

What is statistics?

INTRODUCTION TO STATISTICS IN R



Maggie Matsui

Content Developer, DataCamp

What is statistics?

- **The field of statistics** - the practice and study of collecting and analyzing data
- **A summary statistic** - a fact about or summary of some data

What is statistics?

- **The field of statistics** - the practice and study of collecting and analyzing data
- **A summary statistic** - a fact about or summary of some data

What can statistics do?

- How likely is someone to purchase a product? Are people more likely to purchase it if they can use a different payment system?
- How many occupants will your hotel have? How can you optimize occupancy?
- How many sizes of jeans need to be manufactured so they can fit 95% of the population? Should the same number of each size be produced?
- A/B tests: Which ad is more effective in getting people to purchase a product?

What can't statistics do?

- *Why* is *Game of Thrones* so popular?

Instead...

- Are series with more violent scenes viewed by more people?

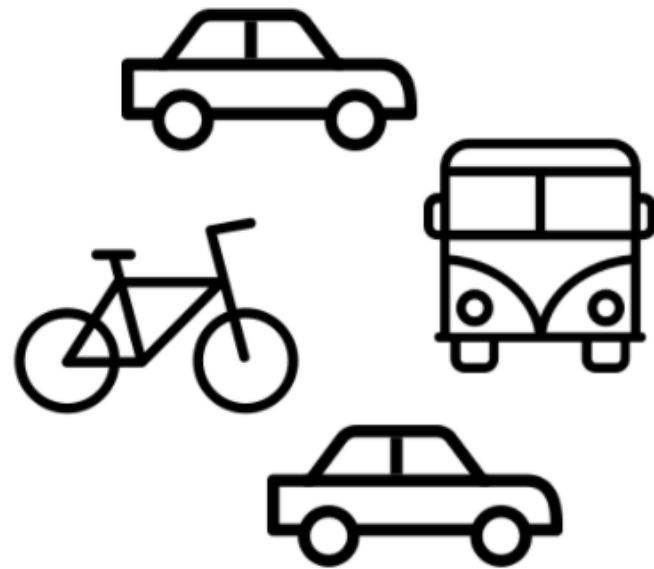
But...

- Even so, this can't tell us if more violent scenes lead to more views

Types of statistics

Descriptive statistics

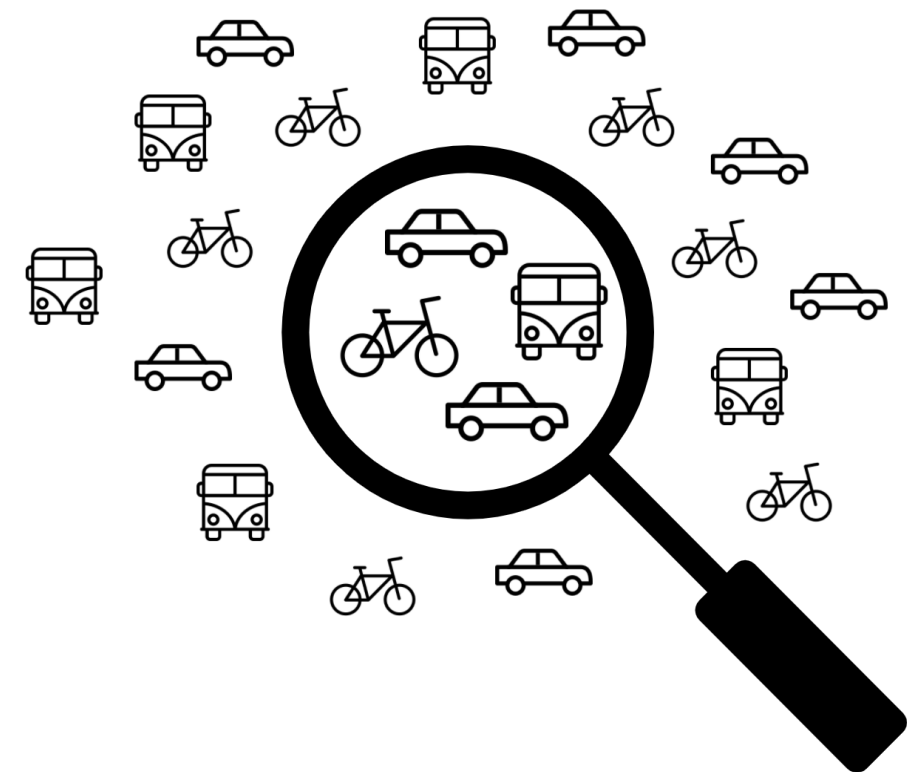
- *Describe* and summarize data



- 50% of friends drive to work
- 25% take the bus
- 25% bike

Inferential statistics

- Use a sample of data to make *inferences* about a larger population



What percent of people drive to work?

Types of data

Numeric (Quantitative)

- **Continuous (Measured)**
 - Airplane speed
 - Time spent waiting in line
- **Discrete (Counted)**
 - Number of pets
 - Number of packages shipped

Categorical (Qualitative)

- **Nominal (Unordered)**
 - Married/unmarried
 - Country of residence
- **Ordinal (Ordered)**
 - ☐ Strongly disagree
 - ☐ Somewhat disagree
 - ☐ Neither agree nor disagree
 - ☒ Somewhat agree
 - ☐ Strongly agree

Categorical data can be represented as numbers

Nominal (Unordered)

- Married/unmarried (1 / 0)
- Country of residence (1, 2, ...)

Ordinal (Ordered)

- Strongly disagree (1)
- Somewhat disagree (2)
- Neither agree nor disagree (3)
- Somewhat agree (4)
- Strongly agree (5)

Why does data type matter?

Summary statistics

```
car_speeds %>%  
  summarize(avg_speed = mean(speed_mph))
```

```
avg_speed  
1  40.09062
```

Plots



Why does data type matter?

Summary statistics

```
demographics %>%  
  count(marriage_status)
```

	marriage_status	n
1	single	188
2	married	143
3	divorced	124

Plots



Let's practice!

INTRODUCTION TO STATISTICS IN R

Measures of center

INTRODUCTION TO STATISTICS IN R



Maggie Matsui

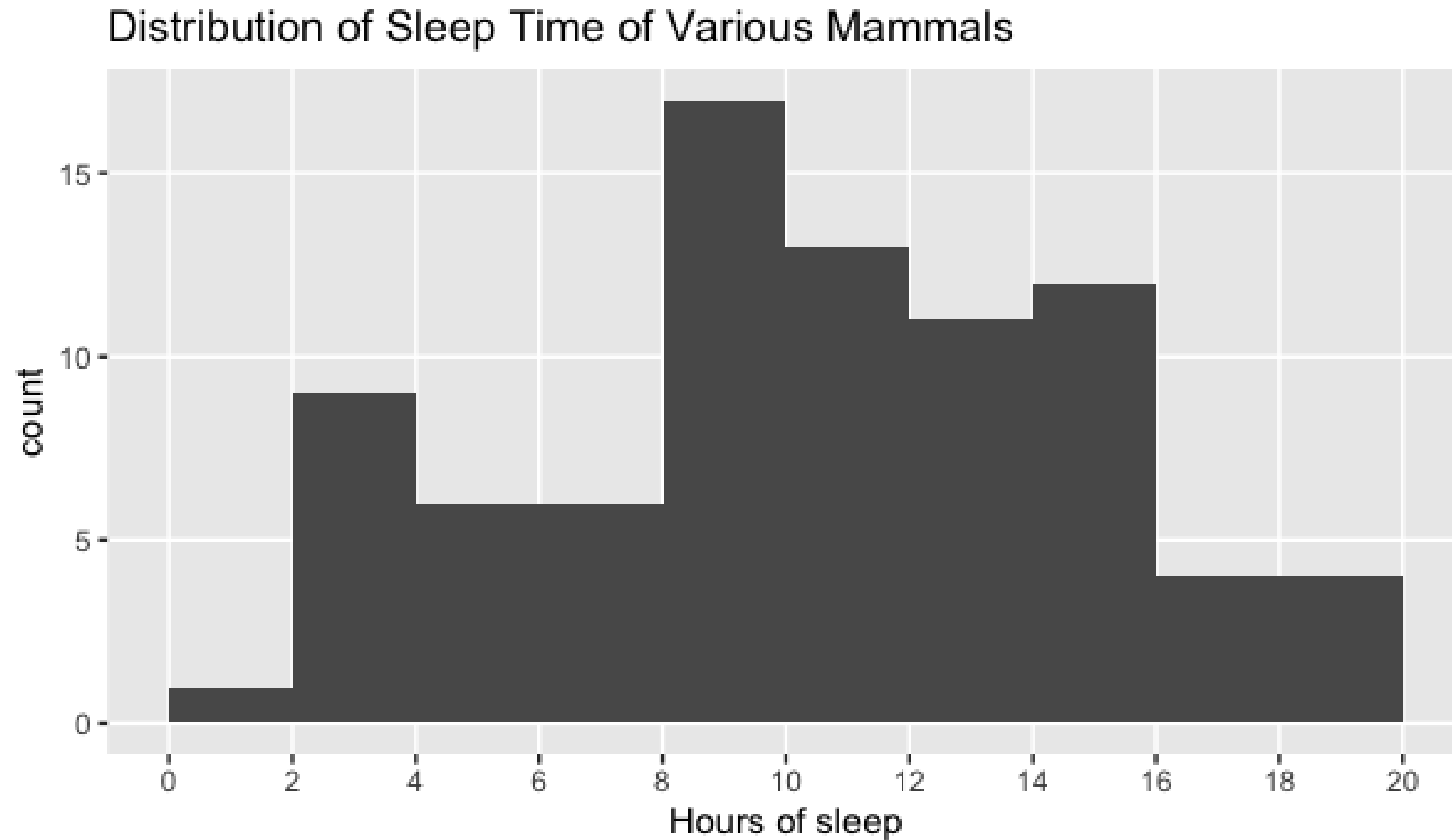
Content Developer, DataCamp

Mammal sleep data

msleep

```
# A tibble: 83 x 11
  name          genus      vore order      sleep_total sleep_rem sleep_cycle awake
  <chr>         <chr>    <chr> <chr>      <dbl>      <dbl>      <dbl> <dbl>
1 Cheetah      Acinonyx  carni Carnivora    12.1        NA        NA    11.9
2 Owl monkey   Aotus     omni  Primates     17          1.8        NA     7
3 Mountain beaver Aplodontia herbi Rodentia    14.4        2.4        NA    9.6
4 Greater short... Blarina   omni  Soricomorpha 14.9        2.3        0.133  9.1
5 Cow          Bos       herbi Artiodactyla  4           0.7        0.667  20
6 Three-toed sloth Bradypus  herbi Pilosa    14.4        2.2        0.767  9.6
7 Northern fur... Callorhinus carni Carnivora    8.7        1.4        0.383  15.3
# ... with 76 more rows, and 2 more variables: brainwt <dbl>, bodywt <dbl>
```

Histograms

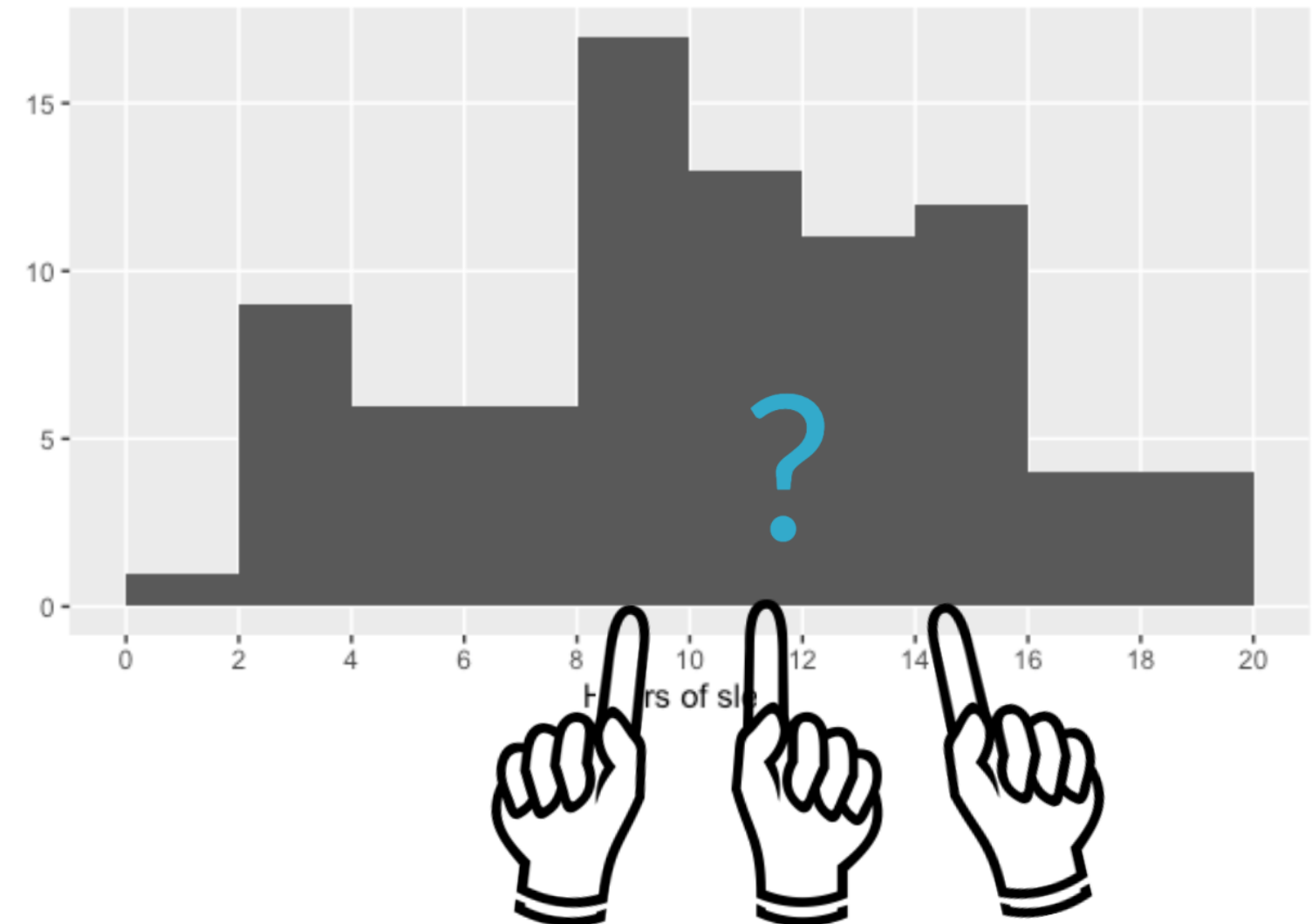


How long do mammals in this dataset typically sleep?

What's a typical value?

Where is the center of the data?

- Mean
- Median
- Mode



Measures of center: mean

```
name                sleep_total
1 Cheetah            12.1
2 Owl monkey         17.0
3 Mountain beaver    14.4
4 Greater short-tailed shrew 14.9
...
```

$$\text{Mean sleep time} = \frac{12.1 + 17.0 + 14.4 + 14.9 + \dots}{83} = 10.43$$

```
mean(msleep$sleep_total)
```

```
10.43373
```

Measures of center: median

```
sort(msleep$sleep_total)
```

```
[1] 1.9 2.7 2.9 3.0 3.1 3.3 3.5 3.8 3.9 4.0 4.4 5.2 5.3 5.3 5.4 5.6 6.2  
...  
[52] 11.5 12.1 12.5 12.5 12.5 12.5 12.8 12.8 13.0 13.5 13.7 13.8 14.2 14.3 14.4 14.4 14.5  
[69] 14.6 14.9 14.9 15.6 15.8 15.8 15.9 16.6 17.0 17.4 18.0 18.1 19.4 19.7 19.9
```

```
sort(msleep$sleep_total)[42]
```

```
10.1
```

```
median(msleep$sleep_total)
```

```
10.1
```


Measures of center: mode

Most frequent value

```
msleep %>% count(sleep_total, sort = TRUE)
```

	sleep_total	n
	<dbl>	<int>
1	12.5	4
2	10.1	3
3	5.3	2
4	6.3	2
...		

```
msleep %>% count(vore, sort = TRUE)
```

	vore	n
	<chr>	<int>
1	herbi	32
2	omni	20
3	carni	19
4	NA	7
5	insecti	5

Adding an outlier

```
msleep %>%  
  filter(vore == "insecti")
```

	name	genus	vore	order	sleep_total
	<chr>	<chr>	<chr>	<chr>	<dbl>
1	Big brown bat	Eptesicus	insecti	Chiroptera	19.7
2	Little brown bat	Myotis	insecti	Chiroptera	19.9
3	Giant armadillo	Priodontes	insecti	Cingulata	18.1
4	Eastern american mole	Scalopus	insecti	Soricomorpha	8.4

Adding an outlier

```
msleep %>%  
  filter(vore == "insecti") %>%  
  summarize(mean_sleep = mean(sleep_total),  
            median_sleep = median(sleep_total))
```

```
mean_sleep median_sleep  
      <dbl>      <dbl>  
1    16.52      18.9
```

Adding an outlier

```
msleep %>%  
  filter(vore == "insecti")
```

	name	genus	vore	order	sleep_total
	<chr>	<chr>	<chr>	<chr>	<dbl>
1	Big brown bat	Eptesicus	insecti	Chiroptera	19.7
2	Little brown bat	Myotis	insecti	Chiroptera	19.9
3	Giant armadillo	Priodontes	insecti	Cingulata	18.1
4	Eastern american mole	Scalopus	insecti	Soricomorpha	8.4
5	Mystery insectivore	0.0

Adding an outlier

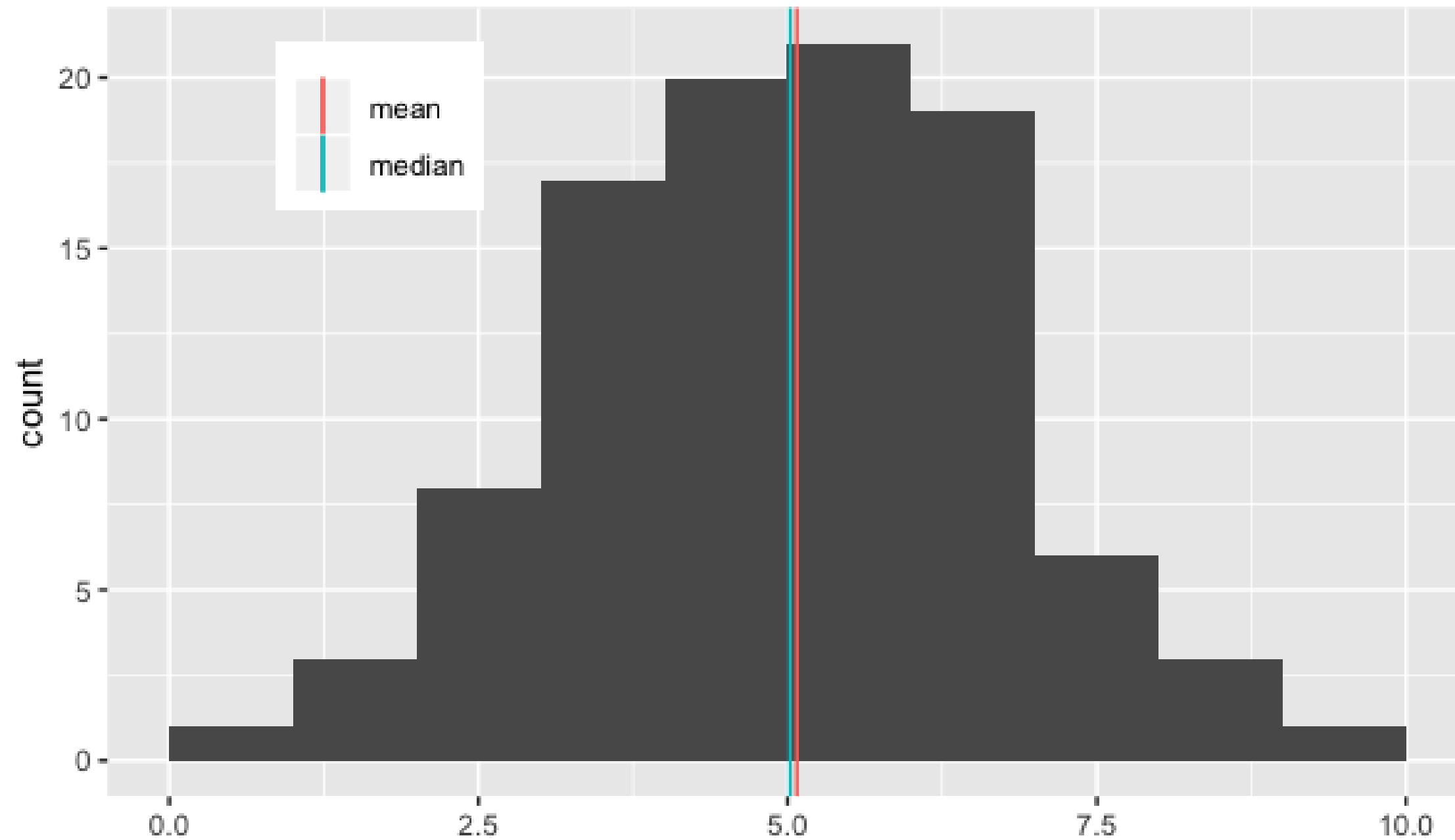
```
msleep %>%  
  filter(vore == "insecti") %>%  
  summarize(mean_sleep = mean(sleep_total),  
            median_sleep = median(sleep_total))
```

```
mean_sleep median_sleep  
      <dbl>      <dbl>  
1    13.22      18.1
```

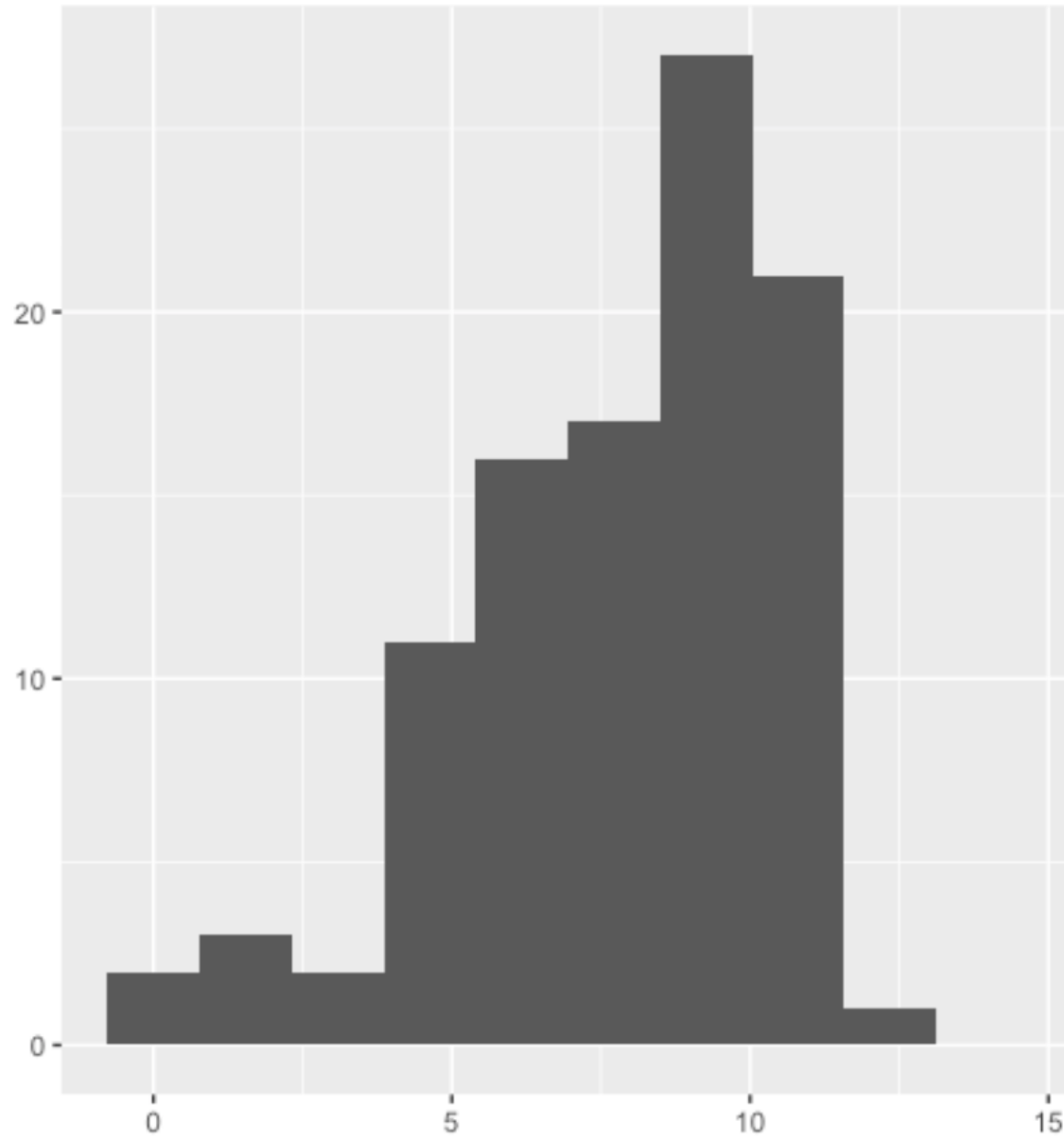
Mean: 16.5 → 13.2

Median: 18.9 → 18.1

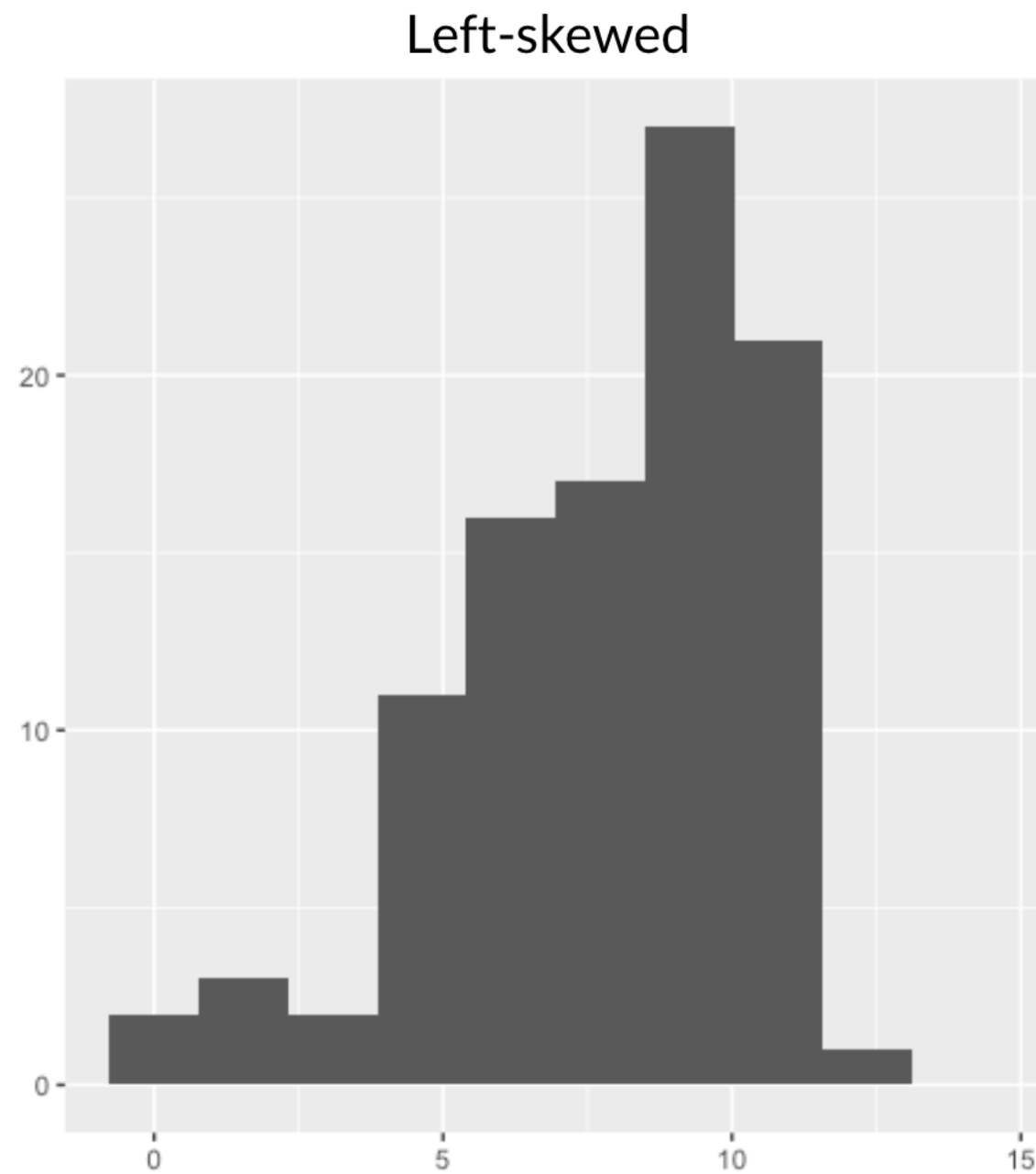
Which measure to use?



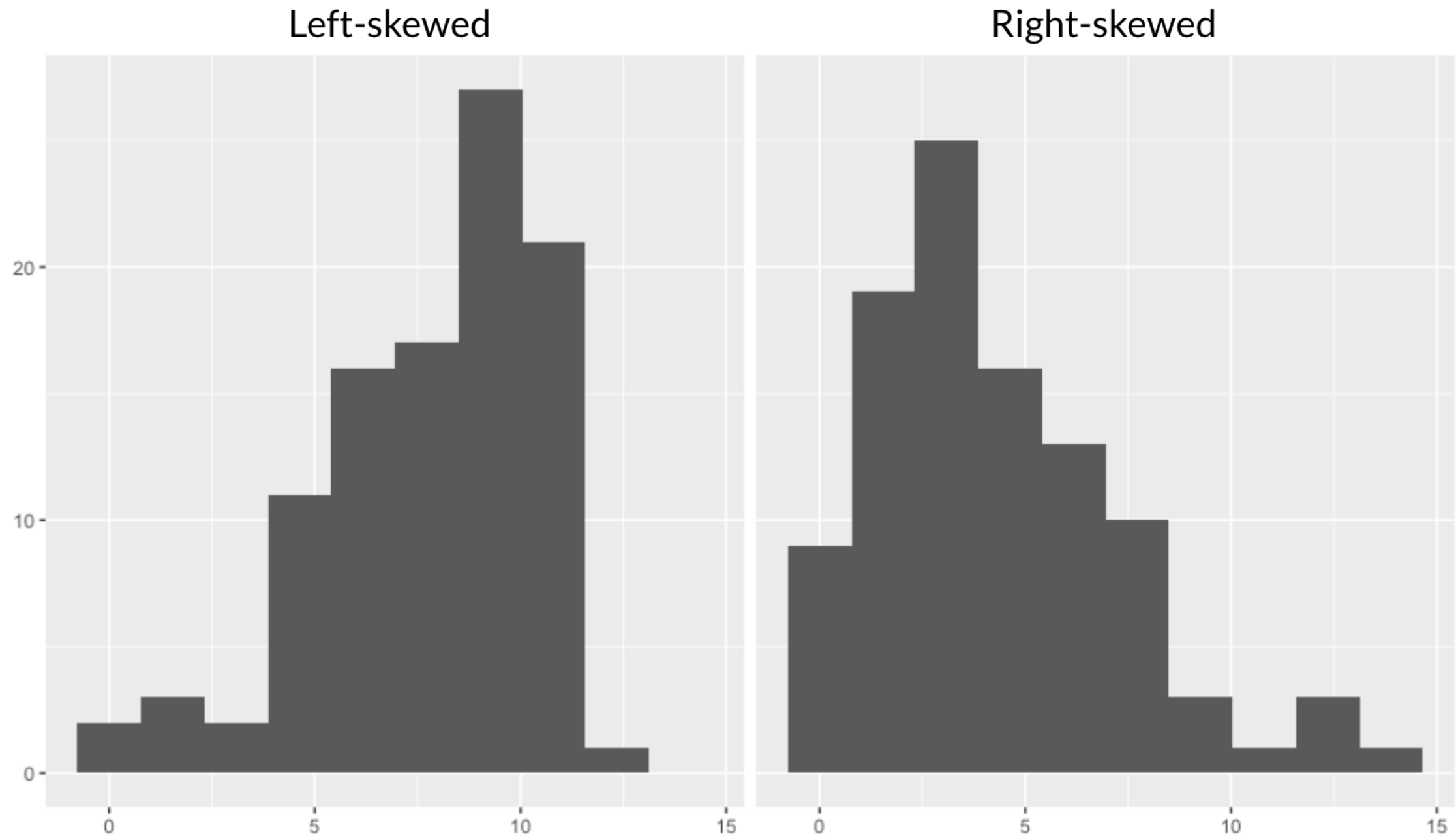
Skew



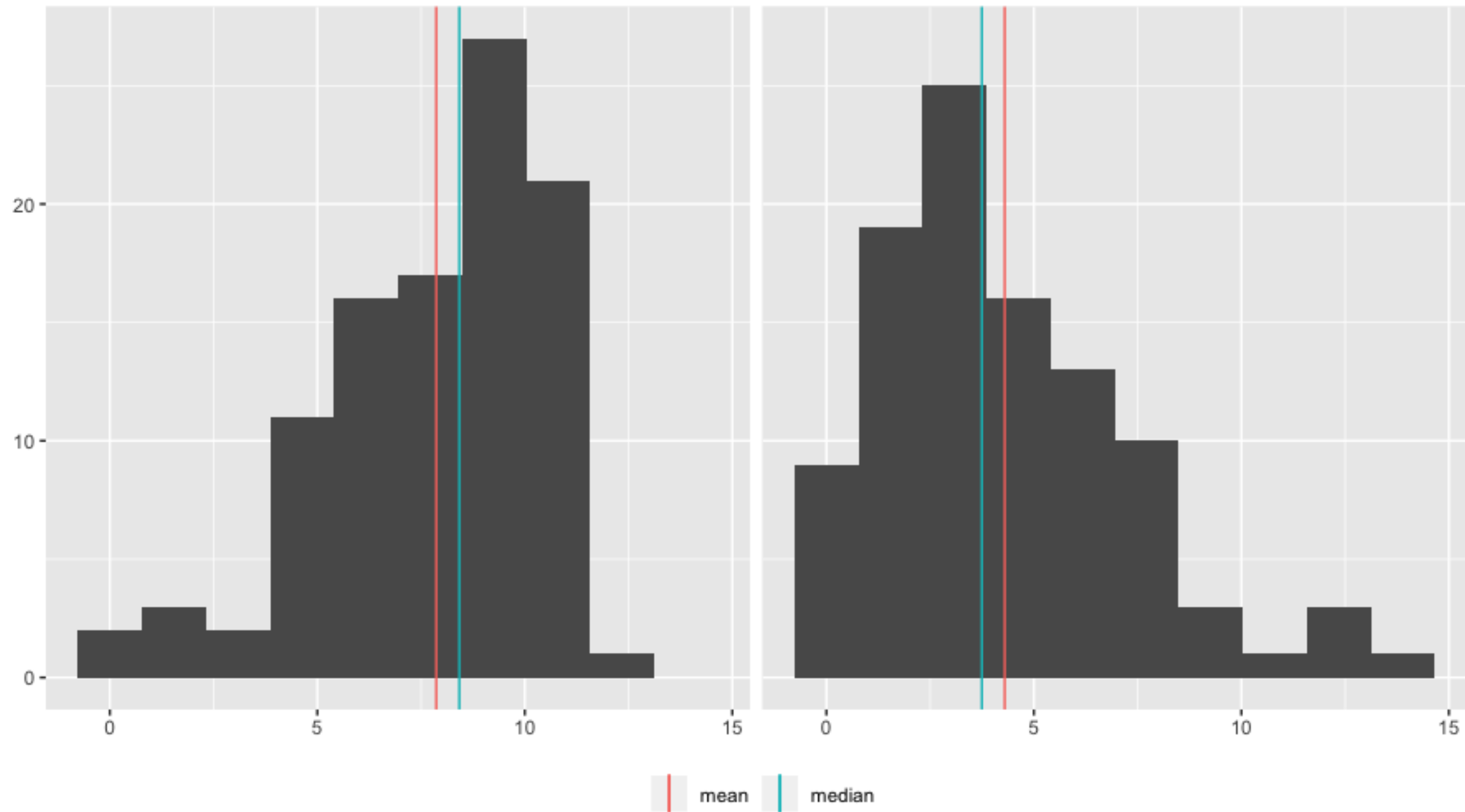
Skew



Skew



Which measure to use?



Let's practice!

INTRODUCTION TO STATISTICS IN R

Measures of spread

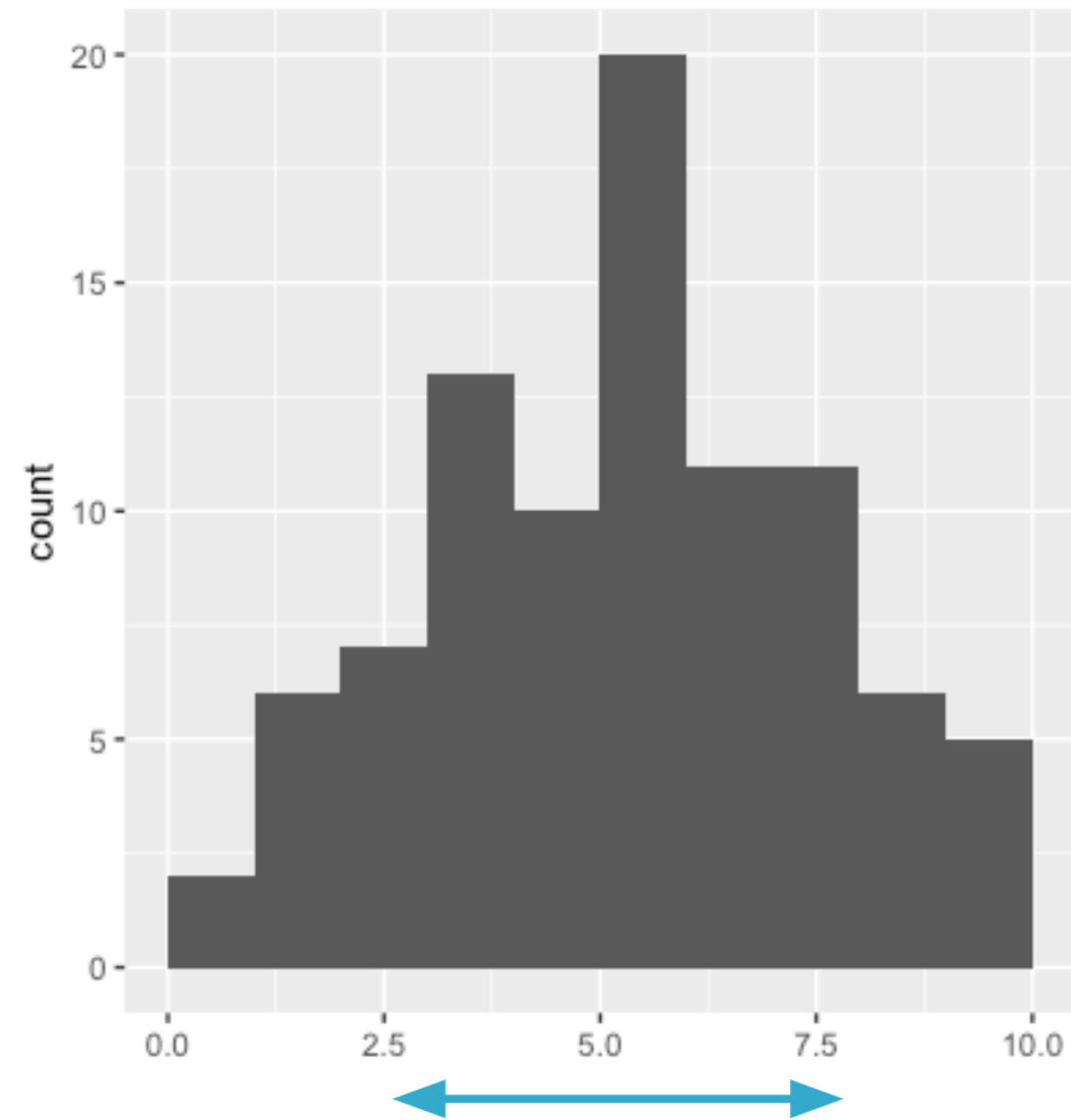
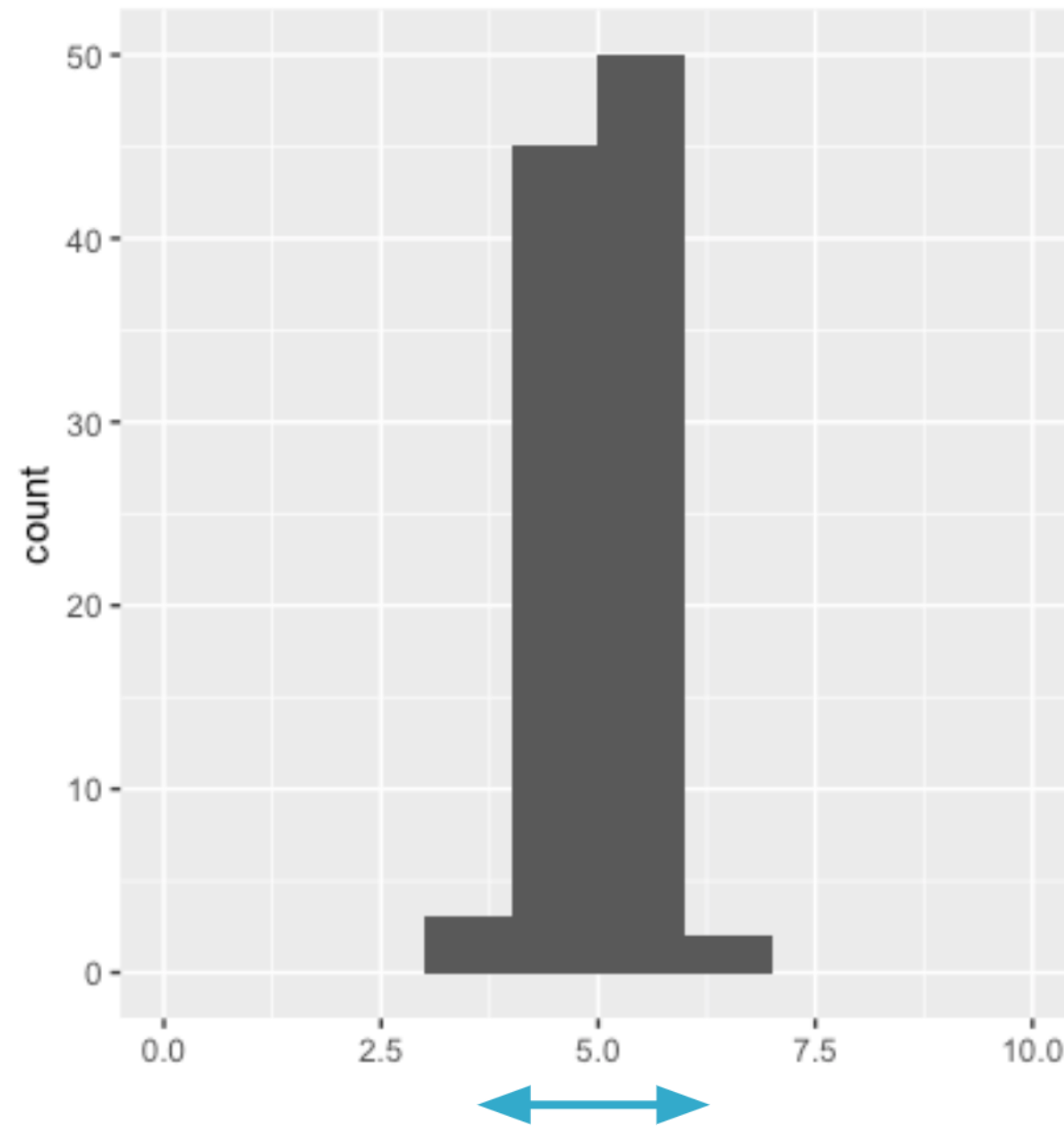
INTRODUCTION TO STATISTICS IN R



Maggie Matsui

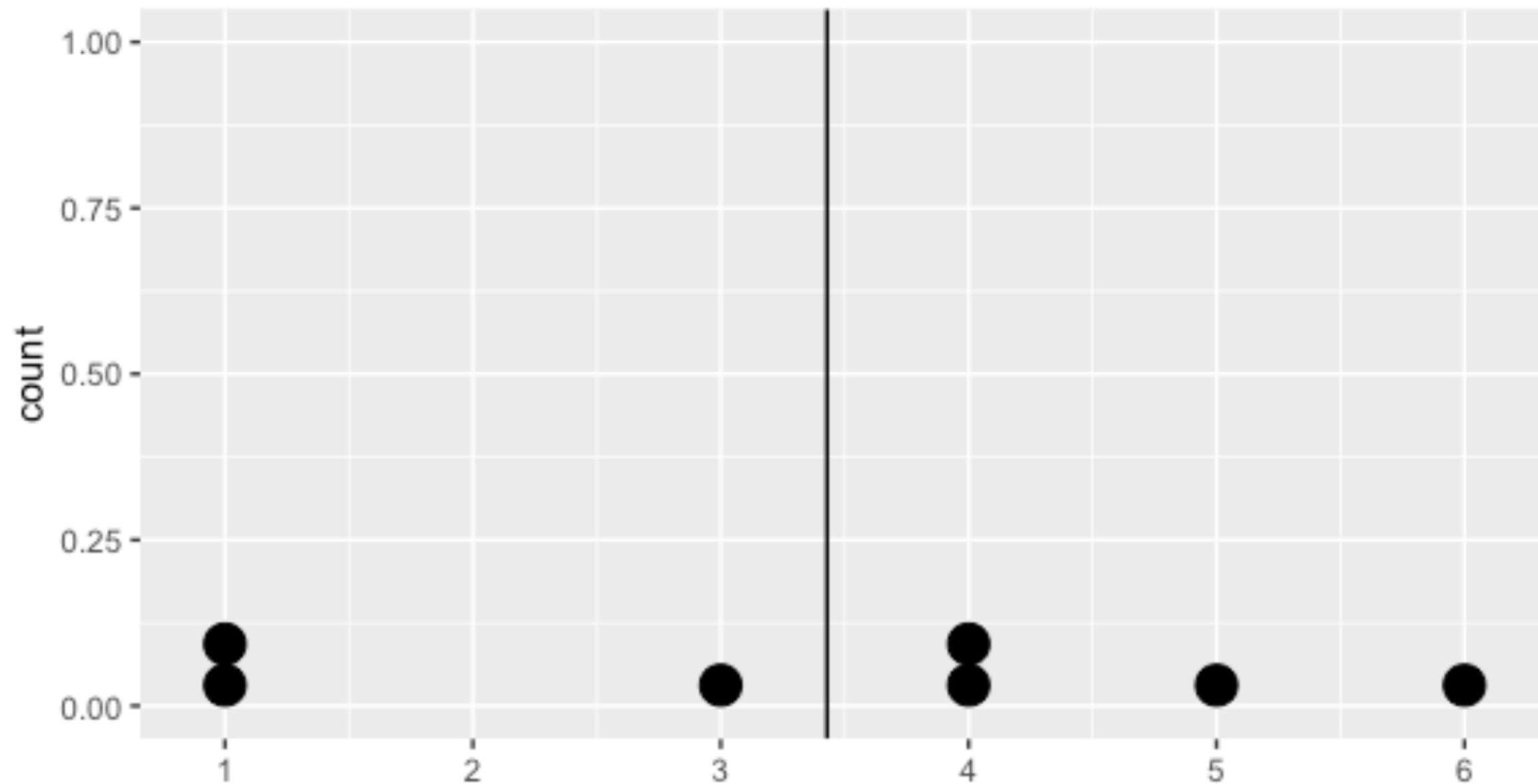
Content Developer, DataCamp

What is spread?

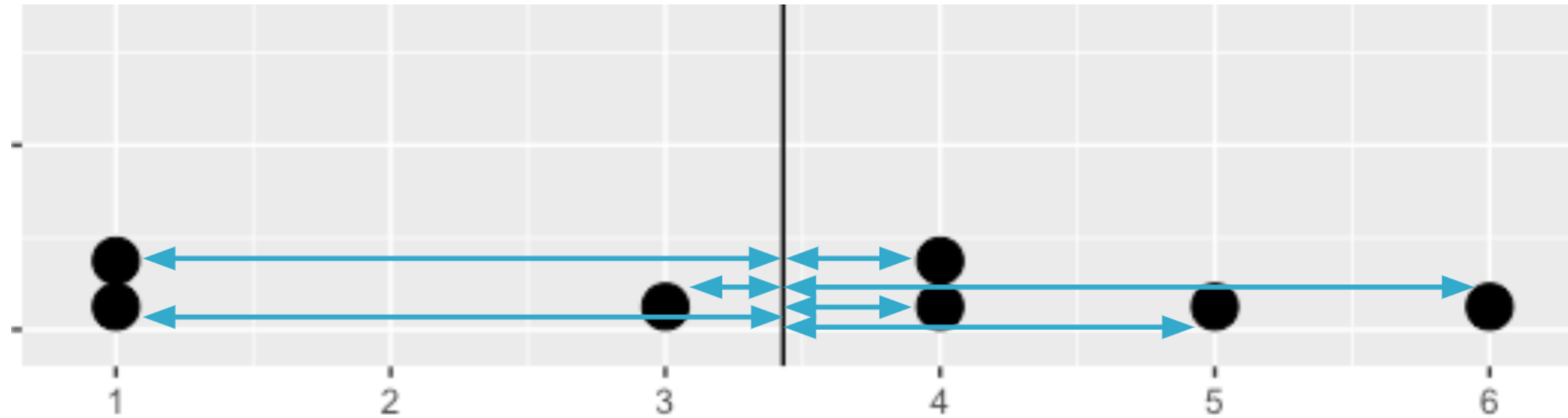


Variance

Average distance from each data point to the data's mean



Variance



```
dists <- msleep$sleep_total - mean(msleep$sleep_total)
dists
```

```
1.66626506  6.56626506 ... -4.13373494  2.06626506 -0.63373494
```

Variance

```
squared_dists <- (dists)^2
```

```
2.776439251 43.115836841 ... 17.087764552 4.269451299 0.401619974
```

```
sum_sq_dists <- sum(squared_dists)  
sum_sq_dists
```

```
1624.066
```


Variance

```
sum_sq_dists/82
```

```
19.80568
```

```
var(msleep$sleep_total)
```

```
19.80568
```

Standard deviation

```
sqrt(var(msleep$sleep_total))
```

```
4.450357
```

```
sd(msleep$sleep_total)
```

```
4.450357
```

Mean absolute deviation

```
dists <- msleep$sleep_total - mean(msleep$sleep_total)
mean(abs(dists))
```

```
3.566701
```

Standard deviation vs. mean absolute deviation

- SD squares distances, penalizing longer distances more than shorter ones.
- MAD penalizes each distance equally.
- One isn't better than the other, but SD is more common than MAD.

Quartiles

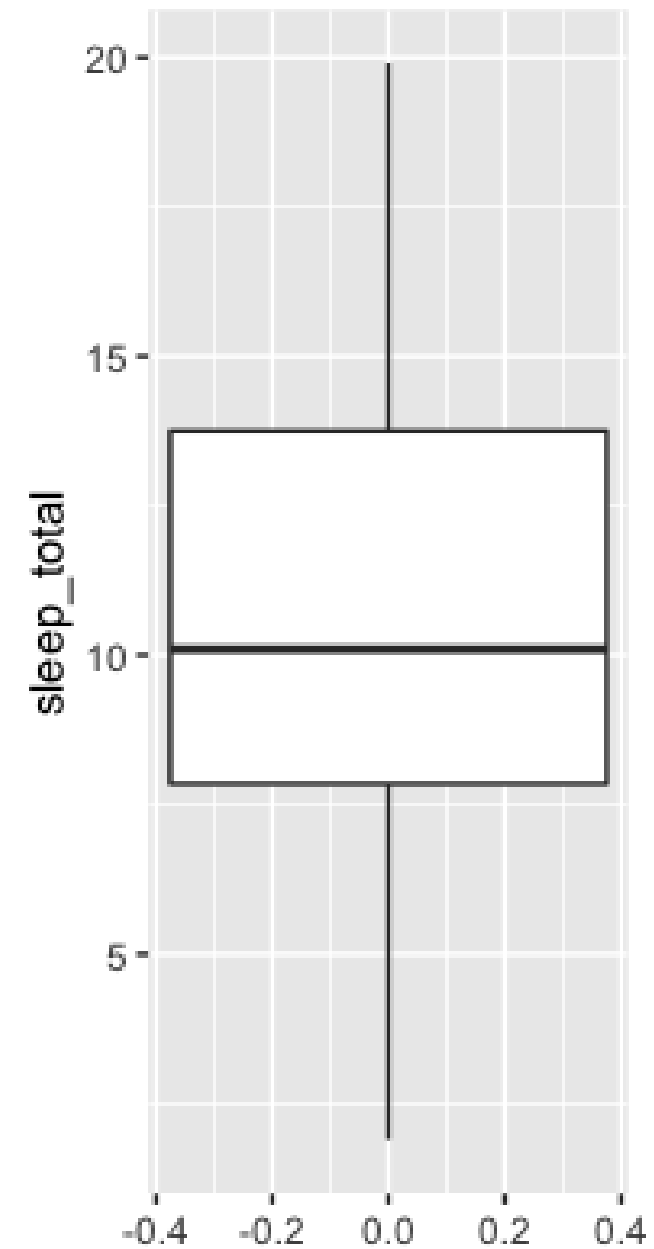
```
quantile(msleep$sleep_total)
```

0%	25%	50%	75%	100%
1.90	7.85	10.10	13.75	19.90

Second quartile/50th percentile = median

Boxplots use quartiles

```
ggplot(msleep, aes(y = sleep_total)) +  
  geom_boxplot()
```



Quantiles

```
quantile(msleep$sleep_total, probs = c(0, 0.2, 0.4, 0.6, 0.8, 1))
```

0%	20%	40%	60%	80%	100%
1.90	6.24	9.48	11.14	14.40	19.90

```
seq(from, to, by)
```

```
quantile(msleep$sleep_total, probs = seq(0, 1, 0.2))
```

0%	20%	40%	60%	80%	100%
1.90	6.24	9.48	11.14	14.40	19.90

Interquartile range (IQR)

Height of the box in a boxplot

```
quantile(msleep$sleep_total, 0.75) - quantile(msleep$sleep_total, 0.25)
```

```
5.9
```

Outliers

Outlier: data point that is substantially different from the others

How do we know what a substantial difference is? A data point is an outlier if:

- $\text{data} < Q1 - 1.5 \times \text{IQR}$ or
- $\text{data} > Q3 + 1.5 \times \text{IQR}$

Finding outliers

```
iqr <- quantile(msleep$bodywt, 0.75) - quantile(msleep$bodywt, 0.25)
lower_threshold <- quantile(msleep$bodywt, 0.25) - 1.5 * iqr
upper_threshold<- quantile(msleep$bodywt, 0.75) + 1.5 * iqr
```

```
msleep %>% filter(bodywt < lower_threshold | bodywt > upper_threshold ) %>%
  select(name, vore, sleep_total, bodywt)
```

```
# A tibble: 11 x 4
  name          vore  sleep_total bodywt
1 Cow          herbi         4      600
2 Asian elephant herbi        3.9    2547
3 Horse        herbi        2.9     521
... 
```

Let's practice!

INTRODUCTION TO STATISTICS IN R