Lecture 4: Probability

STAT 324

Ralph Trane
University of Wisconsin–Madison

Spring 2020

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Today

1. Statistics vs. Probability
2. Definition: what is "probability"?
3. Examples
4. Properties of Probabilities
5. Concepts
    - independence
    - conditional probabilities

Statistics vs. Probability

Statistics seeks to make inference about properties of a population based on a sample.

Probability takes information about a population, and allows us to make statements about random samples from said population.

Statistics vs. Probability

Example 1:

We are trapped in an old underground chamber. There are five doors. We know that two of the doors lead to chambers with poisonous snakes (and immediate death), while the other three doors will take us straight to freedom.

Probability can help us answer questions like "if we randomly pick a door, what is the probability we will survive?

Example 2:

We all decided to get really into ant farms. We have a big farm with 300 ants. Some of them are poisoneous, but we do not know how many. We take a random sample of 40 ants, and determine 8 of them are poisoneous.

Statistics can help us answer questions like "what percentage of the ants are poisoneous?"

Statistics vs. Probability

Important distinction between the two:

- probability deals with the entire population, hence no uncertainty/randomness
    - can compute the exact probability of survival
- statistics deals with a sample, hence uncertainty/randomness
    - we can only provide a "guess" (called an *estimate*) as to what the true percentage of poisoneous ants is
    - new sample -> new *estimate*

Definitions

Will talk about probabilities in regards to outcomes of a *random process*/an *experiment*.

- we do not know the outcome before performing the experiment
- the *outcome* is the result we observe after performing the experiment
- an *event* is a collection of outcomes.

Example:

Consider the experiment "draw a card at random".

Possible outcomes: "Queen of Spades", "King of Hearts", "Three of Clubs", etc.

An event could be "the card is a King" -- this is the collection of four outcomes: "King of Spades", "King of Hearts", "King of Diamonds", "King of Clubs".

Probability Theory

Generally, two ways of thinking about probability:

- Classical/frequentist interpretation
- Subjective/Bayesian/Degree of belief interpretation

We will solely focus on the former.

Definition of Probability

Two ways of thinking about probability in the frequentist framework. We will mainly use the second, but the first is included for completeness:

1. If all outcomes are equally likely, the *probability* of an event is

$$P(\text{event}) = \frac{\text{number of outcomes in event}}{\text{total number of possible outcomes}}$$

2. The long run proportion of times the event occurs if the experiment is repeated an *infinite number of times*

Example: roll a die

Let $A = \{\text{roll is a 4}\}$, $B = \{\text{roll is a 1 or 5}\}$, and $C = \{\text{roll is even}\}$.

Use definition 1:

$$P(\text{event}) = \frac{\text{number of outcomes in event}}{\text{total number of possible outcomes}}.$$

What is

- $P(A)$, $P(B)$, and $P(C)$?
- $P(A \text{ OR } B)$, and $P(A \text{ OR } C)$?
- $P(B \text{ and } C)$?

Example: Student Survey

Consider the experiment: randomly select a student, and ask "do you like snow?"

Possible outcomes: yes, no, maybe.

What is $P(\text{yes})$?

Using definition 1: $P(\text{yes}) = \frac{1}{3}$. Does that sound right?

Of course not! Why does it fail? All outcomes are NOT equally likely.

Using definition 2: must repeat "an infinite number of times". Alright, let's do it!

Example: Student Survey

We don't really have time for that, so let's simply do it "a very large number of times", and let's get a bit of help from `R`.

Example: Student Survey

```
library(tidyverse); theme_set(theme_bw())
snow_sample <- read_rds("snow_samples.Rds"); survey <- read_csv("../../csv_data/survey
```

Our sample of 10,000:

```
snow_sample %>%
  group_by(`Do you like snow?`) %>%
  summarize(n = n()) %>%
  mutate(Proportion = n / sum(n)) %>%
  knitr::kable(format = "markdown",
               digits = 4)
```

Do you like snow?	n	Proportion
maybe	2341	0.2341
no	1245	0.1245
yes	6414	0.6414

The entire population:

```
survey %>%
  group_by(`Do you like snow?`) %>%
  summarize(n = n()) %>%
  mutate(Proportion = n / sum(n)) %>%
  knitr::kable(format = "markdown",
               digits = 4)
```

Do you like snow?	n	Proportion
maybe	24	0.2353
no	13	0.1275
yes	65	0.6373

Example: Student Survey

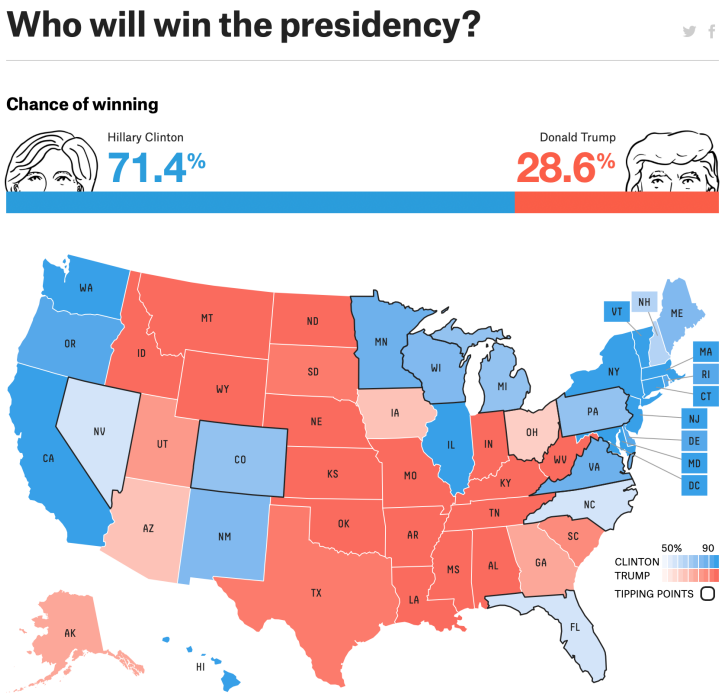
In reality, cannot/would not sample $10,000$ observations from this sample:

1. Too expensive

2. Might as well just ask the entire population, and find the Truth^{TM}.

This is simply to illustrate the idea of "long-run proportion". It is, indeed, an abstract idea, which is impossible to perform in practice, but helps us think about and interpret probabilities.

Example: 2016 Election Prediction



So how do we think about this kind of probability?

I go the "multiple universes" route... Alternatively, Bayesian but we won't go there.

Example: Student Survey

Consider the experiment: randomly select a student, and ask "Are you a Packers fan?"

Possible outcomes: yes, no, maybe. What is $P(\text{yes})$?

Imagine we've asked 20 students, and got these replies:

```
sample_result <- survey %>%
  sample_n(20) %>%
  group_by(`Are you a Packers fan?`) %>%
  summarize(n = n())

sample_result %>%
  knitr::kable(format = "markdown")
```

Are you a Packers fan?	n
maybe	2
no	8
yes	10

How would we estimate $P(\text{yes})$? Simply $\frac{\text{n yes}}{\text{total}} = \frac{10}{20} = 0.5$

Is this the *exact* probability? No, because this comes from a sample, not the entire population.

Example: Student Survey

This is a special case - we actually know the Truth^{TM}, since we have surveyed the entire population. The true probability:

```
survey %>%
  group_by(`Are you a Packers fan?`) %>%
  summarize(n = n()) %>%
  mutate(Probability = n / sum(n)) %>%
  knitr::kable(format = "markdown")
```

Are you a Packers fan?	n	Probability
maybe	10	0.0980392
no	45	0.4411765
yes	46	0.4509804
NA	1	0.0098039

Our estimate is a bit off. How can we do better? Larger sample!

Example: Student Survey

Moral of the story:

- probability = proportion of *entire population*
- "long run proportion" $\approx$ probability
- "infinite run proportion" == probability

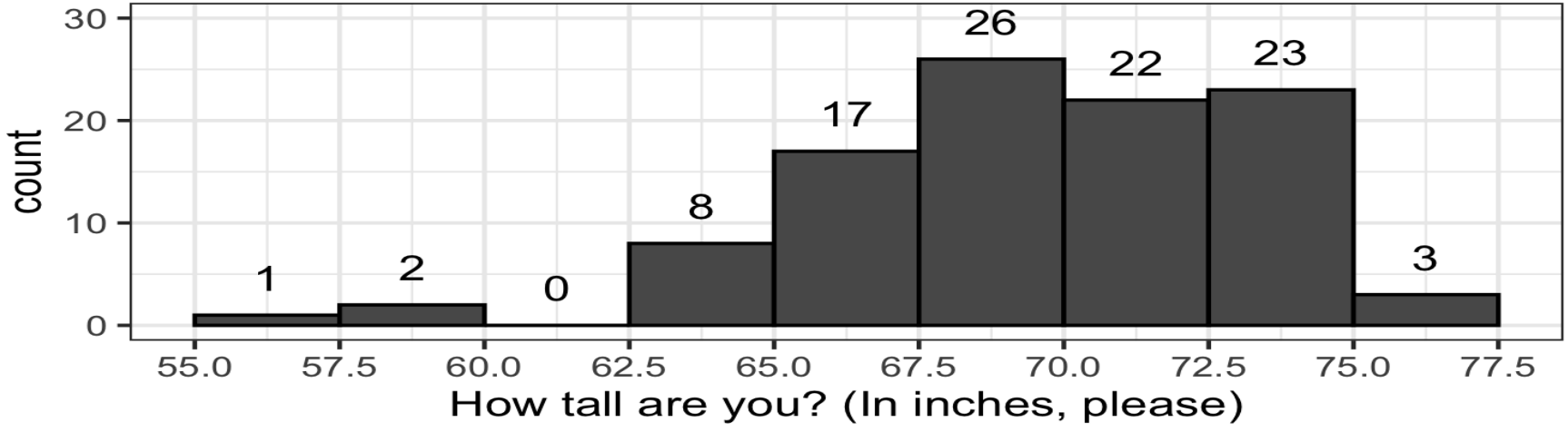
If we wish to *estimate* (i.e. give our "best guess") the probability, repeat experiment many times.

In practice, "number of experiments" is often thought of as "sample size". This is accurate if size of population is infinite. If limited, not quite...

Example: Student Survey

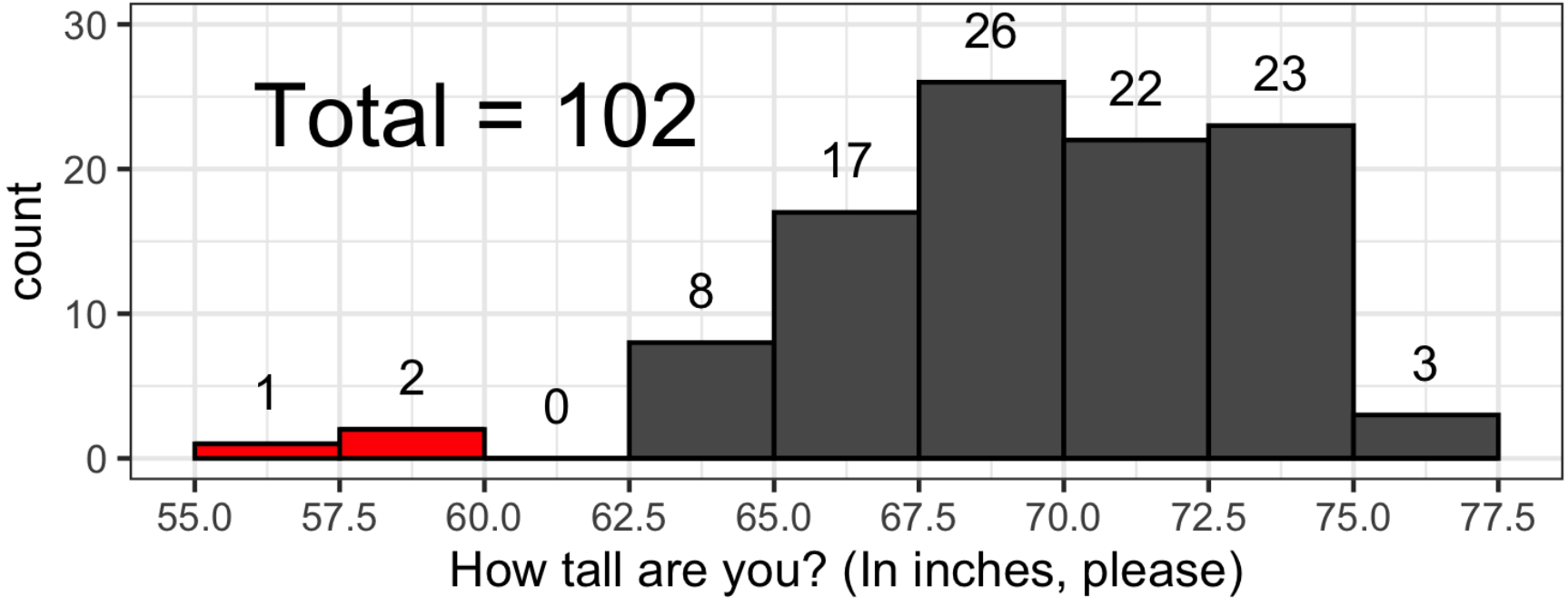
Probabilities from histogram.

```
ggplot(data = survey,
       aes(x = `How tall are you? (In inches, please)`)) +
  geom_histogram(binwidth = 2.5, color = "black",
                 boundary = 55, closed = 'right') +
  geom_text(stat = "bin", binwidth = 2.5, boundary = 55,              ## These two line
            closed = 'right', vjust = -1, aes(label = ..count..)) + ## Arguments binw
  scale_y_continuous(limits = c(0, 30)) +
  scale_x_continuous(breaks = seq(55, 79, by = 2.5))
```



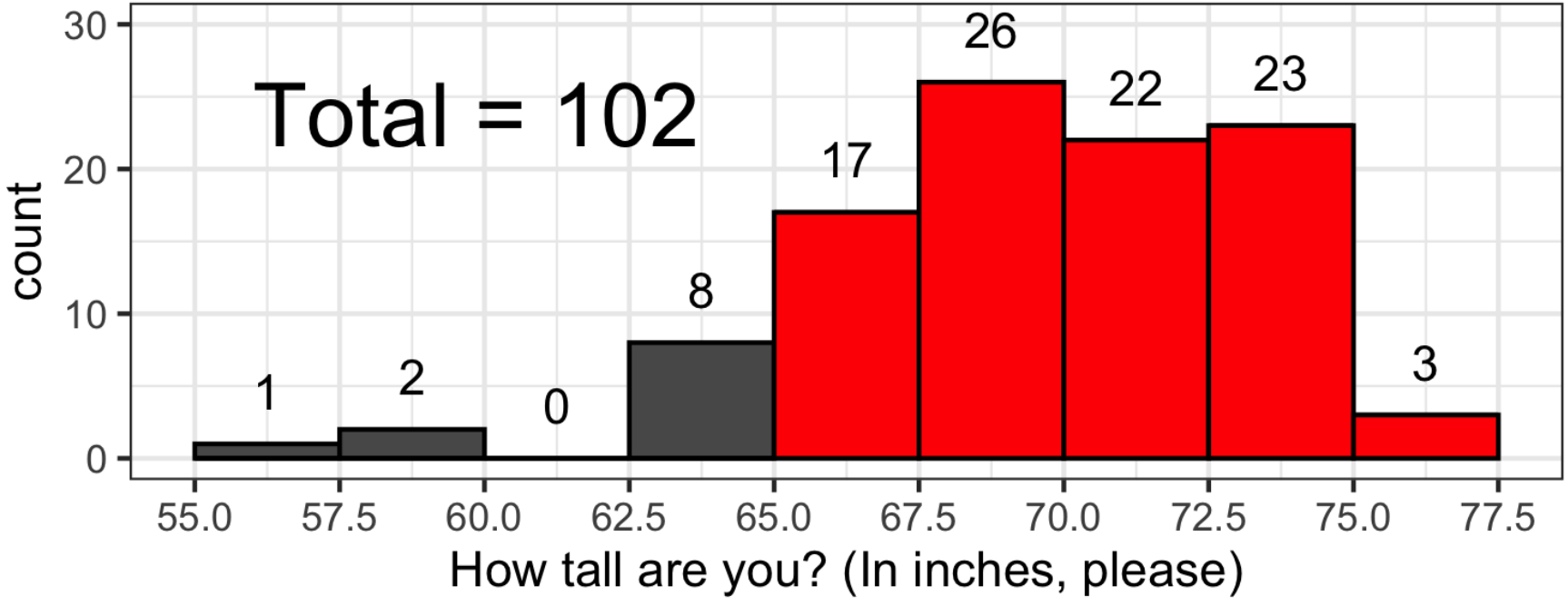
Example: Student Survey

$P(height < 61) = 0.0294$



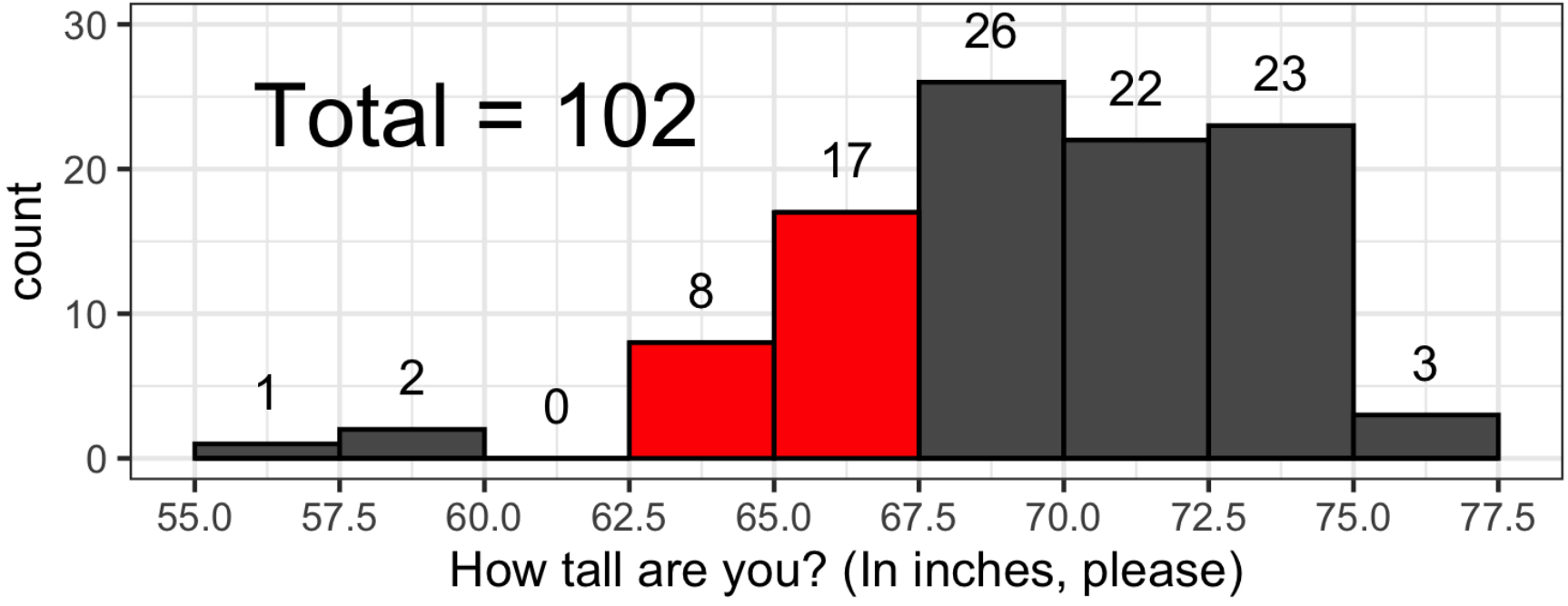
Example: Student Survey

$P(height > 65) = 0.8922$



Example: Student Survey

$P(62 < height \leq 67.5) = 0.2451$



Example: Student Survey

Alternatively

$$\begin{aligned} P(62 < height \leq 67.5) &= P(height \leq 67.5) - P(height < 62) \\ &= 0.2745 - 0.0294 = 0.2451. \end{aligned}$$

Probability Theory

Properties

1. The probability of an **event** is the sum of the probabilities of the outcomes in the event

2. The probability is a number between 0 and 1. **VERY IMPORTANT**

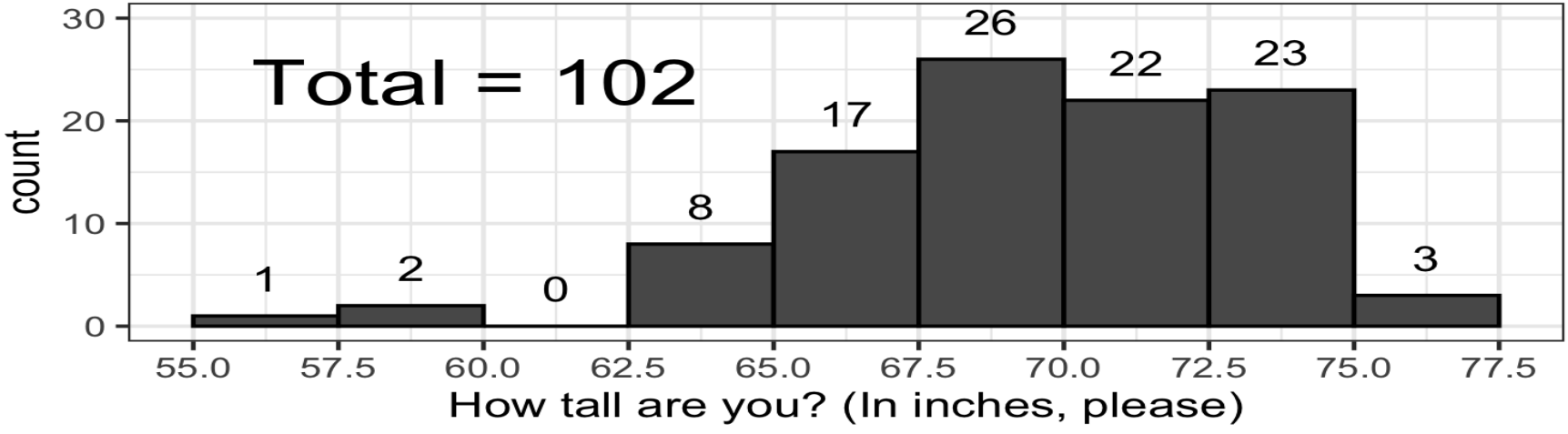
    a. Probability of 0 means it can NEVER happen
    b. Probability of 1 means it ALWAYS happens

3. The sum of the probabilities of all outcomes is 1

    a. The probability of an event A **NOT** happening is $1 - P(A)$. **VERY IMPORTANT**

Example: Student Survey

What is $P(height > 62)$?



$$P(height > 62) = 1 - P(height < 62) = 1 - 0.0294 = 0.9706.$$

Important Concept: Conditional Probability

Recall the experiment: randomly select a student, and ask "Are you a Packers fan?"

Possible outcomes: yes, no, maybe. What is $P(\text{yes})$?

This is a special case - we actually know the Truth^{TM}, since we have surveyed the entire population. The true probability:

```
survey %>%
  group_by(`Are you a Packers fan?`) %>%
  summarize(n = n()) %>%
  mutate(Probability = n / sum(n)) %>%
  knitr::kable(format = "markdown")
```

Are you a Packers fan?	n	Probability
maybe	10	0.0980392
no	45	0.4411765
yes	46	0.4509804
NA	1	0.0098039

Important Concept: Conditional Probability

What if we had a bit more information: the student we just randomly selected grew up in Wisconsin.

```
survey %>%
  filter(`What state did you grow up in?` == "wisconsin") %>%
  group_by(`Are you a Packers fan?`) %>%
  summarize(n = n()) %>%
  mutate(Probability = n/sum(n)) %>%
  knitr::kable(format = "markdown")
```

Are you a Packers fan?	n	Probability
maybe	3	0.0625000
no	5	0.1041667
yes	39	0.8125000
NA	1	0.0208333

Important Concept: Conditional Probability

Or maybe they did not grow up in Wisconsin:

```
survey %>%
  filter(`What state did you grow up in?` != "wisconsin") %>%
  group_by(`Are you a Packers fan?`) %>%
  summarize(n = n()) %>%
  mutate(Probability = n/sum(n)) %>%
  knitr::kable(format = "markdown")
```

Are you a Packers fan?	n	Probability
maybe	6	0.1153846
no	39	0.7500000
yes	7	0.1346154

Important Concept: Conditional Probability

These are examples of "Conditional Probabilities". Notationally, we write $P(A|B)$ for the probability of A given B.

The previous examples:

- $P(\text{Packers fan}|\text{grew up in Wisconsin}) = 0.8125$
- $P(\text{Packers fan}|\text{did NOT grow up in Wisconsin}) = 0.1346154$

Generally, we can calculate these as $P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{\#A \text{ and } B}{\#B}$.

Important Concept: Conditional Probability

What operating system do you use?	Other	Wisconsin	Total
linux	0	1	1
macos	22	9	31
windows	31	37	68
NA	1	1	2
Total	54	48	102

$$P(\text{OS} = \text{macOS} | \text{Grew up in Wisconsin}) = \frac{\#\{\text{OS} = \text{macOS}\} \text{ AND } \{\text{Grew up in Wisconsin}\}}{\#\text{ Grew up in Wisconsin}} = \frac{9}{48} = 0.1875.$$

$$P(\text{Did not grow up in Wisconsin} | \text{OS} = \text{Windows}) = \frac{\#\{\text{Did not grow up in Wisconsin}\} \text{ AND } \{\text{OS} = \text{Windows}\}}{\#\text{OS} = \text{Windows}} = \frac{31}{68} \approx 0.4559.$$

Most Important Concept: Independence

In the "grew up in Wisconsin?" and "Packers fan?" example, knowing the origin of the student drastically changed the probability of getting a "yes" to the question.

When this is the case, we say the two variables are *dependent*.

If this is not the case, we say the two variables are *independent*.

Independence is crucial in many parts of statistics, since it dramatically simplifies the math.

Most Important Concept: Independence

Recall: in practice, "number of experiments" is thought of as "sample size". This is accurate if size of population is infinite. If limited, not quite... Why not?

Consider a deck of cards. Draw a sample of size 20. Is the 20th "experiment" really a repetition of the 1st? No, because the possible outcomes change!

If we have thousands of decks of cards and draw a sample of size 20, then you could argue that the 20th is approximately the same experiment as the 1st.

In real life, never "sample with replacement". So, if population is small, hard to repeat "many times". Generally not a practical problem... unless you are dealing with survey data from a small population.

When we say "repeat the experiment many times", we really mean "repeat the experiment many times independent of each other"! Very, **VERY** important!

Most Important Concept: Independence

Example

Are operating system and origin independent?

$$P(\text{OS} = \text{macOS} | \text{from Wisconsin}) = \frac{9}{48} = 0.1875.$$

$$P(\text{OS} = \text{macOS} | \text{NOT from Wisconsin}) = \frac{22}{54} \approx 0.4074.$$

So, not independent.

Although technically correct, hard to use in practice. Changes are that even truly independent events will seem to be dependent.

Therefore, independence of observations is often an *assumption* we make, and have to defend. To assume independence is a subjective decision.

Most Important Concept: Independence

Remember, in general $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$.

If A and B are independent, then $P(A|B) = P(A)$ and $P(B|A) = P(B)$. In words, "knowing B provides no information about A", and vice versa.

So, if A and B are independent, then

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

So, if A and B are independent: $P(A)P(B) = P(A \text{ and } B)$.

Most Important Concept: Independence

Example

Are the observations independent?

1. A die is rolled. It is a 4. What is the probability the die will be a 4 when rolled again?
    - The two rolls are independent
2. From a standard 52 deck of cards, a card is chosen. It is the 3 of hearts. A second card is chosen. Will the outcome of the second card be independent of the first?
    - No. The probability changes because we do not put the card back in the deck.
3. Randomly pick an NBA player. Ask them to shoot 10 free throws. Are the outcomes of the 10 free throws independent?
    - Maybe... I would argue yes.

Most Important Concept: Independence

Example

Calculating probabilities of independent events:

> A mine safety chamber has a battery operated telephone and a chemical oxygen generator, each of which must work for the chamber to help the miners after an accident. The phone fails 1% of the time and the generator fails 5% of the time, and these failures are independent. What is the probability that the chamber will be helpful after an accident?

$$P(\text{phone works AND generator works}) = P(\text{phone works}) \cdot P(\text{generator works})$$

Since $P(\text{phone works}) = 1 - P(\text{phone fails}) = 1 - 0.01 = 0.99$ and
$P(\text{generator works}) = 1 - P(\text{generator fails}) = 1 - 0.05 = 0.95$,

$$P(\text{phone works} AND \text{generator works}) = 0.99 \cdot 0.95 = 0.9405$$