

# Research Design

## **01: Introduction; Academic Integrity; Concepts of Probability**

# *Meeting your professor*

## **Doctor of 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*

## **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.

## **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

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

A few words on our current mode of course  
discussion: Columbia calls “hy-flex” — and my office hours.

What do we mean by *research design*?

Are these *good* questions:

*What is the difference in height, if any, between male and female graduate students studying in the applied analytics program at Columbia University?*

*What effect might changing our 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 fisheries?*

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

How can we *answer* them? What *assumptions*, if any, might be needed that *limit the scope* of our answers? How can we *explain* our answers to others?

*Academic integrity*, a building block for knowledge



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



Speaking of code, **R** isn't just a letter in an alphabet!?

**Intermission: group hellos!**

*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.

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

```
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, 178.4, 7.6)
height_f <- rnorm(n_females, 164.7, 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

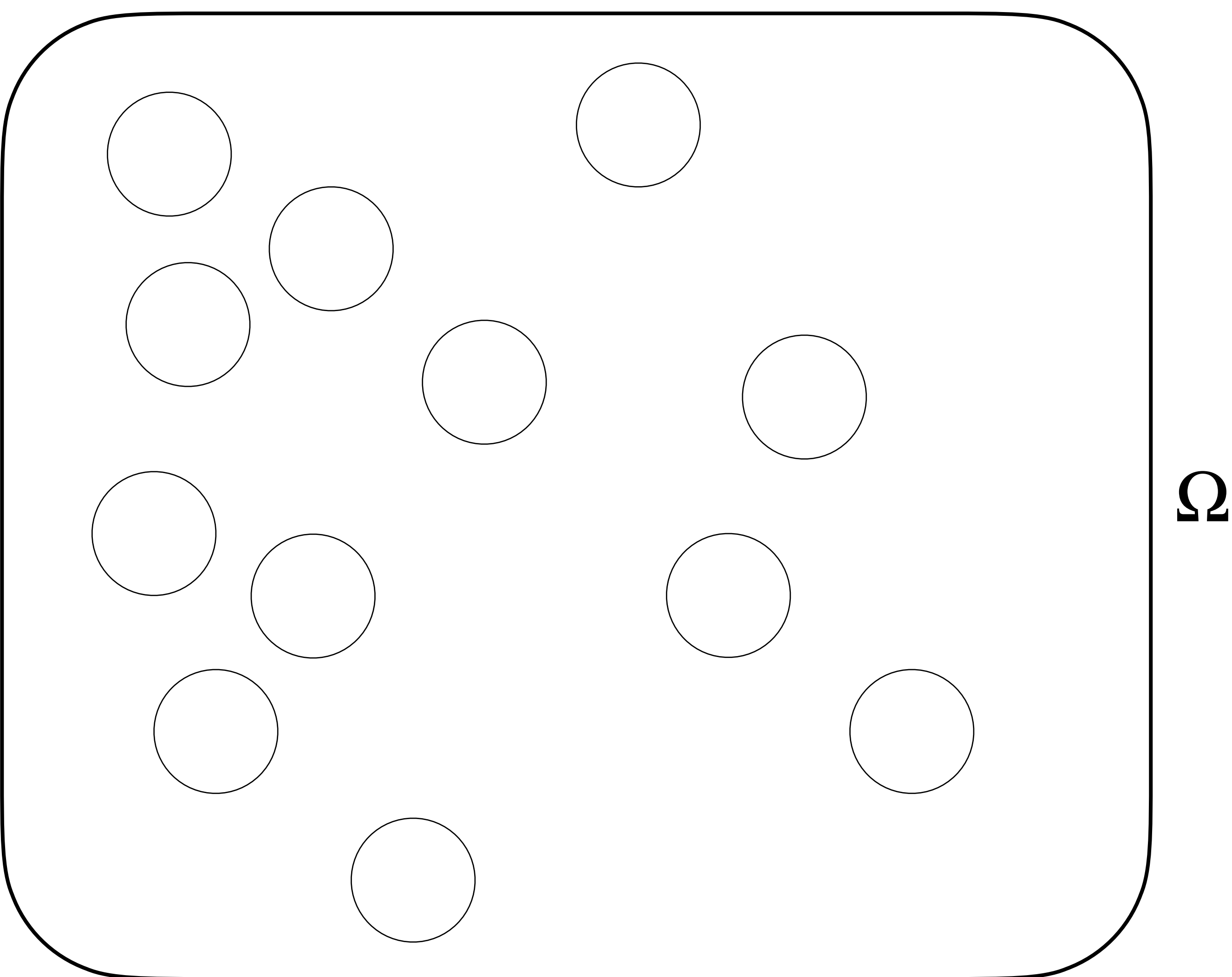
## sample

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.

```
sample_idx <-  
  sample(nrow(population_nyc),  
         size = 100,  
         replace = FALSE)  
  
samples <- population_nyc[sample_idx,]
```

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

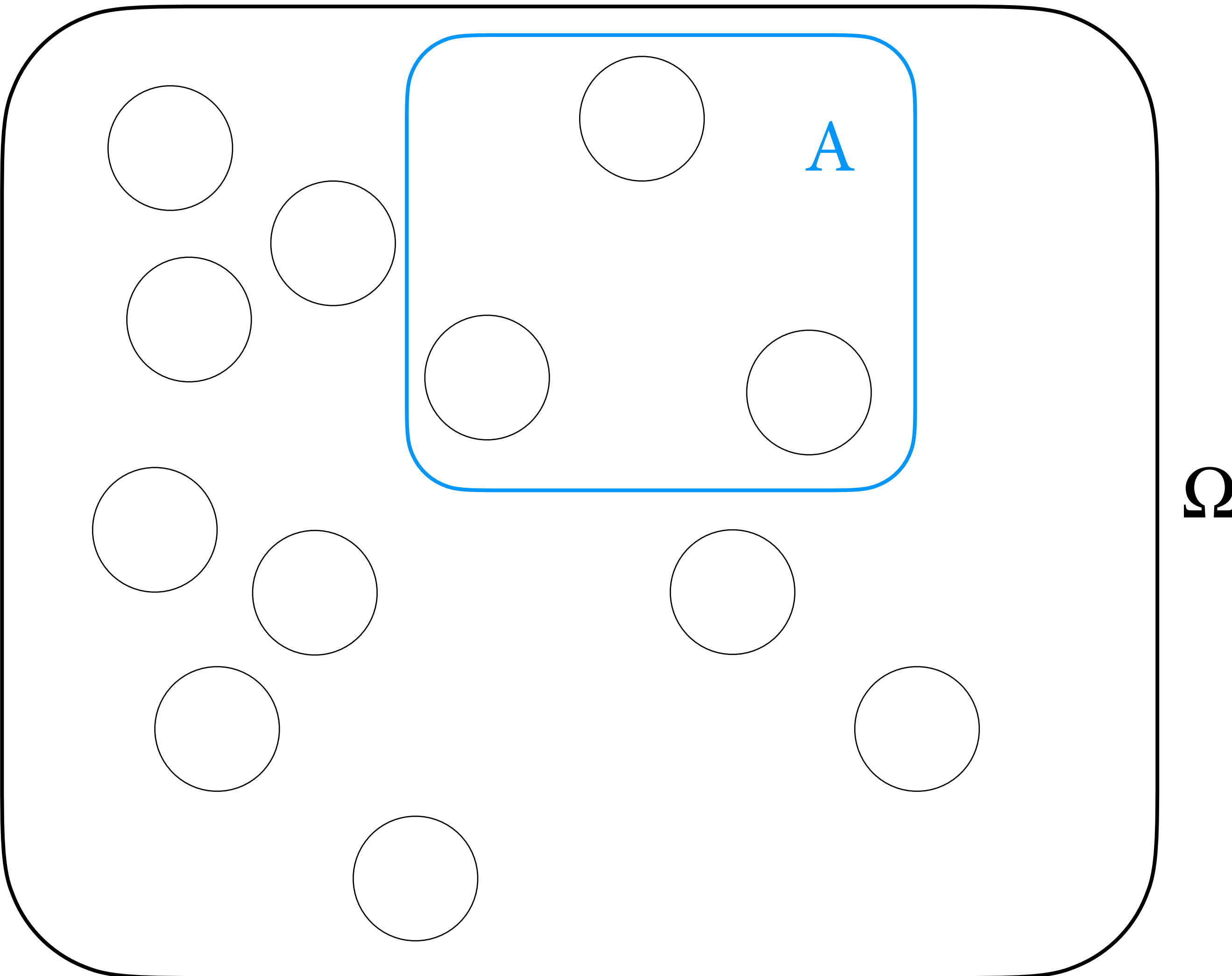


**sample space  $S$  or  $\Omega$**

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

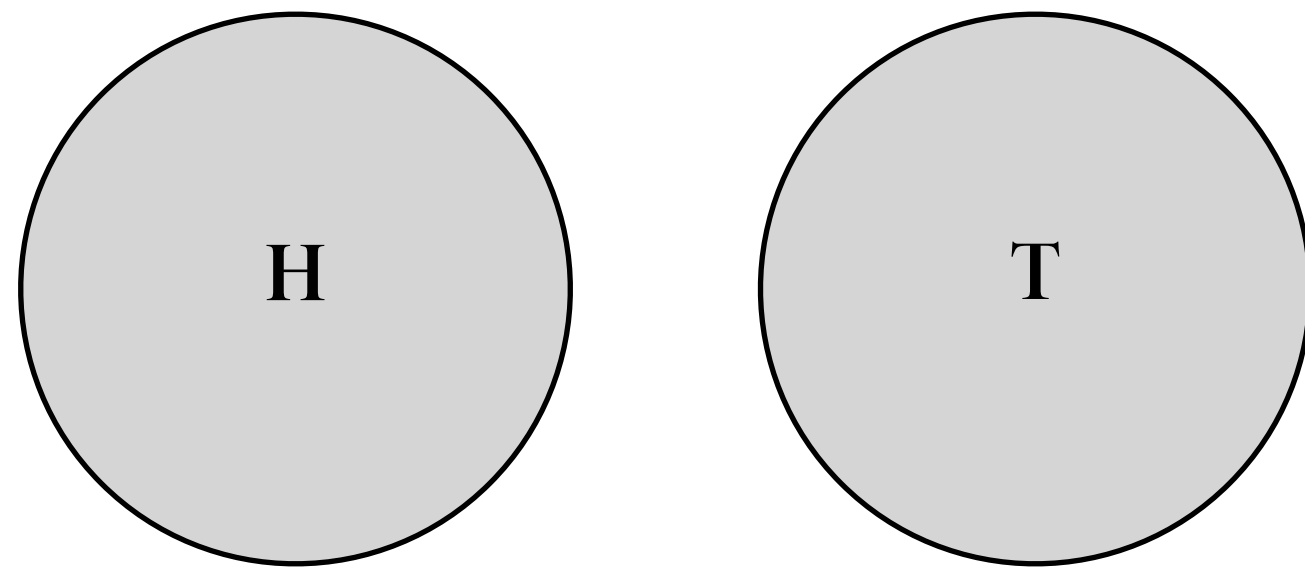
## event

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



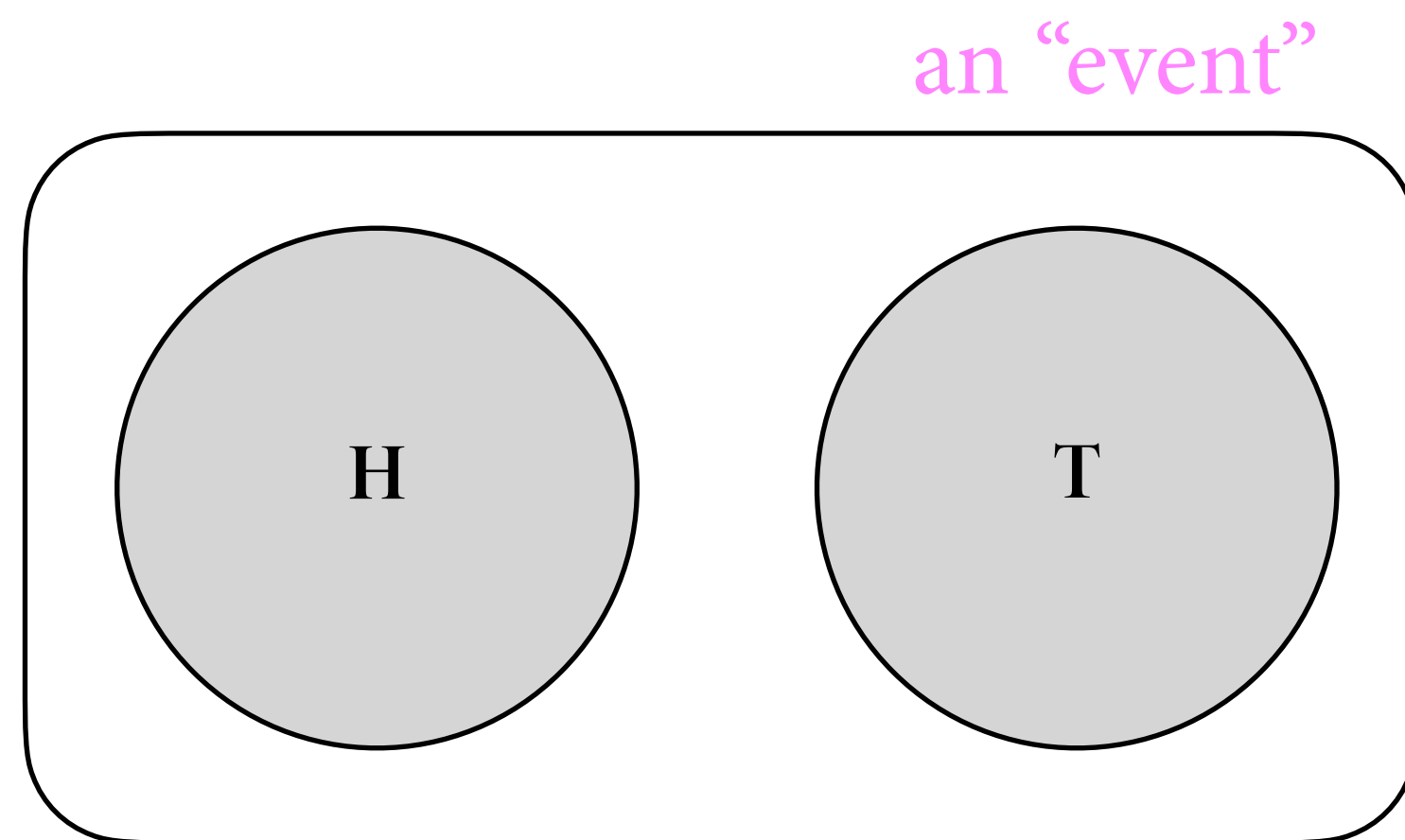


If a coin is flipped twice,



what is the sample space  $\Omega$ ?

If a coin is flipped twice,

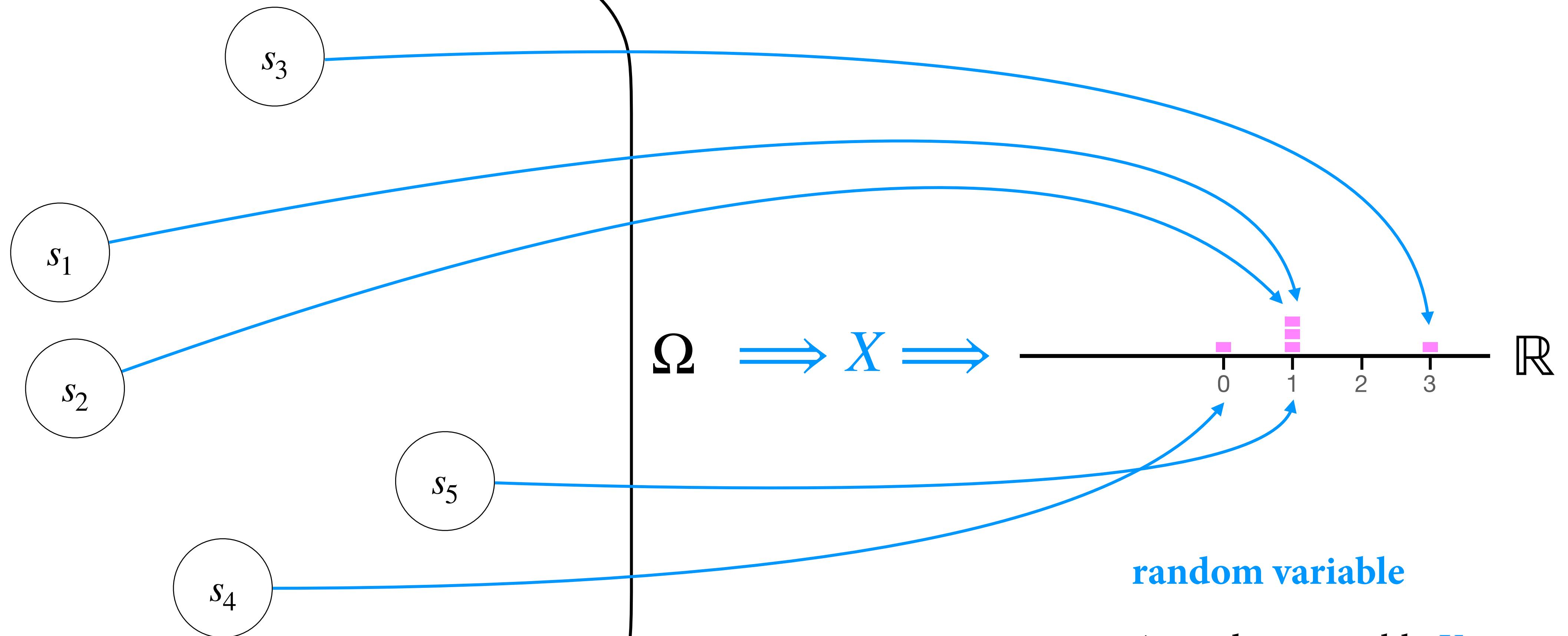


all potential outcomes...

{HH, HT, TH, TT}

what is the sample space  $\Omega$ ?

What might be  $\Omega$  for our toy simulation of heights?



### random variable

A random variable  $X$  is a function that maps each potential outcome  $\{s_1, s_2, \dots, s_5\}$  of the sample space  $\Omega$  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,

$$P(A | B)$$

When,

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

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

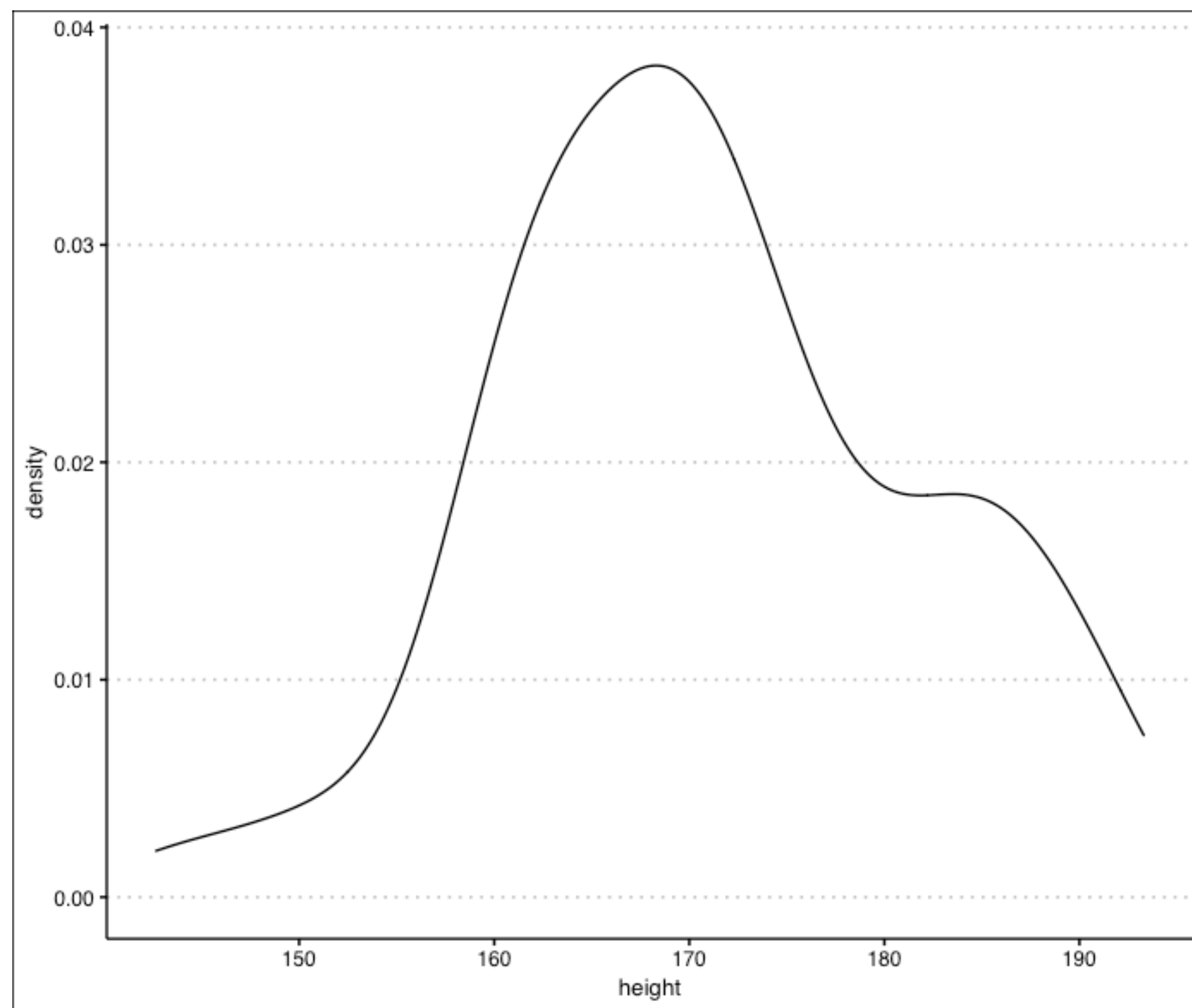
Is our sample height dependent on sex?



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

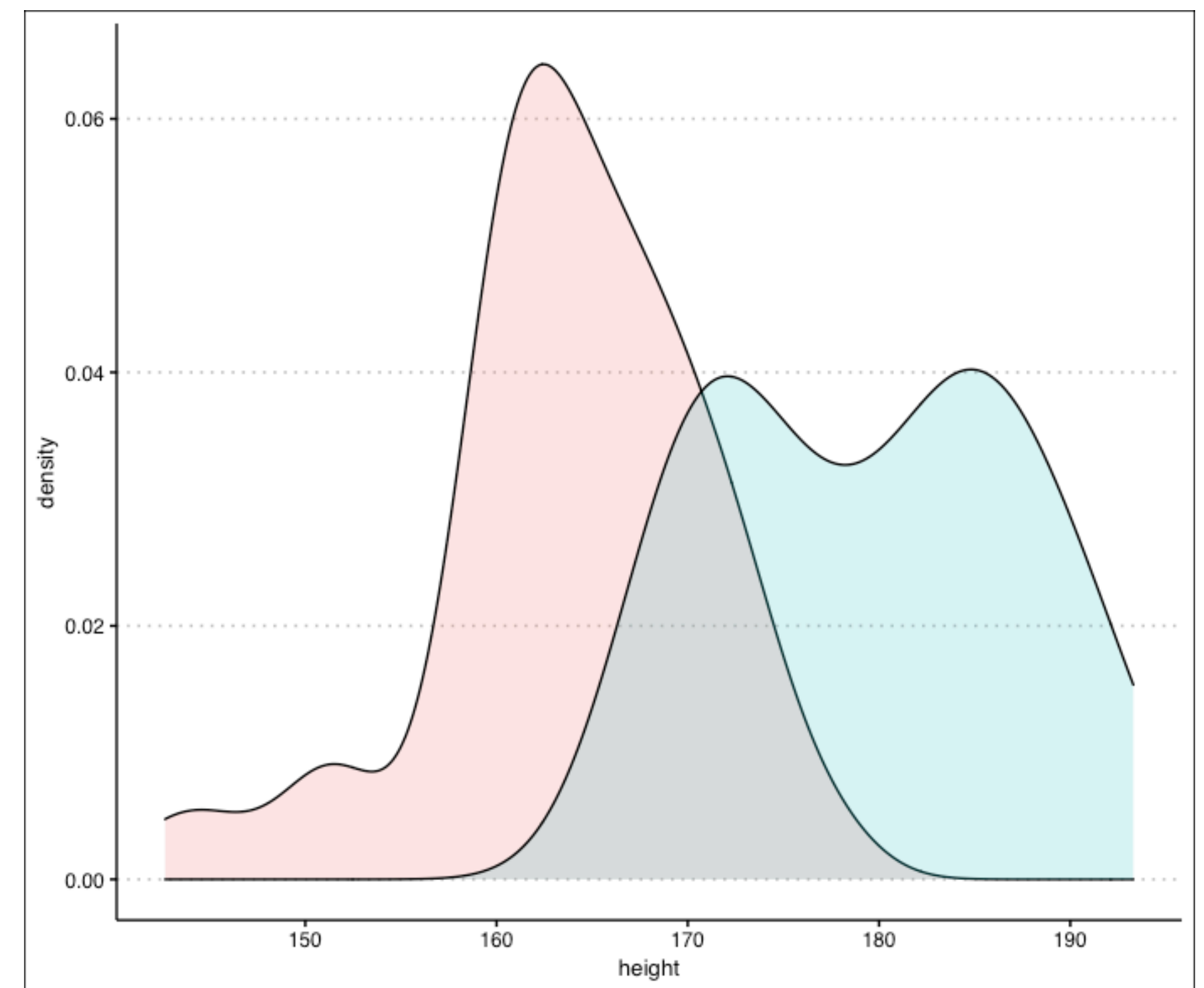
```
library(ggplot2)
library(ggthemes)
theme_set( theme_clean() )

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 = 0.2) +
  scale_color_manual(values = c("pink", "blue")) +
  theme(legend.position = "")
```



How can we read or interpret these? Do these suggest  $P(A | B) = p(A)$  ?

Statistics: sample mean

$$\bar{x} = \frac{\sum_{i=1}^n X_i}{n}$$

Let's code the sample mean for both groups. This,

```
library(dplyr)

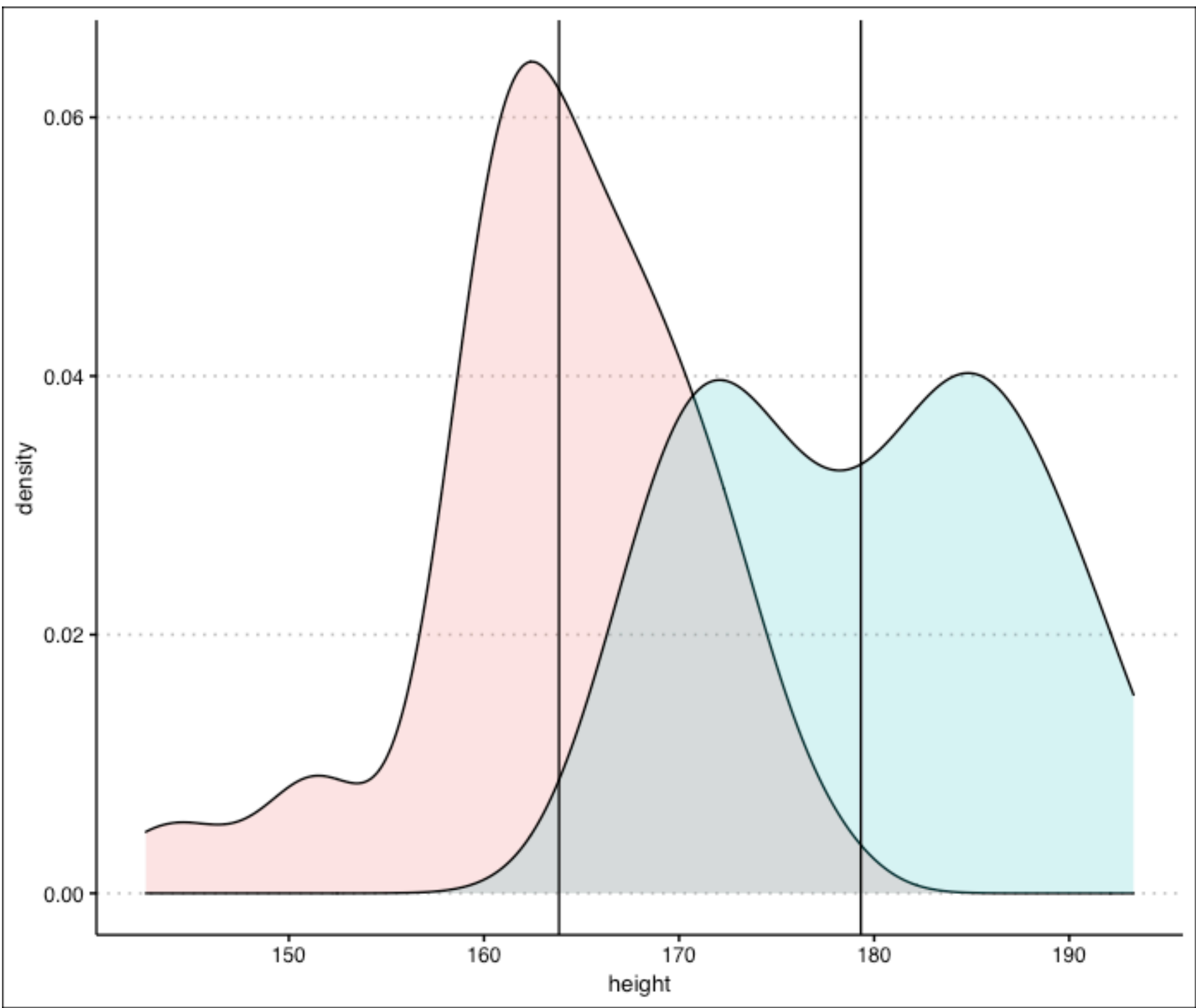
samples %>%
  group_by(male) %>%
  summarise(x_bar = mean(height))
```

returns (in relevant part),

male	x_bar
FALSE	163.8453
TRUE	179.3198

Let's add these to the graph of conditional distributions:

```
ggplot(samples) +
  geom_density(aes(x = height,
                  group = male,
                  fill = male),
              alpha = 0.2) +
  scale_color_manual(values = c("pink", "blue")) +
  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)))
```



# 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>