Research Design, Fall 2021

01: Introduction; Academic Integrity; Concepts of Probability

Scott Spencer | Columbia University

Meeting your professor

Doctorate
Jurisprudence
Honors in research and writing
Focus — analysis

Master of Science
Sports Management
Focus — data science analytics
Won, SABR analytics competition

Bachelor of Science
Chemical Engineering
Focus — numerical methods, statistical process control

Scott Spencer

Columbia University

Faculty, Lecturer, Alumnus

Consultant, Data Scientist

Professional sports
Example — Major-league baseball research and development for player performance & manager decision-making

Data for good
Example — Bayesian, generative modeling effects of climate change on perceived expectations of property values

Innovation
Example — whether invented attributes of an edible oil previously existed or was made or sold by competitor

Teaching and Research

Developing generative models
Building Bayesian, generative models to enable decision-making in complex fields such as sports performance.

Communicating uncertainty
Writing monograph on quantitative persuasion amid uncertainty. Developing R packages to tie human perception to graphical representation of data.

Contributing open-source software
Contribute to interfaces to Stan, a probabilistic programming language. Develop R packages for data visualization.

Meeting your associate

Kristina Arakelyan

Columbia University

Master of Public Health
Expected Fall 2021

Master of Science
Criminology and Criminal Justice

Master of Arts
European & Russian Studies

Bachelor of Arts
Philosophy

Experience

NYC Mayor's Office
Director of Research and Evaluation at the Mayor's Office to End Domestic and Gender-Based Violence (ENDGBV)

Non-profit
Non-profit and public sectors on policy research in human trafficking, migrant smuggling, and labor exploitation

Teaching and Research

Along with this course, teaches courses in the humanities and social sciences at local universities.

Mixed-methods research to identify data-informed solutions to health and social policy issues.

Who are your fellow students and future colleagues? Say hello.

weekly, in-person discussion

Class meets Wednesdays 6:10-8PM
Faculty House: Seminar 1 & Reception

office hours

Professor Scott Spencer
Click to schedule appointment

Associate Kristina Arakelyan
Email ka2544@columbia.edu for appointment

What is ***research design***?

research, n.1

1. The act of searching carefully for or pursuing a specified thing or person; an instance of this.

2.a. Systematic investigation or inquiry aimed at contributing to knowledge of a theory, topic, etc., by careful consideration, observation, or study of a subject. In later use also: original critical or scientific investigation carried out under the auspices of an academic or other institution. Occasionally with of; now frequently with into, on.

...

design, n.

I. A plan conceived in the mind, and related senses.
1. A plan or scheme conceived in the mind and intended for subsequent execution; the preliminary conception of an idea that is to be carried into effect by action; a project.

- Oxford English Dictionary.

- 1. Introduction to Research Design; Academic Integrity; Statistics and Probability
- 2. Populations, Research Questions, and Hypotheses
- 3. Experiments and Variables
- 4. Statistical Sampling and Tests, Part 1
- 5. Statistical Sampling and Tests, Part 2
- 6. Observational Studies
- 7. Survey Design
- 8. Multifactorial Experiments
- 9. Power, Sample Size, and Simulations
- 10. Operationalizing Research
- 11. Communicating Research Plans

Are these *good* questions:

How do heights differ, if at all, between male and female graduate students studying in the applied analytics program at Columbia University?

Do emotions related to alcohol consumption differ by alcohol type?

What effect might changing a marketing message have on consumer response?

Are ocean levels over time associated with changes in frequency or severity of flooding on coastal properties and in turn associated with a change in those property values?

What oceanic properties, weather, or events may be associated with the probability of losses at salmon fisheries?

Which characteristics of baseball pitches may affect — or be associated with — the number of runs scored by the opposing team?

How might someone *answer* them? What *assumptions*, if any, may *limit the scope* of the answers? How may the answers be *explained*?

R isn’t just a letter in an alphabet!?

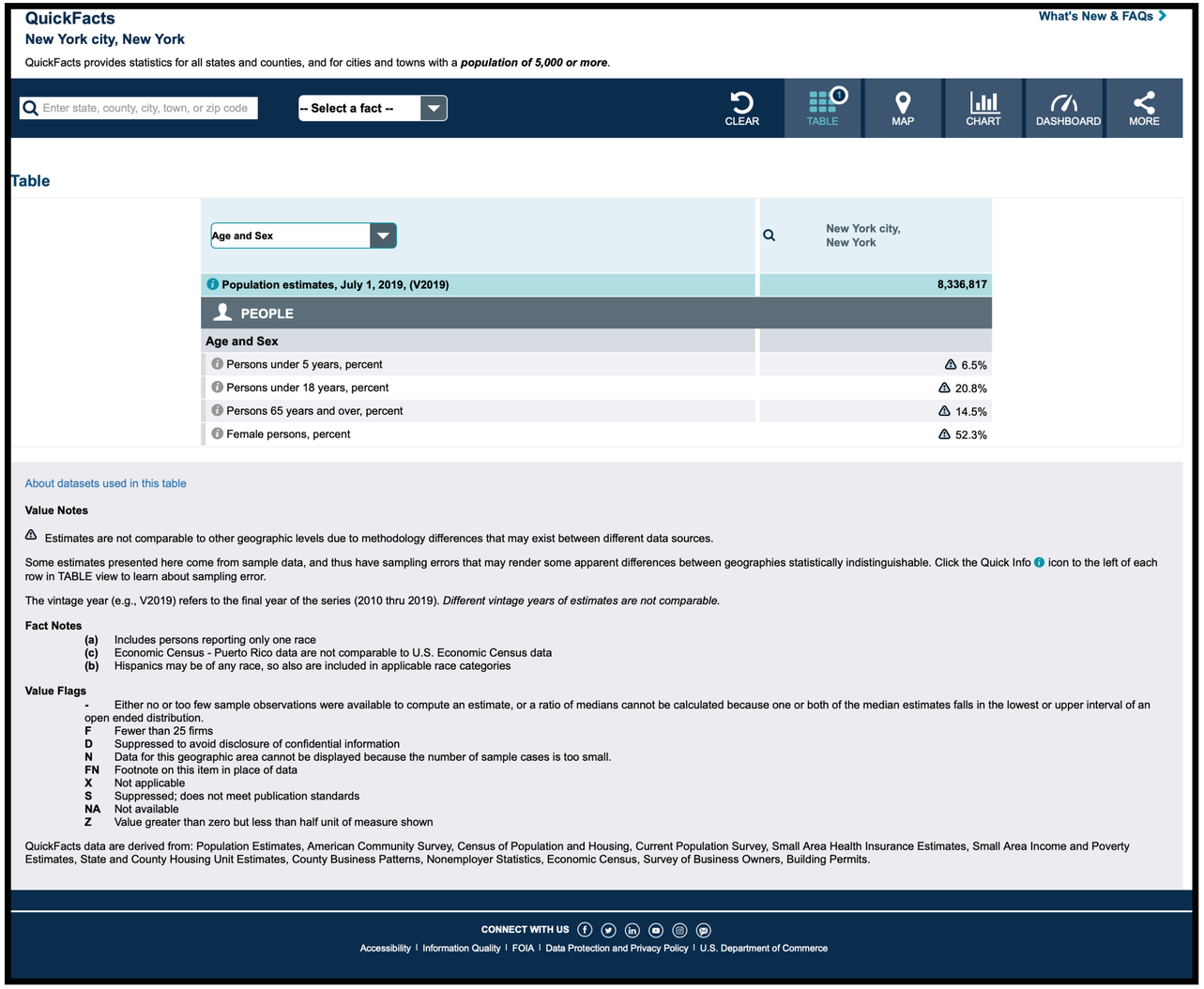
Probability, a foundational tool for research design and — more generally— for data science

population

A population consists of the set of all items or attributes of interest. The population may consist of a group of people or some other kind of object.

population

A population consists of the set of all items or attributes of interest. The population may consist of a group of people or some other kind of object.



QuickFacts

New York city, New York

What's New & FAQs

QuickFacts provides statistics for all states and counties, and for cities and towns with a ***population of 5,000 or more***.

Enter state, county, city, town, or zip code | -- Select a fact -- | CLEAR | TABLE | MAP | CHART | DASHBOARD | MORE

Table

Age and Sex	New York city, New York
Population estimates, July 1, 2019, (V2019)	8,336,817
PEOPLE	
Age and Sex	
Persons under 5 years, percent	6.5%
Persons under 18 years, percent	20.8%
Persons 65 years and over, percent	14.5%
Female persons, percent	52.3%

About datasets used in this table

Value Notes

Estimates are not comparable to other geographic levels due to methodology differences that may exist between different data sources.

Some estimates presented here come from sample data, and thus have sampling errors that may render some apparent differences between geographies statistically indistinguishable. Click the Quick Info icon to the left of each row in TABLE view to learn about sampling error.

The vintage year (e.g., V2019) refers to the final year of the series (2010 thru 2019). *Different vintage years of estimates are not comparable.*

Fact Notes

- **(a)** Includes persons reporting only one race
- **(c)** Economic Census - Puerto Rico data are not comparable to U.S. Economic Census data
- **(b)** Hispanics may be of any race, so also are included in applicable race categories

Value Flags

- **-** Either no or too few sample observations were available to compute an estimate, or a ratio of medians cannot be calculated because one or both of the median estimates falls in the lowest or upper interval of an open ended distribution.
- **F** Fewer than 25 firms
- **D** Suppressed to avoid disclosure of confidential information
- **N** Data for this geographic area cannot be displayed because the number of sample cases is too small.
- **FN** Footnote on this item in place of data
- **X** Not applicable
- **S** Suppressed; does not meet publication standards
- **NA** Not available
- **Z** Value greater than zero but less than half unit of measure shown

QuickFacts data are derived from: Population Estimates, American Community Survey, Census of Population and Housing, Current Population Survey, Small Area Health Insurance Estimates, Small Area Income and Poverty Estimates, State and County Housing Unit Estimates, County Business Patterns, Nonemployer Statistics, Economic Census, Survey of Business Owners, Building Permits.

CONNECT WITH US

Accessibility | Information Quality | FOIA | Data Protection and Privacy Policy | U.S. Department of Commerce

— U.S. Census Bureau QuickFacts - Population estimates, July 1, 2019, (V2019)
https://www.census.gov/quickfacts/fact/table/newyorkcitynewyork/PST045219

population

A population consists of the set of all items or attributes of interest. The population may consist of a group of people or some other kind of object.

Height is normally distributed

Adult heights within a population are approximately normally distributed due to genetic and environmental variance.[44]

Height is partly determined by the interaction of 423 genes with 697 variants.[45]

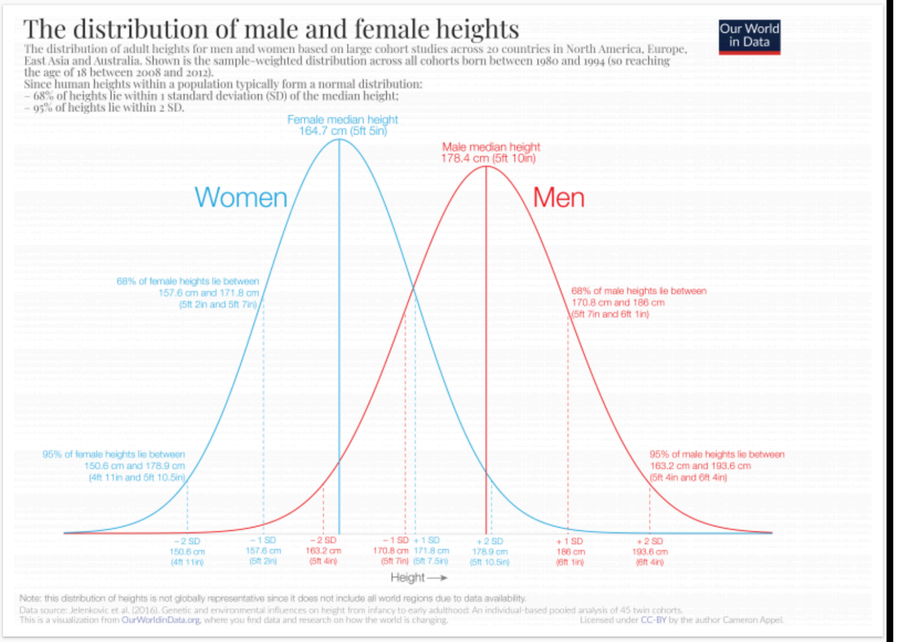
One of the basic rules of probability (known as the Central Limit Theorem) says the distribution of a trait that is determined by independent random variables, like height and genes, roughly follows a bell curve. This means the range of human heights in a population fall centrally around the mean height. In statistical terms, it's also the case that the mean and median height are the same – they fall right in the middle of the distribution.[46]

The normal distribution of heights allows us to make inferences about the range. Around 68% of heights will fall within one standard deviation of the mean height; 95% within two standard deviations; and 99.7% within three. If we know the mean and standard deviation of heights, we have a good understanding of how heights vary across a population.

Drawing upon height data from almost 150,000 twinned pairs born between 1886 and 1994, one study investigated the variance in heights across populations through time, and tried to explain how much could be explained by genetics versus environmental differences.[47]

We see this distribution of heights in the chart. As an aggregate of the regions with available data – Europe, North America, Australia, and East Asia – they found the mean male height to be 178.4 centimeters (cm) in the most recent cohort (born between 1980 and 1994).[48] The standard deviation was 7.59 cm. This means 68% of men were between 170.8 and 186 cm tall; 95% were between 163.2 and 193.6 cm. Women were smaller on average, with a mean height of 164.7 cm, and standard deviation of 7.07 cm. This means 68% of women were between 157.6 and 171.8 cm; and 95% between 150.6 and 178.84 cm.

Regionally, the standard deviation of male heights is largest in North America and Australia, at 7.49 cm, and smallest in East Asia, at 6.37 cm. The pattern is the same for women, with 6.96 cm in North America and Australia, and 5.74 cm in East Asia. Some of the distribution of heights within a population is likely to reflect the degree of genetic variance.[49]



— Roser, Max, Cameron Appel, and Hannah Ritchie. "Human Height." Our World in Data, 2013. https://ourworldindata.org/human-height#height-is-normally-distributed

population

A population consists of the set of all items or attributes of interest. The population may consist of a group of people or some other kind of object.

Let's *simulate* an example population, heights of males and females *in New York City*:

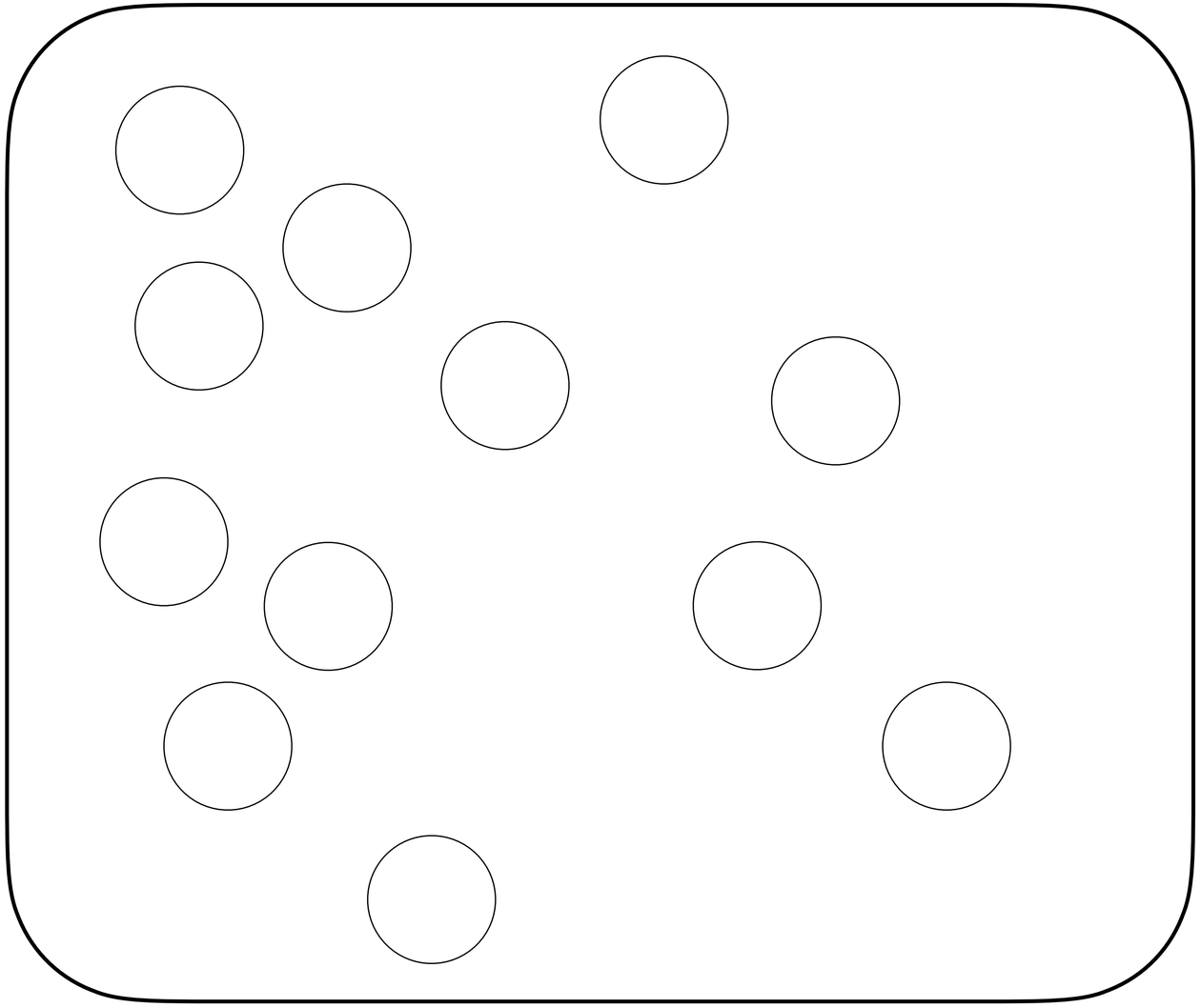
```
library(tidyverse); library(ggthemes)
theme_set( theme_clean() )
set.seed(1)
# (U.S. Census, 2019)
n_nyc     <- 8336817
n_females <- floor(n_nyc * 0.523)
n_males   <- n_nyc - n_females

# (Rosner, 2013)
height_m <- rnorm(n_males,   mean = 178.4, sd = 7.6)
height_f <- rnorm(n_females, mean = 164.7, sd = 7.1)

population_nyc <-
  data.frame(
    height = c(height_m, height_f),
    male   = c(rep(TRUE, n_males), rep(FALSE, n_females))
  )
```

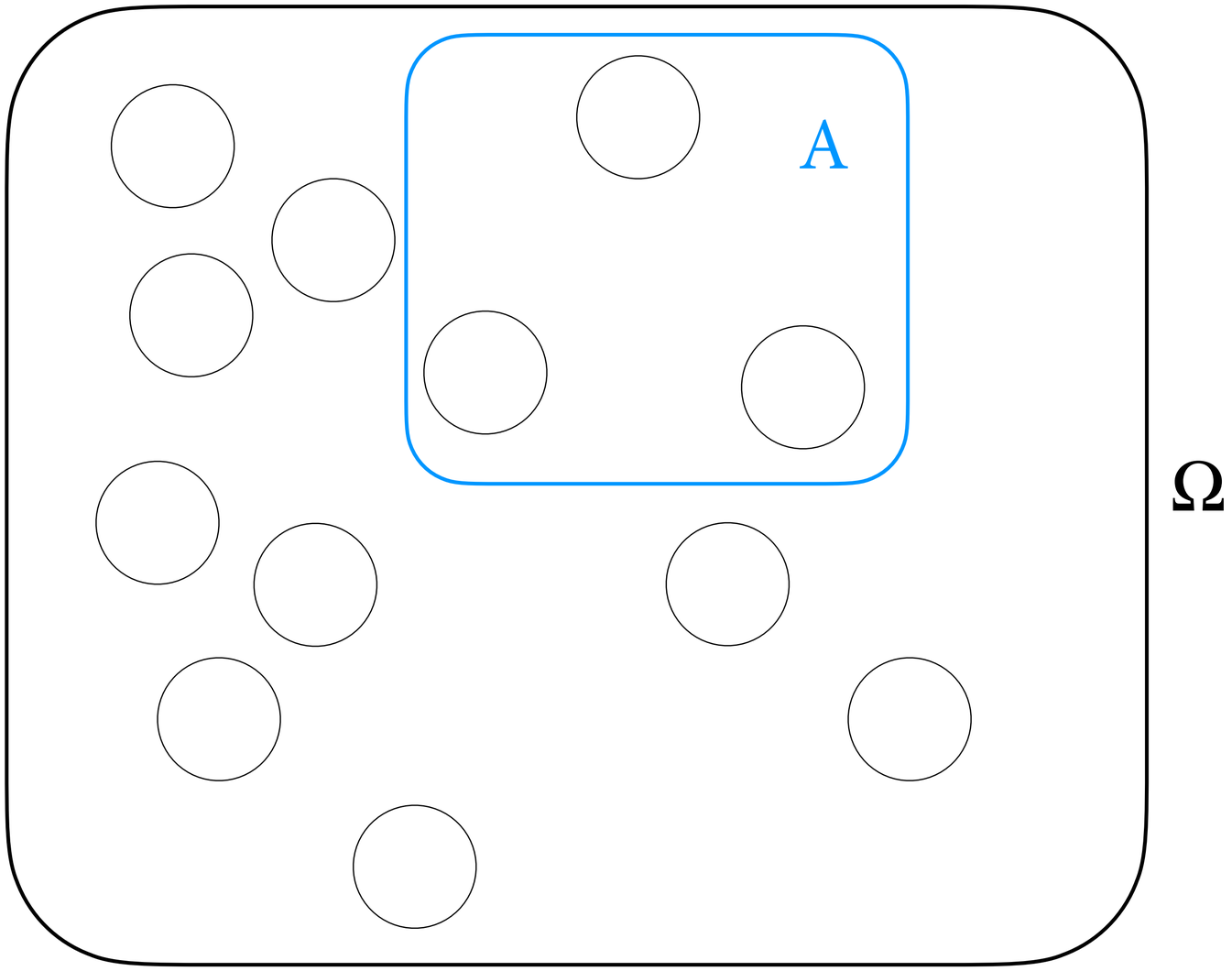
Here are the first and last five simulated observations:

```
|         |   height|male  |
|:-------|--------:|:-----|
|1        | 173.6390|TRUE  |
|2        | 179.7957|TRUE  |
|3        | 172.0492|TRUE  |
|4        | 190.5241|TRUE  |
|5        | 180.9043|TRUE  |
...
|8336813 | 172.3524|FALSE |
|8336814 | 160.6757|FALSE |
|8336815 | 159.3852|FALSE |
|8336816 | 165.0408|FALSE |
|8336817 | 162.7466|FALSE |
```



sample space S or Ω

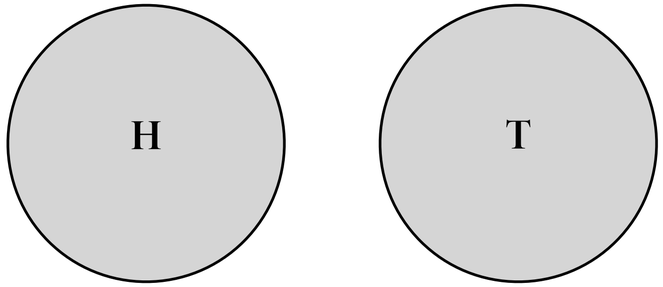
A sample space is the set of all possible outcomes taken from a population for a probability experiment.



event

An event A is a subset of the sample space Ω, and we say that A occurred if the actual outcome is in A.

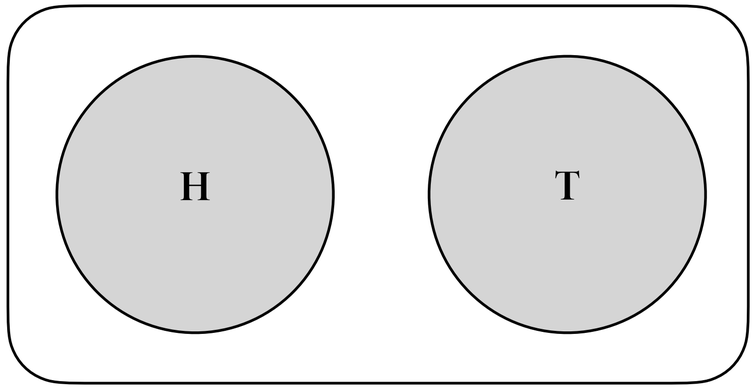
If a coin is flipped twice, what is the event?



And what is the sample space Ω?

If a coin is flipped twice, what is the event?

an “event”



all potential outcomes…

{HH, HT, TH, TT}

And what is the sample space Ω?

What might be Ω for our toy simulation of heights?

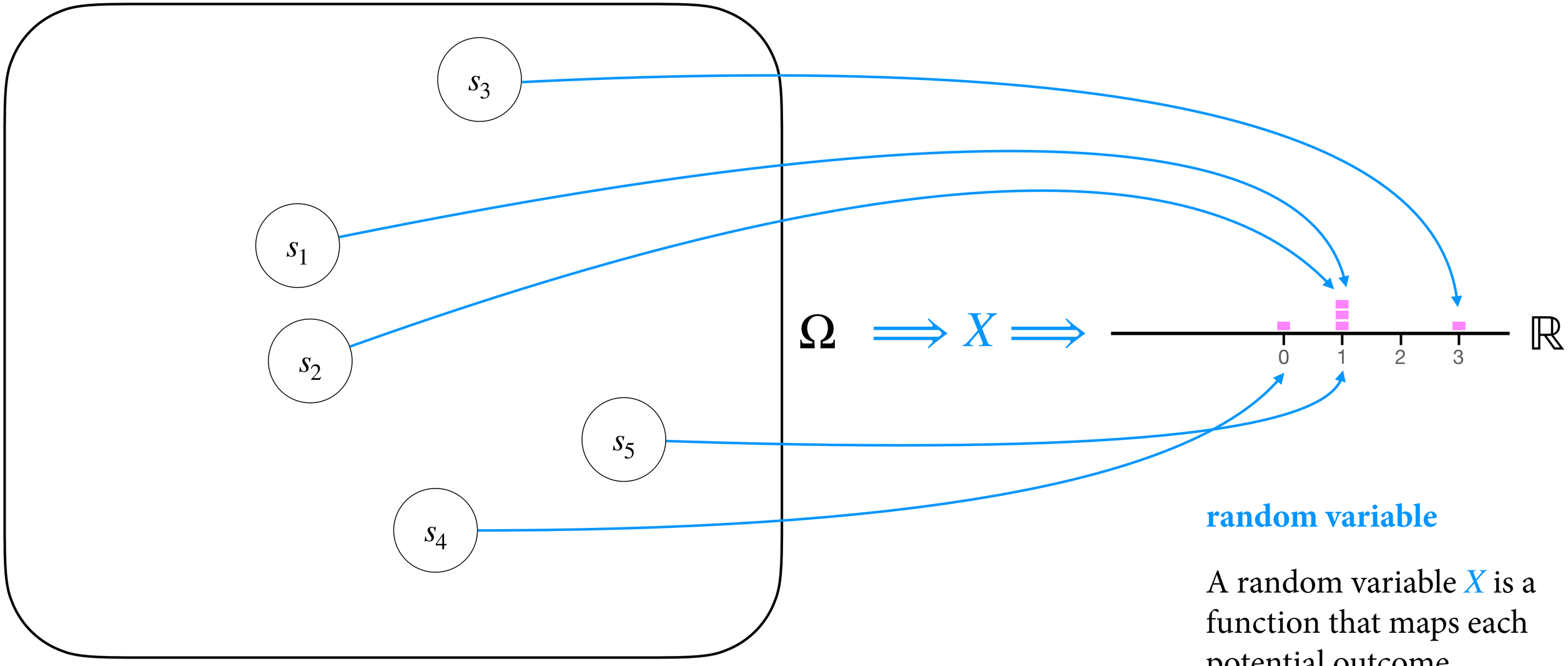
sample

A sample is a group of items chosen from a population. The characteristics of the sample are used to estimate the characteristics of the population. (See sampling…)

Here is *one way* to sample, say, 100 observations from our toy, simulated population of heights.

```
samples <-
  population_nyc %>%
  slice_sample(n = 100, replace = FALSE)
```

How might we ***judge the quality*** of this sample?



random variable

A random variable X is a function that maps each potential outcome $\{s_1, s_2, \ldots, s_5\}$ of the sample space Ω onto the real number line $\mathbb{R}$.

A random variable represents a *distribution* of outcomes $X(s)$ — each its own *probability* of occurring. There are two types of random variables: *discrete* (as shown in this example) and *continuous*, depending on the sample space.

distribution functions

We can use distribution functions to answer questions, like what's the probability that a random variable results in a value or range of values?

One such function for *discrete* sample spaces is called a *probability mass function*:

$$p_x(x) = P(X = x)$$

Denotes an *event*, consisting of all outcomes s to which the random variable X assigns the number x.

$$p_x(x) = P(\text{Two flips} = \text{HH})$$

continuous sample space

Unlike with discrete sample spaces, we cannot always list out the possible values with continuous spaces.

The *continuous* sample space may include any real value, but the probability of an outcome having a specific real value is 0, so we get probabilities differently, using a *probability density function*.

Instead, we get the probability that an outcome has a value within an interval by integrating the function over that interval:

$$P(a < X < b) = \int_a^b f(x)dx$$

Conditional probability and (in)dependence

Let A be our attribute of interest, and B other information. Then we say the probability of A occurring, conditional or given that B has occurred is written in math notation as,

$$P(A \mid B)$$

When,

$$P(A \mid B) = P(A) \text{ and } P(B \mid A) = P(B)$$

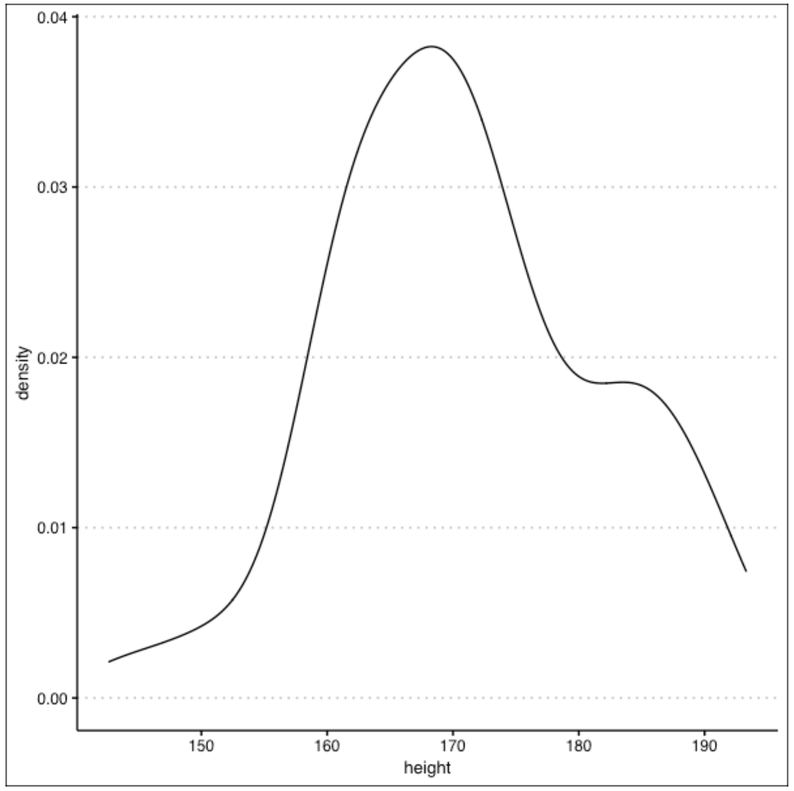
we can say that they are *independent*, one does not depend on the other.

Is our sample height dependent on sex?

Let's graph the marginal distribution of *sample* heights,

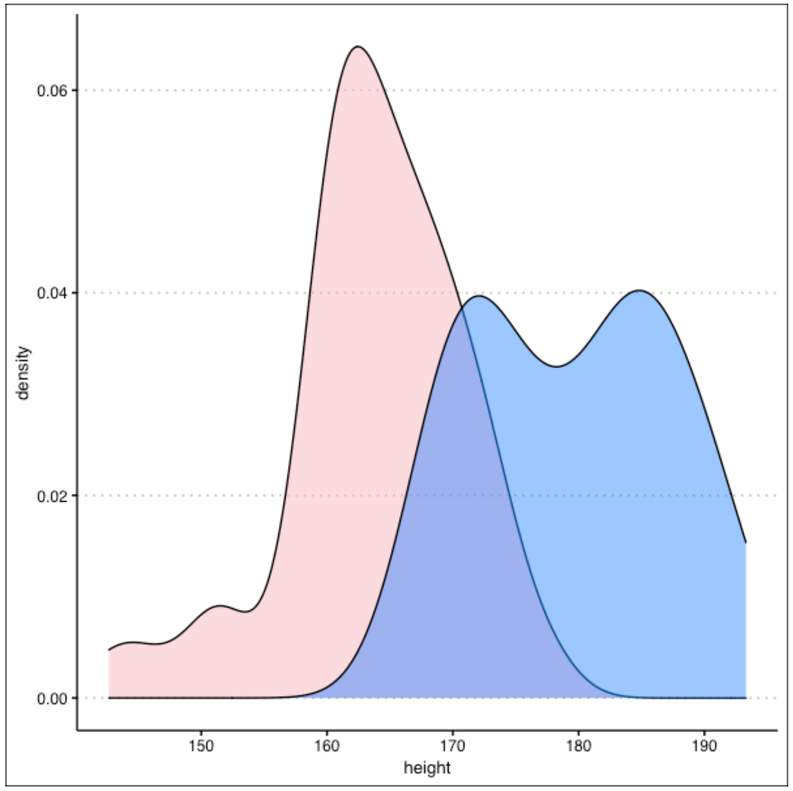
```
library(ggplot2)
library(ggthemes)
theme_set( theme_clean() )

ggplot(samples) +
  geom_density(aes(height))
```



Let's graph the distributions of heights conditional on sex,

```
ggplot(samples) +
  geom_density(aes(x = height,
                   group = male,
                   fill = male),
               alpha = 0.5) +
  scale_fill_manual(
    breaks = c(FALSE, TRUE),
    values = c("lightpink", "dodgerblue")) +
  theme(legend.position = "")
```



How can we read or interpret these? Do these suggest $P(A \mid B) = p(A)$, where A is height, B is sex?

Statistics: sample mean, variance, standard deviation

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$S = \sqrt{S^2}$$

Let's code these statistics for both groups. This,

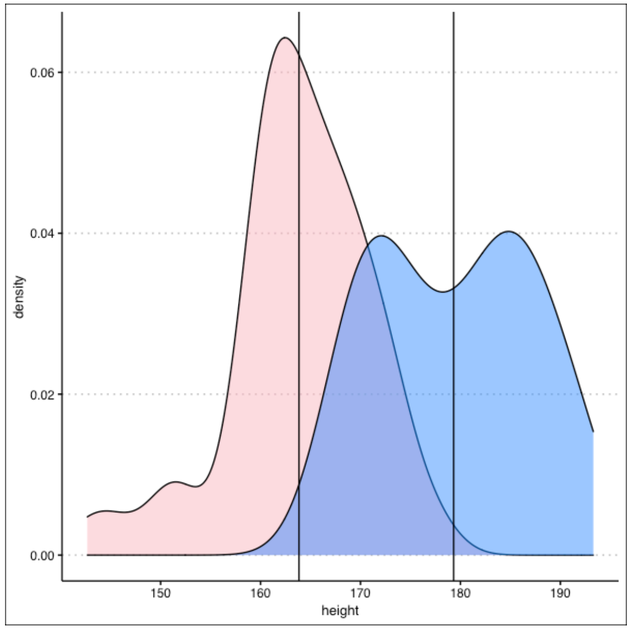
```
samples %>%
  group_by(male) %>%
  summarise(
    x_bar = mean(height),
    var   = var(height),
    sd    = sd(height)
  )
```

returns (in relevant part),

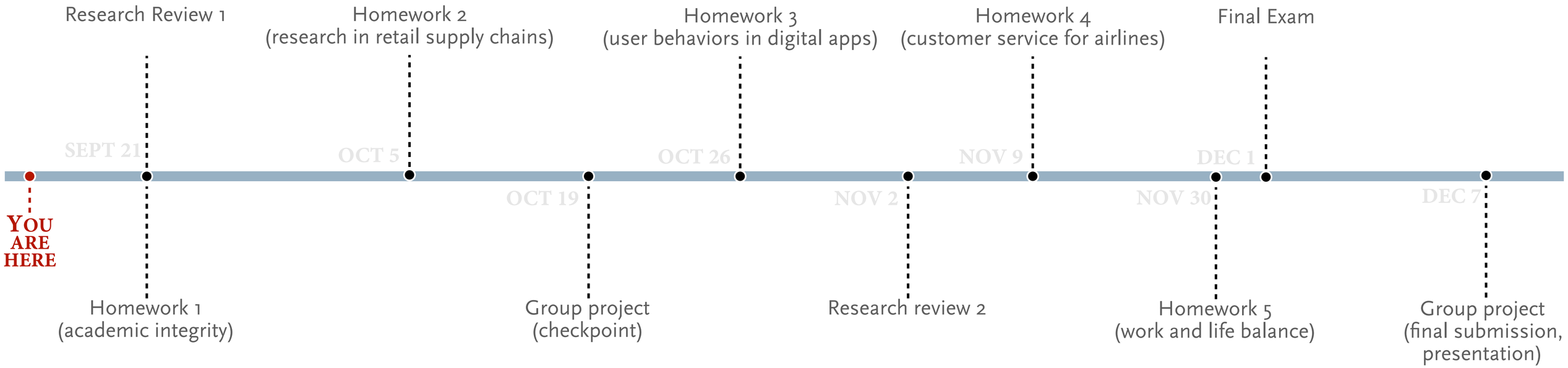
```
|male  |    x_bar|      var|       sd|
|:-----|--------:|--------:|--------:|
|FALSE | 163.8453| 48.84046| 6.988595|
|TRUE  | 179.3198| 61.36701| 7.833710|
```

Let's graph the mean with the conditional distributions:

```
ggplot(samples) +
  geom_density(aes(x = height,
                   group = male,
                   fill = male),
               alpha = 0.5) +
  scale_fill_manual(
    breaks = c(FALSE, TRUE),
    values = c("lightpink", "dodgerblue")) +
  theme(legend.position = "") +
  geom_vline(data = filter(samples, male == FALSE),
             aes(xintercept = mean(height))) +
  geom_vline(data = filter(samples, male == TRUE),
             aes(xintercept = mean(height)))
```



Course deliverables


Research Review 1
Homework 2
(research in retail supply chains)
Homework 3
(user behaviors in digital apps)
Homework 4
(customer service for airlines)
Final Exam
SEPT 21
OCT 5
OCT 26
NOV 9
DEC 1
YOU ARE HERE
OCT 19
NOV 2
NOV 30
DEC 7
Homework 1
(academic integrity)
Group project
(checkpoint)
Research review 2
Homework 5
(work and life balance)
Group project
(final submission, presentation)

References

Baker, Monya. “Is There a Reproducibility Crisis?” Nature 533, no. 26 (May 2016): 452–54.

Blitzstein, Joseph K., and Jessica Hwang. *Introduction to Probability*. Second edition. Boca Raton: Taylor & Francis, 2019.

Booth, Wayne C, Gregory G Columb, Joseph M Williams, Joseph Bizup, and William T Fitzgerald. “14. Incorporating Sources.” In *The Craft of Research*, Fourth. University of Chicago Press, 2016.

Downing, Douglas. *Dictionary of Mathematics Terms* Third Edition. Barron’s, 2009.

Durrett, Richard. *Probability: Theory and Examples*. Fifth edition. Cambridge Series in Statistical and Probabilistic Mathematics 49. Cambridge ; New York, NY: Cambridge University Press, 2019.

Gelman, Andrew. *Ethics and Statistics: Honesty and Transparency Are Not Enough*. CHANCE 30, no. 1 (April 2017): 1–3.

Roser, Max, Cameron Appel, and Hannah Ritchie. “Human Height.” Our World in Data, 2013. https://ourworldindata.org/human-height#height-is-normally-distributed

U.S. Census Bureau QuickFacts - Population estimates, July 1, 2019, (V2019) https://www.census.gov/quickfacts/fact/table/newyorkcitynewyork/PST045219

Wickham, Hadley, and Garrett Grolemund. *R for Data Science*. Online, First. O’Reilly. Accessed September 15, 2021. https://r4ds.had.co.nz/.

Wickham, Hadley, Danielle Navarro, and Thomas Lin. *ggplot2: Elegant Graphics for Data Analysis*. Third. Springer, 2021. https://ggplot2-book.org/ .

Save area

random, n., adv., and adj.

b. *Statistics*. Governed by or involving equal chances for each of the actual or hypothetical members of a population; (also) produced or obtained by a such a process, and therefore unpredictable in detail.

distribution, n.

c. *Statistics*. The way in which a particular measurement or characteristic is spread over the members of a class.

- Oxford English Dictionary.

Academic integrity, a building block for knowledge

A code word for *honesty* and *transparency*?