# STA220H1: The Practice of Statistics I

Elena Tuzhilina

January 17, 2023

# Please turn on your videos :)


Figure 1: [picture source]

# Learning strategy

**Wednesday-Friday**: watch modules at https://sta220.utstat.utoronto.ca

**Wednesday-Friday**: do practice sets, attend TAs office hours if something is not clear

**Friday-Monday**: do Quiz, attend my office hours on Monday if something is still not clear

**Get help**: post your questions on Piazza (not my personal email, pls!) or attend office hours

# Agenda for today

- Recap: summary statistics, boxplots
- Summarizing one quantitative variable: histogram, standard deviation
- Summarizing relationship between two variables: barplot, scatterplot, correlation

# Recap: data

```
sta220.data
```

```
##              student grade
## 1      Jenny Holder    77
## 2        Tammy Snow    88
## 3     Victoria Hall    90
## 4    Saoirse Spence    86
## 5       Raja Cooper    94
## 6 Nicolas Roberson    68
## 7    Finnley Wright    85
## 8      Nate Mcgrath    93
## 9    Joshua Pollard    82
```

# Recap: mean

- Measures the **central tendency** of a data set

```
sta220.data$grade
```

```
## [1] 77 88 90 86 94 68 85 93 82
```

```
mean(sta220.data$grade)
```

```
## [1] 84.77778
```

# Recap: median

- Also measures the **central tendency** of a data set
- If we were to sort all of the values, then the **median** is the value in the middle

```
sort(sta220.data$grade)
```

```
## [1] 68 77 82 85 86 88 90 93 94
```

```
median(sta220.data$grade)
```

```
## [1] 86
```

# Recap: median

- Sometimes we need to use **interpolation** (when $n$ even)

```
sort(sta220.data$grade[1:8])
```

```
## [1] 68 77 85 86 88 90 93 94
```

```
median(sta220.data$grade[1:8])
```

```
## [1] 87
```

# Recap: first and third quartiles

- To find the **first quartile** we travel quarter (1/4) way through the sorted list
- To find the **third quartile** we travel three quarters (3/4) way through the sorted list

```
sort(sta220.data$grade)
```

```
## [1] 68 77 82 85 86 88 90 93 94
```

```
quantile(sta220.data$grade)
```

```
##   0%  25%  50%  75% 100%
##   68   82   86   90   94
```

# Recap: first and third quartiles

- Sometimes we need to use **interpolation** (when $n-1$ is not divisible by 4)

```
sort(sta220.data$grade[1:8])
```

```
## [1] 68 77 85 86 88 90 93 94
```

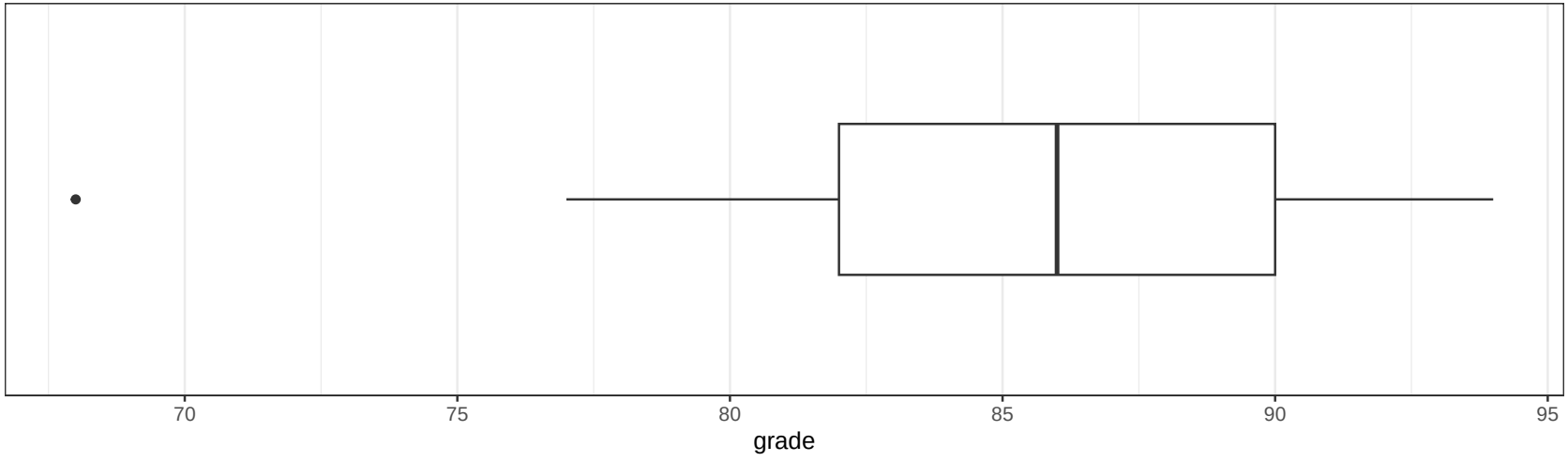
```
quantile(sta220.data$grade[1:8])
```

```
##     0%    25%    50%    75%   100%
## 68.00 83.00 87.00 90.75 94.00
```

# Recap: boxplot

```
quantile(sta220.data$grade)
```

```
##   0%  25%  50%  75% 100%
##   68   82   86   90   94
```

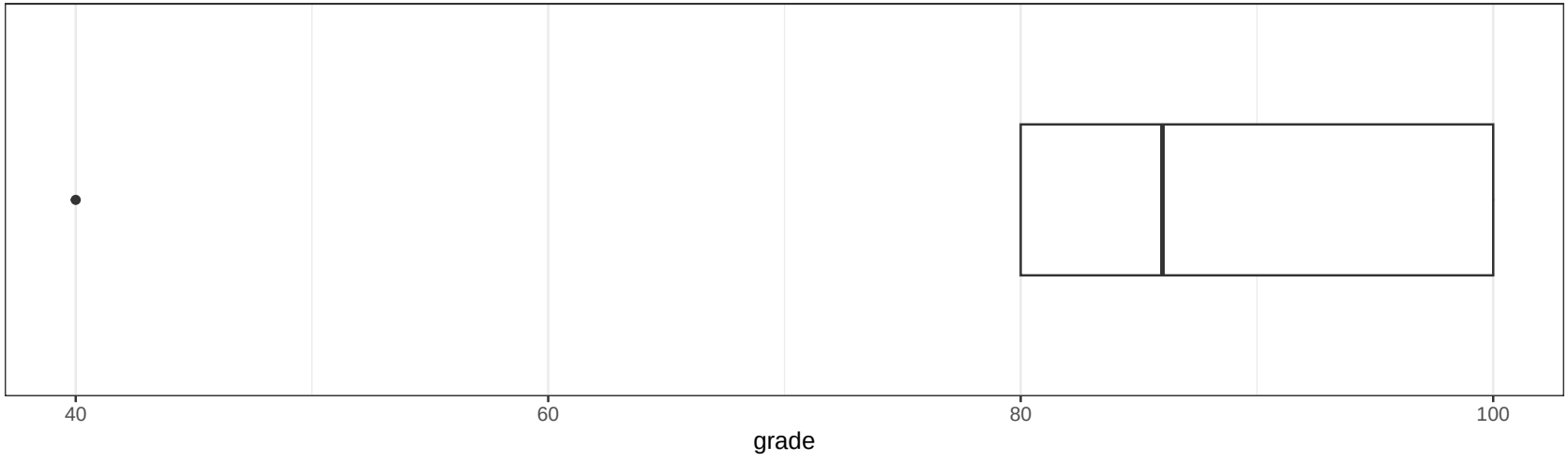


# Recap: boxplot

- No observations in $[LF, Q_1]$ range $\Rightarrow$ no lower whisker
- No observations in $[Q_3, UF]$ range $\Rightarrow$ no upper whisker

```
grade = c(40,80,80,80,80,85,86,94,95,100,100,100,100)
quantile(grade)
```

```
##   0%  25%  50%  75% 100%
##   40   80   86  100  100
```



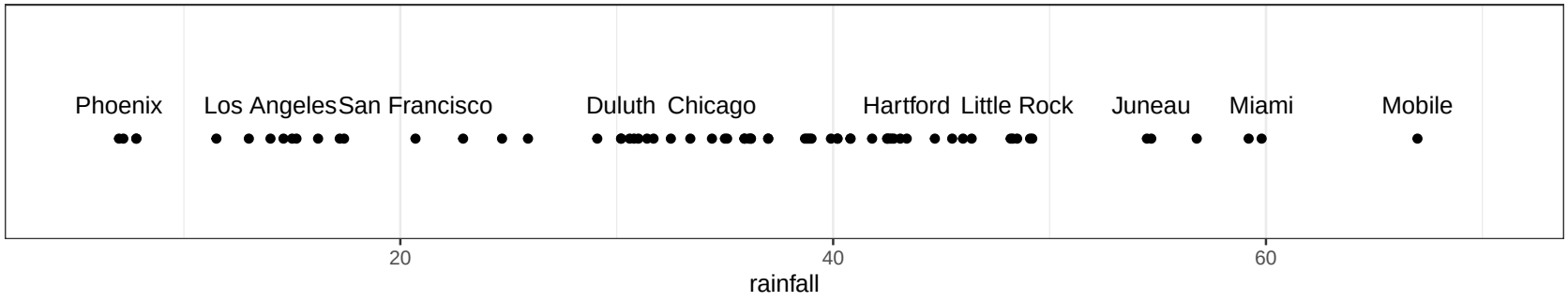
# Plots: histogram

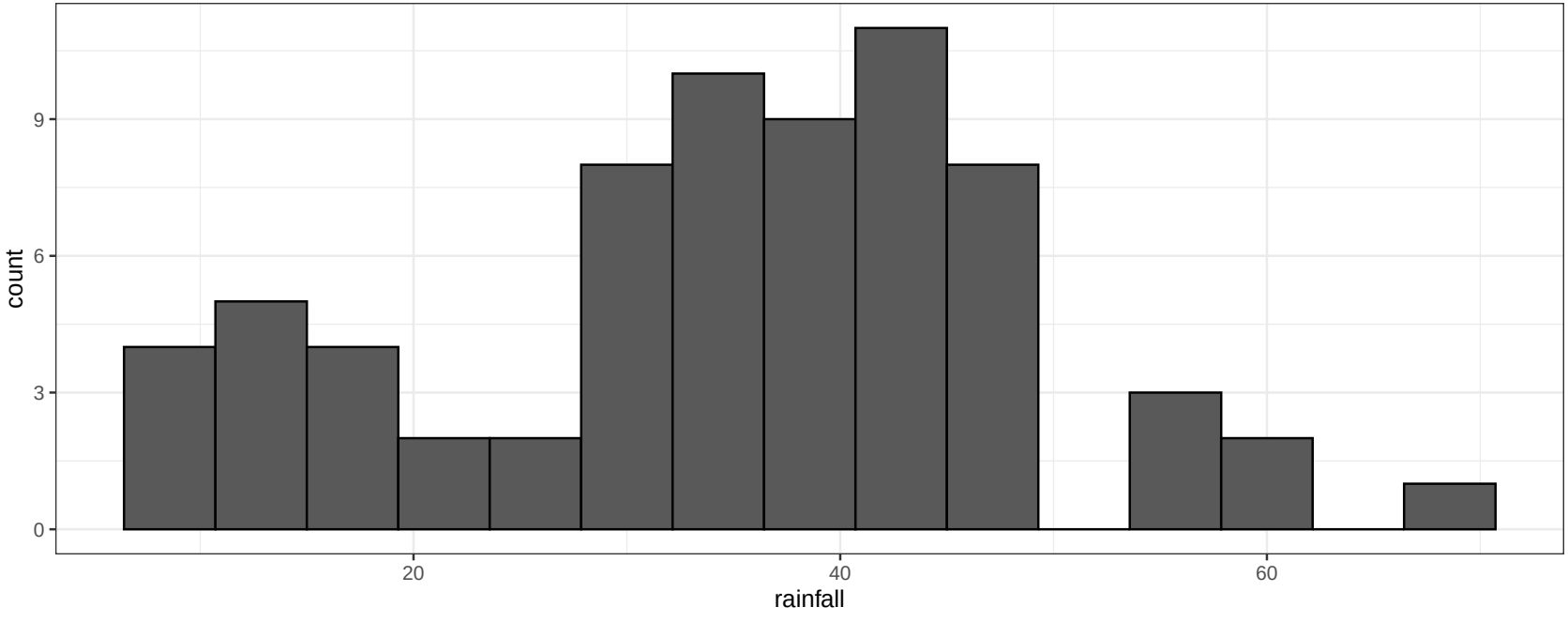
*Data set:* the rainfall level in inches for 69 United States cities

| | rainfall |
|---|---|
| Mobile | 67.0 |
| Juneau | 54.7 |
| Phoenix | 7.0 |
| Little Rock | 48.5 |
| Los Angeles | 14.0 |
| Sacramento | 17.2 |
| San Francisco | 20.7 |
| Denver | 13.0 |
| Hartford | 43.4 |
| Wilmington | 40.2 |
| Washington | 38.9 |
| Jacksonville | 54.5 |
| Miami | 59.8 |
| Atlanta | 48.3 |
| Honolulu | 22.9 |

# Plots: histogram

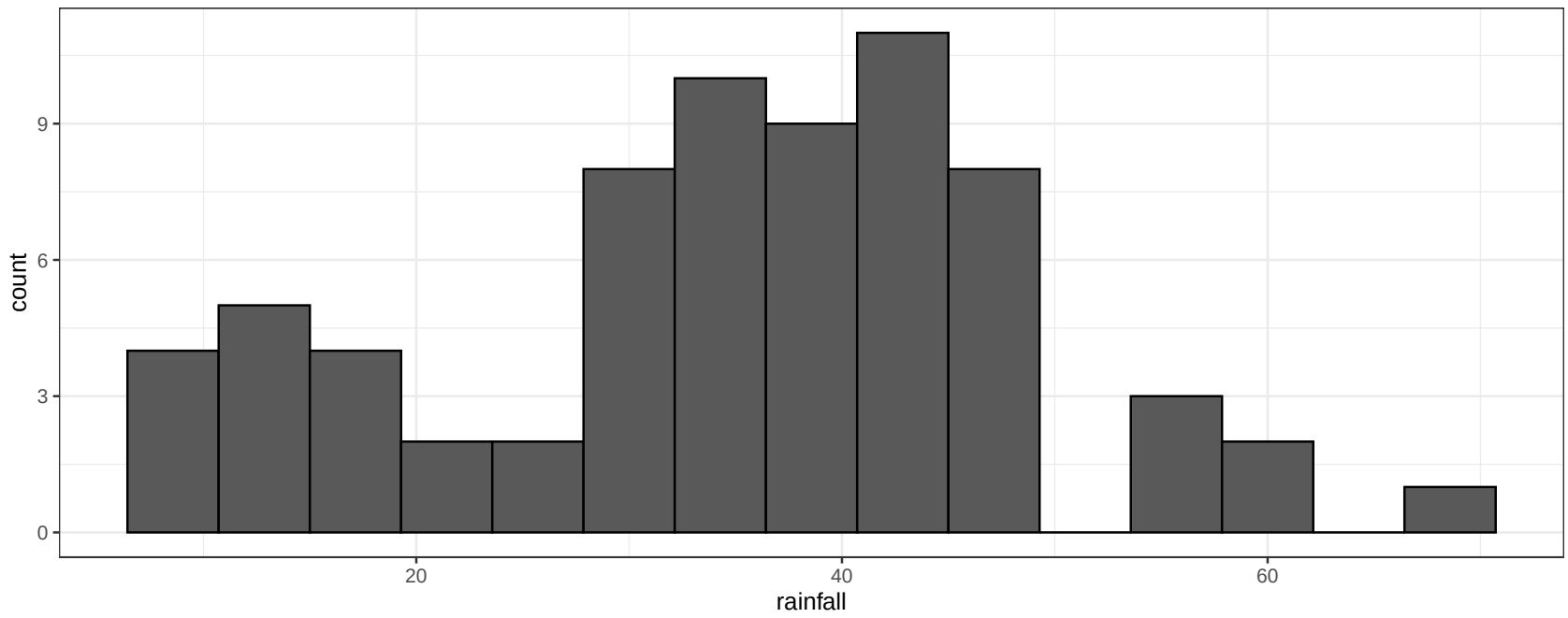
- **Histogram** is used for visualizing data **distibution**



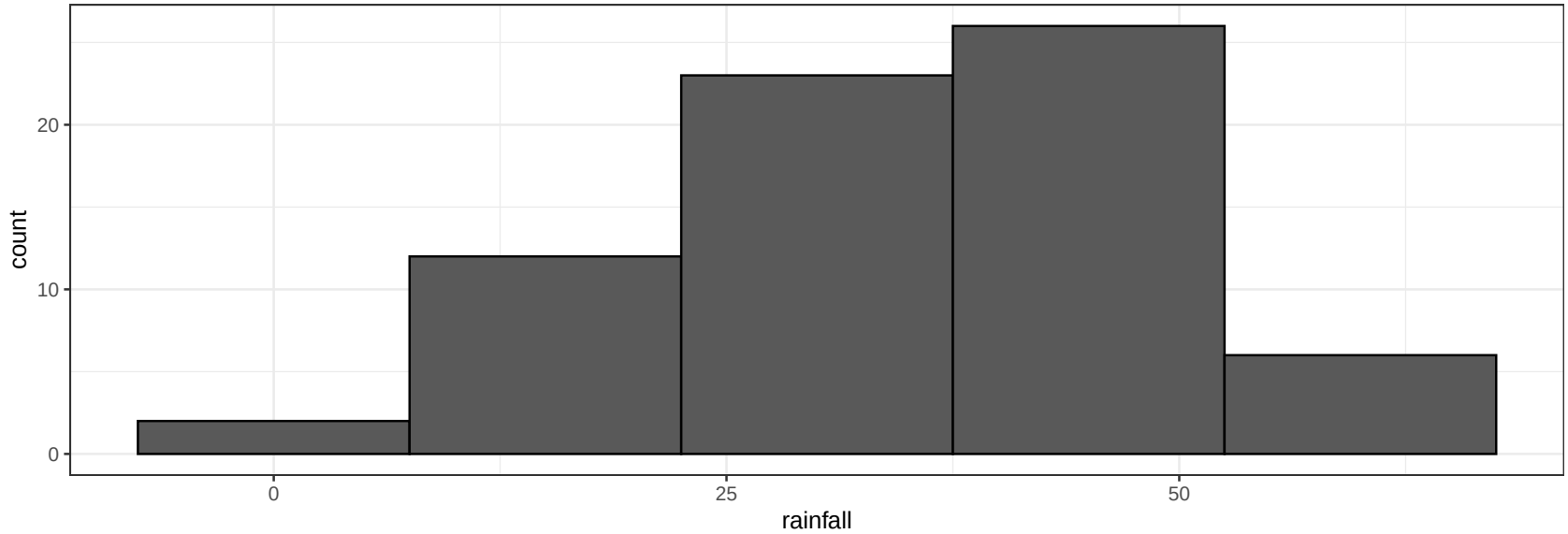
# Plots: histogram

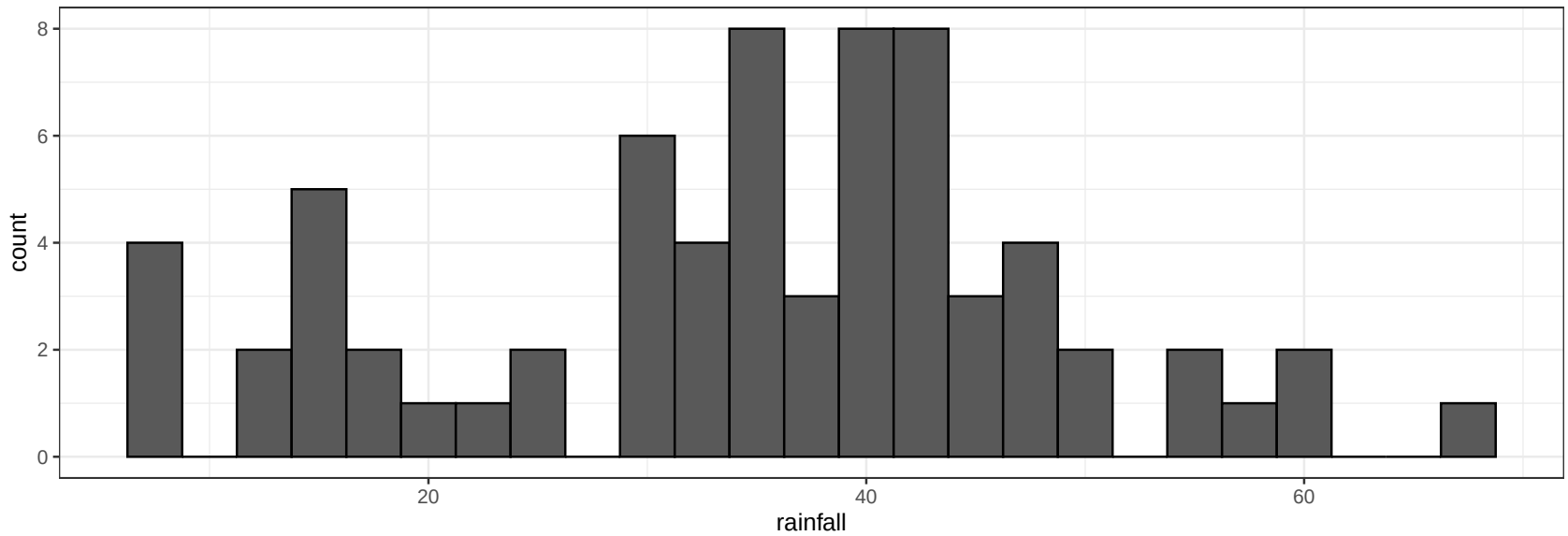
- X-axis is split in **bins**, they should be mutually exclusive and exhaustive
- **Breaks (cutpoints)** are the values that define the beginnings and the ends of the bins
- **Counts (frequencies)** are numbers of data points in each bin (height of each bar)



# Plots: histogram

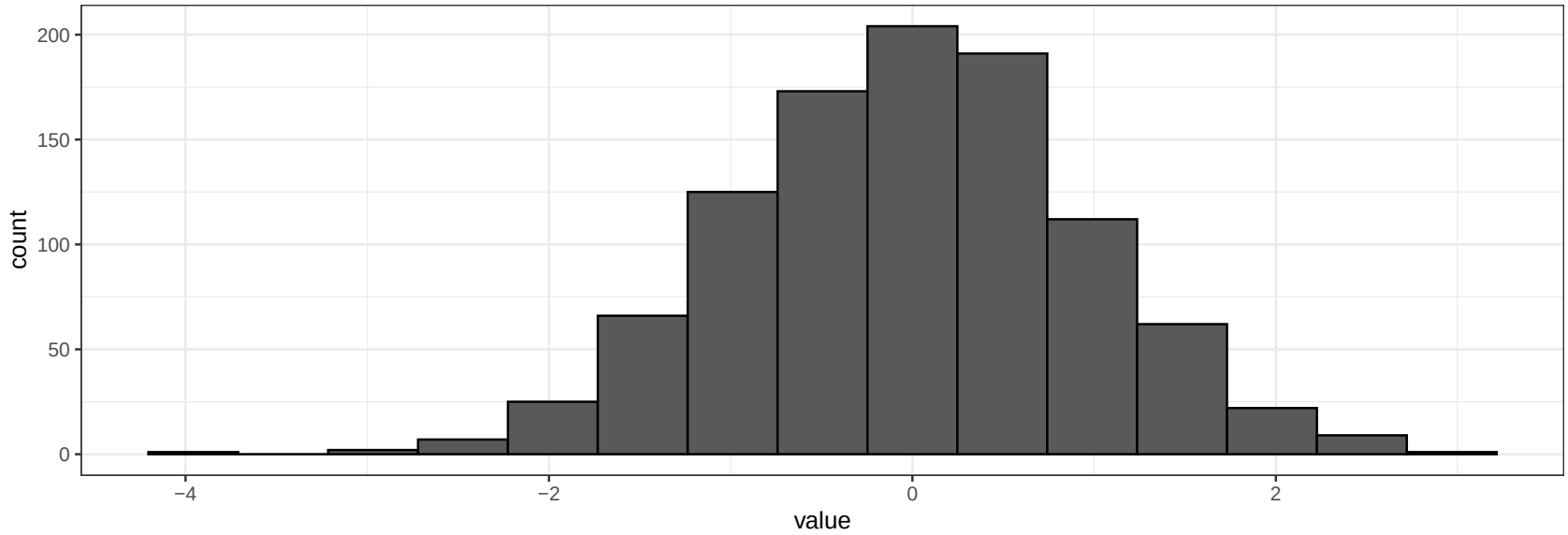
- The appearance of histogram **depends on the cutpoints**

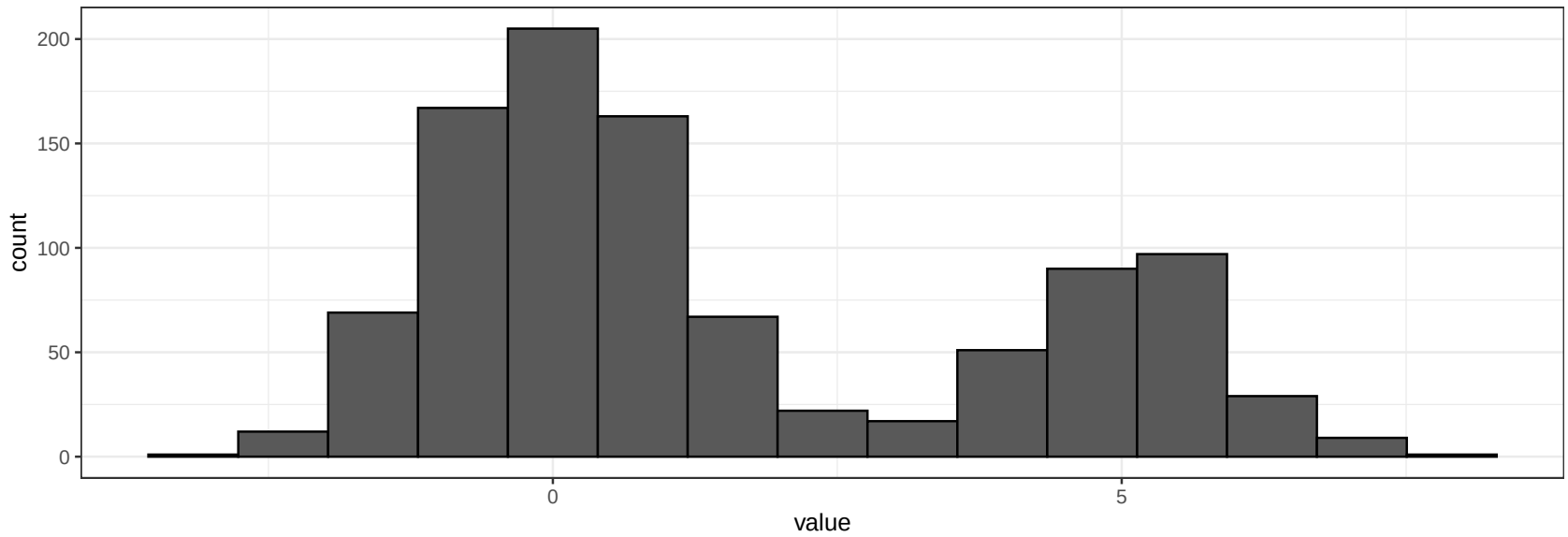


# Plots: histogram

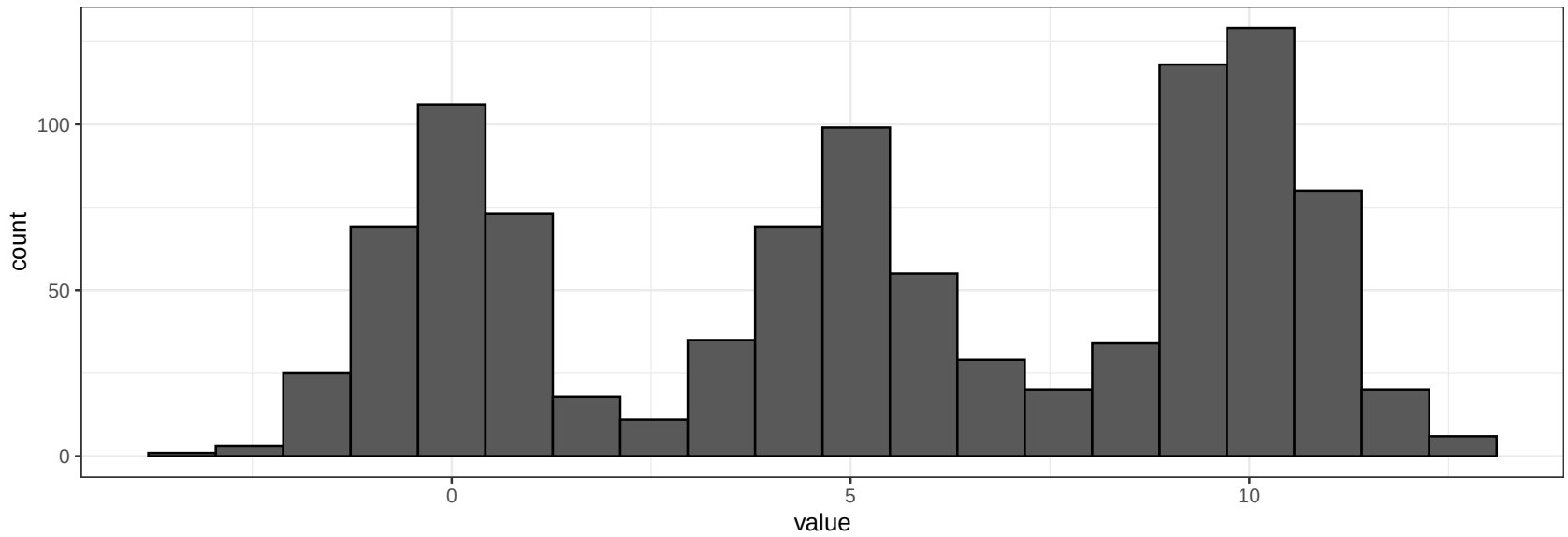
- **Mode** - the peak of the distribution
- Histogram can be **unimodal**, **bimodal**, **multimodal**, **uniform**

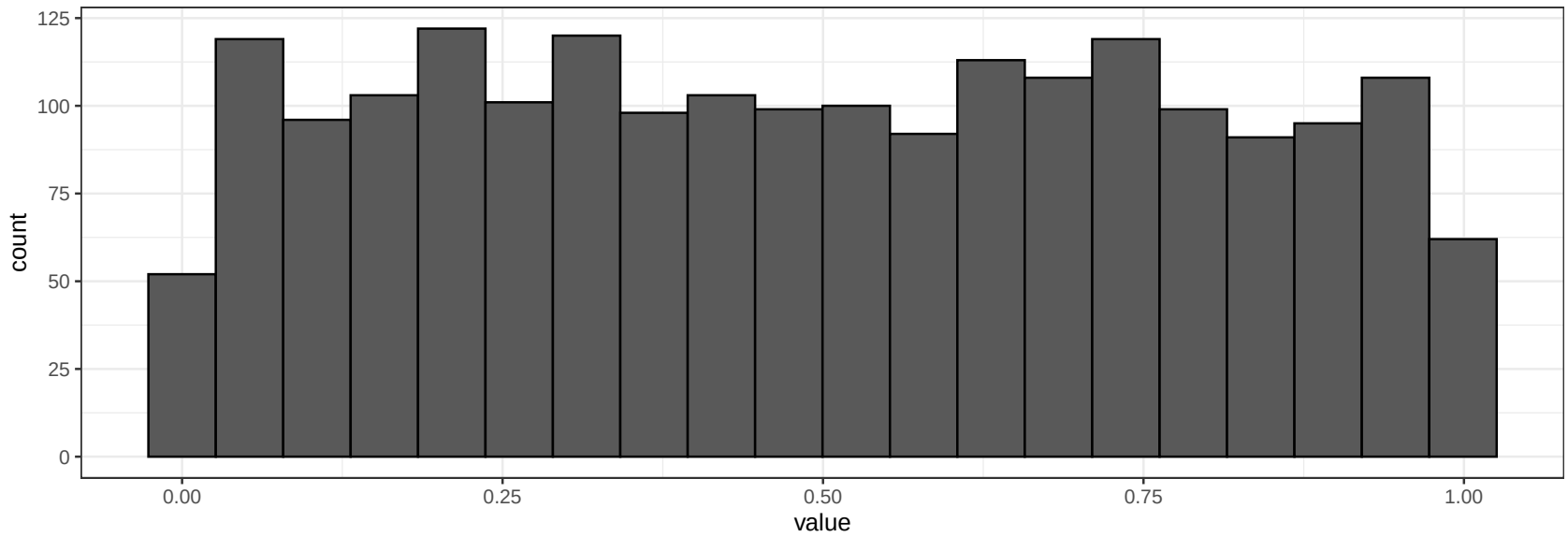


# Plots: histogram

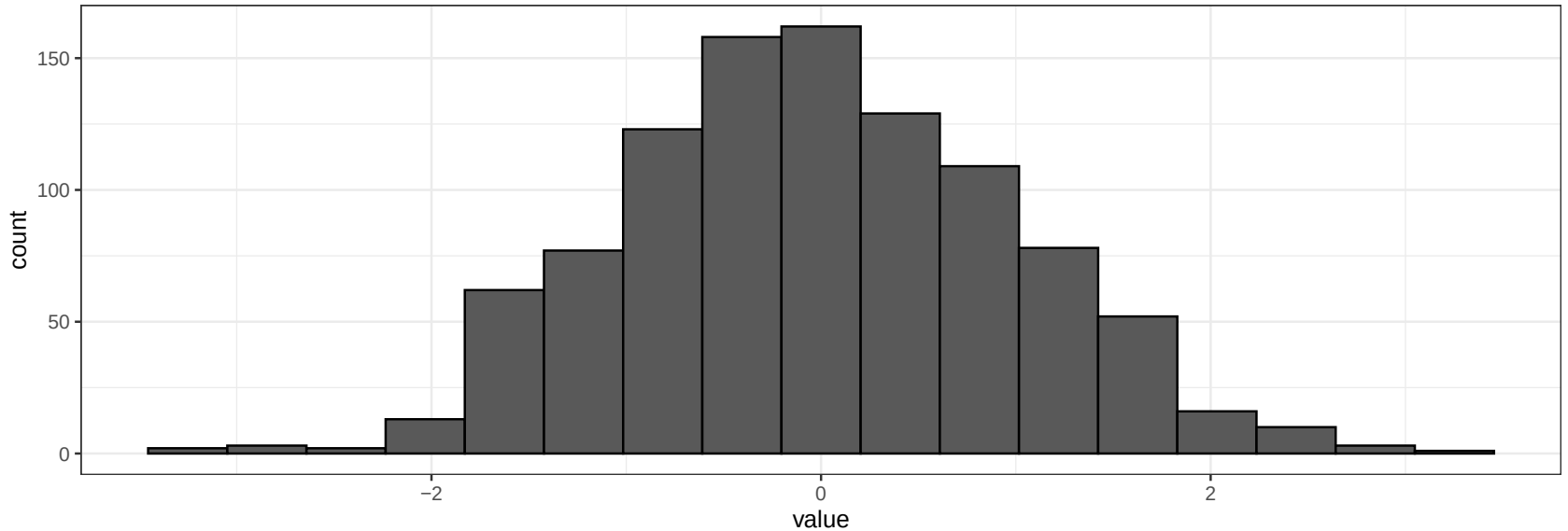
- **Mode** - the peak of the distribution
- Histogram can be **unimodal**, **bimodal**, **multimodal**, **uniform**

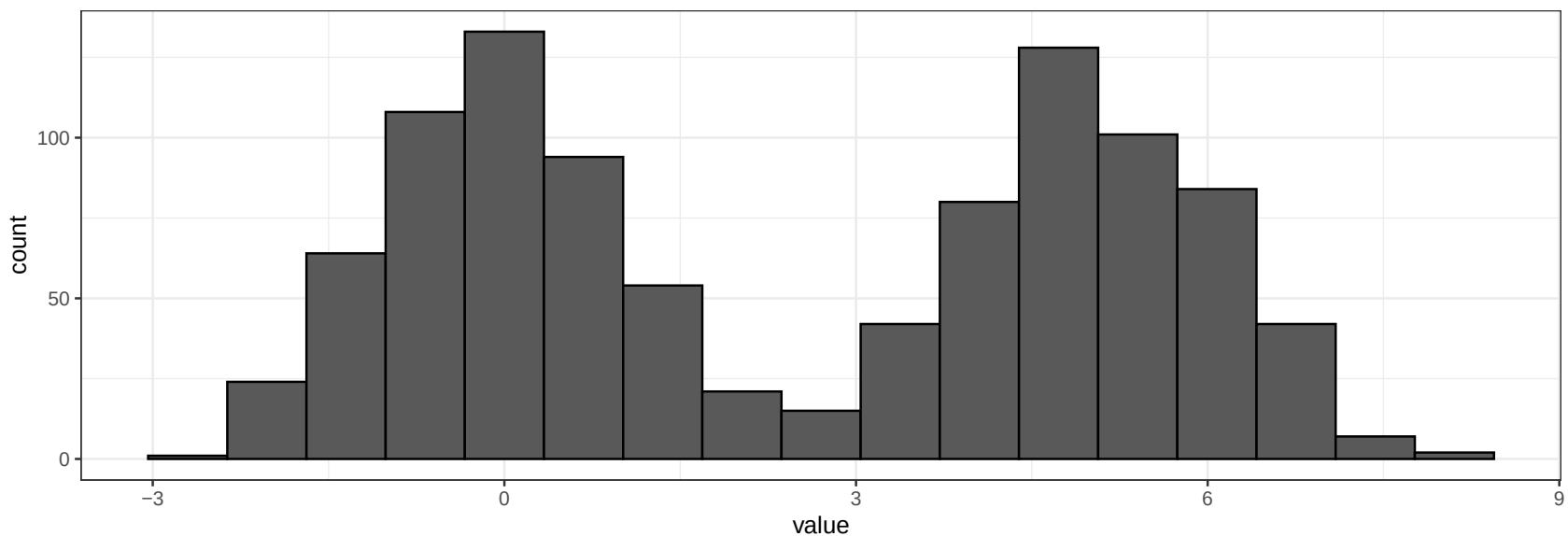


# Plots: histogram

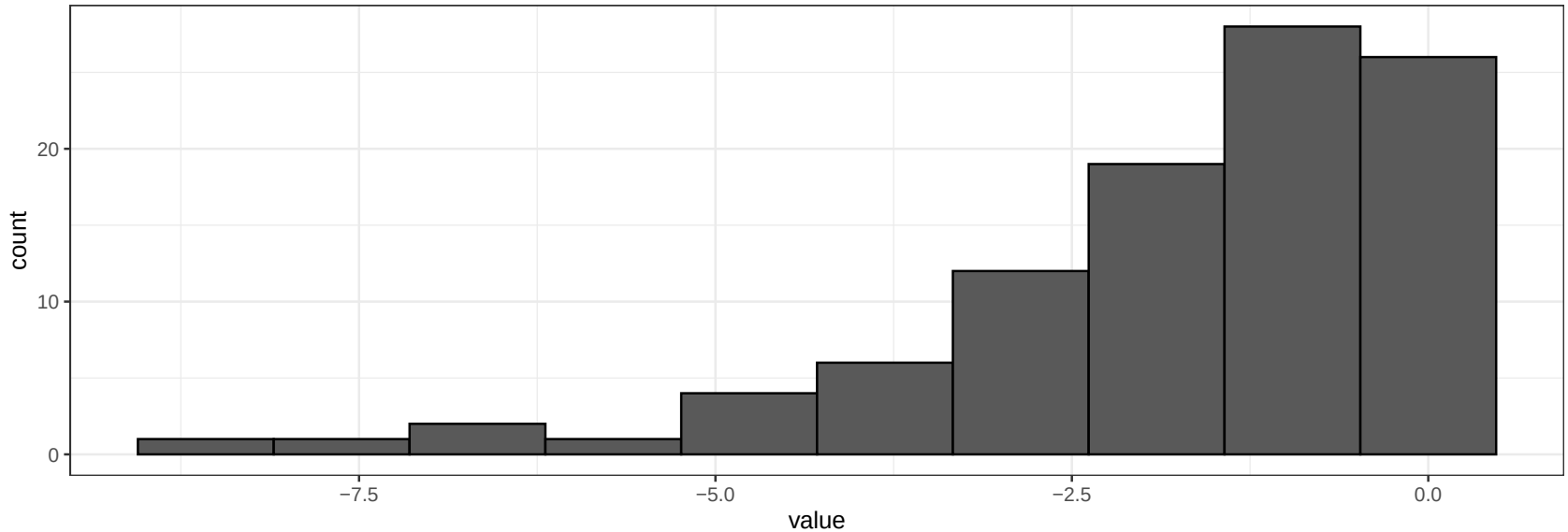
- Histogram can be **symmetric**, **left-skewed** (long left tail), **right-skewed** (long right tail)

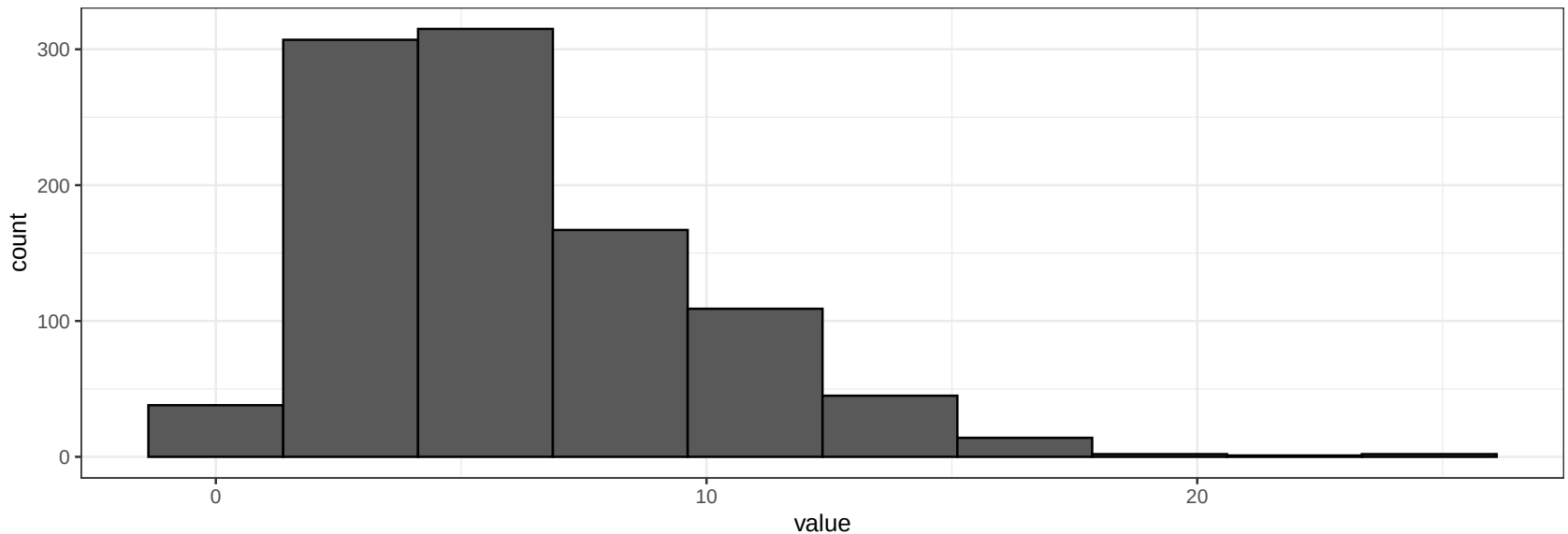


# Plots: histogram

- Histogram can be **symmetric**, **left-skewed** (long left tail), **right-skewed** (long right tail)



# Exercise

For a sample 11,1,2,6,6,6 plot the histogram with cutpoints 0,3,10,15. How many bars are there? How tall is each bar?

# Summary statistics: standard deviation

There are several ways to measure the **spread of the data**

$$IQR = Q_3 - Q_1$$

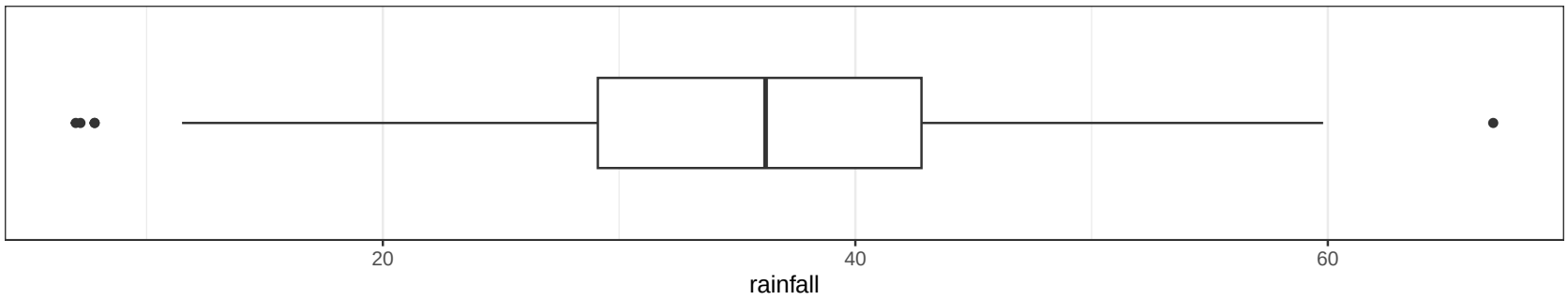
$$range = x_{(n)} - x_{(1)}$$

```
IQR(precip.data$rainfall)
```

```
## [1] 13.7
```

```
max(precip.data$rainfall) - min(precip.data$rainfall)
```

```
## [1] 60
```



# Summary statistics: standard deviation

$$variance = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s_x^2$$

$$standard\ \ deviation = \sqrt{variance} = s_x$$

```
var(precip.data$rainfall)
```

```
## [1] 190.5252
```

```
sd(precip.data$rainfall)
```

```
## [1] 13.80309
```

# Exercise

*Compute standard deviation of the following values:*

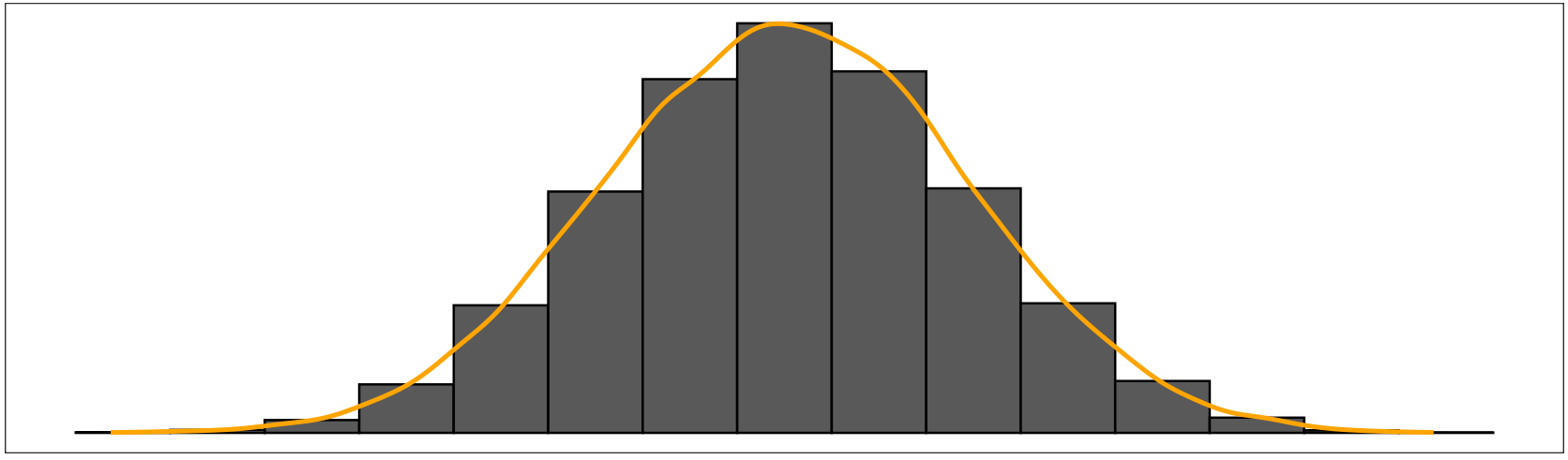
*3, 10, 5, 6, 10, 8?*

```
vec = c(3, 10, 5, 6, 10, 8)
summary(vec)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    5.25    7.00    7.00    9.50   10.00
```

# Summary statistics: standard deviation

There is an **empirical rule** for **symmetric, unimodal, bell-shaped** distributions.



# Summary statistics: standard deviation

- **68%** of the data lies in $[\bar{x} - s_x, \bar{x} + s_x]$
- **95%** of the data lies in $[\bar{x} - 2 \cdot s_x, \bar{x} + 2 \cdot s_x]$
- **99.7%** of the data lies in $[\bar{x} - 3 \cdot s_x, \bar{x} + 3 \cdot s_x]$

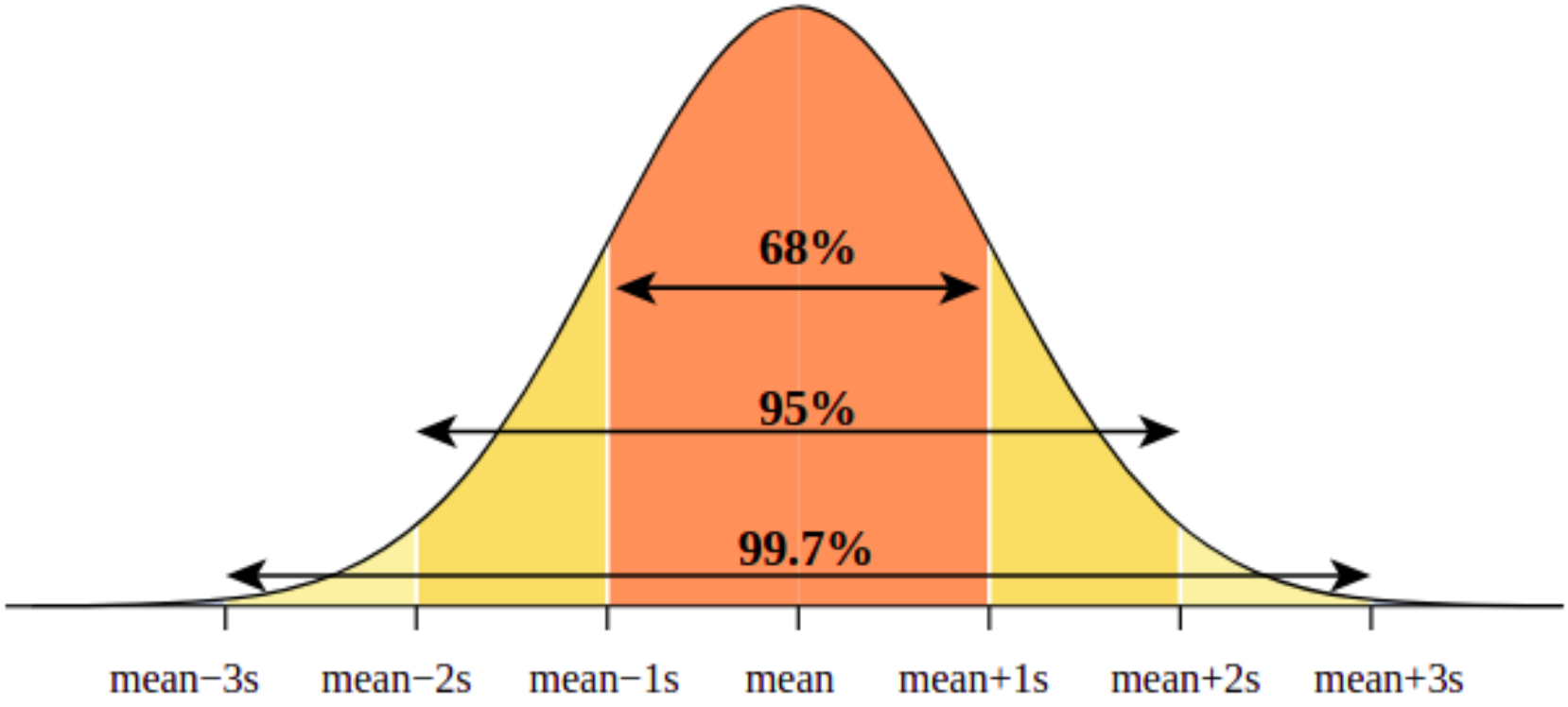


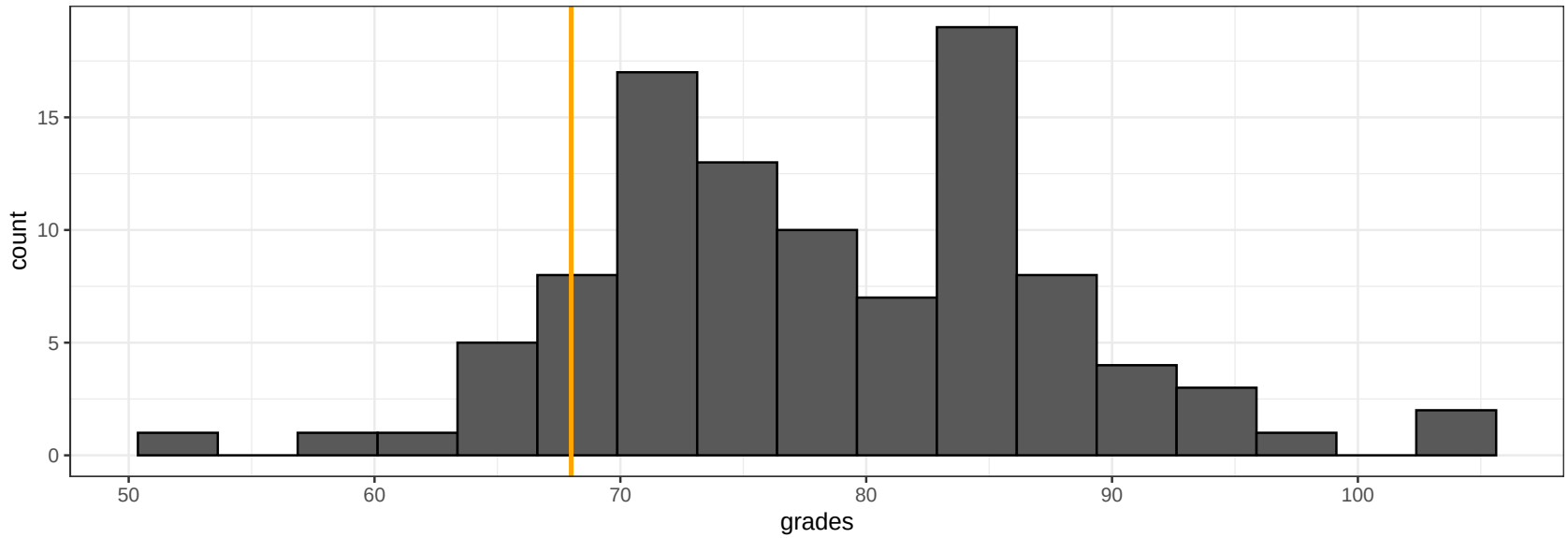

Figure 2: [picture source]

# How bad is my midterm score of 68?

*Option 1*: use a histogram to compare your score to other students.

```
summary(grades)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   51.00   72.00   77.50   78.27   84.25  103.00
```



# How bad is my midterm score of 68?

*Option 2*: quantify your relative performance using z-score.

- **z-score** is an adjustment of a data value to get its position in a data set
- It tells you how many standard deviations a data value is away from its mean

$$z = \frac{x - \bar{x}}{s_x}$$

```
(mygrade - mean(grades))/sd(grades)
```

```
## [1] -1.126134
```

# Data summary: one quantitative variable

- Compute **numerical summary (summary statistics)** - mean, minimum, maximum, range, median, quartiles, IQR, standard deviation
- Summarize using **plots** - histogram and boxplot

# Data summary: one categorical variable

- **Numerical summary** is very limited - frequencies, relative frequencies
- Summarize using **plots** - barplot, piechart

# Data summary: one categorical variable

*Data set*: an experiment was conducted to measure effectiveness of various feed supplements on the growth rate of 71 chickens

| weight | feed |
|---|---|
| 179 | soybean |
| 160 | soybean |
| 136 | soybean |
| 227 | soybean |
| 217 | soybean |
| 168 | soybean |
| 108 | soybean |
| 124 | soybean |
| 143 | soybean |
| 140 | soybean |
| 309 | linseed |
| 229 | linseed |
| 181 | linseed |
| 141 | linseed |
| 260 | linseed |

# Numerical summary: distribution

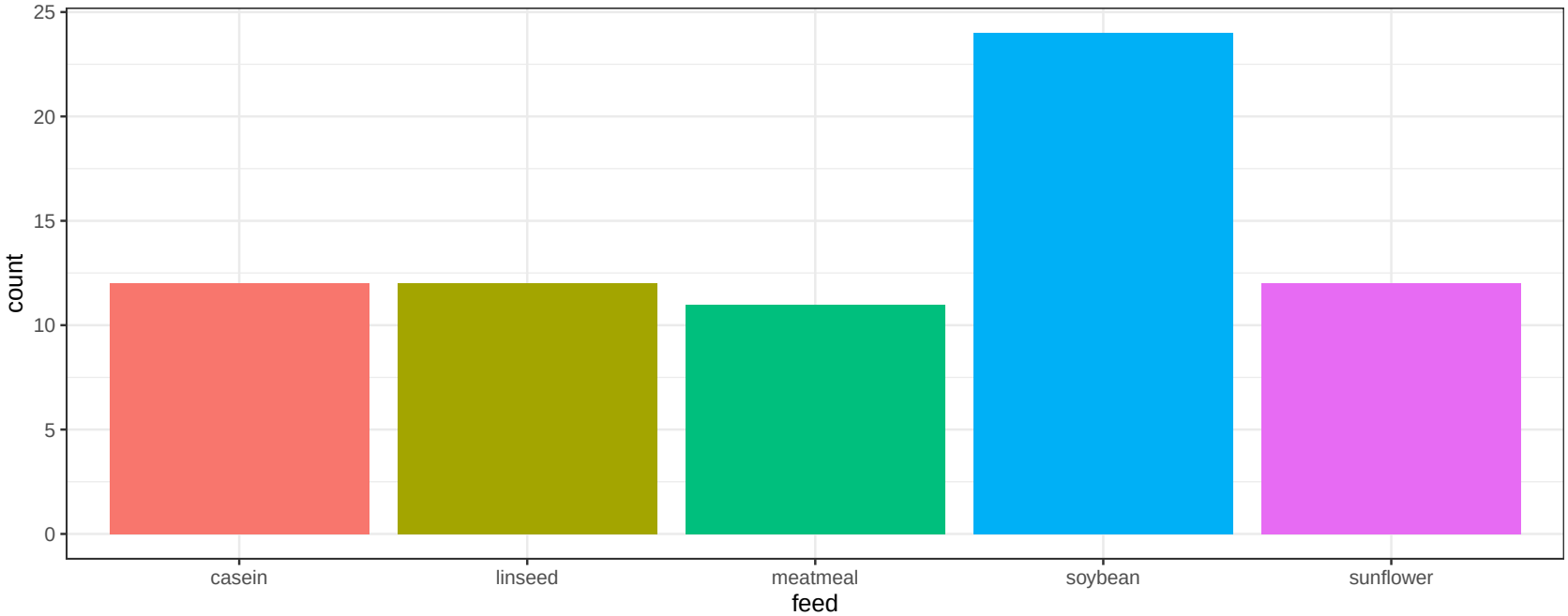
- **Distribution** describes how data are divided between different possible values
- **Frequencies** measure how many observations are in each category

```
tab = table(chick.data$feed)
tab
```

```
##
##    casein   linseed  meatmeal   soybean sunflower
##        12        12        11        24        12
```

# Plots: barplot

- In a sense, this is an analogue of a histogram



# Numerical summary: distribution

- **Distribution** describes how data are divided between different possible values
- **Relative frequencies** measure proportion of observations in each category

```
prop.table(tab)
```

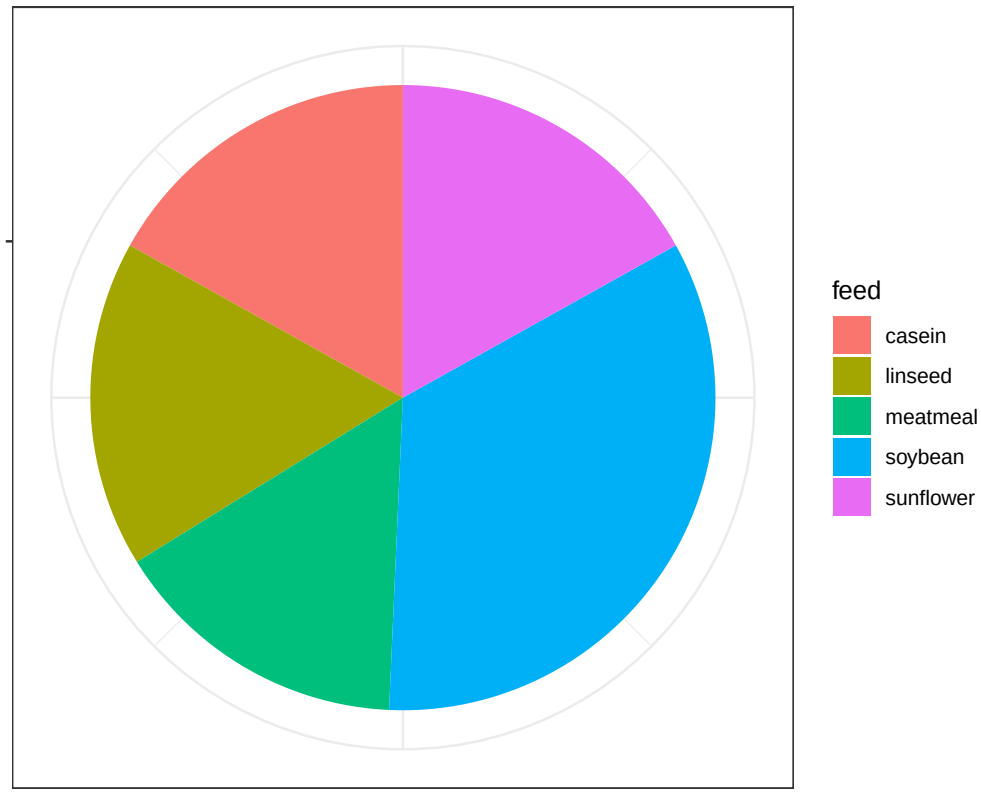
```
##
##    casein   linseed  meatmeal   soybean sunflower
## 0.1690141 0.1690141 0.1549296 0.3380282 0.1690141
```

# Plots: stacked barplot

- All proportions add up to one!

# Plots: piechart

- Size of each slice illustrates the proportion of a category



# Exercise

You get the distribution (frequencies) of pets in the building you live. The information was collected among $n$ students. Can you estimate $n$?

```
##
##      cat      dog     fish hamster  iguana     none
##       15       12        1        4        3       15
```

# Data summary: quantitative vs quantitative variables

- **Summary statistics** - correlation
- Use **plots** - scatterlplot

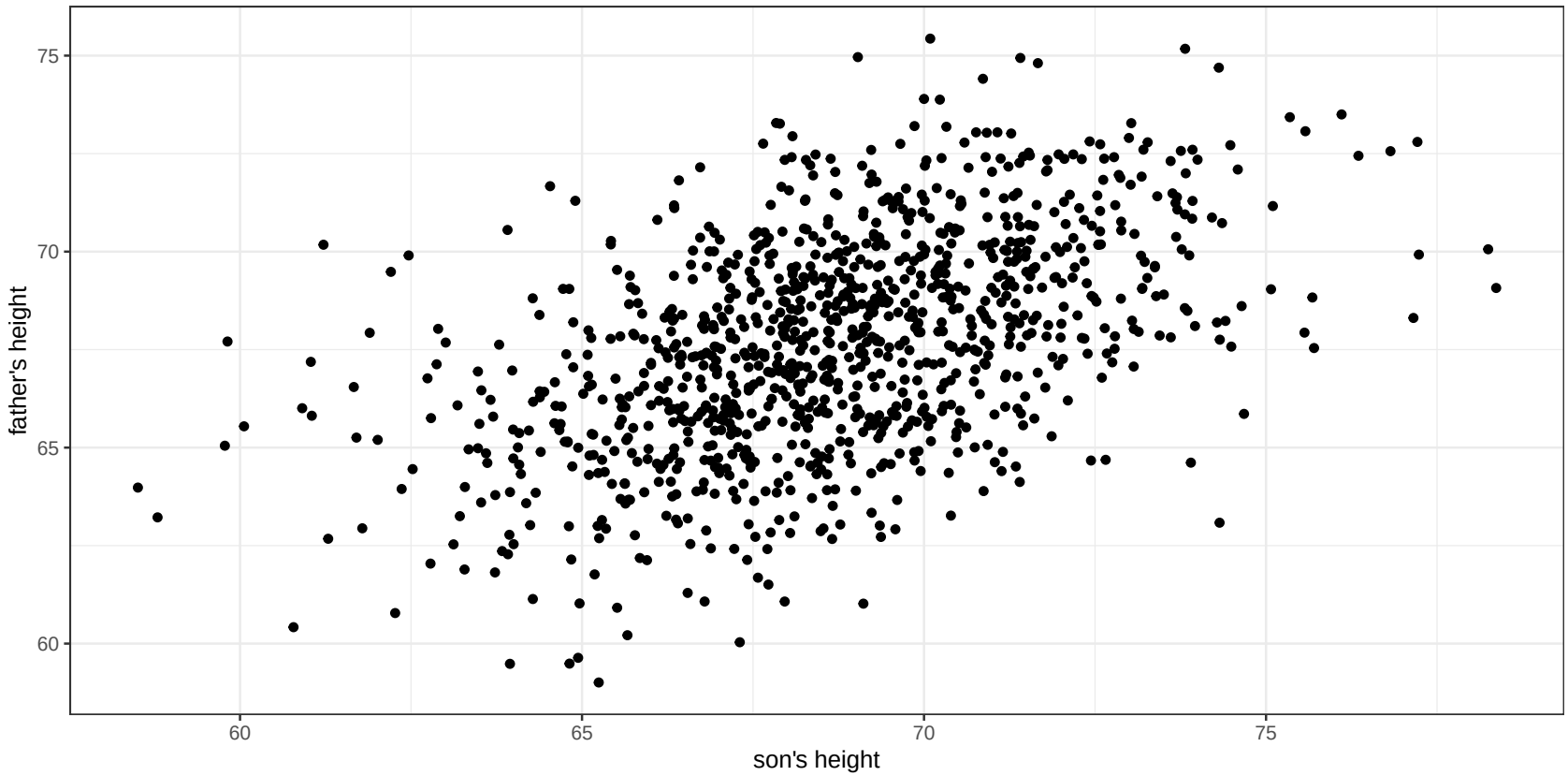
# Data summary: quantitative vs quantitative variables

*Data set*: 1078 measurements of a father's height and his son's height.

| fheight | sheight |
|---|---|
| 65.04851 | 59.77827 |
| 63.25094 | 63.21404 |
| 64.95532 | 63.34242 |
| 65.75250 | 62.79238 |
| 61.13723 | 64.28113 |
| 63.02254 | 64.24221 |
| 65.37053 | 64.08231 |
| 64.72398 | 63.99574 |
| 66.06509 | 64.61338 |
| 66.96738 | 63.97944 |
| 59.00800 | 65.24451 |
| 62.93203 | 65.35102 |
| 63.67063 | 65.67992 |
| 64.07386 | 65.43664 |
| 64.68851 | 65.29391 |

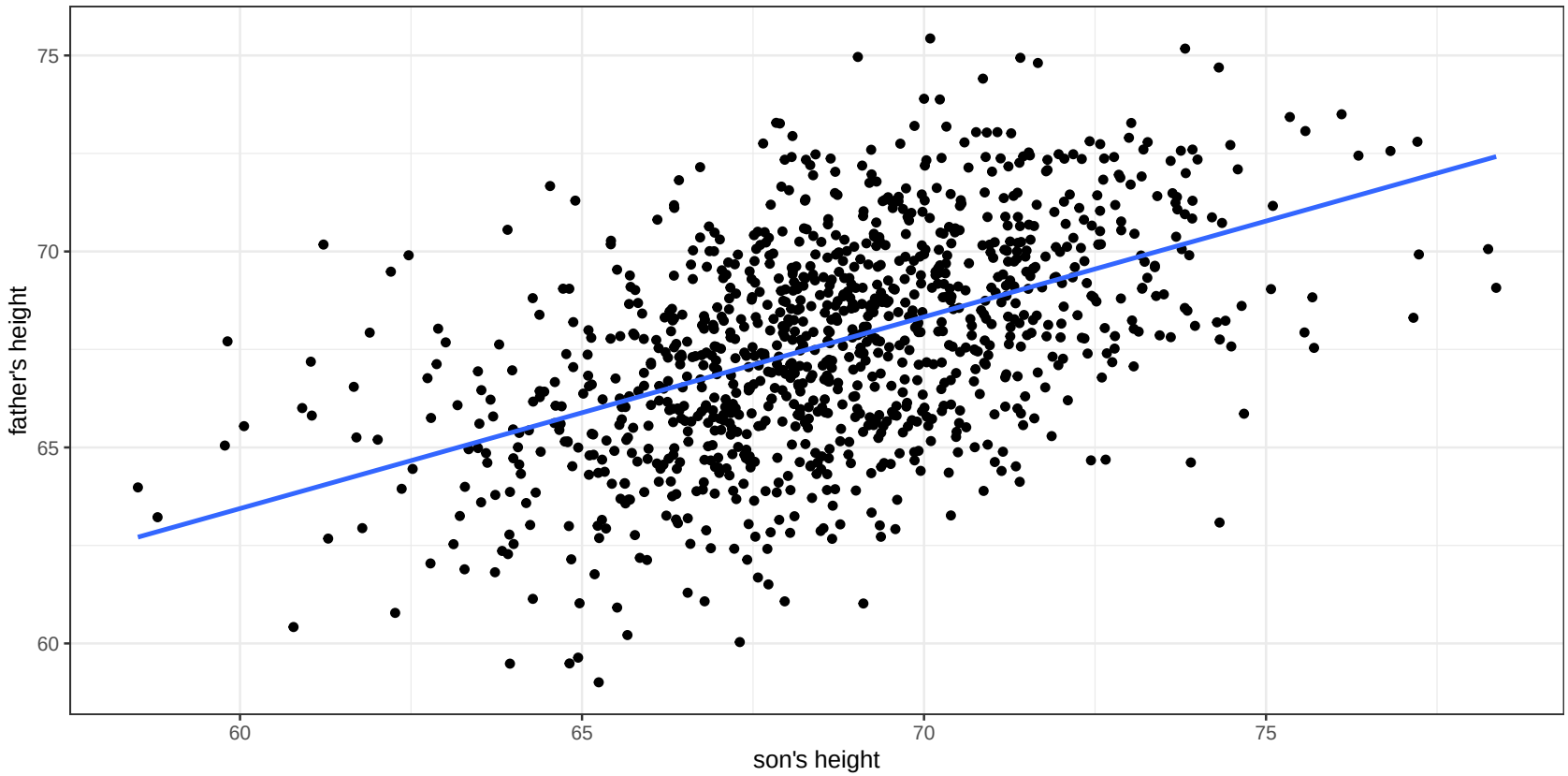
# Plots: scatterplot

- Is there any relation between variables?



# Plots: scatterplot

- There seems to be a positive relationship: taller father $\Rightarrow$ taller son



# Summary statistics: covariance

*Can we quantify the trend?*

- $n$ will denote the number of observations
- $x_1, x_2, ..., x_n$ will denote the observations for the first variable
- $y_1, y_2, ..., y_n$ will denote the observations for the second variable

$$covariance = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = cov_{xy}$$

# Summary statistics: covariance

- Positive **covariance** $\Rightarrow$ the variables tend to both increase together
- Negative **covariance** $\Rightarrow$ one variable tends to increase when the other decreases
- But it depends on the scale of variables!

```
cov(father.son.data$sheight, father.son.data$fheight)
```

```
## [1] 3.873333
```

# Summary statistics: correlation

- **Correlation** refers to the scaled form of covariance
- Correlation value is between -1 and 1

$$correlation = \frac{cov_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = r_{xy}$$

# Summary statistics: correlation

*Can we quantify the trend?*

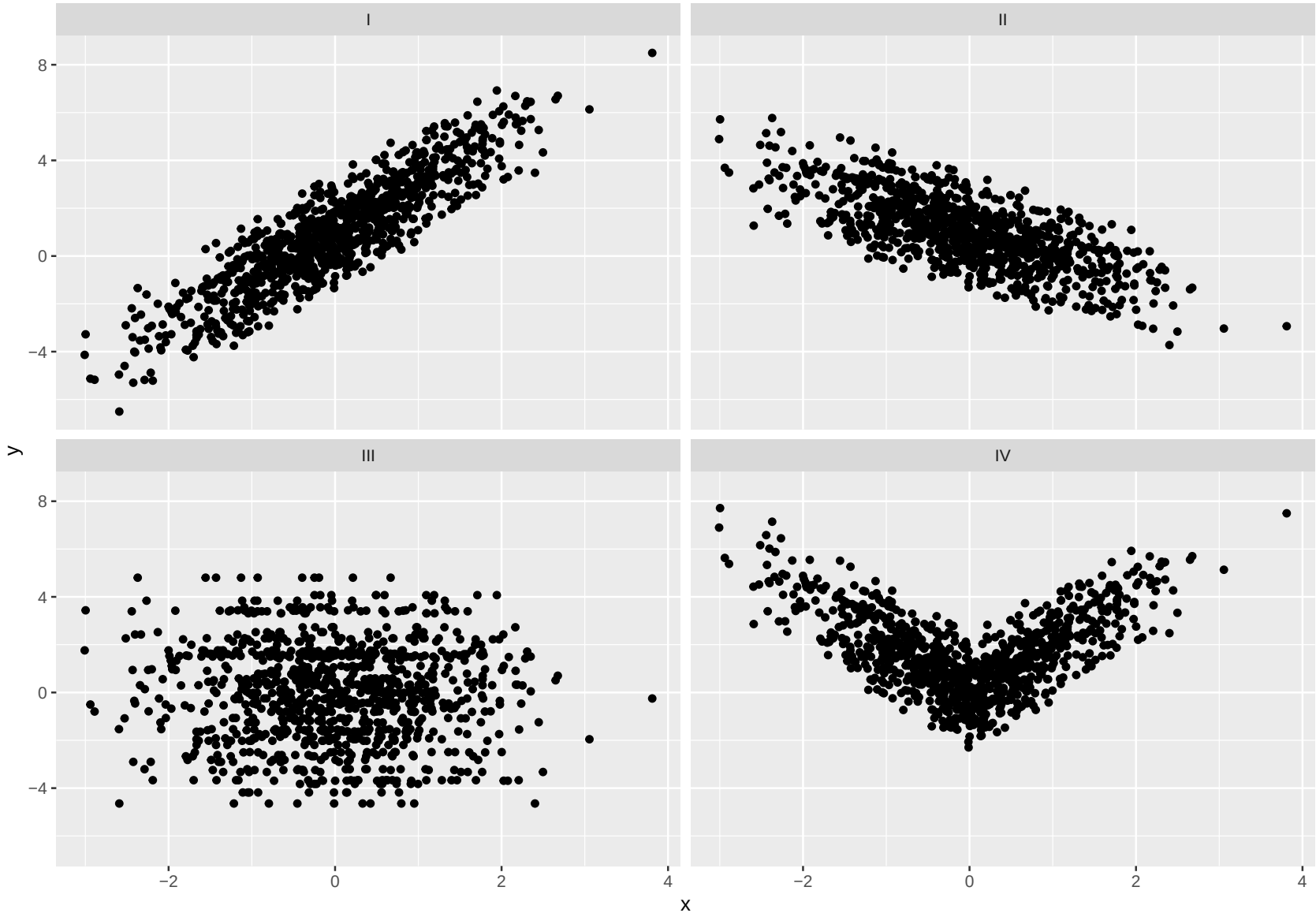
- If there is a perfect linear relationship, e.g. $y_i = a \cdot x_i + b$, then correlation is 1 (if $a > 0$) or $-1$ (if $a < 0$)

```
cor(father.son.data$sheight, father.son.data$fheight)
```

```
## [1] 0.5013383
```

# Exercise

What is the correlation (close to 1,-1 or 0)?



# Data summary: categorical vs quantitative variables

- Compute **summary statistics** - within each category
- Use **plots** - boxplot

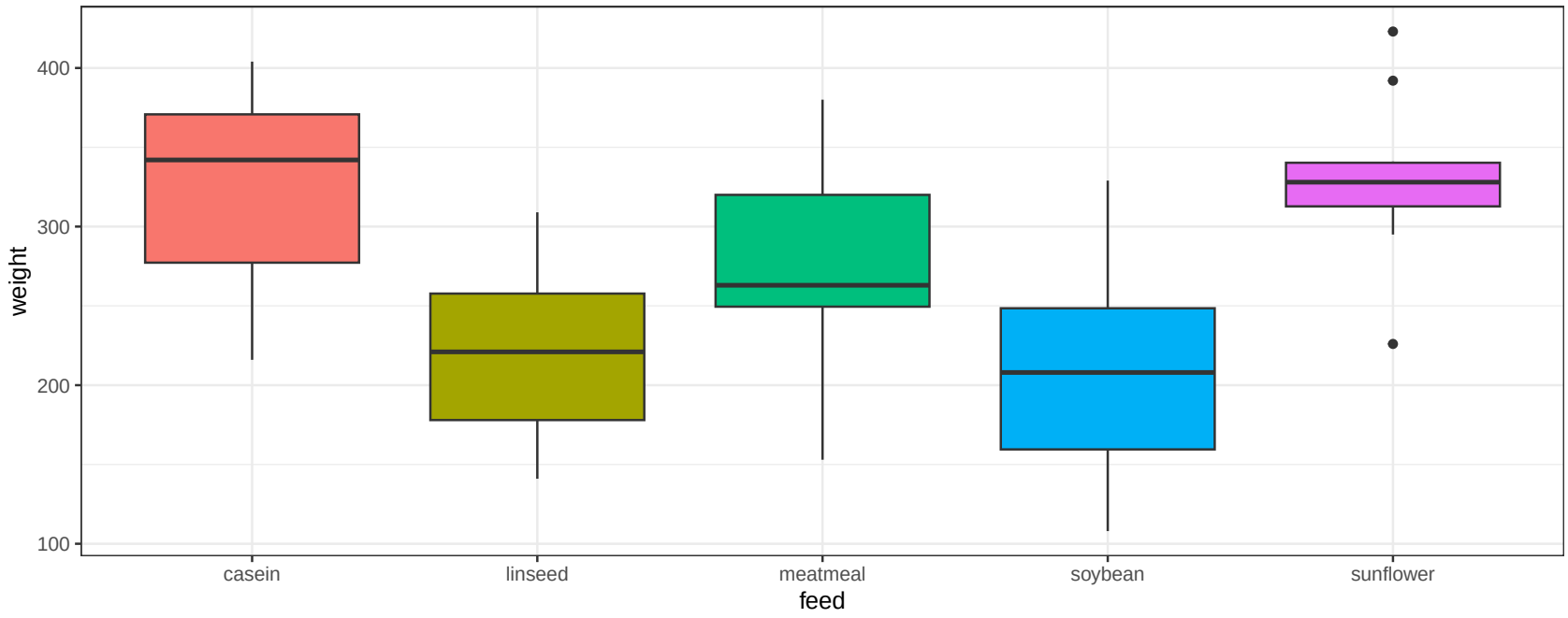
# Summary statistics

- You can compute summary statistics, e.g. mean, median and sd, within each category

| feed | min | max | mean | median | Q1 | Q3 | sd |
|---|---|---|---|---|---|---|---|
| casein | 216 | 216 | 323.5833 | 342 | 277.25 | 277.25 | 64.43384 |
| linseed | 141 | 141 | 218.7500 | 221 | 178.00 | 178.00 | 52.23570 |
| meatmeal | 153 | 153 | 276.9091 | 263 | 249.50 | 249.50 | 64.90062 |
| soybean | 108 | 108 | 210.5000 | 208 | 159.50 | 159.50 | 64.23124 |
| sunflower | 226 | 226 | 328.9167 | 328 | 312.75 | 312.75 | 48.83638 |

# Plots: boxplot

- Use x-axis for different categories
- This method is good, but sometimes it is really hard to say if the difference is significant



# Data summary: categorical vs categorical variables

- **Numerical summary** is very limited - frequencies and relative frequencies
- Use **plots** - barplot

# Data summary: categorical vs categorical variables

*Data set*: provides information on the fate of 891 passengers on the fatal maiden voyage of the ocean liner "Titanic", summarized according to economic status (class), sex, age and survival.

| PassengerId | Sex | Age | Class | Survived |
|---|---|---|---|---|
| 1 | male | 22 | 3 | No |
| 2 | female | 38 | 1 | Yes |
| 3 | female | 26 | 3 | Yes |
| 4 | female | 35 | 1 | Yes |
| 5 | male | 35 | 3 | No |
| 6 | male | NA | 3 | No |
| 7 | male | 54 | 1 | No |
| 8 | male | 2 | 3 | No |
| 9 | female | 27 | 3 | Yes |
| 10 | female | 14 | 2 | Yes |
| 11 | female | 4 | 3 | Yes |
| 12 | female | 58 | 1 | Yes |
| 13 | male | 20 | 3 | No |
| 14 | male | 39 | 3 | No |
| 15 | female | 14 | 3 | No |

# Numerical summary: joint distribution

*Is it true that rich people (e.g. 1st class passengers) survived more often that poor people (e.g. 3rd class passengers)?*

```
table(titanic.data$Class)
```

```
##
##   1   2   3
## 216 184 491
```

```
table(titanic.data$Survived)
```

```
##
##  No Yes
## 549 342
```

# Numerical summary: joint distribution

- **Joint distribution** is the frequency/relative frequency of observations for a combination of two variables

```
tab = table(titanic.data$Class, titanic.data$Survived)
tab
```

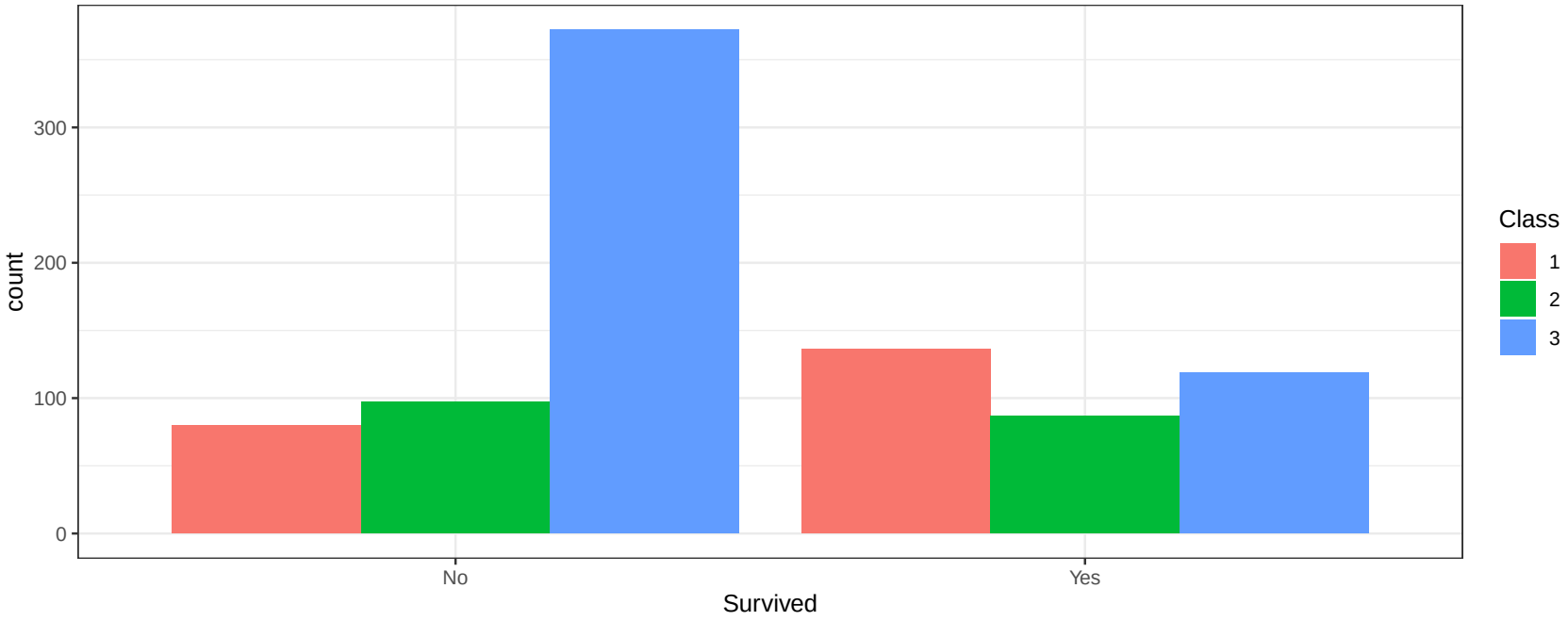
```
##
##       No Yes
##    1  80 136
##    2  97  87
##    3 372 119
```

```
ptab = prop.table(tab)
ptab
```

```
##
##              No        Yes
##    1 0.08978676 0.15263749
##    2 0.10886644 0.09764310
##    3 0.41750842 0.13355780
```

# Plots: barplot

- There are many 3rd class passengers that did not survive
- But it is hard to compare as there were many people who did not survive



# Numerical summary: marginal distribution

- **Marginal distribution** is the frequency/relative frequency of only one variable

```
addmargins(tab)
```

```
##
##         No Yes Sum
##    1    80 136 216
##    2    97  87 184
##    3   372 119 491
##    Sum 549 342 891
```

# Numerical summary: conditional distribution

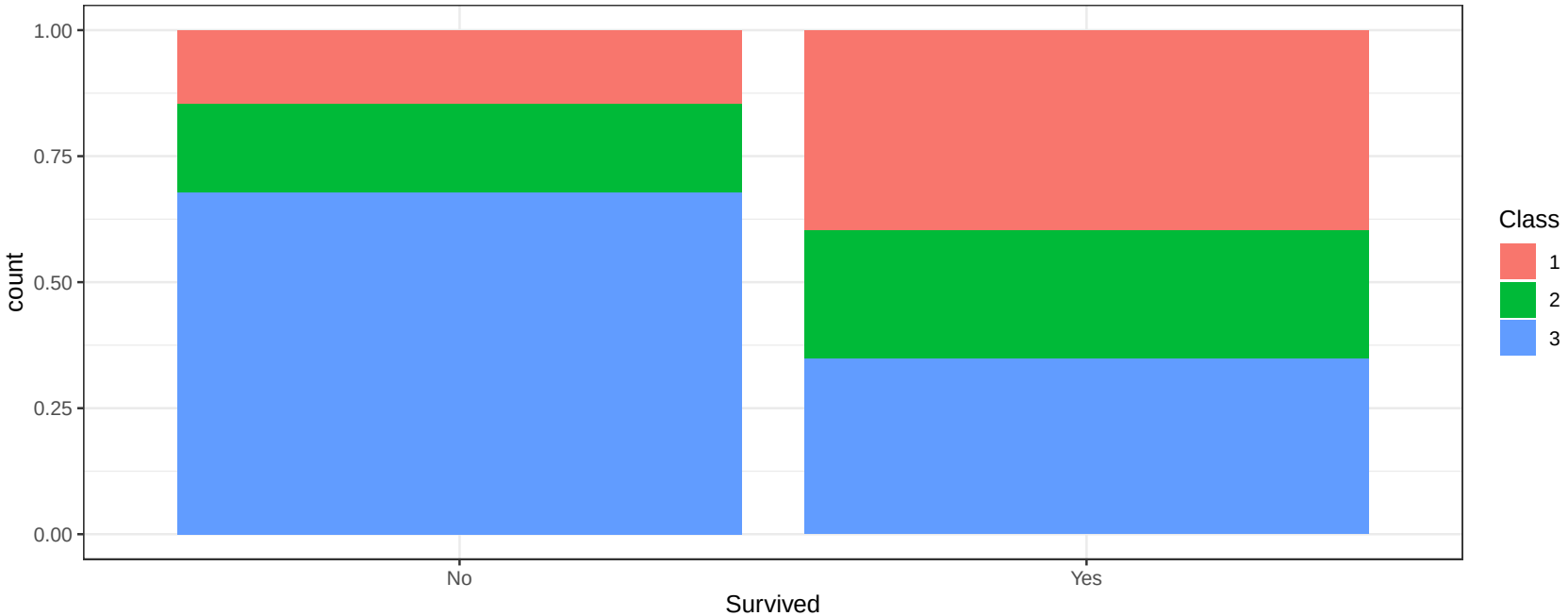
- **Conditional distribution** is the distribution of one variable within a fixed value of a second value
- Comparing conditional distributions for each cetegory can tell if there is any relationship between two variables

```
##
##         No Yes
##    1    80 136
##    2    97  87
##    3   372 119
##    Sum 549 342
```

```
##
##                 No       Yes
##    1   0.1457195 0.3976608
##    2   0.1766849 0.2543860
##    3   0.6775956 0.3479532
##    Sum 1.0000000 1.0000000
```

# Plots: stacked barplot

- Two variables are **independent** if conditional distribution of one variable is the same for all values of the other variable



# Exersice

Find conditional distribution of Sex and Survived variables. Do you think there is any relationship?

```
##
##             No Yes
##   female  81 233
##   male   468 109
```

# TO DO

1. Module 1. Summarizing Data: One variable and Module 1. Summarizing Data: Relationships Between Variables
2. Quiz 2 due Monday (January 23) @ 11:59 PM (EST)
3. Practice Problem Set 2