Lecture 3: Descriptive Statistics

STAT 324

Ralph Trane
University of Wisconsin–Madison

Spring 2020

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Objectives

1. "Previously on STAT 324..."
2. Descriptive Statistics
    - Types of data/variables
    - Numerical summaries
    - Graphical summaries

Previously on STAT 324...

- The general setup
- Three parts of this class:
    1. Descriptive Statistics
    2. Probability
    3. Inferential Statistics

Descriptive Statistics

- What: the art of describing data with few important measures ('summary statistics')
- Why:
  - know your population!!
  - explore your data
- How: try to get an idea of the distributions of variables included
  - what's a distribution?!
  - what's a variable?!

Descriptive Statistics

- Discrete data
  - categorical
    - no natural ordering
    - examples: sex, race, blood type, political orientation, etc.
  - ordinal
    - naturally ordered
    - educational level, age groups, disease severity scales, counts, etc.
  - summarized by
    - frequency counts
    - relative frequencies

- Continuous data
  - numerical
    - examples: age, height, weight, BMI, proportions, etc.
  - infinite (uncountable, actually...) number of potential values
  - summarized by
    - location measures
    - spread/variation measures
    - shape

Descriptive Statistics

Data example: STAR WARS

First, some R setup:

```
library(tidyverse)
library(DT); library(kableExtra)

# Change theme for plots
theme_set(theme_bw())

# Change color scheme for plots
scale_color_continuous <- scale_color_viridis_c
scale_color_discrete <- scale_color_viridis_d

scale_fill_continuous <- scale_fill_viridis_c
scale_fill_discrete <- scale_fill_viridis_d
```

Descriptive Statistics

Data example: STAR WARS

Let's load some data, and take a look:

```
starwars <- read_csv("../../csv_data/starwars.csv")

## datatable and formatStyle are from the DT package
datatable(starwars, options = list(pageLength = 4, lengthChange = FALSE,
                                   scrollX = TRUE, dom = "tp")) %>%
  formatStyle(columns = 1:ncol(starwars), fontSize = "12pt")
```

	name	height	mass	hair_color	skin_color	eye_color	birth_year	ge
1	Luke Skywalker	172	77	blond	fair	blue	19	male
2	C-3PO	167	75		gold	yellow	112	
3	R2-D2	96	32		white, blue	red	33	
4	Darth Vader	202	136	none	white	yellow	41.9	male

Previous 1 2 3 4 5 … 22 Next

Descriptive Statistics

Data example: STAR WARS

How many films have the characters been in?

Ordinal variable. Summarized using frequency counts, and relative frequencies.

Numerical summary:

```
n_films <- starwars %>%
  group_by(n_films) %>%
  summarize(frequency = n()) %>%
  mutate(relative = frequency / sum(freq

kable(n_films,
      format = "html") %>%
  kable_styling(font_size = 12)
# kable_styling is from the kableExtra p
```

n_films	frequency	relative
1	46	0.5287356
2	18	0.2068966
3	13	0.1494253
4	2	0.0229885
5	5	0.0574713
6	2	0.0229885
7	1	0.0114943

Descriptive Statistics

Data example: STAR WARS

How many films have the characters been in?

Ordinal variable. Summarized using frequency counts, and relative frequencies.

Graphical summary:

```
ggplot(data = starwars,
       aes(x = n_films)) +
  geom_bar() +
  labs(x = "Number of films",
       y = "Frequencies")
```

Descriptive Statistics

Data example: STAR WARS

How many films have the characters been in?

Ordinal variable. Summarized using frequency counts, and relative frequencies.

Graphical summary:

```
ggplot(data = starwars,
       aes(x = n_films)) +
  geom_bar(aes(y = ..count../sum(..count
  labs(x = "Number of films",
       y = "Relative frequencies")
```

Descriptive Statistics

Data example: STAR WARS

When would we use relative frequencies rather than frequencies? When comparing groups.

```
n_films_gender <- starwars %>%
  group_by(gender, n_films) %>%
  summarize(frequencies = n()) %>%
  mutate(relative = frequencies / sum(fr

kable(n_films_gender,
             format = "html") %>%
  kable_styling(font_size = 10)
```

gender	n_films	frequencies	relative
female	1	9	0.4736842
female	2	6	0.3157895
female	3	3	0.1578947
female	5	1	0.0526316
hermaphrodite	3	1	1.0000000
male	1	34	0.5483871
male	2	12	0.1935484
male	3	9	0.1451613
male	4	2	0.0322581
male	5	4	0.0645161
male	6	1	0.0161290
none	1	2	1.0000000
NA	1	1	0.3333333
NA	6	1	0.3333333
NA	7	1	0.3333333

Data example: STAR WARS

When would we use relative frequencies rather than frequencies? When comparing groups.

```
ggplot(data = n_films_gender,
       aes(x = n_films, y = frequencies, fill = gender)) +
  geom_bar(stat = "identity") +
  facet_grid(~gender) +
  labs(x = "Number of films")
```

Descriptive Statistics

Data example: STAR WARS

When would we use relative frequencies rather than frequencies? When comparing groups.

```
ggplot(data = n_films_gender,
       aes(x = n_films, y = relative, fill = gender)) +
  geom_bar(stat = "identity") +
  facet_grid(~gender) +
  labs(x = "Number of films")
```

Descriptive Statistics

For continuous data, measures of location and spread

Measures of location:

- mean/average: $\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$
- median: the observation in the middle of the data.
- minimum and maximum: the smallest and largest observations, respectively

Simple example data: $1, 5, 2, -3, 7, -12, 0$.

Mean: 0

Median: 1

Minimum and maximum: -12 and 7

Descriptive Statistics

Measures of spread:

- variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- standard deviation: $s = \sqrt{s^2}$
- range: $\max - \min$
- percentiles (sometimes referred to as quantiles):

> The *p'th percentile* of a set of measurements is the one that has $p\%$ of the data below it.

- Inter Quartile Range (IQR): 75th percentile - 25th percentile
  - 25th percentile = first quartile = median of 1st half including median
  - median = second quartile
  - 75th percentile = third quartile = median of 2nd half including median

Simple example data: $1, 5, 2, -3, 7, -12, 0$.

Variance: 38.6666667. Standard deviation: 6.2182527. Range: 19.
1st and 3rd quartiles (i.e. 25th and 75th percentiles): -1.5, 3.5. IQR: 5

Why don't we use these summaries for discrete data? Doesn't make sense for categorical data. Could maybe use these in some cases when working with ordinal data, although interpretation might be hard.

Descriptive Statistics

Data example: STAR WARS

```
starwars %>%
  filter(!is.na(height), gender %in% c('male', 'female')) %>%
  group_by(gender) %>%
  summarize(n = n(),
            mean = mean(height),
            var = var(height),
            sd = sd(height),
            median = median(height),
            Q1 = quantile(height, p = 0.25),
            Q2 = quantile(height, p = 0.50),
            Q3 = quantile(height, p = 0.75),
            IQR = IQR(height),
            min = min(height),
            max = max(height),
            range = max - min) %>%
  kable(format = "markdown", digits = 2)
```

Descriptive Statistics

Data example: STAR WARS

gender	n	mean	var	sd	median	Q1	Q2	Q3	IQR	min	max	range
female	17	165.47	530.39	23.03	166	163	166	178	15	96	213	117
male	59	179.24	1252.56	35.39	183	174	183	193	19	66	264	198

Descriptive Statistics

Data example: STAR WARS

```
starwars %>%
  filter(!is.na(height), gender %in% c('male', 'female')) %>%
  group_by(gender) %>%
  summarize(n = n(),
            '10th percentile' = quantile(height, p = 0.1),
            '57th percentile' = quantile(height, p = 0.57),
            '82nd percentile' = quantile(height, p = 0.82)) %>%
  kable(format = "markdown", digits = 2)
```

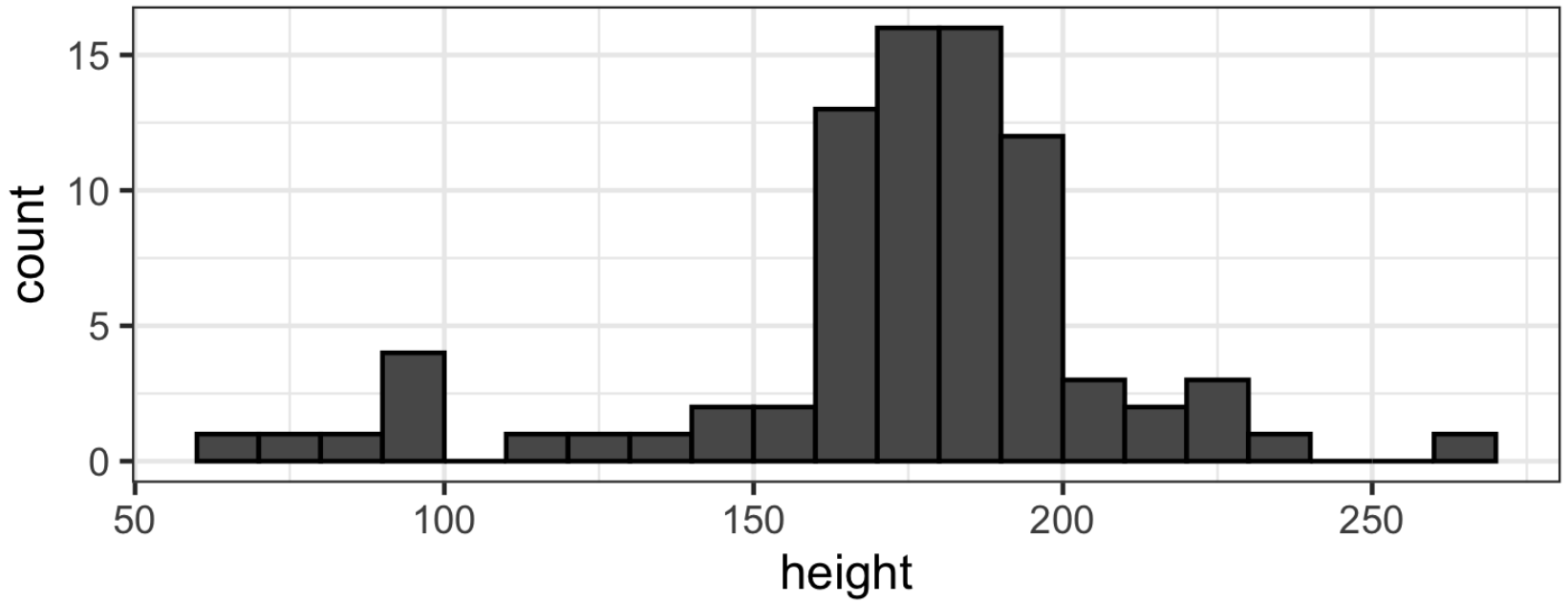
gender	n	10th percentile	57th percentile	82nd percentile
female	17	150	167.12	178.00
male	59	134	188.00	197.12

Descriptive Statistics

Data example: STAR WARS

Visually, continous data can be summarized using a *histogram*.

```
ggplot(data = starwars,
       aes(x = height)) +
  geom_histogram(binwidth = 10, boundary = 60,
                 na.rm = TRUE, color = 'black')
```

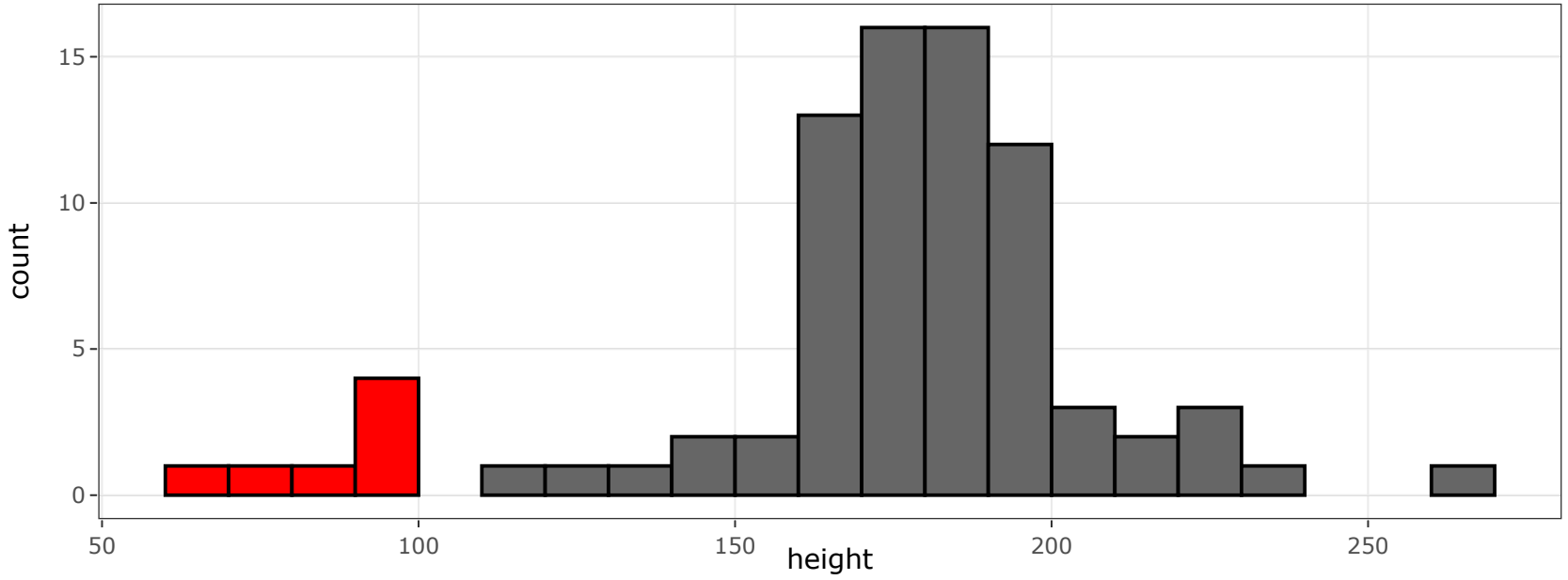


Descriptive Statistics

Data example: STAR WARS

Proportion of an area is the proportion of data in the corresponding interval: proportion of characters with height less than $100 = \frac{7 \cdot 10}{81 \cdot 10} \approx 0.086$.

This will be **very** important to us later.



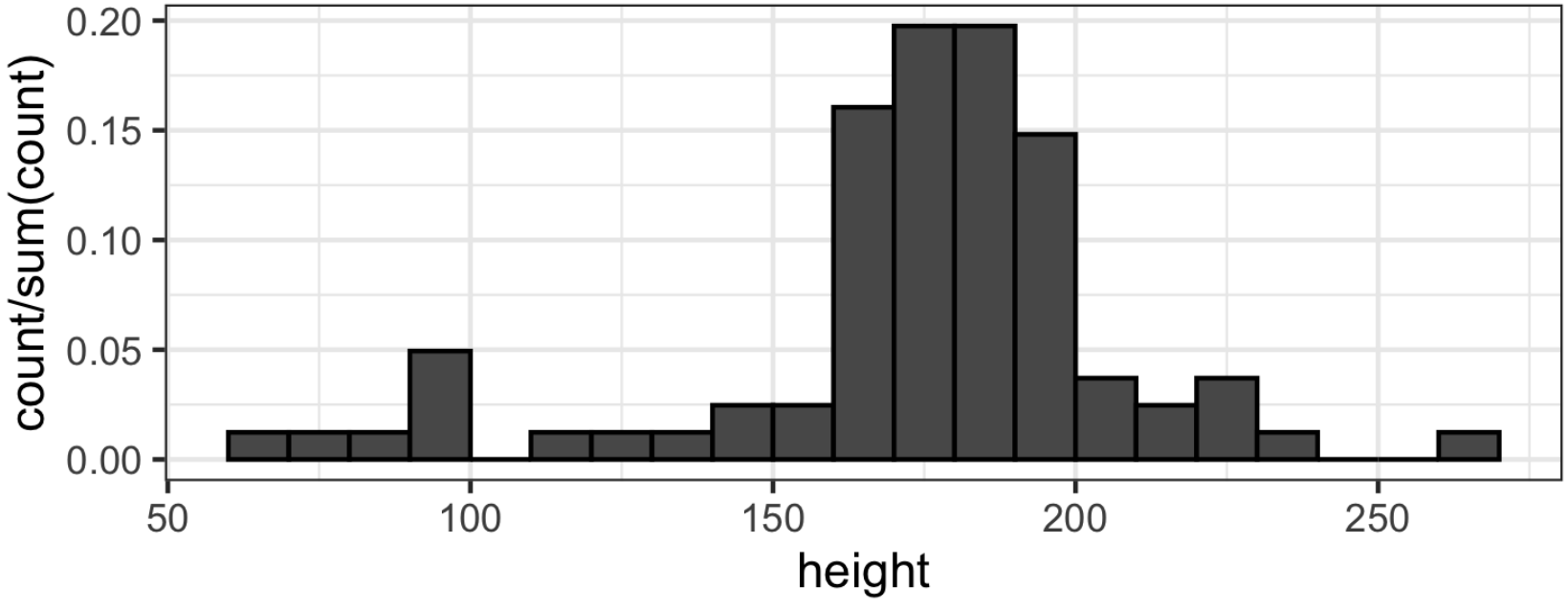
Descriptive Statistics

Data example: STAR WARS

Descriptive Statistics

Data example: STAR WARS

Histograms can also be used to depict *relative frequencies*. To do so, you simply divide the counts by the total number of observations.



Data example: STAR WARS

Another useful figure for numerical data: the boxplot (also known as box-and-whisker plot, or box-and-whisker diagram).

The key to decipher a box plot:

Outlier if

$$\text{obs} < \text{Q1} - 1.5 \cdot \text{IQR}$$

or

$$\text{obs} > \text{Q3} + 1.5 \cdot \text{IQR}$$

25% of the data are above the box

50% of the data are in the the box

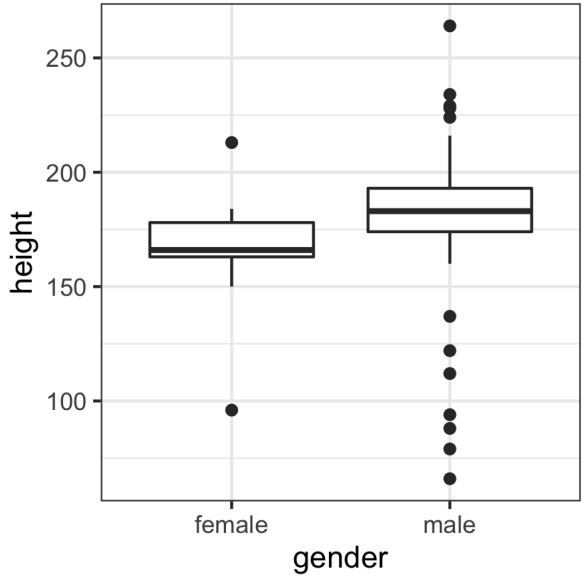
25% of the data are below the box

Descriptive Statistics

Data example: STAR WARS

```
boxplots <- ggplot(data = starwars %>% filter(gender %in% c('male', 'female')),
       aes(x = gender, y = height)) +
  geom_boxplot()

boxplots
```

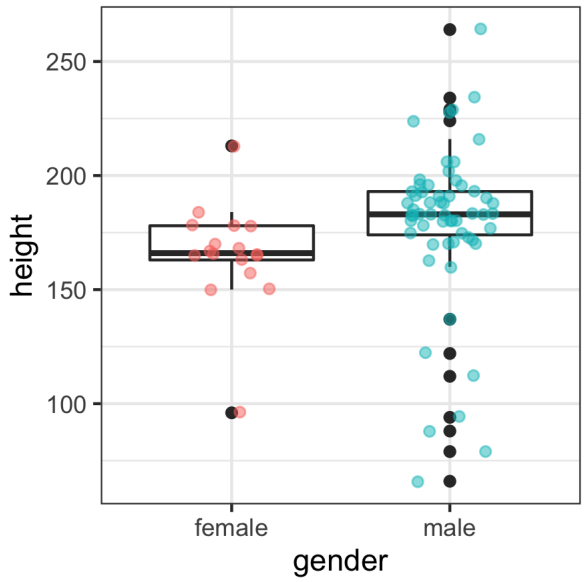


Descriptive Statistics

Data example: STAR WARS

My personal favorite when data set not too large: boxplot + points!

```
boxplots +
  geom_point(aes(color = gender), alpha = 0.5,
             position = position_jitter(width = 0.2)) +
  guides(color = "none")
```

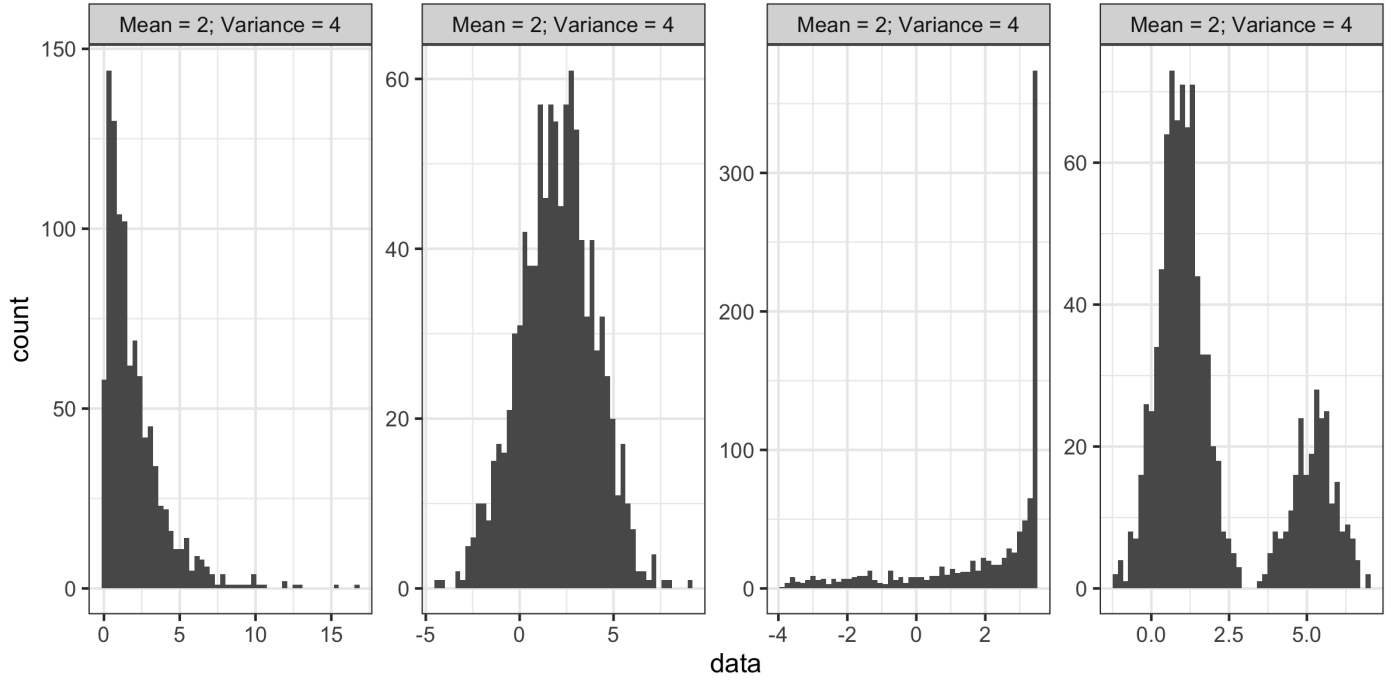


Descriptive Statistics

Continuous Data: shape

So far, only talked about *location* and *spread* of data.

However, this does not provide the full picture:

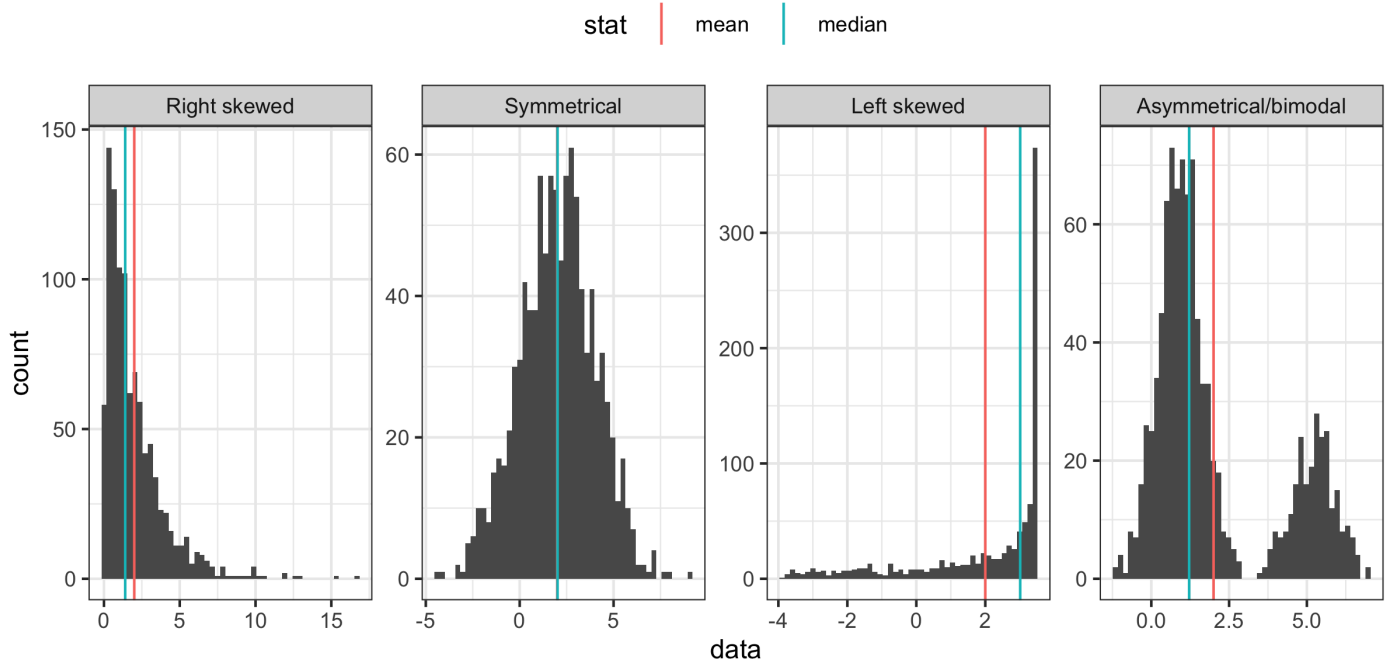


Descriptive Statistics

Continuous Data: shape

The *shape* of the data is generally described as either *symmetrical* or *non-symmetrical*, and if non-symmetrical as either *right skewed* or *left-skewed*.

When symmetrical, mean = median. Right skewed, mean > median. Left skewed, mean < median.



Descriptive Statistics

Continuous Data: shape

Boxplots don't give us the entire shape, but they let us determine if the data is symmetrical:

For symmetrical data

- median is in the middle of the data.
- median = mean.

Right skewed data => median closer to bottom of box.

Left skewed data => median closer to top of box.