EDS 212: Day 4, Lecture 2

Essential summary statistics and exploration

August 8th, 2024

Data types

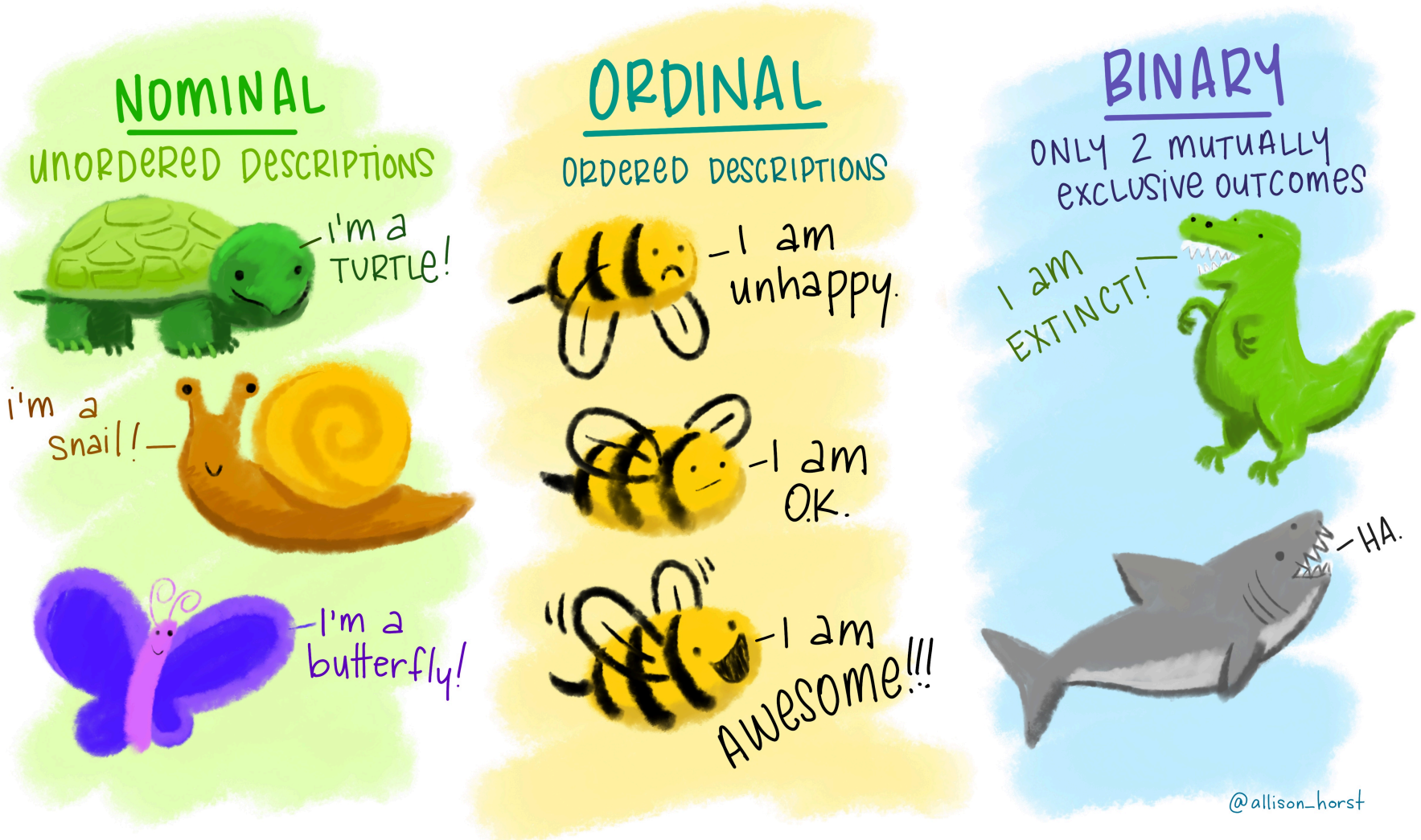
- **Quantitative:** numeric information
- **Qualitative:** descriptions (usually words)

A bit deeper:

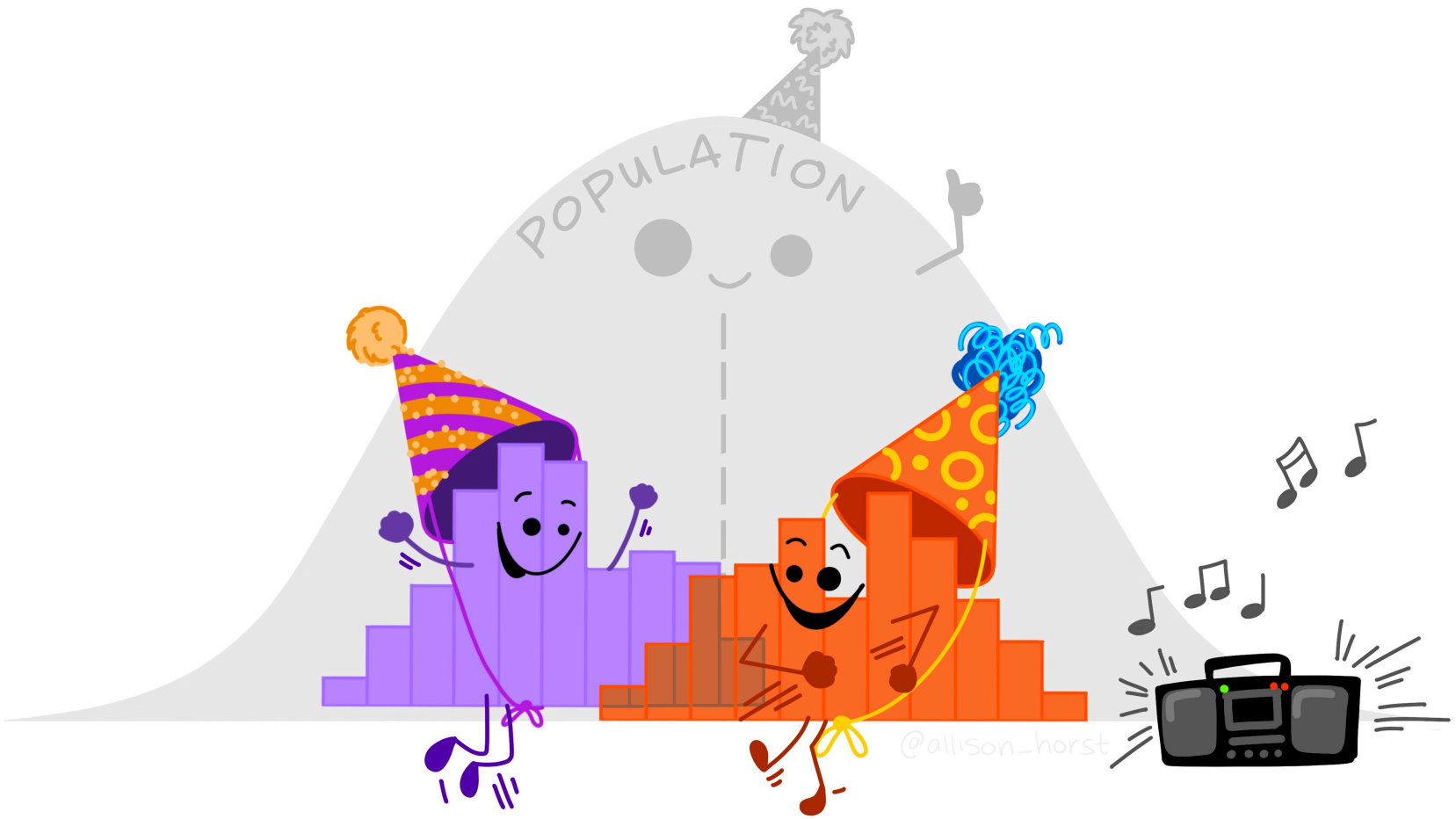
- **Continuous:** measured values, can take an infinite possible values for a variable
- **Discrete:** can only have certain values (e.g. counts)
- **Ordinal:** order matters, but the difference between values isn't known or equal (e.g. Likert Scale)
- **Binary:** only two possible outcomes (yes/no, true/false, 1/0)

Quantitative data: continuous & discrete

Nominal, ordinal, binary data:



Data distributions



How can we describe how data are distributed?

Our starting points:

- Shape / patterns / clusters (data vizualization)
- Central tendency (mean / median)
- Spread & uncertainty (standard deviation / standard error / confidence interval)

Useful data visualizations

- Histograms
- Boxplots
- Scatterplots

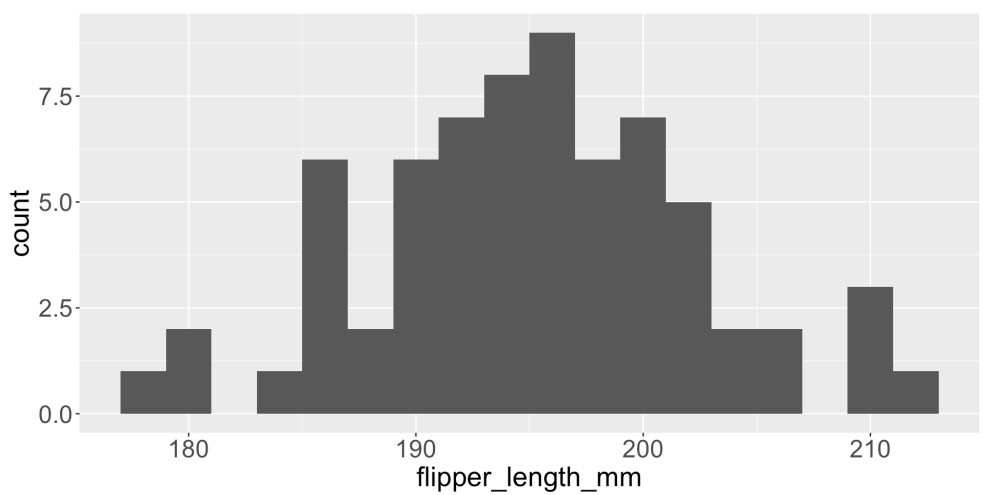
...then get even more involved:

- Beeswarm
- Marginal plots
- Raincloud plots
- Pairs plots

Histogram

A histogram is a graph of the frequency of observations within a series of bins (usually of equal size) for a variable.

Example: distribution of penguin flipper lengths for chinstrap penguins

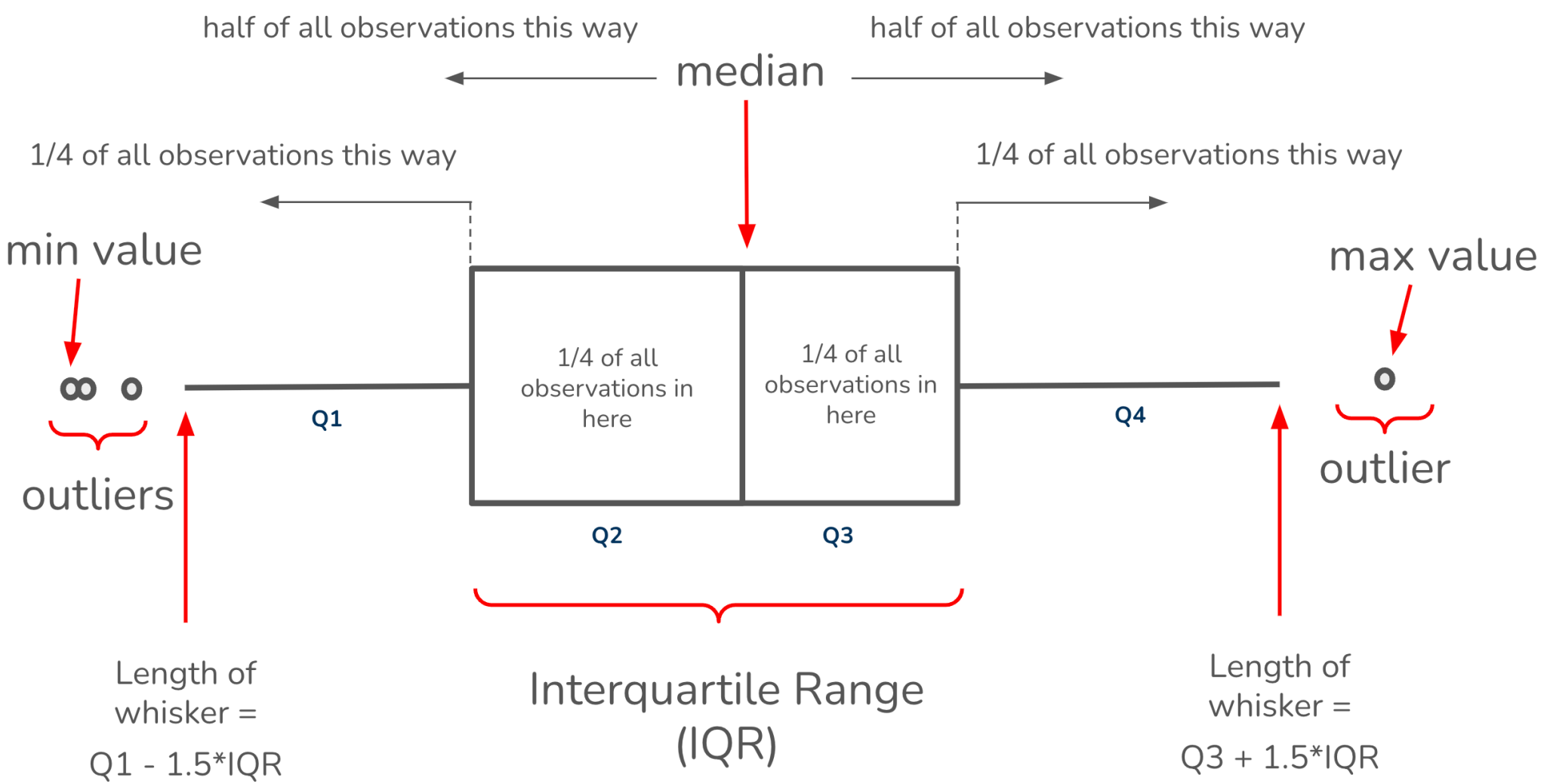


Boxplot

Most often:

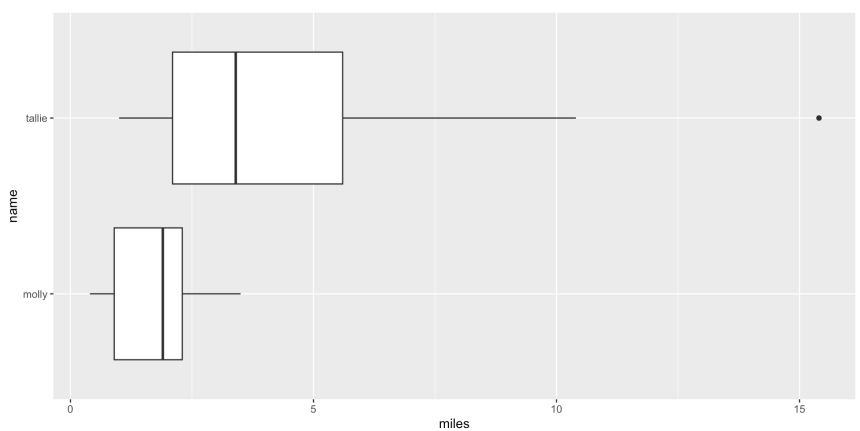
- Box extends to 1st and 3rd quartile observation values
- Line at the median value
- Whiskers extend to last observation within 1 step (1 step = 1.5*interquartile range)
- Anything beyond whiskers indicated with a dot at the observation value

Boxplot


half of all observations this way
half of all observations this way
median
1/4 of all observations this way
1/4 of all observations this way
min value
max value
1/4 of all observations in here
1/4 of all observations in here
Q1
Q4
Q2
Q3
outliers
outlier
Length of whisker = Q1 - 1.5*IQR
Interquartile Range (IQR)
Length of whisker = Q3 + 1.5*IQR

Boxplot example:

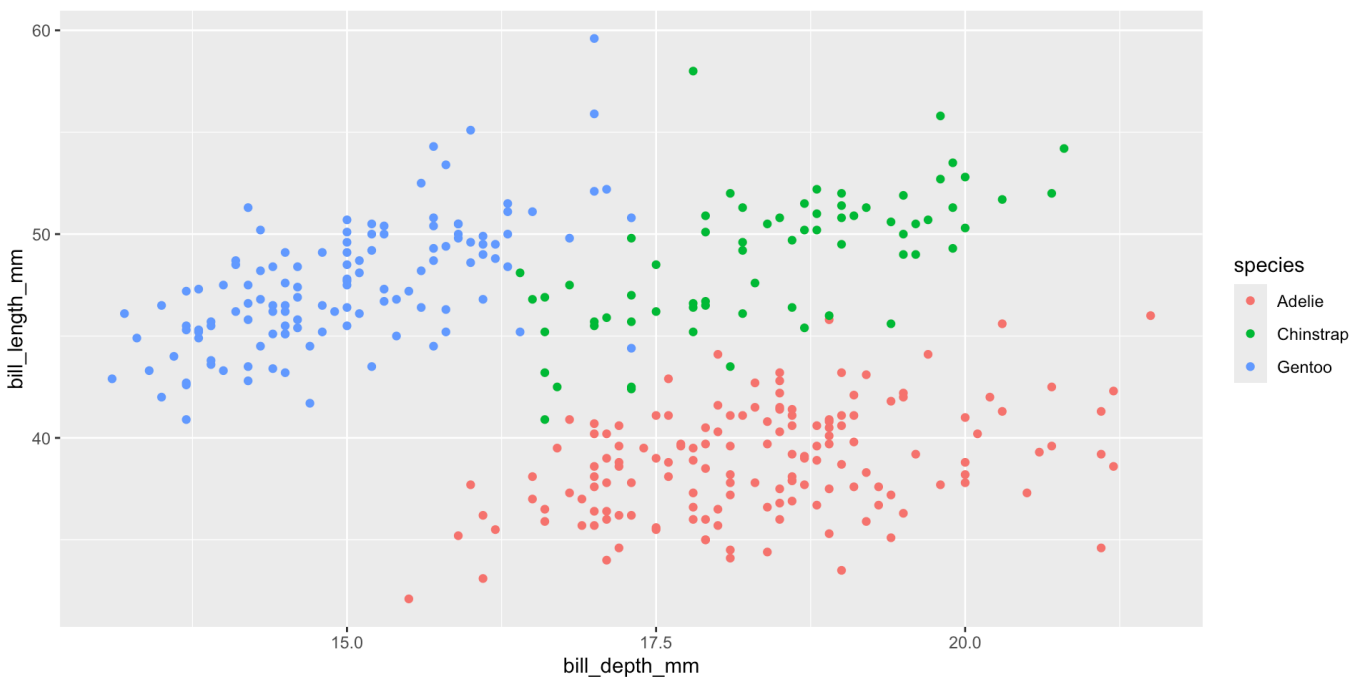
```r
# create vectors of Tallie & Molly miles logged ----
tallie <- c(1.0, 1.2, 1.8, 2.1, 2.4, 2.9, 3.4, 4.7, 5.1, 5.6, 7.8, 10.4, 15.4)
molly <- c(0.5, 0.4, 1.1, 1.2, 3.2, 2.1, 3.3, 2.3, 0.7, 0.9, 1.9, 3.5, 1.9)

# turn vectors into a data frame that can be plotted ----
dog_miles <- data.frame(tallie, molly) |>
  pivot_longer(cols = c(tallie, molly), names_to = "name", values_to = "miles")

# make boxplot of Tallie vs. Molly miles ----
ggplot(data = dog_miles, aes(x = miles, y = name)) +
  geom_boxplot()
```



Scatterplots

Always, always, always look at your data. It is the only way to make a responsible decision about an appropriate type of analysis.

```
1  ggplot(data = palmerpenguins::penguins, aes(x = bill_depth_mm, y = bill_length_mm))
2    geom_point(aes(color = species))
```



Summarizing data numerically

- Central tendency
- Variance and standard deviation
- Standard error
- Confidence interval

Mean

Average value of sample observations, calculated by summing all observation values and dividing by the number of observations. E.g. $mean\ of\ 3, 7, 17 = \frac{3+7+17}{3} = 9$

Pros:

- Average value is often useful metric
- Commonly reported

Cons:

Median

Middle value when all observations are arranged in order. If you have an even number of values, the median is calculated as the average of the middle two values. E.g. $median\ of\ 3, 7, 17 = 7$

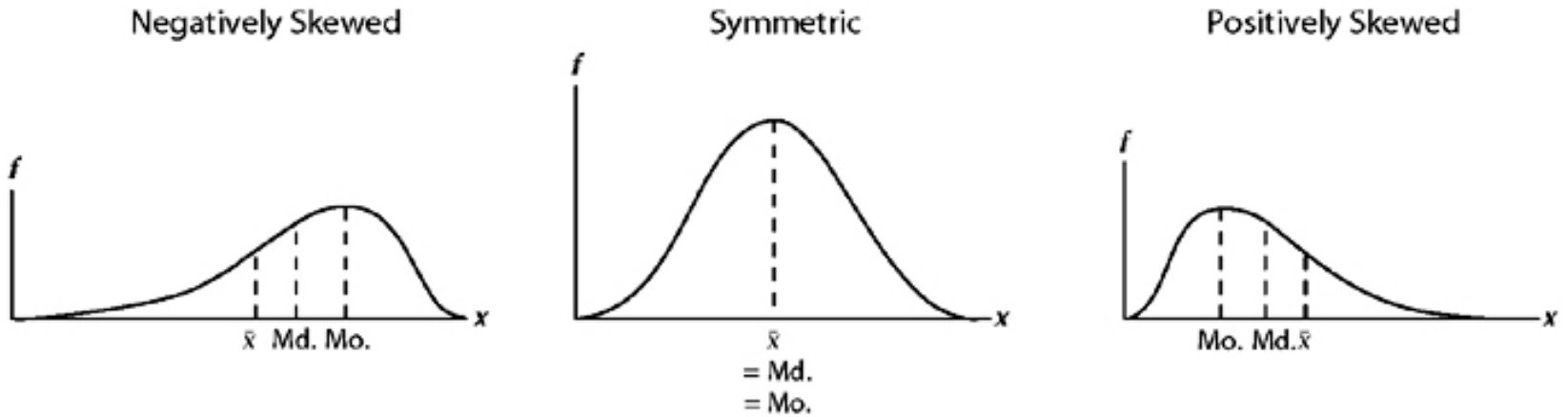
Pros:

- Less susceptible to skew and outliers
- Better as sample size increases

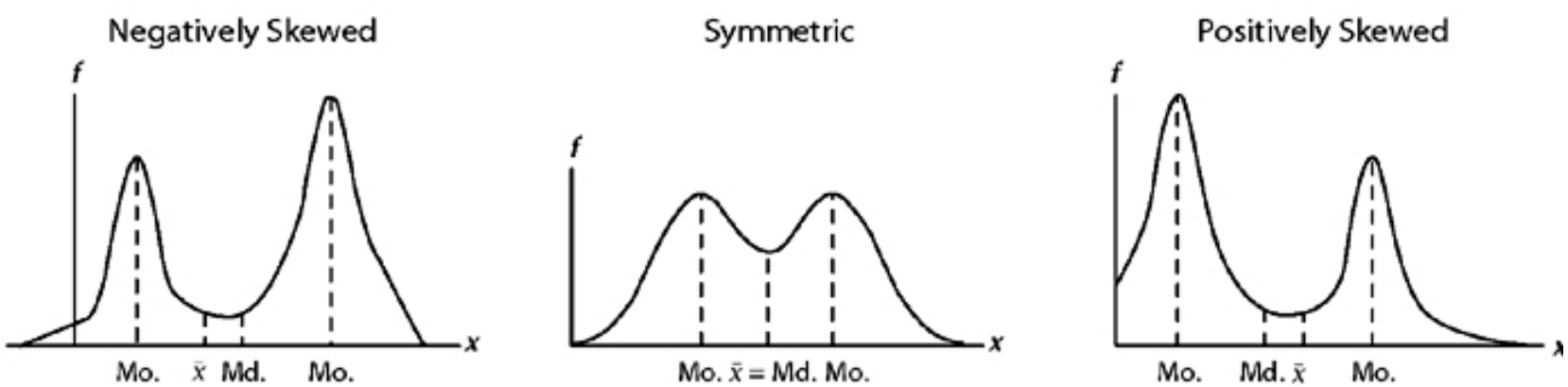
Cons:

- Doesn't take into account the magnitude of all values

Unimodal



Bimodal



The best way to describe the distribution of the data is to present the data itself.

Variance and standard deviation

Both are measures of **data spread**.

Variance

Reported in units of measurement squared

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

Standard deviation

Reported in units of measurement

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n-1}}$$

Variance

Variance: Mean squared distance of observations from the mean

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

Where s^2 is the sample variance, x_i is a sample observation value, $\bar{x}$ is the sample mean, and n is the number of observations.

Calculate variance by hand (1/2)

Given these data: $2, 4, 4, 6, 9$

1. Calculate the mean:

$$mean = \frac{2 + 4 + 4 + 6 + 9}{5} = \frac{25}{5} = 5$$

2. Subtract the mean from each data point and square the result:

$$(2 - 5)^2 = (-3)^2 = 9$$

$$(4 - 5)^2 = (-1)^2 = 1$$

$$(4 - 5)^2 = (-1)^2 = 1$$

$$(6 - 5)^2 = (1)^2 = 1$$

$$(9 - 5)^2 = (4)^2 = 16$$

Calculate variance by hand (2/2)

Given these data: $2, 4, 4, 6, 9$

3. Sum the squared differences:

$$9 + 1 + 1 + 1 + 16 = 28$$

4. Divide by the number of data points minus 1 (n - 1)

$$Variance = \frac{28}{5 - 1} = 7$$

Alternatively, in R:

```
1 var(c(2, 4, 4, 6, 9))
```

```
[1] 7
```

Standard deviation

Also a measure of data spread, calculated by taking the square root of the variance.

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

In R:

```r
sqrt(var(c(2, 4, 4, 6, 9)))
```

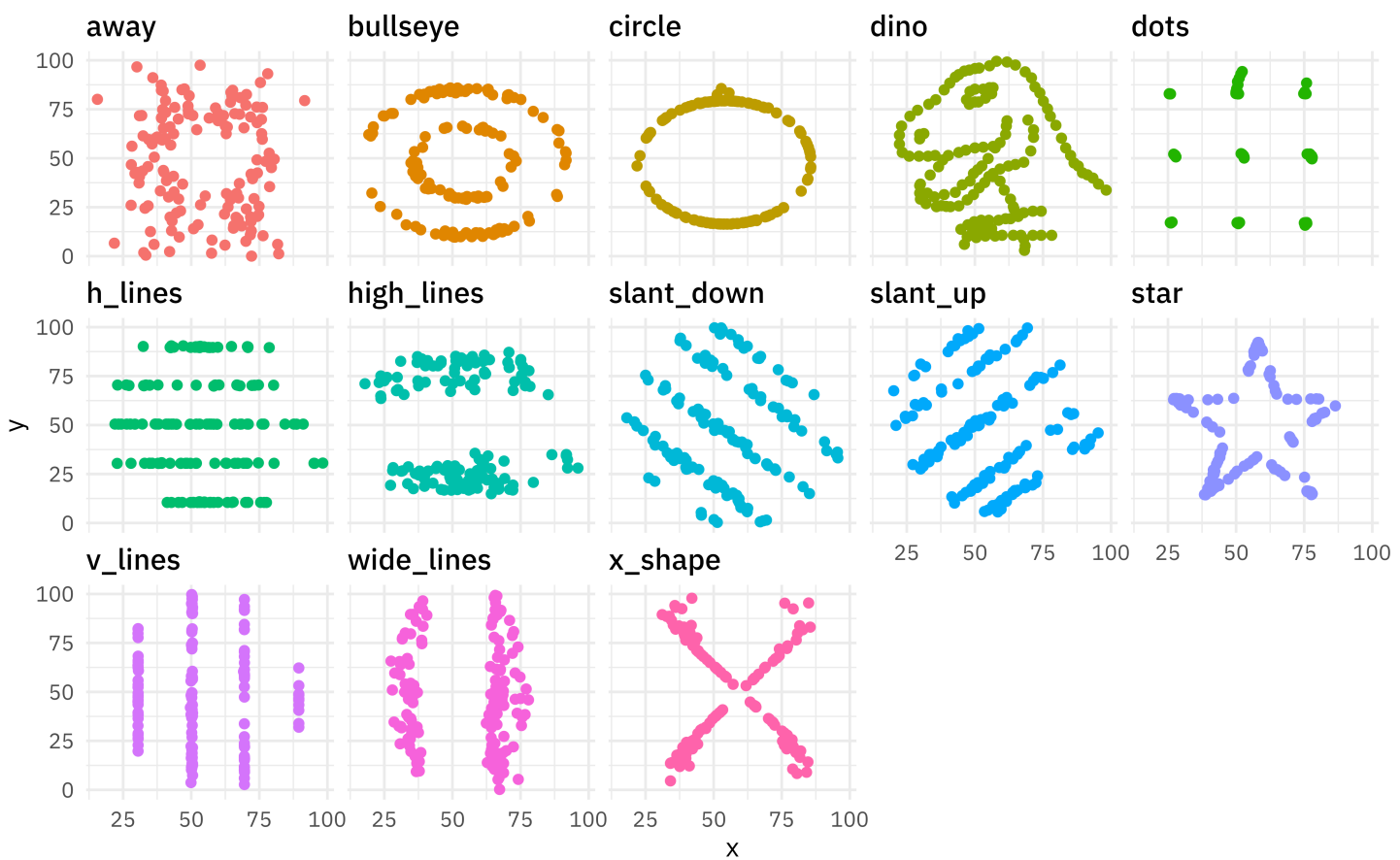
```
[1] 2.645751
```

```r
# or alternatively, just use sd()
sd(c(2, 4, 4, 6, 9))
```

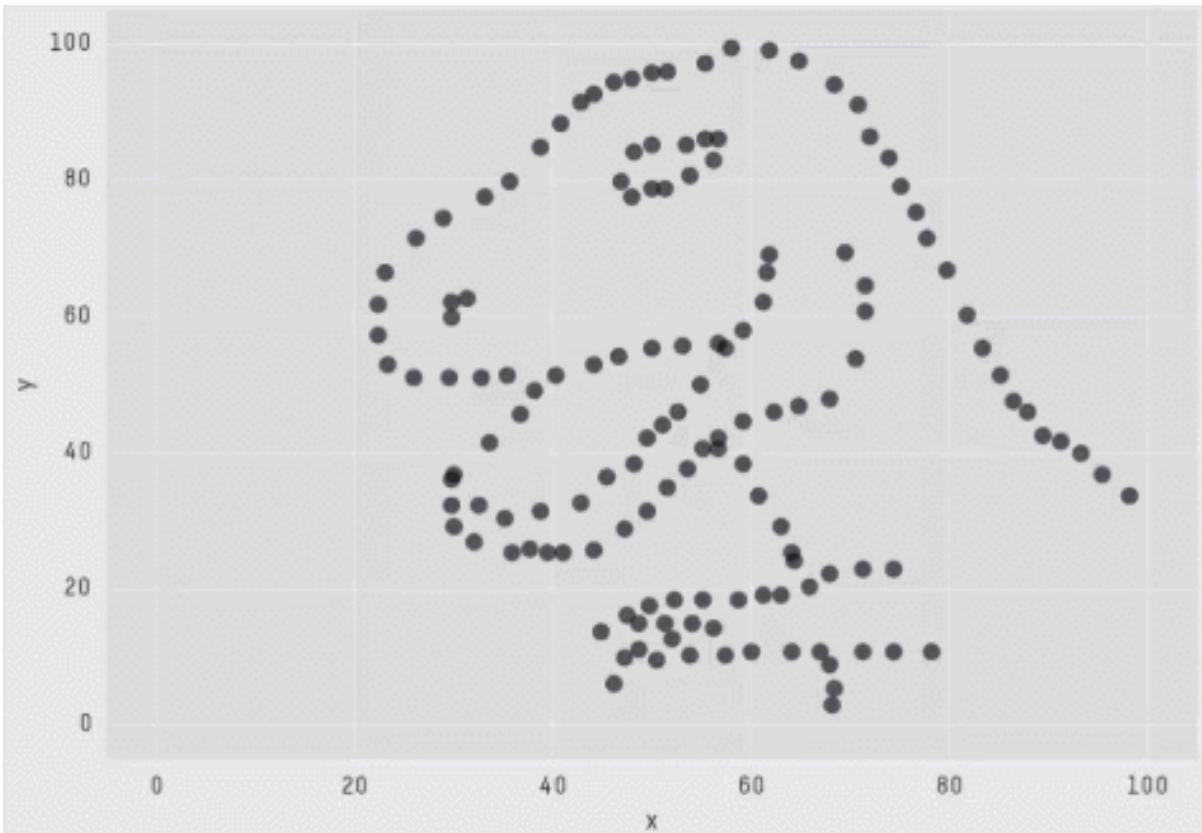
```
[1] 2.645751
```

Beware summary statistics alone . . .

Meet the Datasaurus Dozen



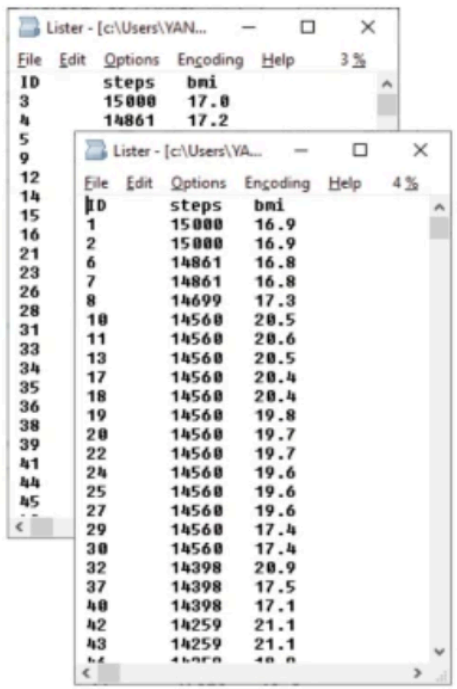
Same summary statistics, different distributions



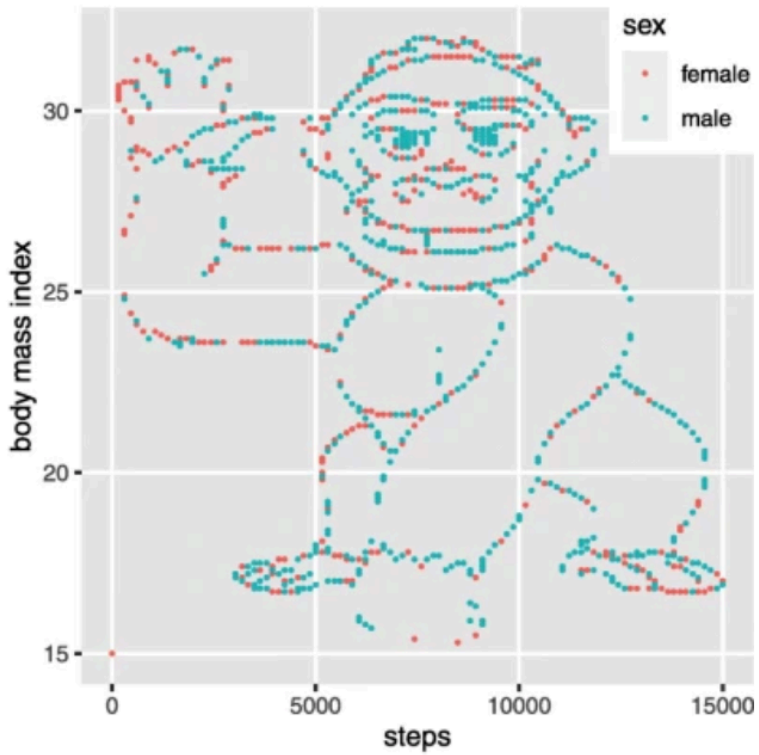
X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

selective attention test

a



b



c

	Gorilla not discovered	Gorilla discovered
Hypothesis-focused	14	5
Hypothesis-free	5	9

Confidence interval

Confidence interval: a range of values (based on a sample) that, if we were to take multiple samples from the population and calculate the confidence interval from each, would contain the true population parameter X percent of the time.

What it's NOT:

"There is a 95% chance that the true population parameter is between values X and Y."

Confidence interval example

Mean shark length is 8.42 ± 3.55 ft (mean ± standard deviation), with a 95% confidence interval of [6.45, 10.39 ft] (n = 15).

What this **DOES NOT** mean: There is a 95% chance that the true population mean length is between 6.45 and 10.39 feet.

The true population mean is a *fixed value* and does not change – the CI either contains this true mean or it does not. There is no probability involved.

What this **DOES** mean: If we took a bunch of sets of samples from the population (all n = 15), then 95% of the time, the calculated mean would fall within this range.

This statement correctly describes the frequency with which we would expect CIs to capture the mean over many samples.

Communicating data summaries