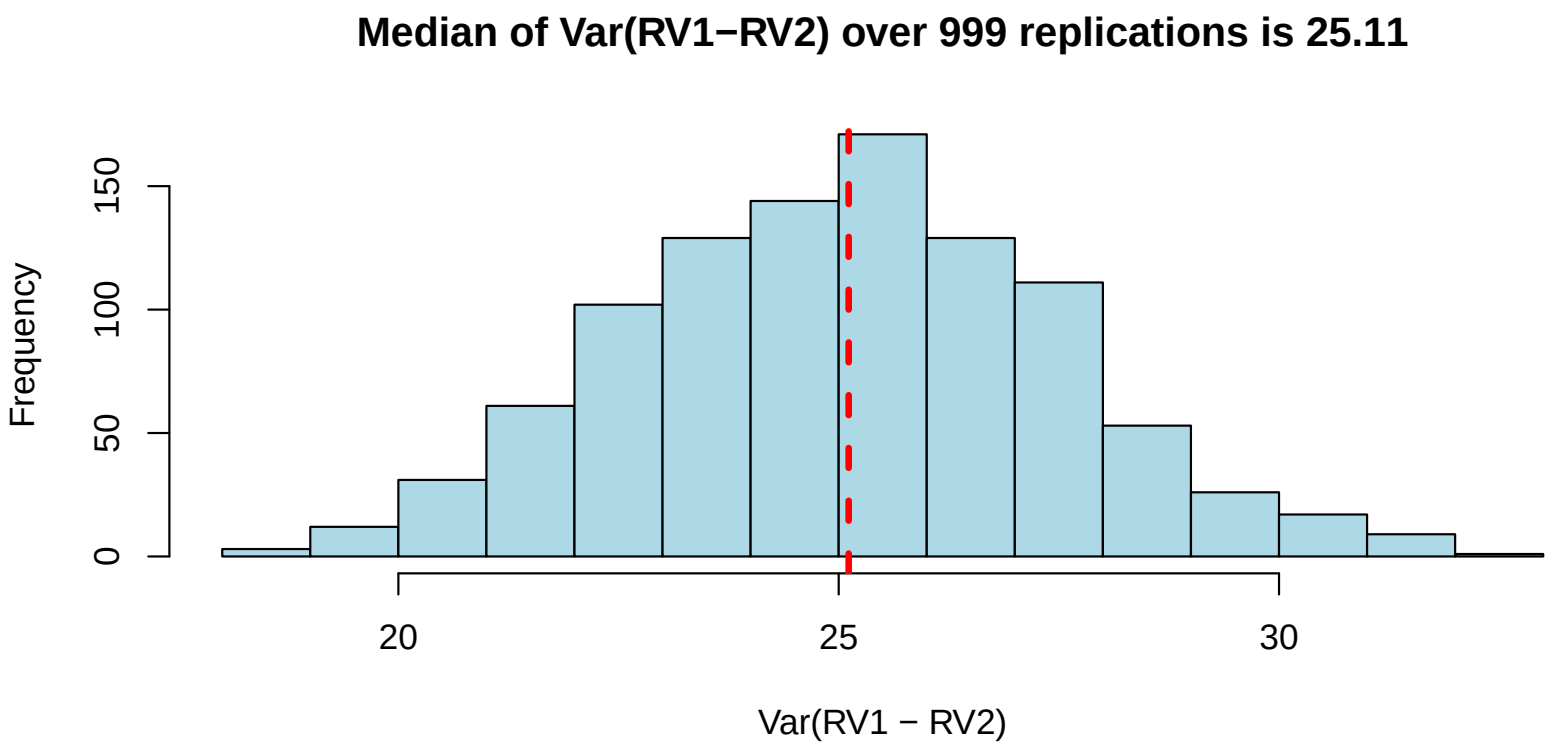
- $$SD\left(\frac{RV_1 + RV_2 + \cdots + RV_n}{n}\right) = \frac{\sqrt{n}\ \times \text{ each } SD}{n} = \frac{\text{common } SD}{\sqrt{n}}$$

- $$\text{Var}\left(\frac{RV_1 + RV_2 + \cdots + RV_n}{n}\right) = \frac{\text{common Var}}{n}$$

Difference of 2 Random Variables via Simulation

```r
set.seed(12)
B <- 999; N <- 200
var_diff <- replicate(B, {
RV1 <- rnorm(N, mean = 2, sd = 3)
RV2 <- rnorm(N, mean = 4, sd = 4)
var(RV1 - RV2)
})

hist(var_diff, col = "lightblue", xlab = "Var(RV1 - RV2)",
main = sprintf("Median of Var(RV1-RV2) over 999 replications is %0.2f",median(var_diff)))
abline(v = median(var_diff), col = "red", lty = 2, lwd = 3)
```



Exam Details

Data visualization

Tidy Data

Descriptive statistics

Grammar of graphics

Parameters, probability, random variables

Sampling Distributions, CLT, Confidence Intervals and p-values

Bootstrap

One sample mean

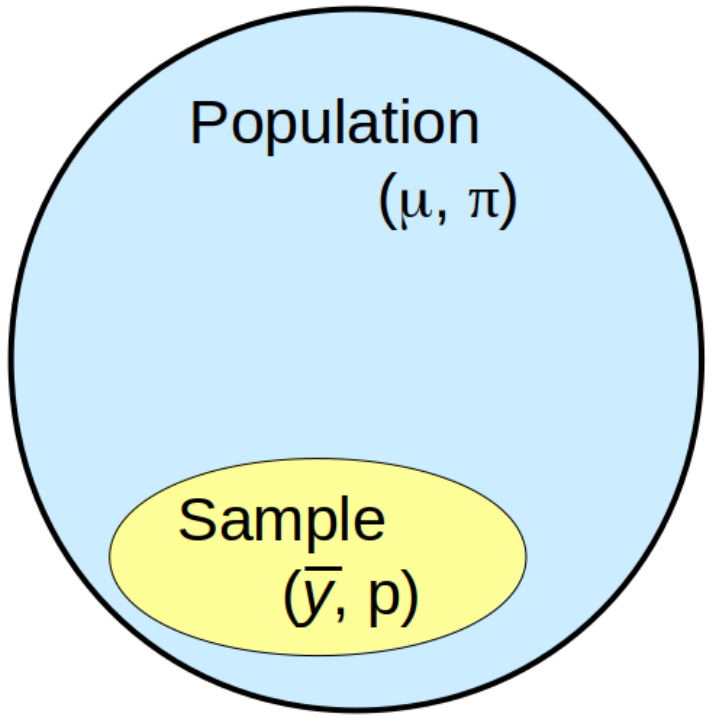
p-values

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

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).

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 2 (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

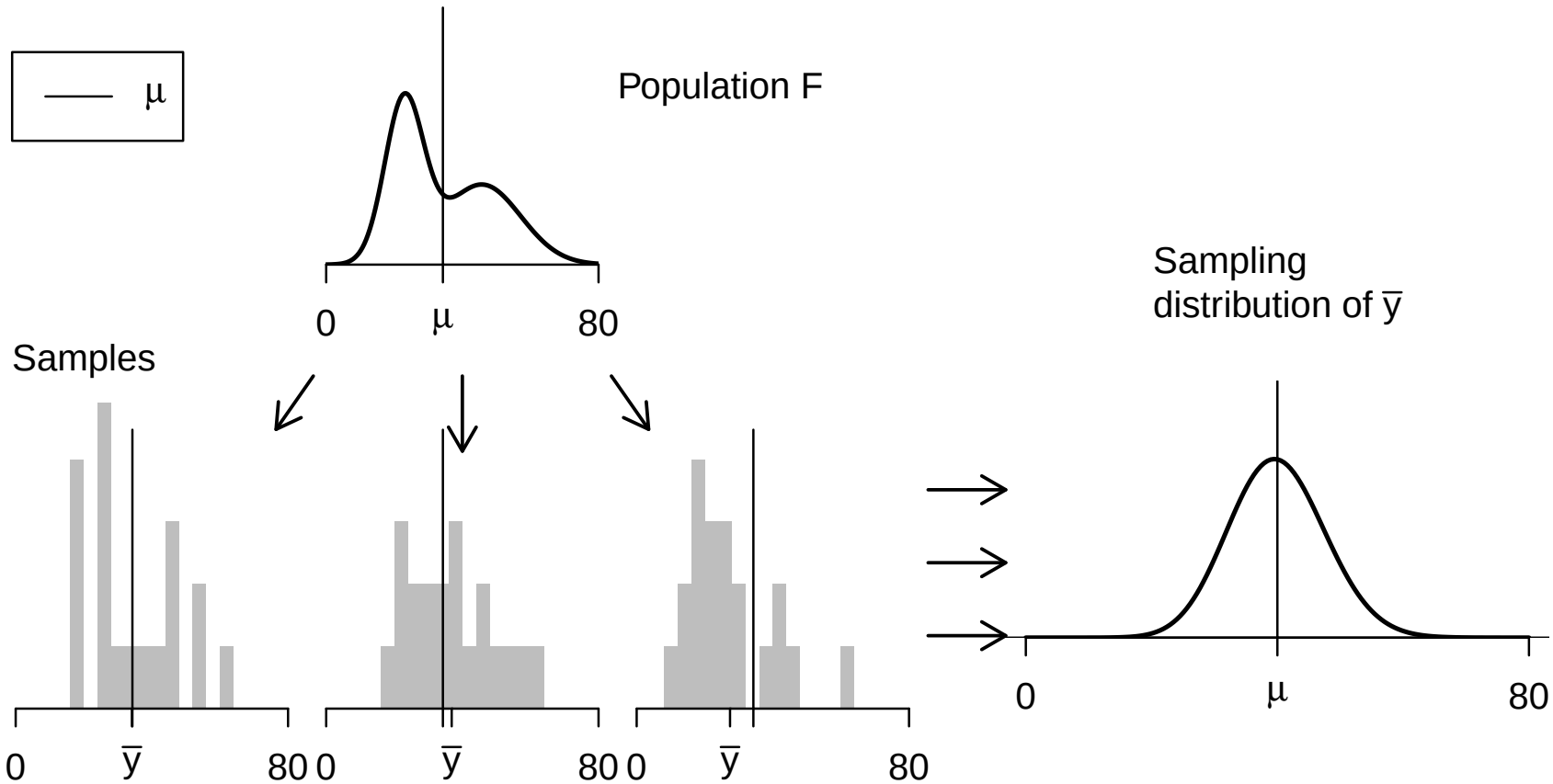


Figure: Ideal world. Sampling distributions are obtained by drawing repeated samples from the population, computing the statistic of interest for each, and collecting (an infinite number of) those statistics as the sampling distribution

Why are sampling distributions important?

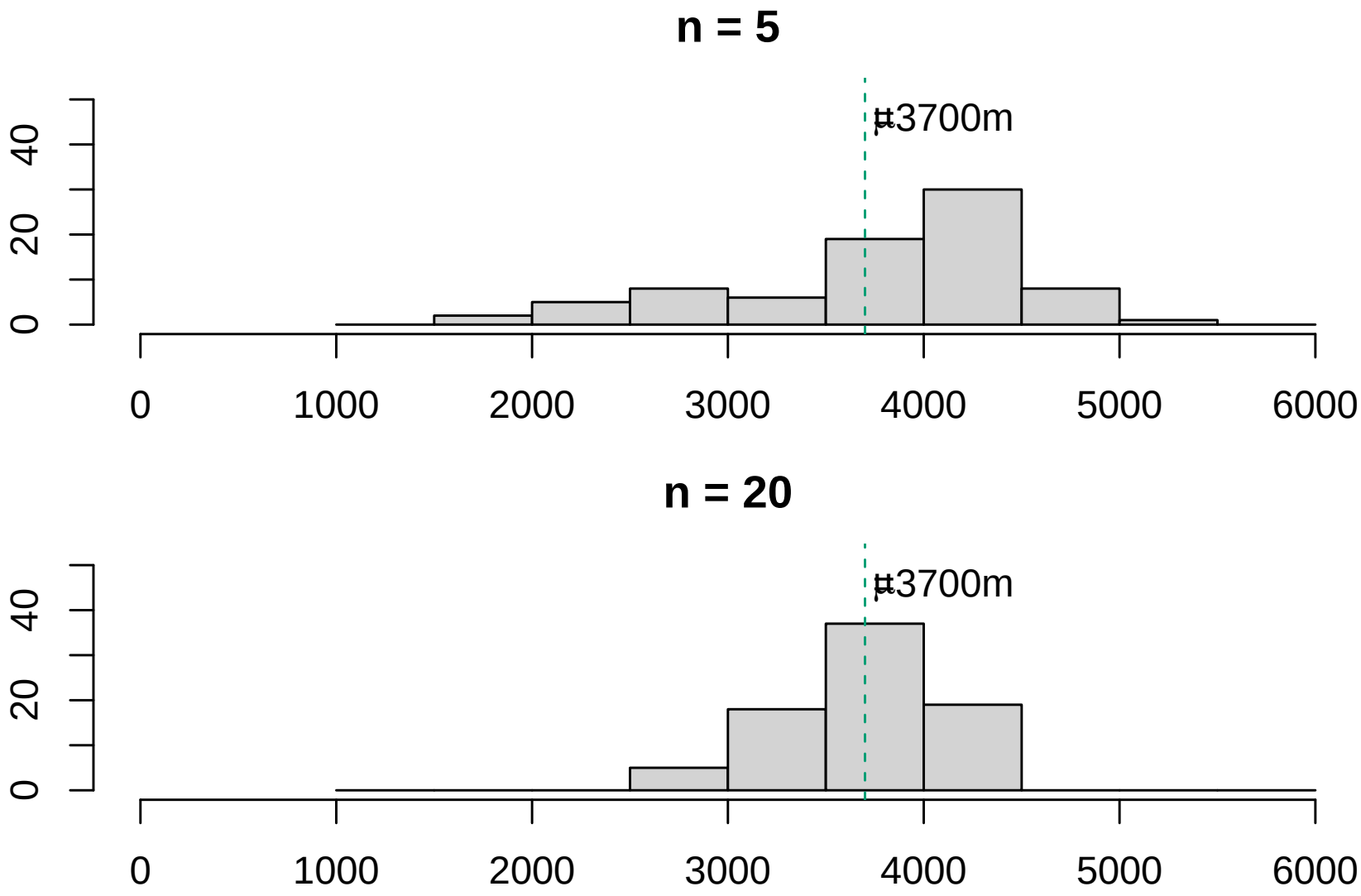
- They tell us how far from the target (true value of the parameter) our statistical shot at it (i.e. the statistic calculated form a sample) is likely to be, or, to have been.

Why are sampling distributions important?

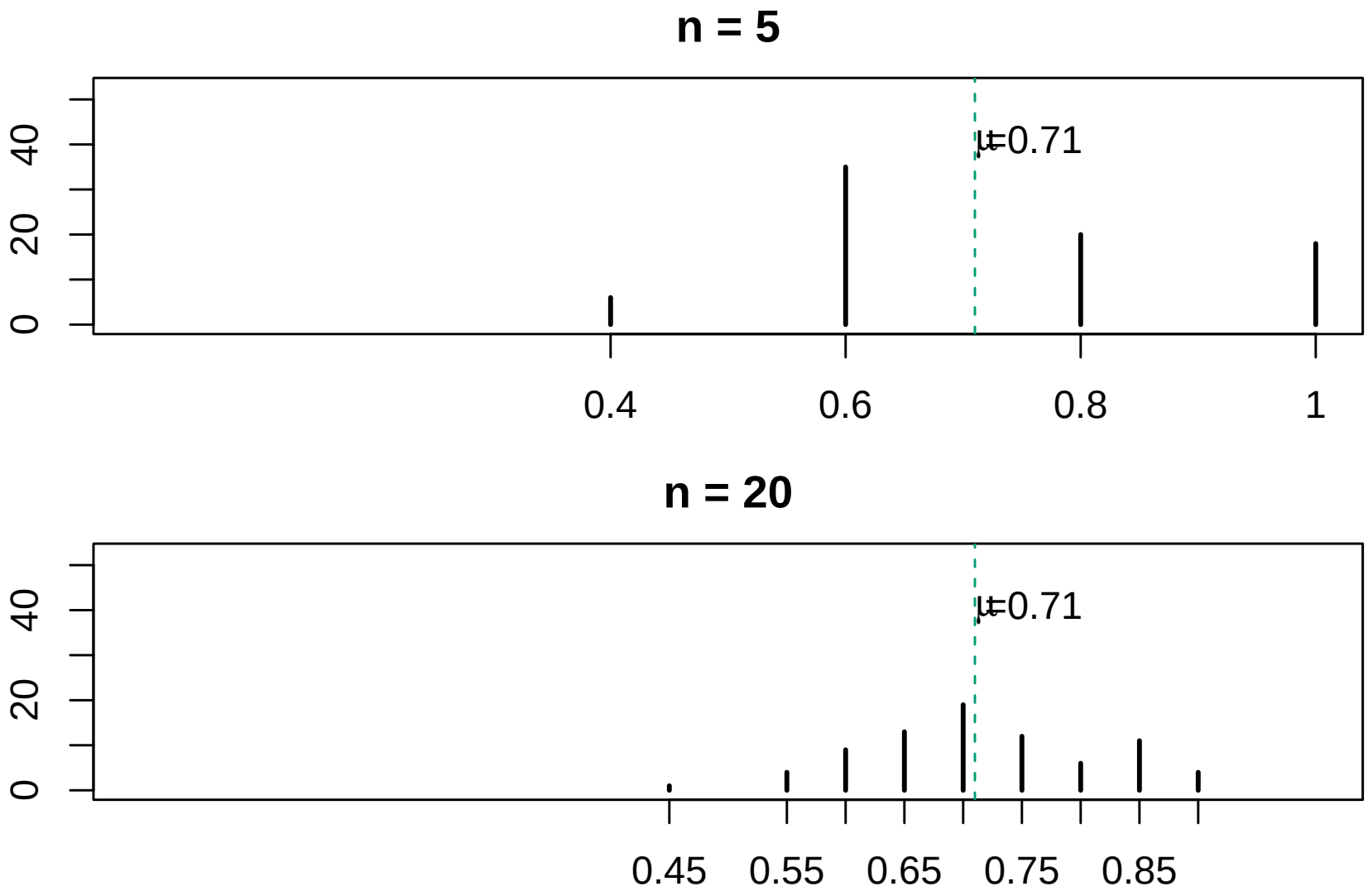
- They tell us how far from the target (true value of the parameter) our statistical shot at it (i.e. the statistic calculated form a sample) is likely to be, or, to have been.

- 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

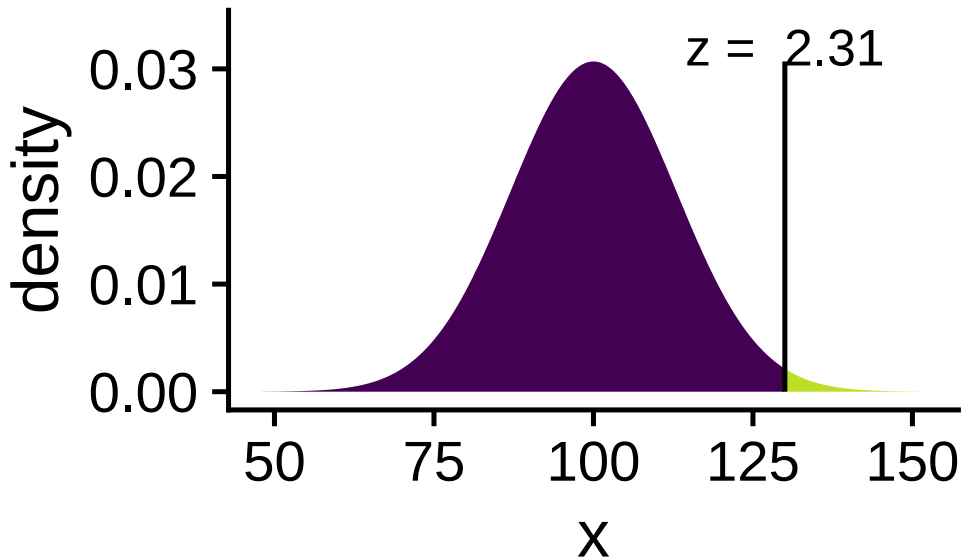


Normal Distribution: For probabilities we use *pnorm*

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

```
## [1] 0.9894919
```

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




```
## [1] 0.9894919
```

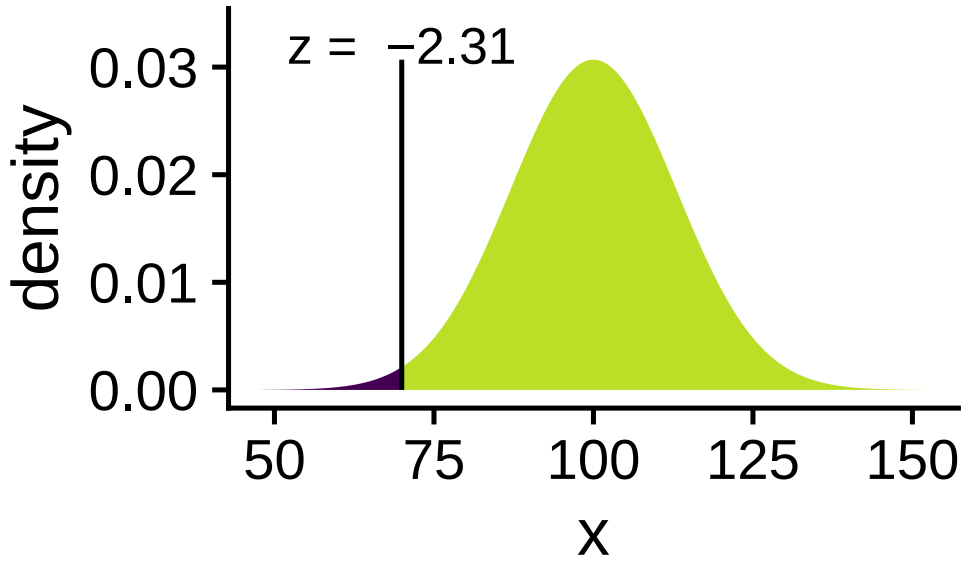
- `pnorm` returns the integral from $-\infty$ to q for a $\mathcal{N}(\mu, \sigma)$
- `pnorm` goes from *quantiles* (think Z scores) to probabilities

Normal Distribution: For quantiles we use *qnorm*

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

```
## [1] 69.94926
```

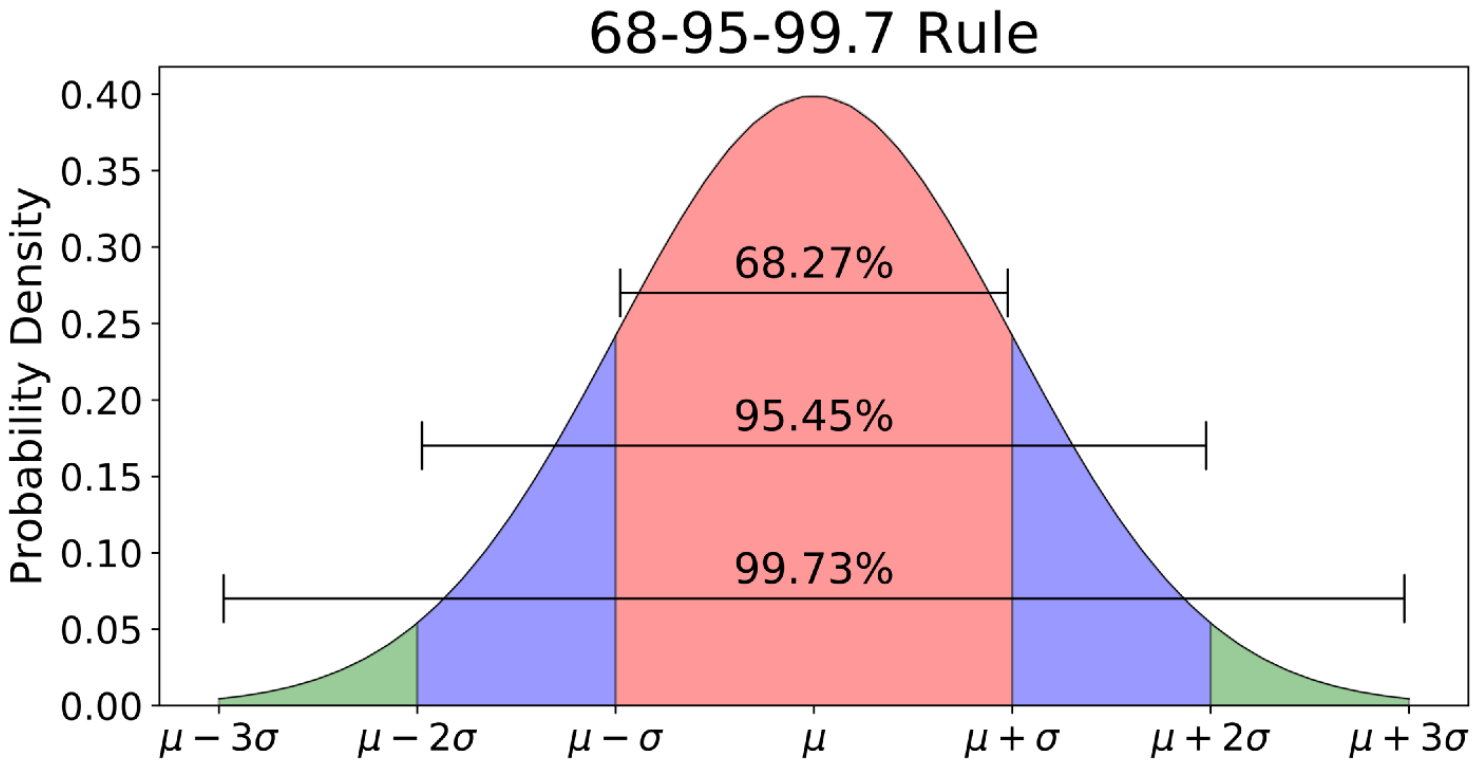
```
mosaic::xqnorm(p = 0.0104, mean = 100, sd = 13)
```




```
## [1] 69.94926
```

- `qnorm` answers the question: What is the Z-score of the pth 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

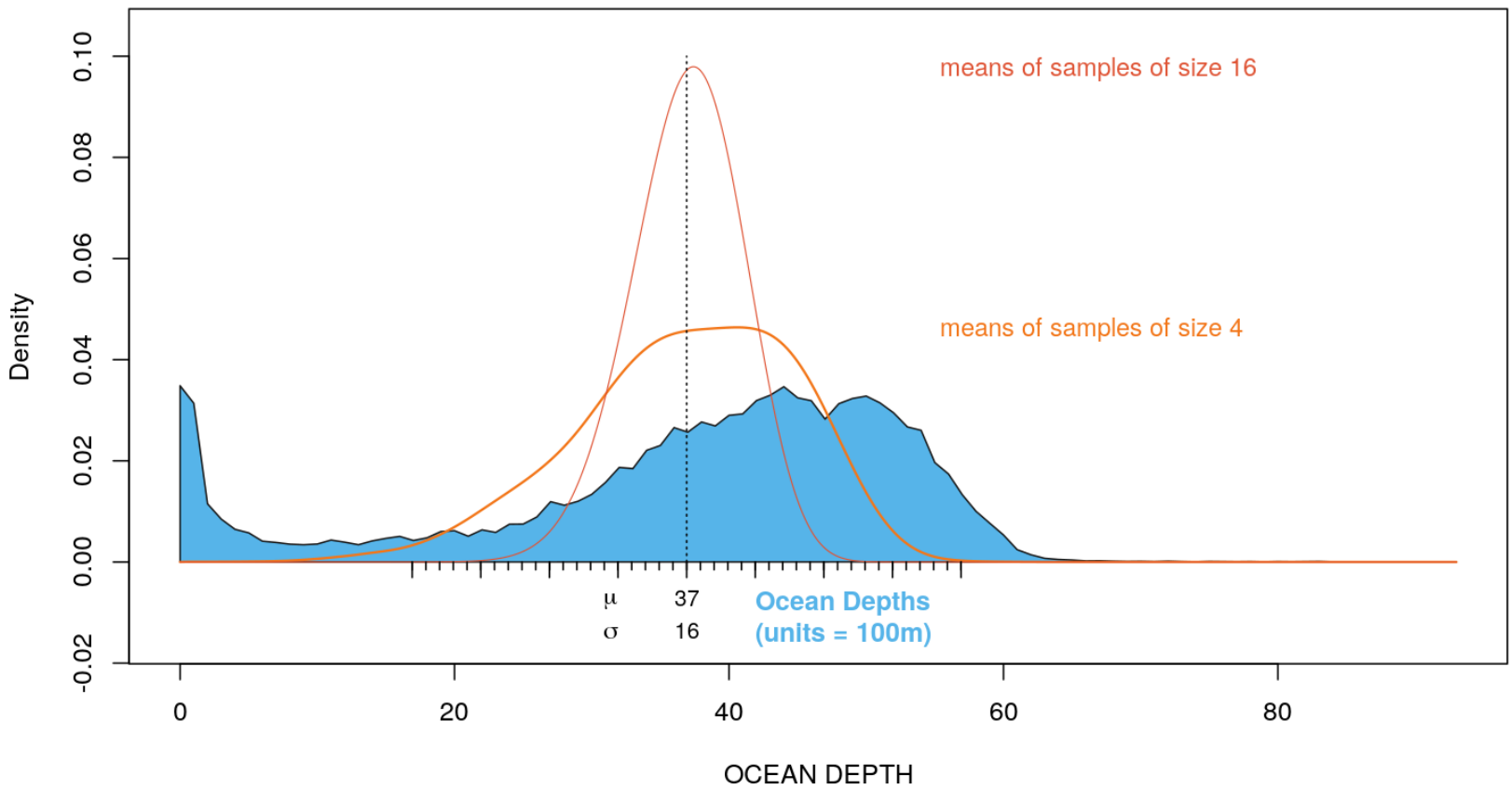


Figure: When the sample size increases from 4 to 16, the spread of the sampling distribution for the mean is reduced by a half, i.e., the range is cut in half. This is known as the curse of the $\sqrt{n}$

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 sample proportion, slope, correlation, or any other statistic created by aggregation of individual observations

Theorem 3 (Central Limit Theorem).

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

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

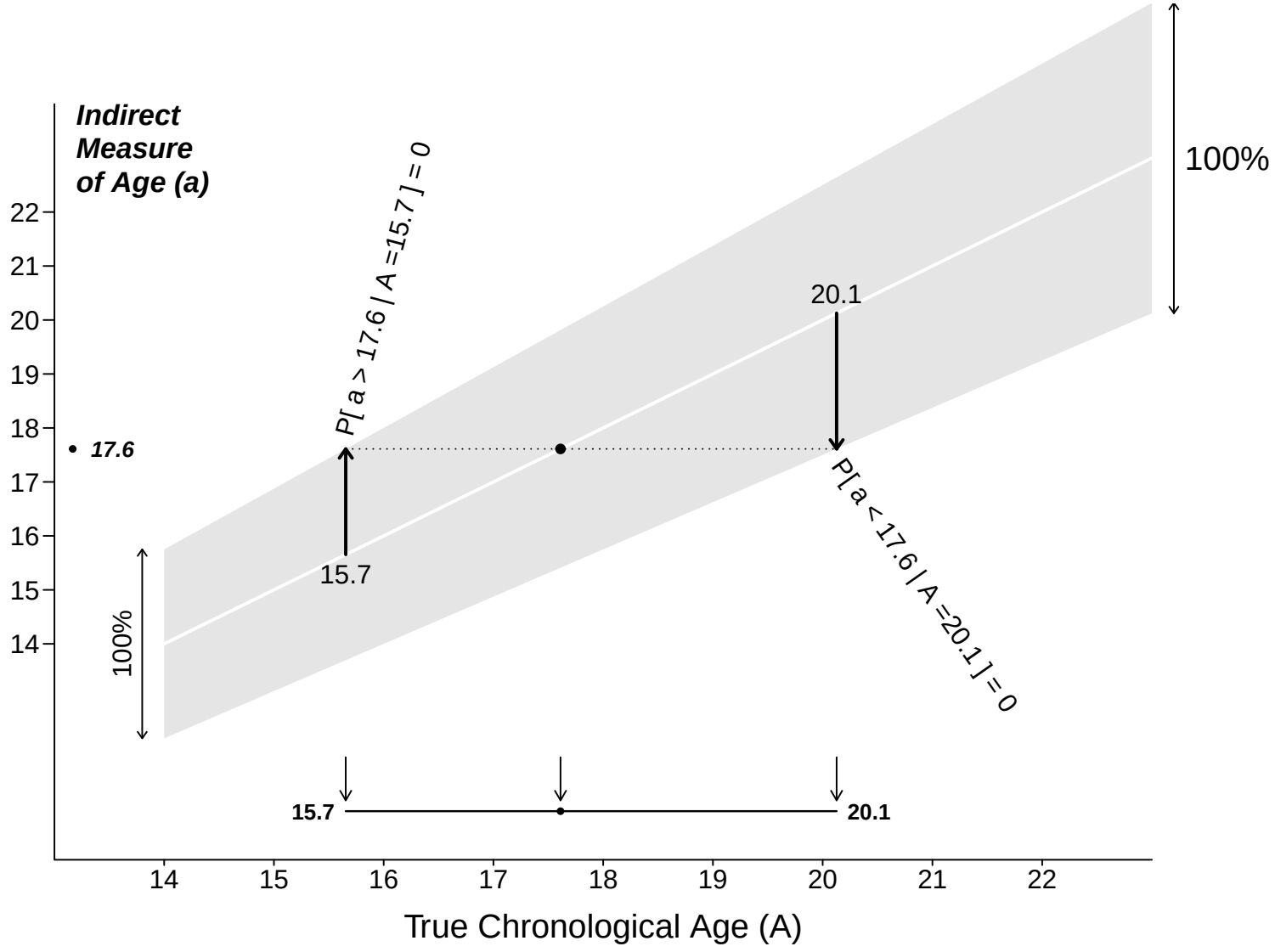


Figure: 100% Confidence Intervals for a person's chronological age when error distributions (that in this example are wider at the older ages) are 100% confined within the shaded ranges.

Confidence Interval

Definition 4 (Confidence Interval).

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

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

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

   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

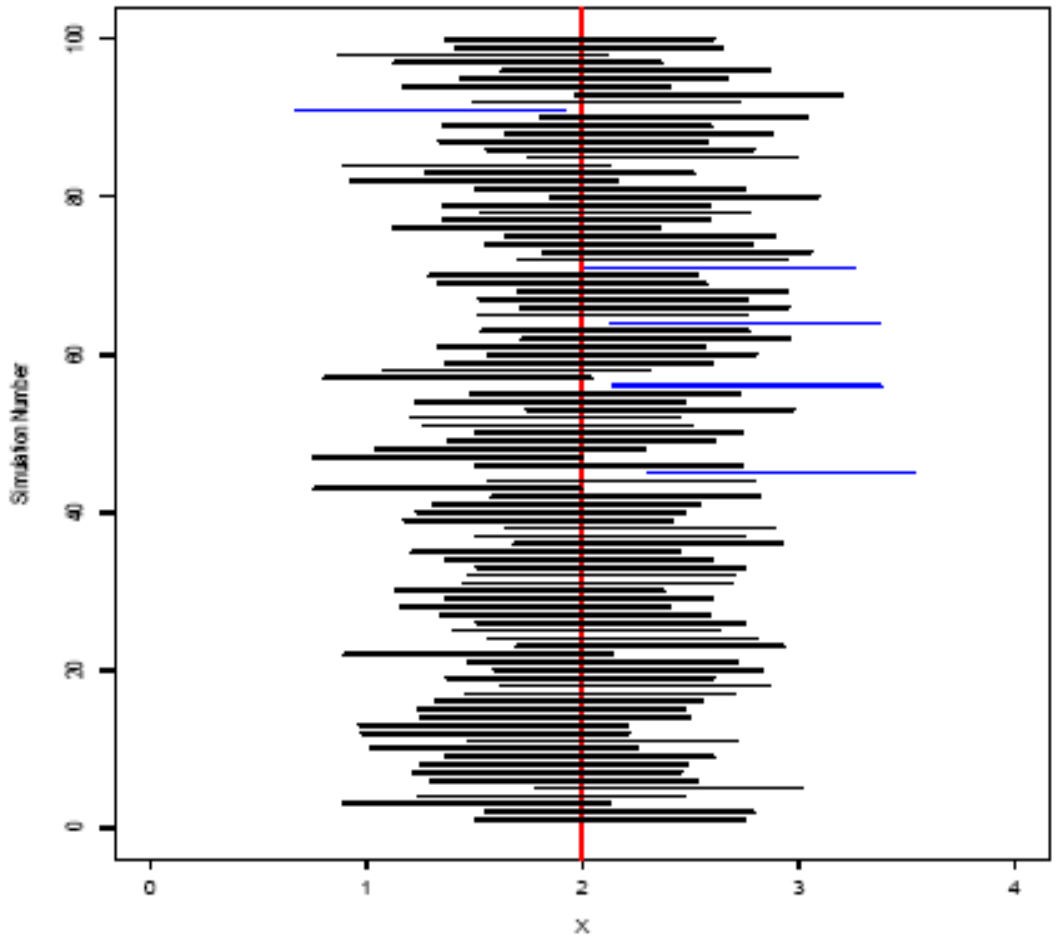


Figure: True parameter value is 2 (red line). Each horizontal black line represents a 95% CI from a sample and contains the true parameter value. The blue CIs do not contain the true parameter value. 95% of all samples give an interval that contains the population parameter.

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

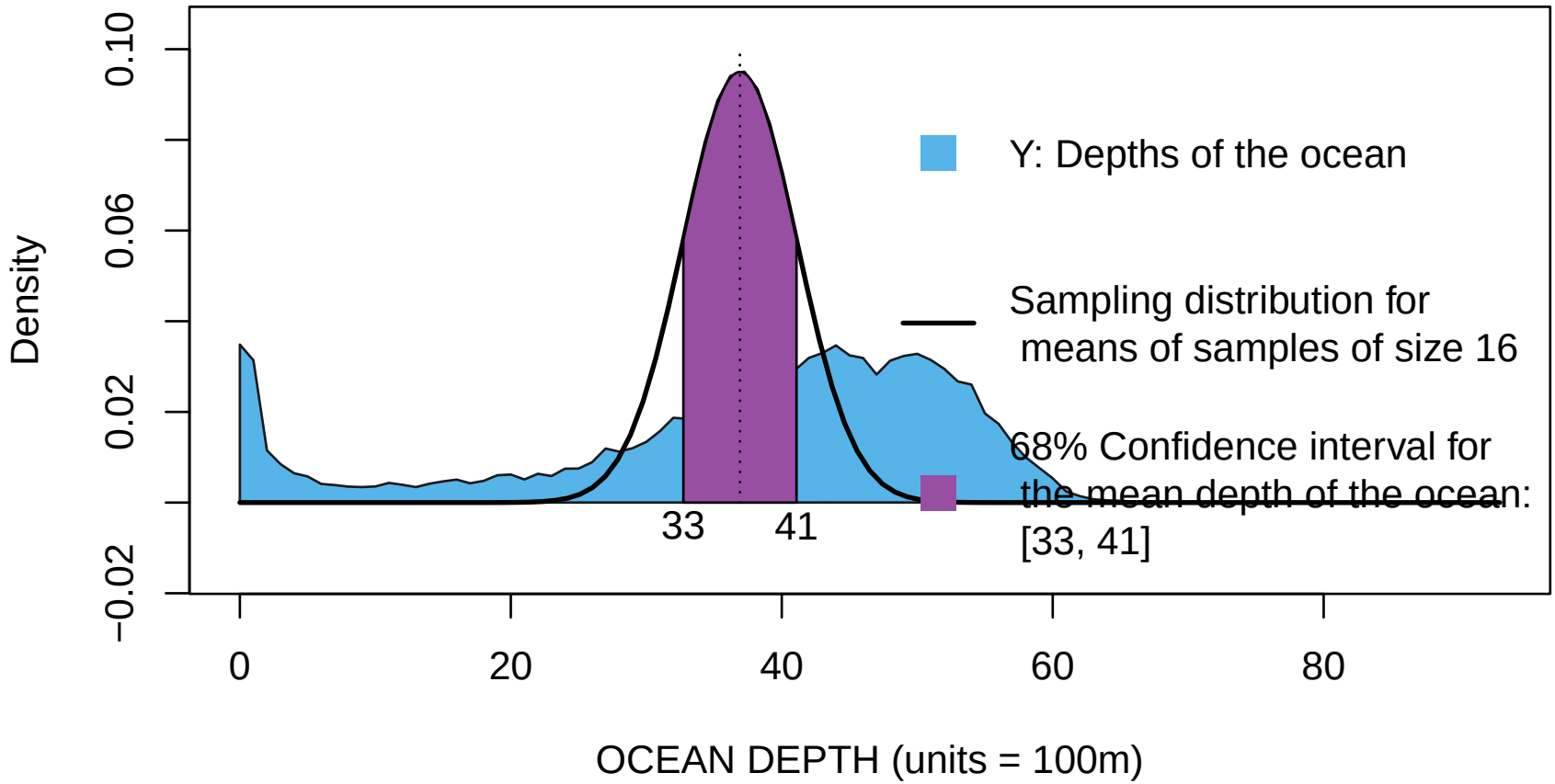


Figure: 68% Confidence interval calculated using
`qnorm(p = c(0.16,0.84), mean = 37, sd = 4.2)`

95% Confidence interval using qnorm

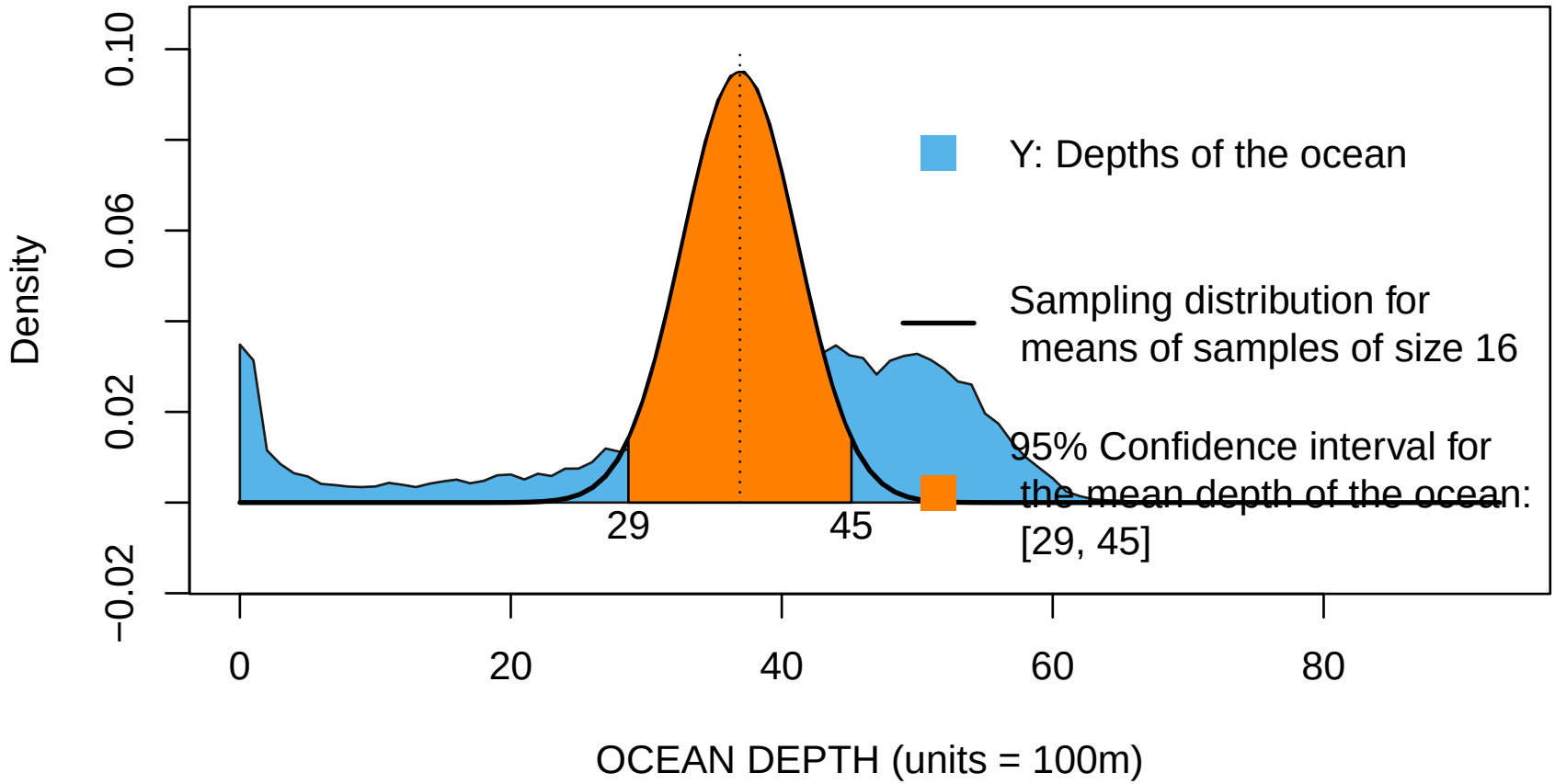


Figure: 95% Confidence interval calculated using
`qnorm(p = c(0.025,0.975), mean = 37, sd = 4.2)`

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.

Motivation for the Bootstrap

- The $\pm$ and `qnorm`/`qt` 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?

Motivation for the Bootstrap

- The $\pm$ and `qnorm`/`qt` 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?

Motivation for the Bootstrap

- The $\pm$ and `qnorm`/`qt` 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

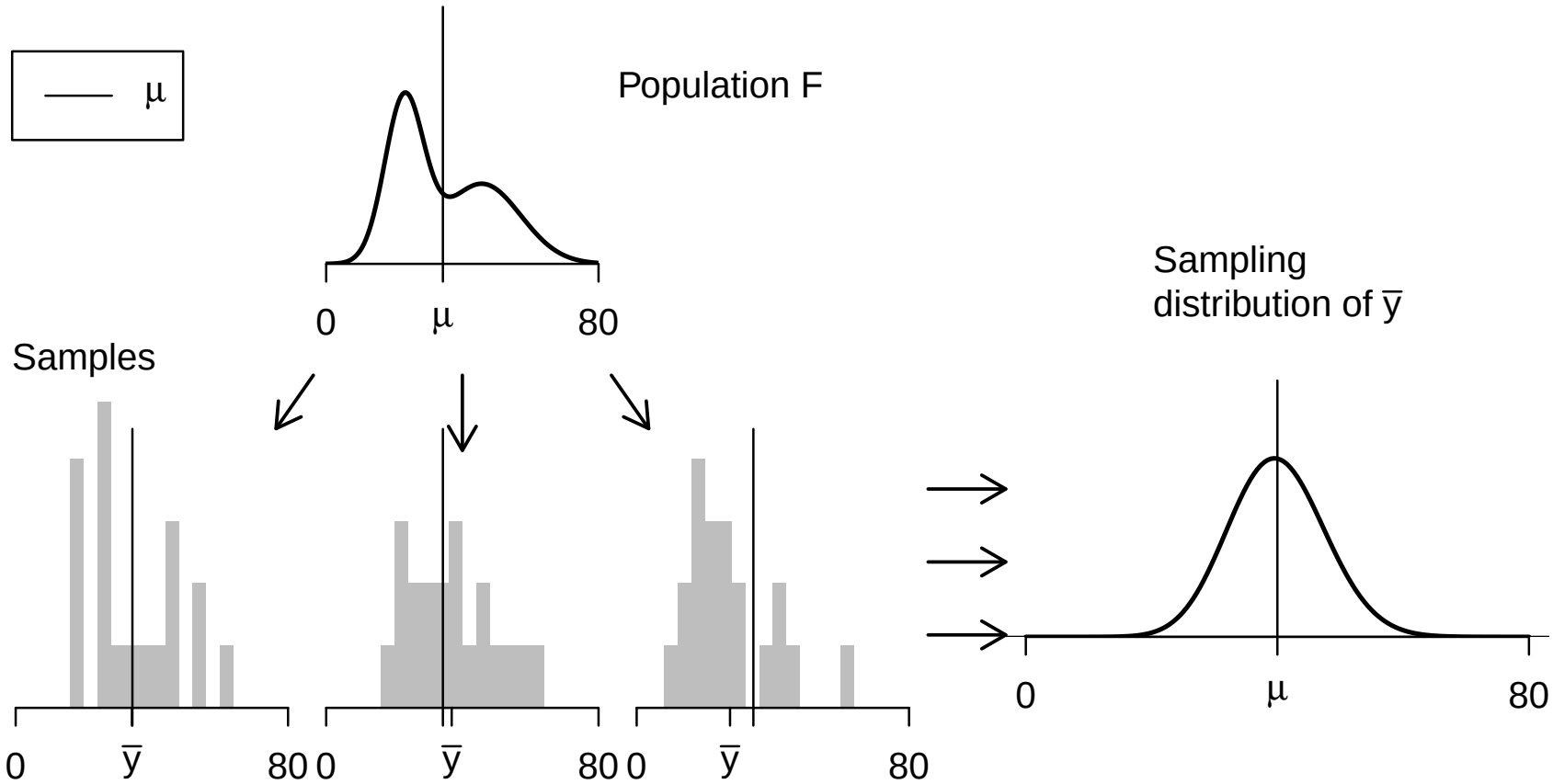


Figure: Ideal world. Sampling distributions are obtained by drawing repeated samples from the population, computing the statistic of interest for each, and collecting (an infinite number of) those statistics as the sampling distribution

Reality: use the bootstrap distribution instead

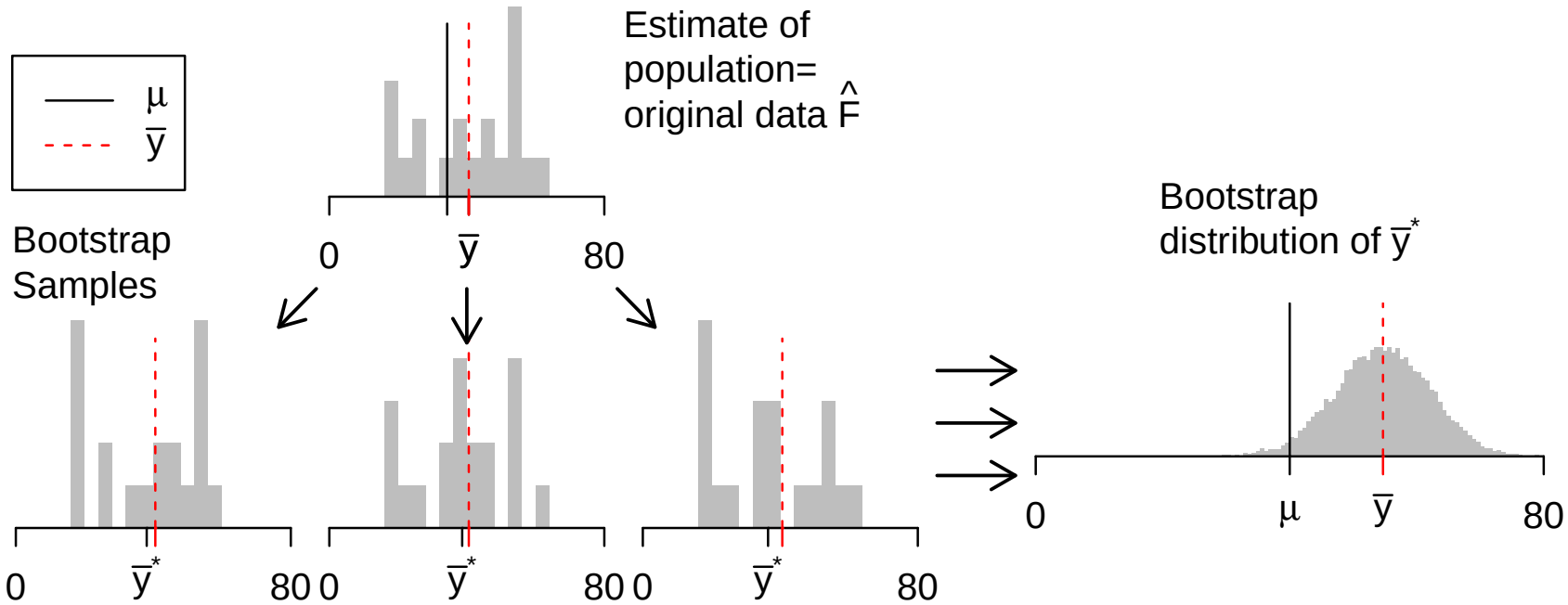
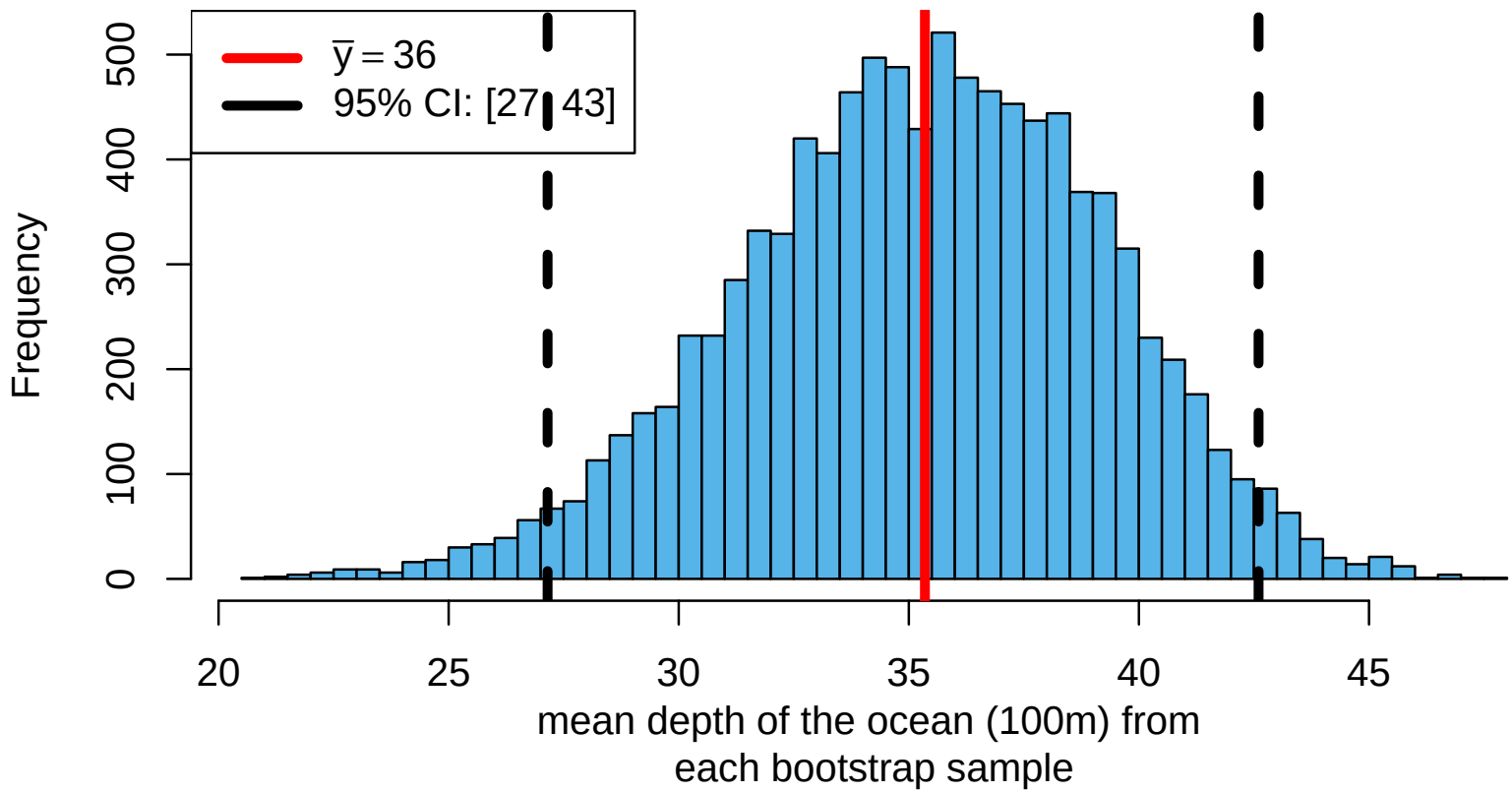


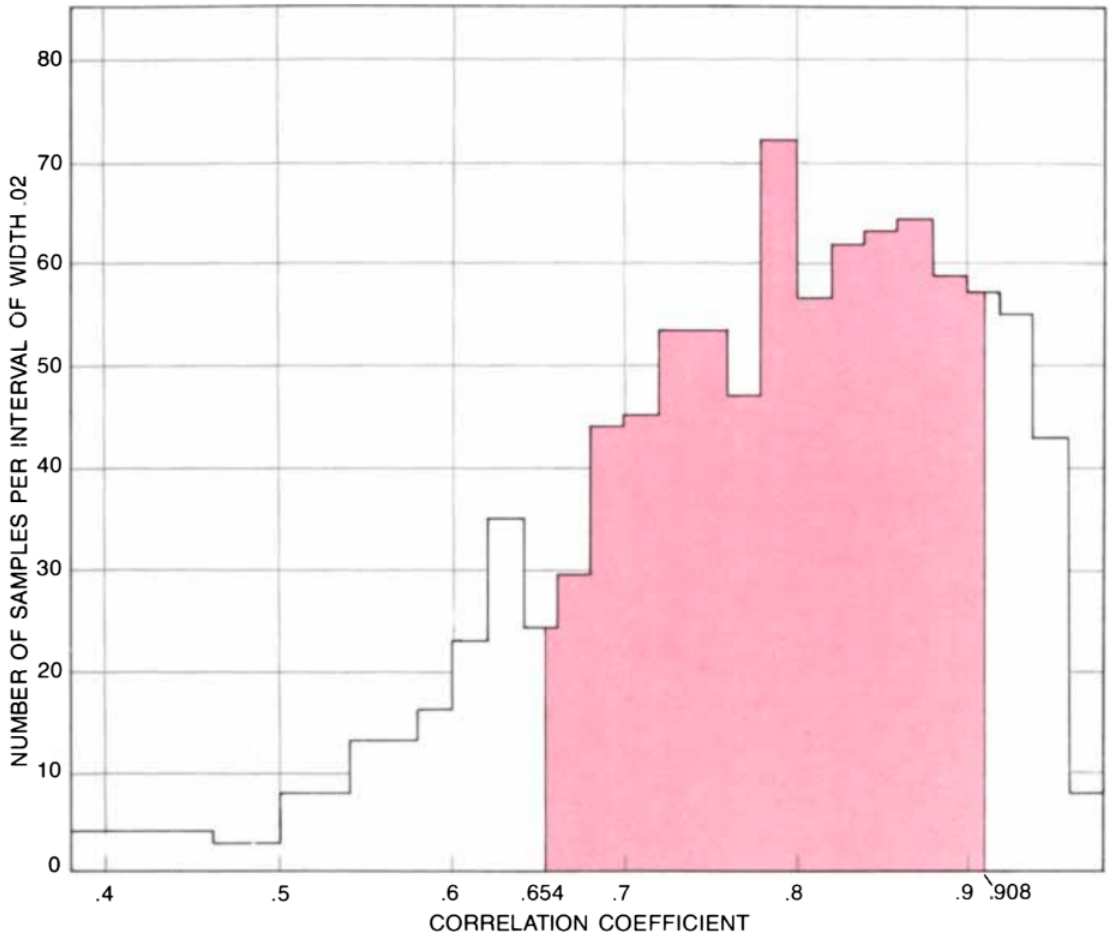
Figure: Bootstrap world. The bootstrap distribution is obtained by drawing repeated samples from an estimate of the population, computing the statistic of interest for each, and collecting those statistics. The distribution is centered at the observed statistic ($\bar{y}$), not the parameter (μ).

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

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.

Application: How fast is your reaction time?

```
reaction.times <- c(325,327,357,299,378)/1000
summary(reaction.times)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2990  0.3250  0.3270  0.3372  0.3570  0.3780

round(sd(reaction.times),3)

## [1] 0.031

length(reaction.times)

## [1] 5
```

5 ways of calculating a confidence interval

We are interested in calculating a 95% confidence interval for the mean reaction time based on the sample of 5 reaction times.

5 ways of calculating a confidence interval

We are interested in calculating a 95% confidence interval for the mean reaction time based on the sample of 5 reaction times.

Five ways of doing this:

1. By hand (using the $\pm$ formula and `R` as a calculator)
2. Using the quantile function for the t distribution `stats::qt`
3. Fitting an intercept-only regression model ($y = \beta_0 + \varepsilon$)
4. Using a canned function (`mosaic::t.test`, `stats::t.test`)
5. Bootstrap

1. By hand using the $\pm$ formula

```
n <- length(reaction.times)
SEM <- sd(reaction.times)/sqrt(n)

## [1] 0.01372734

ybar <- mean(reaction.times)

## [1] 0.3372

multiple.for.95pct <- stats::qt(p = c(0.025, 0.975), df = n-1)

## [1] -2.776445  2.776445

by_hand_CI <- ybar + multiple.for.95pct * SEM

## [1] 0.29909 0.37531
```

2. Using `stats::qt`

Note: R only provides the standard t distribution. In order to get a scaled version we must define our own function.

```
n <- length(reaction.times)
SEM <- sd(reaction.times)/sqrt(n)
ybar <- mean(reaction.times)

# scaled version of the standard t distribution
qt_ls <- function(p, df, mean, sd) qt(p = p, df = df) * sd + mean

qt_ls(p = c(0.025, 0.975), df = n - 1, mean = ybar, sd = SEM)

## [1] 0.2990868 0.3753132
```

3. Fitting an intercept-only regression model

```
fit <- stats::lm(reaction.times ~ 1)
summary(fit)

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.33720    0.01373   24.56 1.63e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0307 on 4 degrees of freedom

stats::confint(fit)

##                 2.5 %    97.5 %
## (Intercept) 0.2990868 0.3753132
```

3. Fitting an intercept-only regression model

In the regression output:

- `Estimate`: the mean reaction time (an estimate of the intercept β_0)
- `t value`: the test statistic
- `Std. Error`: the standard error of the mean (SEM)
- `Pr(>|t|)`: is the p-value

3. Fitting an intercept-only regression model

These are based on the (useless) null hypothesis $H_0 : \mu_0 = 0$

- `t value` = $\frac{\bar{y}-\mu_0}{s/\sqrt{n}} = \frac{0.33720-0}{0.01373} = 24.56$
- `Pr(>|t|)`

  = $P(\text{t value} > t_{(n-1)}) + P(-\text{t value} < t_{(n-1)})$

  = `pt(q = 24.56, df = n-1, lower.tail = FALSE)` + `pt(q = -24.56, df = n-1)`

  = $8.155 \times 10^{-6} + 8.155 \times 10^{-6} = 1.631 \times 10^{-5}$

4. Canned function

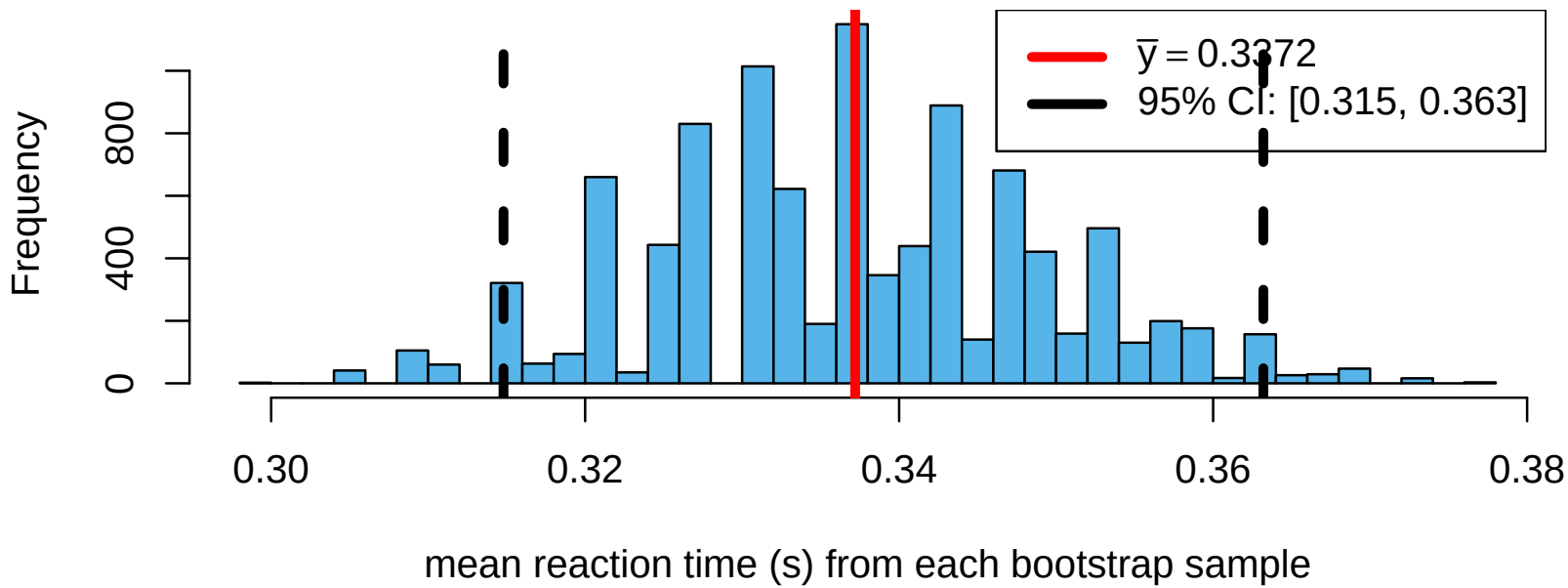
```
stats::t.test(reaction.times)

## One Sample t-test with reaction.times
## t = 24.6, df = 4, p-value = 1.63e-05
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.299 0.375
## sample estimates:
## mean of x
##     0.337
```

5. Bootstrap

```
df_react <- data.frame(reaction.times) # need data.frame to bootstrap
B <- 10000 ; N <- nrow(df_react)
R <- replicate(B, {
dplyr::sample_n(df_react, size = N, replace = TRUE) %>%
dplyr::summarize(r = mean(reaction.times)) %>%
dplyr::pull(r)
})

##  2.5% 97.5%
## 0.315 0.363
```



Exam Details

Data visualization

Tidy Data

Descriptive statistics

Grammar of graphics

Parameters, probability, random variables

Sampling Distributions, CLT, Confidence Intervals and p-values

Bootstrap

One sample mean

p-values

p-values and statistical tests

Definition 5 (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**.

The prosecutor's fallacy [1]

- Restating this both more succinctly, and in terms better suited to a statistically literate readership, the prosecutor's fallacy is to calculate P(evidence | innocence) and interpret it as P(innocence | evidence).
- It may be true that if the accused were innocent, there is only one chance in 3 million of a DNA match. But the DNA match does not necessarily imply that there is only one chance in 3 million of the accused being innocent.
- Stated more generally, the prosecutor's fallacy is

$$P(A|B) = P(B|A)$$

- We know, from Bayes' rule, that

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

[1] The Bayesian flip Correcting the prosecutor's fallacy. Significance. August 2015.

The Bayesian Flip

- In many investigations we may be presented with P(data | theory), but what we would really like to know is P(theory | data): the probability that our theory is correct, given what we have observed
- To move from P(data | theory) to P(theory | data), we need to do the Bayesian flip.

The Bayesian Flip

- In many investigations we may be presented with P(data | theory), but what we would really like to know is P(theory | data): the probability that our theory is correct, given what we have observed
- To move from P(data | theory) to P(theory | data), we need to do the Bayesian flip.
- Every year in the United States 38 million women are tested for breast cancer with mammograms. Of these, 140 000 have cancer. Mammograms have been determined to be 90% accurate for women with breast cancer.

The Bayesian Flip

- In many investigations we may be presented with P(data | theory), but what we would really like to know is P(theory | data): the probability that our theory is correct, given what we have observed
- To move from P(data | theory) to P(theory | data), we need to do the Bayesian flip.
- Every year in the United States 38 million women are tested for breast cancer with mammograms. Of these, 140 000 have cancer. Mammograms have been determined to be 90% accurate for women with breast cancer.
- This figure was calculated by tallying all of the women who were eventually determined to have breast cancer and looking back to see if their initial mammograms were positive, thus:

$$P(+mammogram|cancer) = 0.90$$

and, using a similar empirical investigation,

$$P(+mammogram|nocancer) = 0.10$$

The Bayesian Flip

- It is important to know that a test is both powerful and has a relatively low rate of false positives. But when one is faced with a positive mammogram result, these are hardly useful. We administer a mammogram because we do not know whether or not someone has cancer.
- What we want to know is

$$P(cancer| + mammogram)$$

The Bayesian Flip

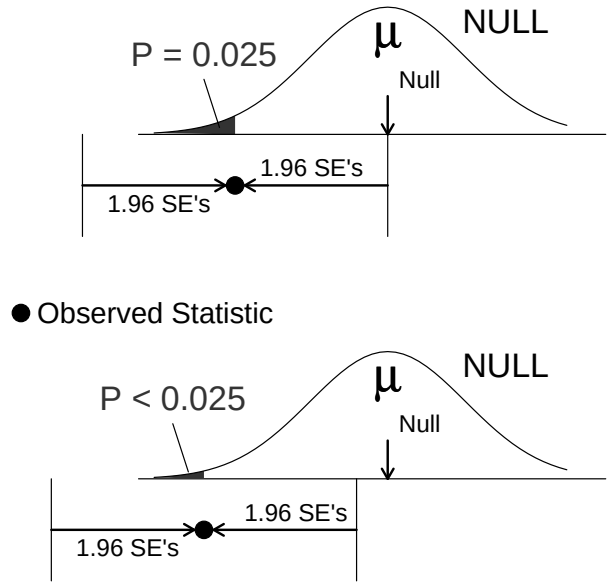
- It is important to know that a test is both powerful and has a relatively low rate of false positives. But when one is faced with a positive mammogram result, these are hardly useful. We administer a mammogram because we do not know whether or not someone has cancer.
- What we want to know is

$$P(cancer|+mammogram)$$

- This probability is a fraction that has as its numerator the number of women annually diagnosed with breast cancer via mammograms, or 140 000, and as its denominator the number of positive mammograms (including both true cancer cases and false positives):

$$\begin{aligned} P(cancer|+mammogram) &= \frac{True\,positives}{(True\,positives + False\,positives)} \\ &= 140000/(140000 + 0.1{\times}38\,million) \\ &= 140000/(140000 + 3800000) \\ &= 140000/3940000 = 0.036 = 3.6\% \end{aligned}$$

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.

Session Info

```
R version 4.1.1 (2021-08-10)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Pop!_OS 21.04

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.13.so

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_2.0-2   oibiostat_0.2.0
 [5] DT_0.16            mosaic_1.7.0       Matrix_1.3-2       mosaicData_0.20.1
 [9] ggformula_0.9.4    ggstance_0.3.4     lattice_0.20-41    kableExtra_1.2.1
[13] socviz_1.2         gapminder_0.3.0    here_0.1           NCStats_0.4.7
[17] FSA_0.8.30         forcats_0.5.1      stringr_1.4.0      dplyr_1.0.7
[21] purrr_0.3.4        readr_1.4.0        tidyr_1.1.3        tibble_3.1.5
[25] ggplot2_3.3.5      tidyverse_1.3.0    knitr_1.36

loaded via a namespace (and not attached):
  [1] readxl_1.3.1       backports_1.2.1    plyr_1.8.6
  [4] splines_4.1.1      crosstalk_1.1.1    leaflet_2.0.3
  [7] TH.data_1.0-10     digest_0.6.28      htmltools_0.5.2
 [10] fansi_0.5.0        magrittr_2.0.1     mosaicCore_0.8.0
 [13] openxlsx_4.1.5     modelr_0.1.8       sandwich_2.5-1
 [16] blob_1.2.1         rvest_1.0.0        ggrepel_0.8.2
 [19] haven_2.3.1        xfun_0.26          crayon_1.4.1
 [22] jsonlite_1.7.2     lme4_1.1-23        survival_3.2-13
 [25] zoo_1.8-8          glue_1.4.2         polyclip_1.10-0
 [28] gtable_0.3.0       emmeans_1.5.1      webshot_0.5.2
 [31] sjstats_0.18.0     sjmisc_2.8.5       car_3.0-9
 [34] abind_1.4-5        scales_1.1.1       mvtnorm_1.1-1
 [37] DBI_1.1.1          rstatix_0.6.0      ggeffects_0.16.0
 [40] Rcpp_1.0.7         viridisLite_0.4.0  xtable_1.8-4
 [43] performance_0.7.3  foreign_0.8-81     datawizard_0.2.0.1
 [46] htmlwidgets_1.5.3  httr_1.4.2         ellipsis_0.3.2
 [49] pkgconfig_2.0.3    farver_2.1.0       dbplyr_1.4.4
 [52] utf8_1.2.2         tidyselect_1.1.1   labeling_0.4.2
 [55] rlang_0.4.11       effectsize_0.4.5   munsell_0.5.0
```