Revised: 1/12/2021

Summarizing Data I

Michael C. Hackett

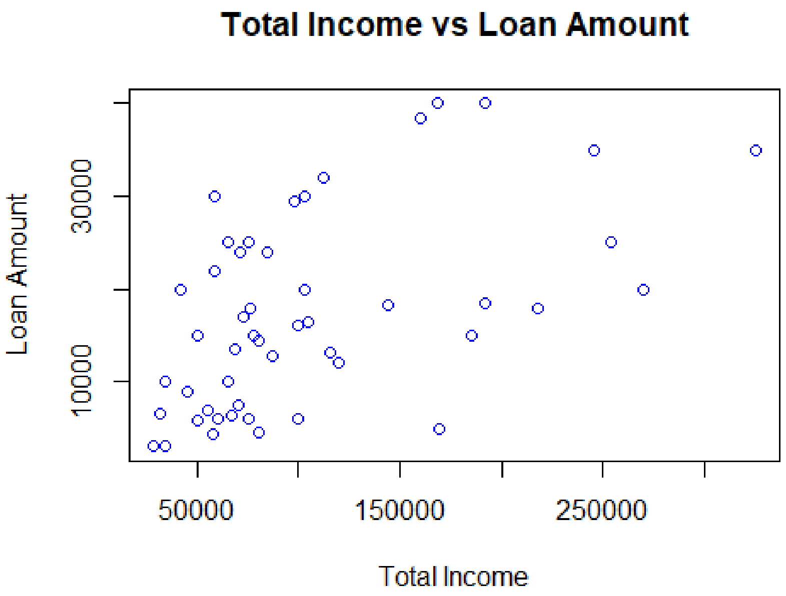
Assistant Professor, Computer Science

Community College of Philadelphia

Scatterplots

- Scatterplots show the relationship between two numerical variables.

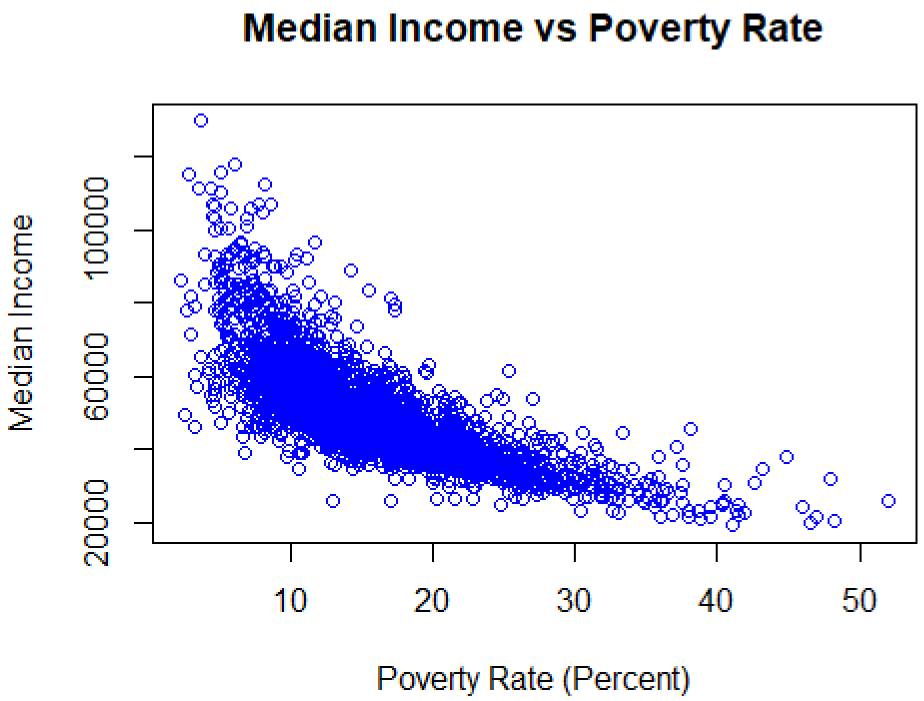
```
library(readr)
loan50 <- read_csv("loan50.csv")
plot(x=loan50$total_income,
     y=loan50$loan_amount,
     main="Total Income vs Loan Amount",
     xlab="Total Income",
     ylab="Loan Amount",
     type="p",
     col="blue")
```



[plot function]
[readr package]

Scatterplots

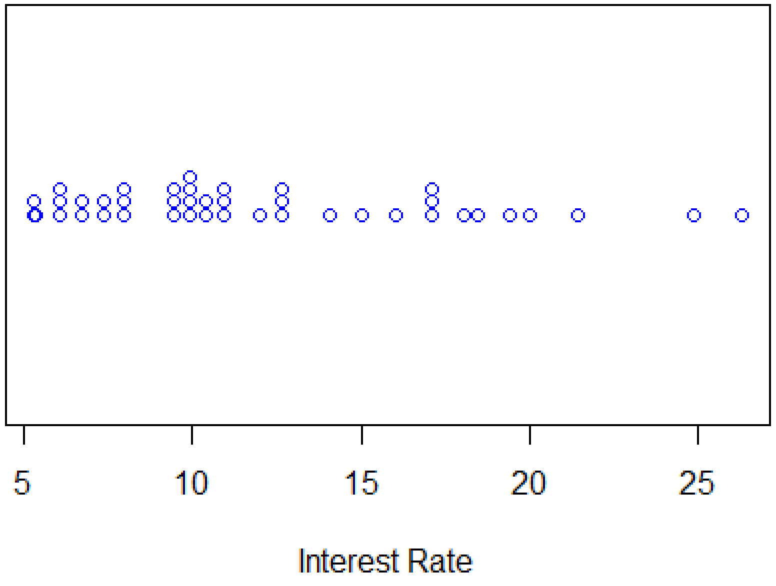
```
library(readr)
county <- read_csv("county.csv")
plot(x=county$poverty,
     y=county$median_hh_income,
     main="Median Income vs Poverty Rate",
     xlab="Poverty Rate (Percent)",
     ylab="Median Income",
     type="p",
     col="blue")
```



Dot Plots

- A one variable scatterplot.
  - Best used with small data sets

```
library(readr)
loan50 <- read_csv("loan50.csv")
stripchart(x=loan50$interest_rate,
     xlab="Interest Rate",
     method="stack",
     pch=21,
     col="blue")
```



[stripchart function]()

Mean

- The **mean** (or average) is one method to find the center of a distribution.
  - The sum of the observed values divided by the total number of observed values.

- The mean is denoted by $\bar{x}$

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Mean

- More specifically, the *sample mean* is denoted by $\bar{x}$
- The *population mean* is denoted by μ

```
library(readr)
loan50 <- read_csv("loan50.csv")
mean(loan50$annual_income)
[1] 86170
```

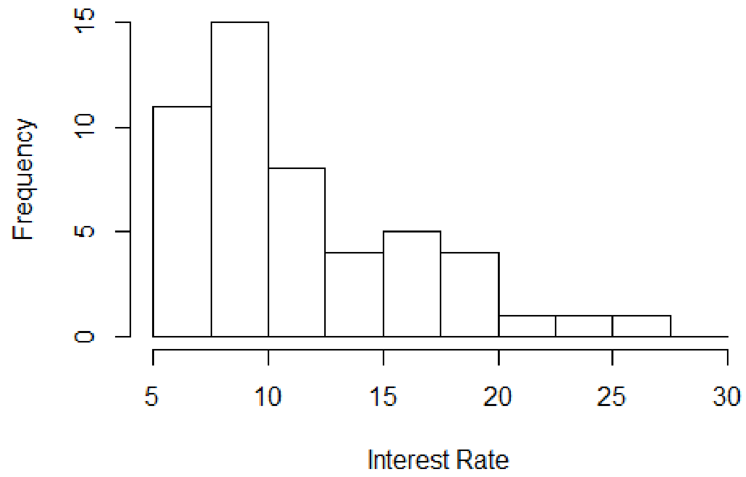
```
> library(readr)
> loan50 <- read_csv("loan50.csv")
Parsed with column specification:
cols(
  state = col_character(),
  emp_length = col_double(),
  term = col_double(),
  homeownership = col_character(),
  annual_income = col_double(),
  verified_income = col_character(),
  debt_to_income = col_double(),
  total_credit_limit = col_double(),
  total_credit_utilized = col_double(),
  num_cc_carrying_balance = col_double(),
  loan_purpose = col_character(),
  loan_amount = col_double(),
  grade = col_character(),
  interest_rate = col_double(),
  public_record_bankrupt = col_double(),
  loan_status = col_character(),
  has_second_income = col_logical(),
  total_income = col_double()
)
> mean(loan50$annual_income)
[1] 86170
```

[mean function]

Histograms

- In a histogram, observed values are placed into “bins”.
  - Histograms show **data density**; Higher bars = fuller bins

```
library(readr)
loan50 <- read_csv("loan50.csv")
hist(x=loan50$interest_rate,
          breaks=seq(5, 30, 2.5),
          xlab="Interest Rate",
          main="")
```



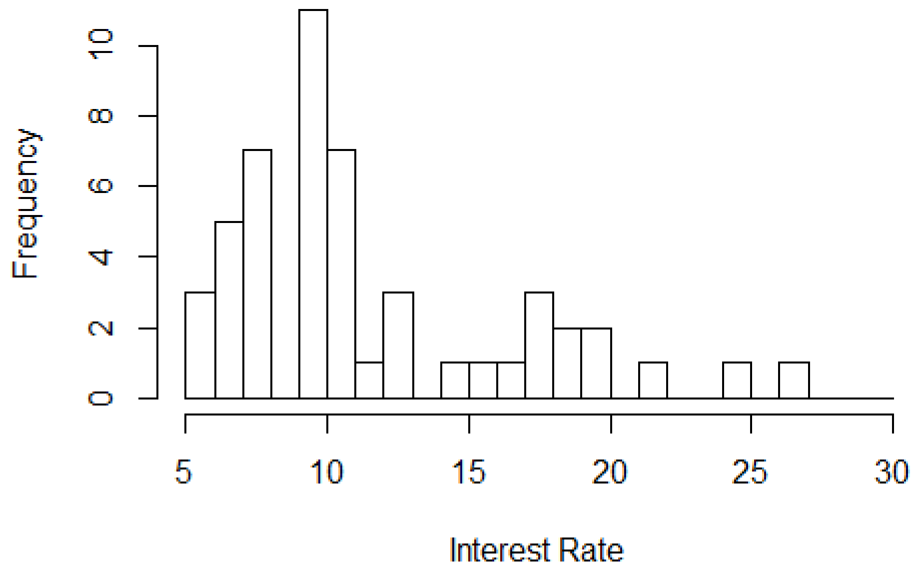
One bin at every 2.5 steps between 5 and 30

[hist function](#)
[seq function](#)

Histograms

- One bin at every step

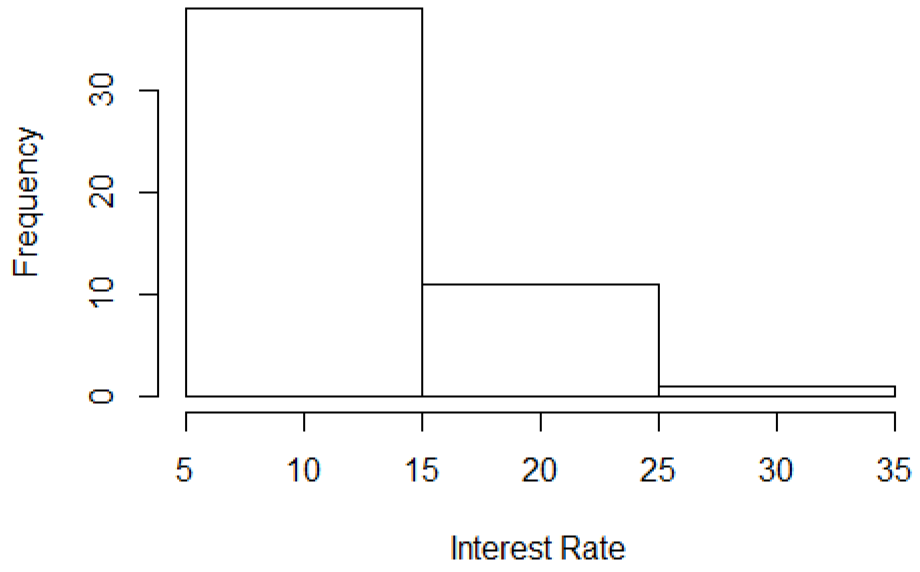
```
library(readr)
loan50 <- read_csv("loan50.csv")
hist(x=loan50$interest_rate,
          breaks=seq(5, 30, 1),
          xlab="Interest Rate",
          main="")
```



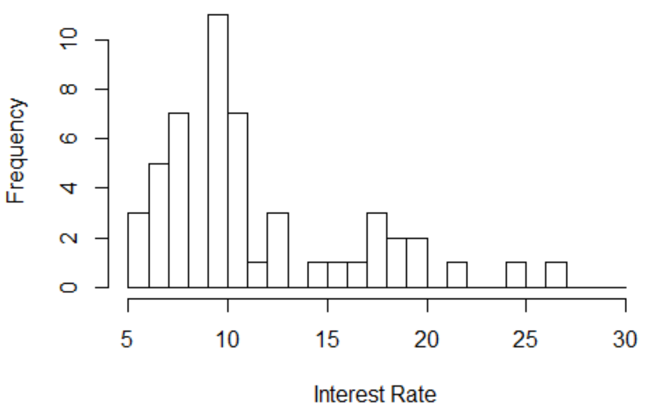
Histograms

- One bin at every ten steps

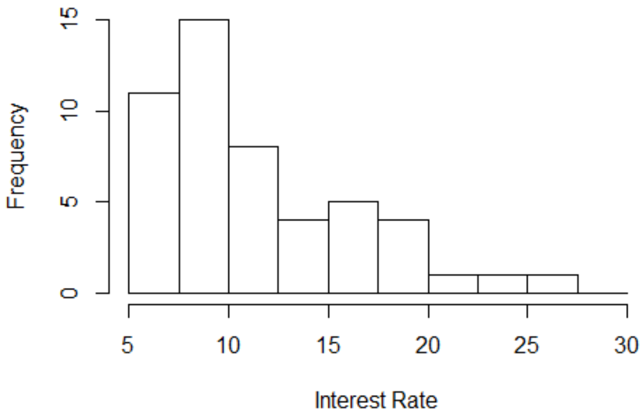
```
library(readr)
loan50 <- read_csv("loan50.csv")
hist(x=loan50$interest_rate,
          breaks=seq(5, 35, 10),
          xlab="Interest Rate",
          main="")
```



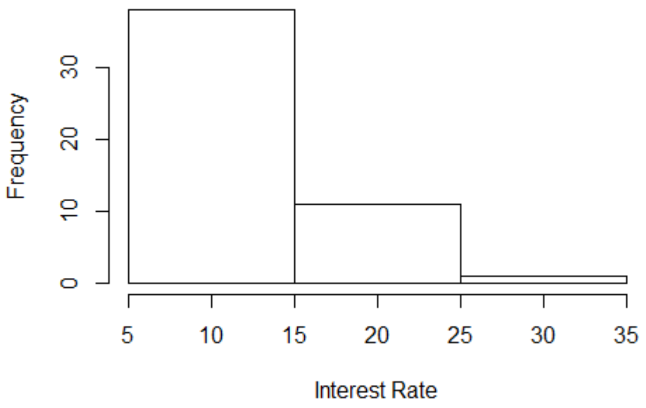
Histograms



Too much detail



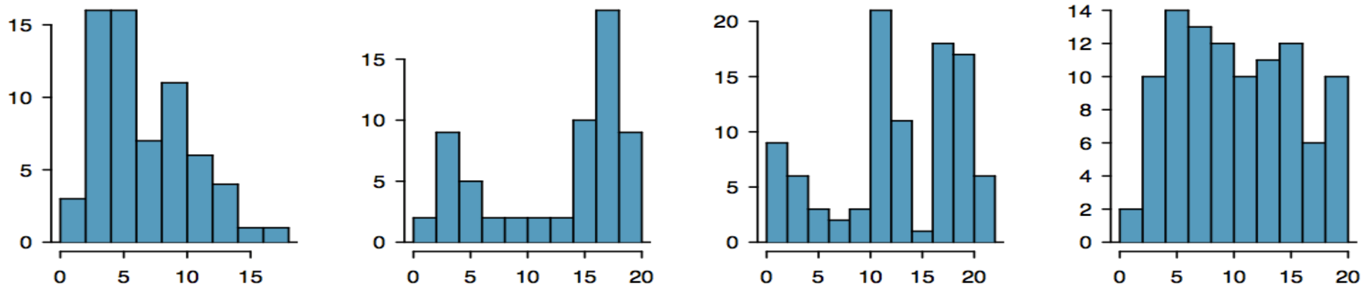
Just right



Too little detail

Modality

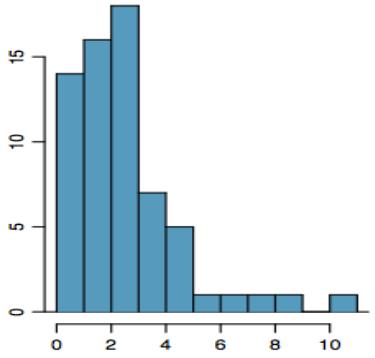
- The **modality** of a distribution is one way to describe its shape
  - *Unimodal*: One prominent peak
  - *Bimodal*: Two prominent peaks
  - *Multimodal*: More than two prominent peaks
  - *Uniform*: No prominent peaks



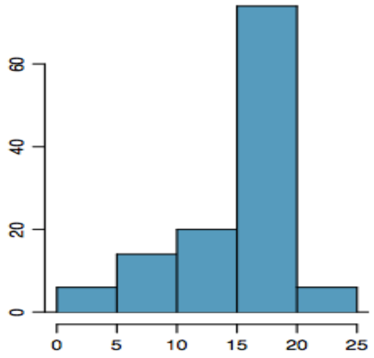
Unimodal Bimodal Multimodal Uniform

Skew

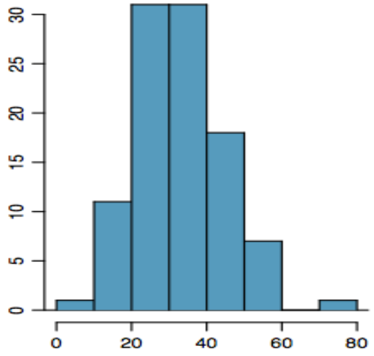
- The **skew** of a distribution is another way to describe its shape
  - *Right Skewed*: The data trails off to the right
  - *Left Skewed*: The data trails off to the left
  - *Symmetric*: The data trails off in both directions (roughly) equally



Right skewed



Left skewed



Symmetric

Variance and Standard Deviation

- The distance of an observation from the mean is called **deviation**.

$$deviation = x_n - \bar{x}$$

```
> mean(loan50$annual_income)
[1] 86170
> sample_mean <- mean(loan50$annual_income)
> x6 <- loan50$annual_income[6]
> deviation <- x6 - sample_mean
> deviation
[1] -19170
```

Variance and Standard Deviation

- The average of the squared deviations from the mean is called the **variance** ($\boldsymbol{s^2}$).

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}$$

[var function]

```
> variance <- var(loan50$interest_rate)
> variance
[1] 25.52387
```

- Measures how the data is dispersed around the mean
  - The greater the spread, the higher the variance is in relation to the mean

Variance and Standard Deviation

- The square root of the variance is called the **standard deviation (s)**.

$$s = \sqrt{s^2}$$

```
> variance <- var(loan50$interest_rate)
> variance
[1] 25.52387
> standard_dev <- sd(loan50$interest_rate)
> standard_dev
[1] 5.052115
```

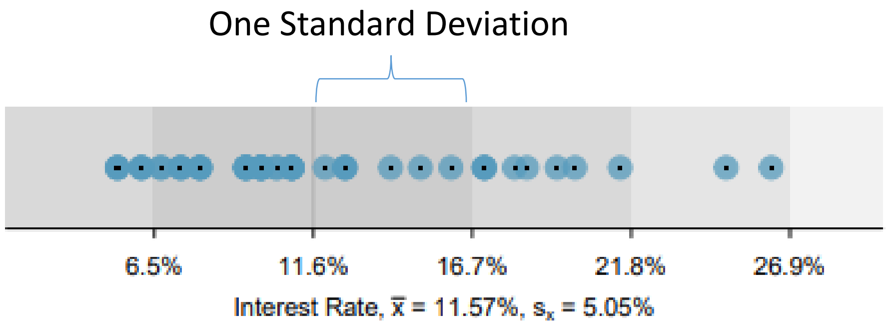
[sd function]

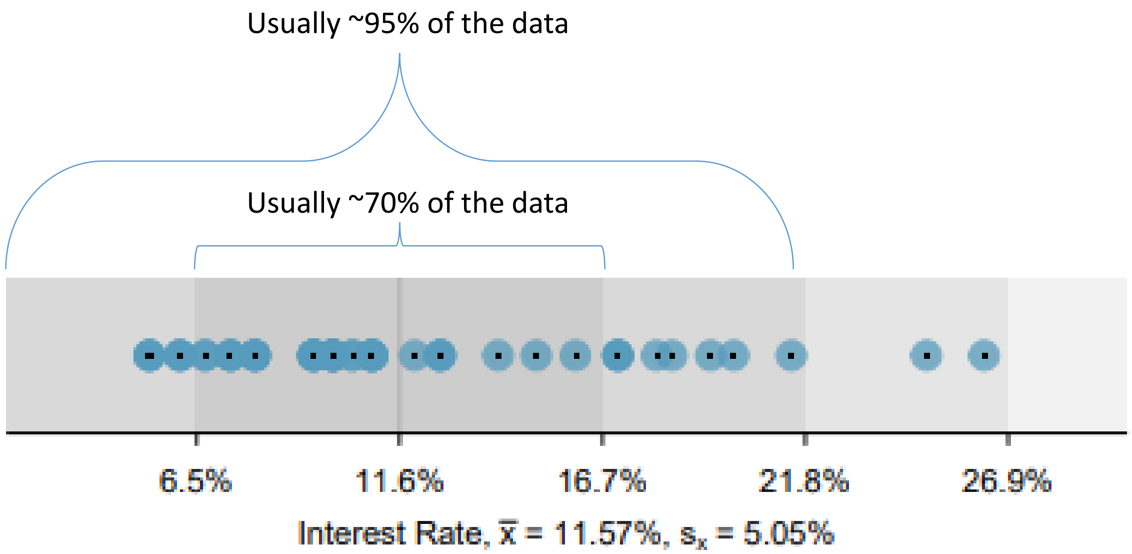
- Represents the typical deviation of observations from the mean
  - 70% of data will typically be within one standard deviation of the mean; 95% will be within two standard deviations

Variance and Standard Deviation

- Symbols:
  - Sample variance: s^2
  - Sample standard deviation: s
  - Population variance: σ^2
  - Population standard deviation: σ

Variance and Standard Deviation



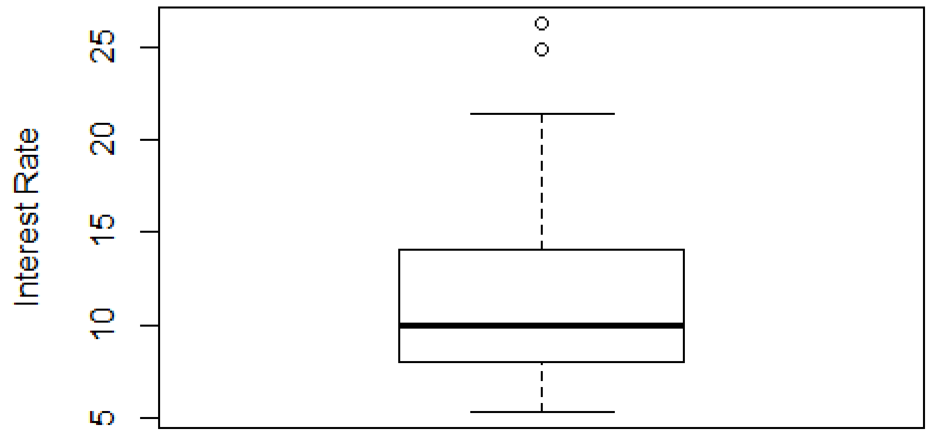



```
> mean(loan50$interest_rate)
[1] 11.5672
> sd(loan50$interest_rate)
[1] 5.052115
```

Box Plots

- The box plot summarizes a data set with five statistics.

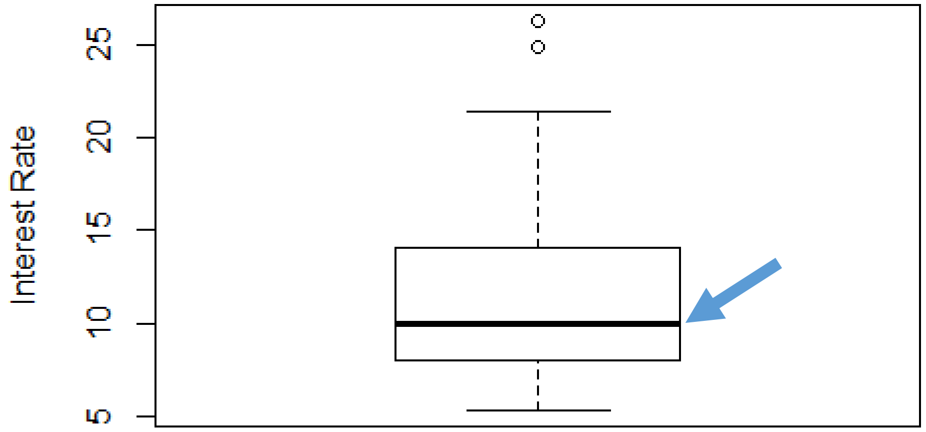
```
library(readr)
loan50 <- read_csv("loan50.csv")
boxplot(x=loan50$interest_rate,
        ylab="Interest Rate",
        main="")
```



[Boxplot function]

Box Plots

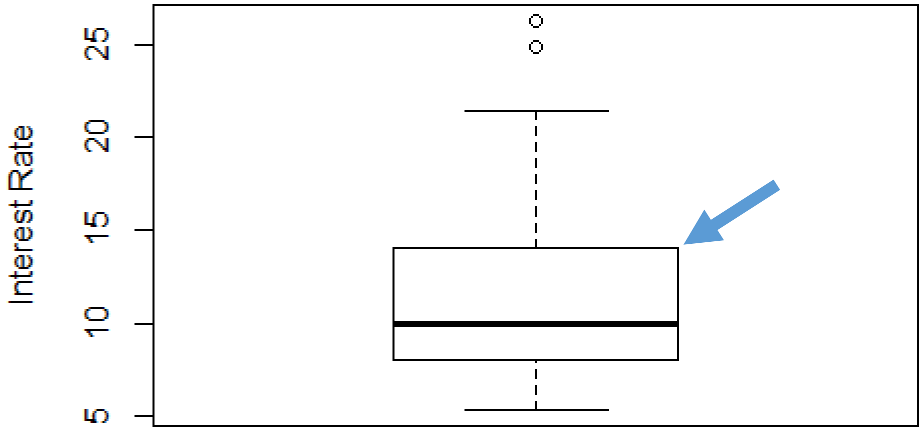
1. The **median** is the observation in the middle of all observations
    - If there are an even number of observations, the average of the two middle observations is used.
    - 50% of data fall above the median; the other 50% falls below it



Box Plots

2. The **third quartile** (Q_3 or "75th percentile") indicates where 75% of values in the data set fall under

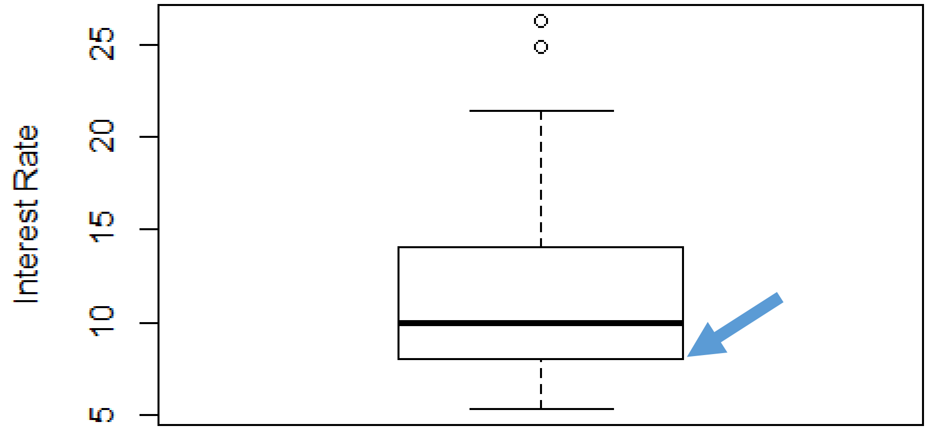
- 75% of observations fall below that line



Box Plots

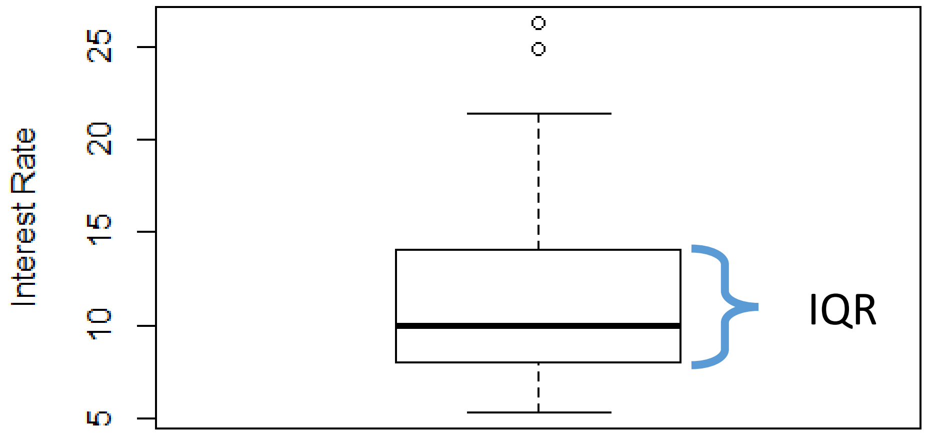
3. The **first quartile** (Q_1 or "25th percentile") indicates where 25% of values in the data set fall under

- 25% of observations fall below that line



Box Plots

- Together, they mark the boundaries of the **interquartile range** or **IQR**.
  - 75% of observations fall below to top line
  - 25% of observations fall below the bottom line
  - Thus, 50% of all observations will fall between them (in the box)

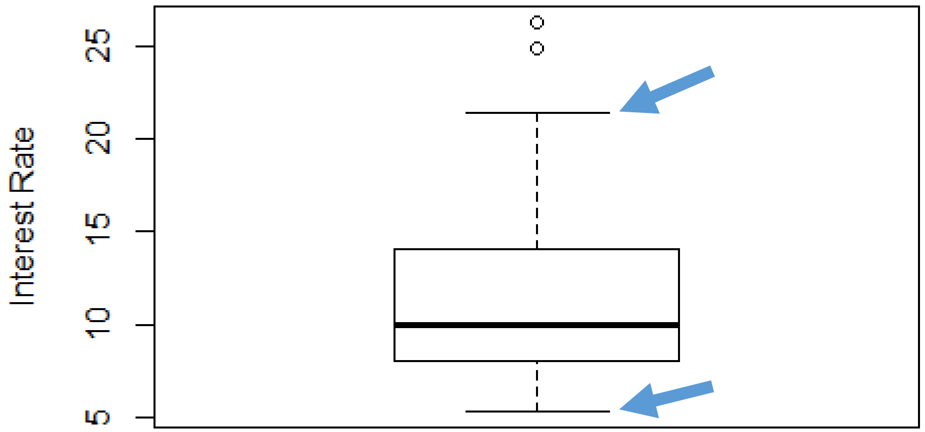


$$IQR = Q_3 - Q_1$$

Box Plots

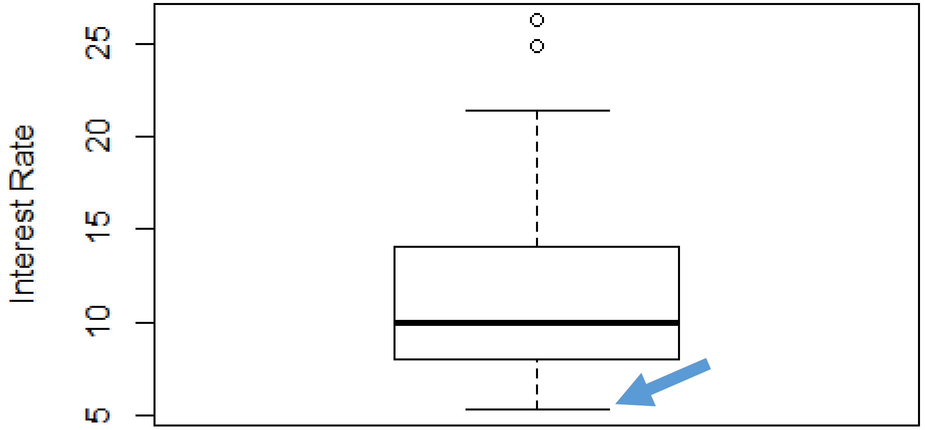
4 and 5. The **whiskers** try to capture the data outside of the IQR

- At most, they can extend $1.5 \times IQR$
- Max upper whisker = $Q_3 + 1.5 \times IQR$
- Max lower whisker = $Q_1 - 1.5 \times IQR$



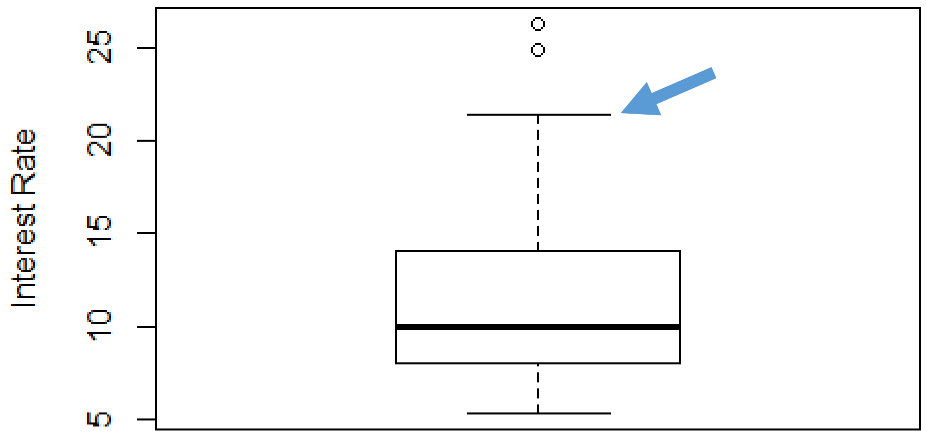
Box Plots

- The lower whisker does not need to extend that far to capture the data below Q_1



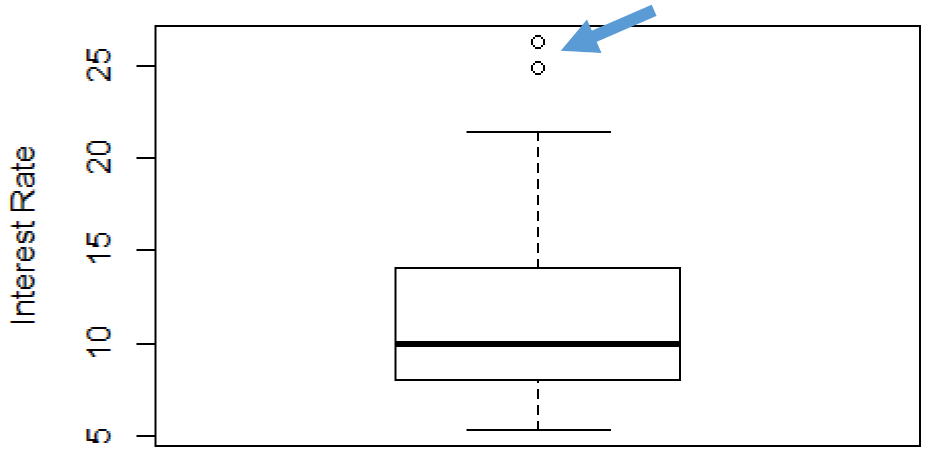
Box Plots

- The upper whisker extends as far as it can go ($Q_3 + 1.5 \times IQR$)
- We can see there are data points still outside of its reach.
  - These two data points (distant from the rest of the data) could be classified as **outliers**



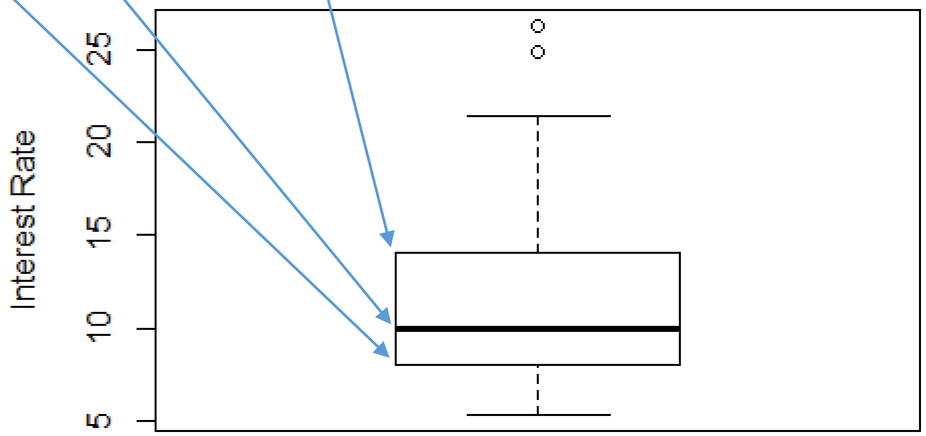
Box Plots

- Looking for outliers is useful for:
  - Identifying strong skew
  - Identifying data collection or data entry errors
  - Offering insight into interesting properties of the data



Box Plots

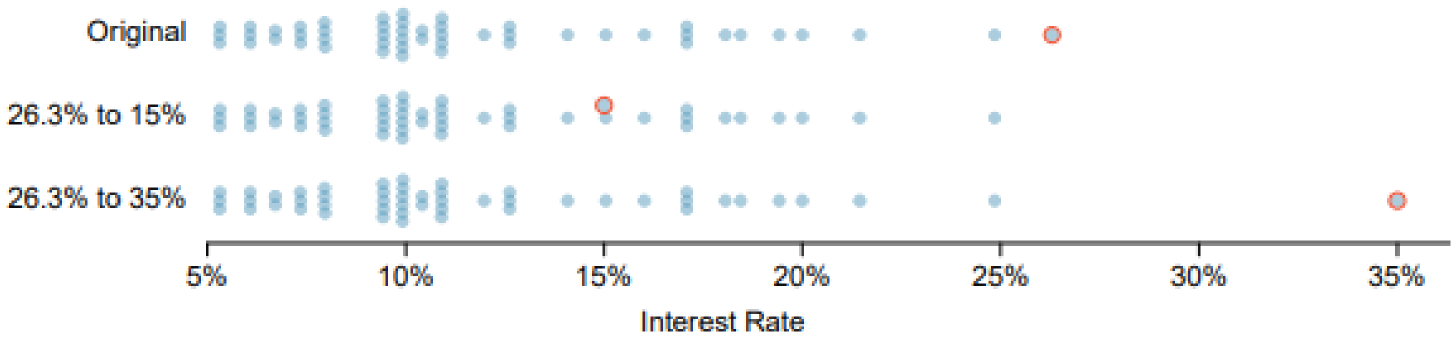
```
> summary(loan50$interest_rate)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   5.31    7.96    9.93   11.57   13.71   26.30 
```



[Summary function]

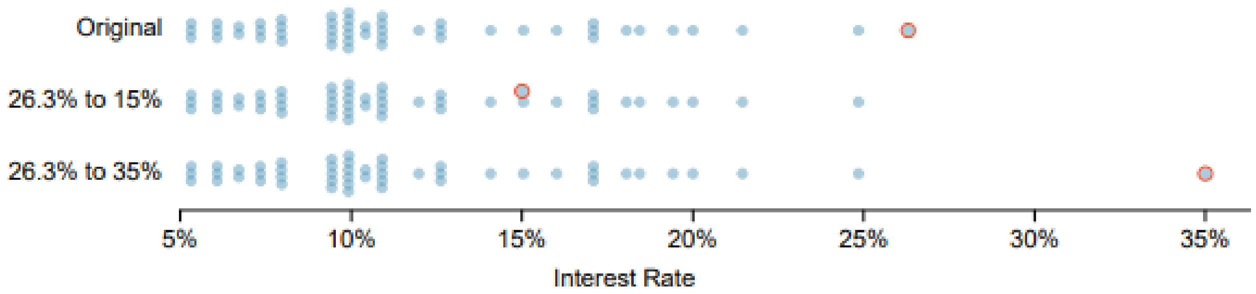
Robust Statistics

- Median and IQR are **robust statistics** in that extreme outliers have little effect on their values.
- This example shows an observation being changed three times.
  - What effect will this have on the sample statistics?



Robust Statistics

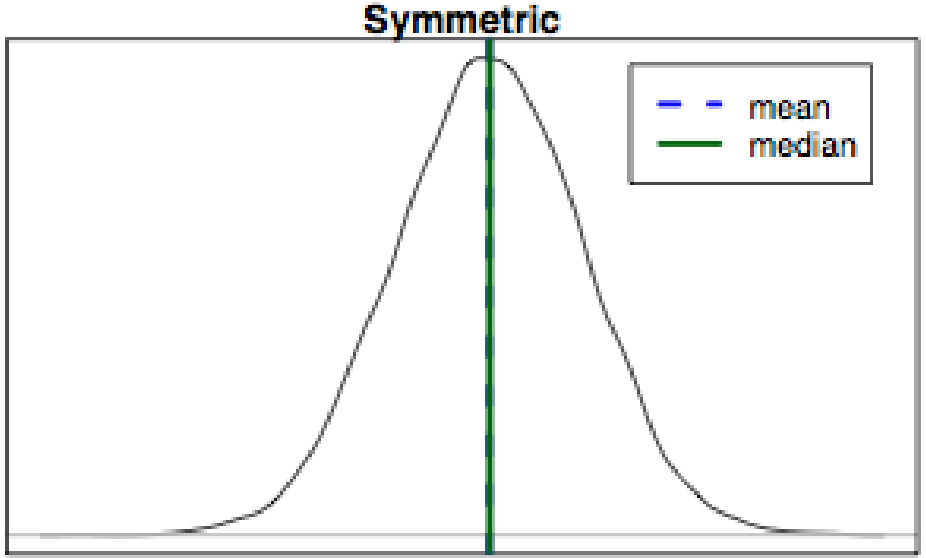
- No impact on median and IQR measurements
- Did impact the mean and standard deviation measurements



	robust		not robust	
scenario	median	IQR	$\bar{x}$	s
original `interest_rate` data	9.93%	5.76%	11.57%	5.05%
move 26.3% → 15%	9.93%	5.76%	11.34%	4.61%
move 26.3% → 35%	9.93%	5.76%	11.74%	5.68%

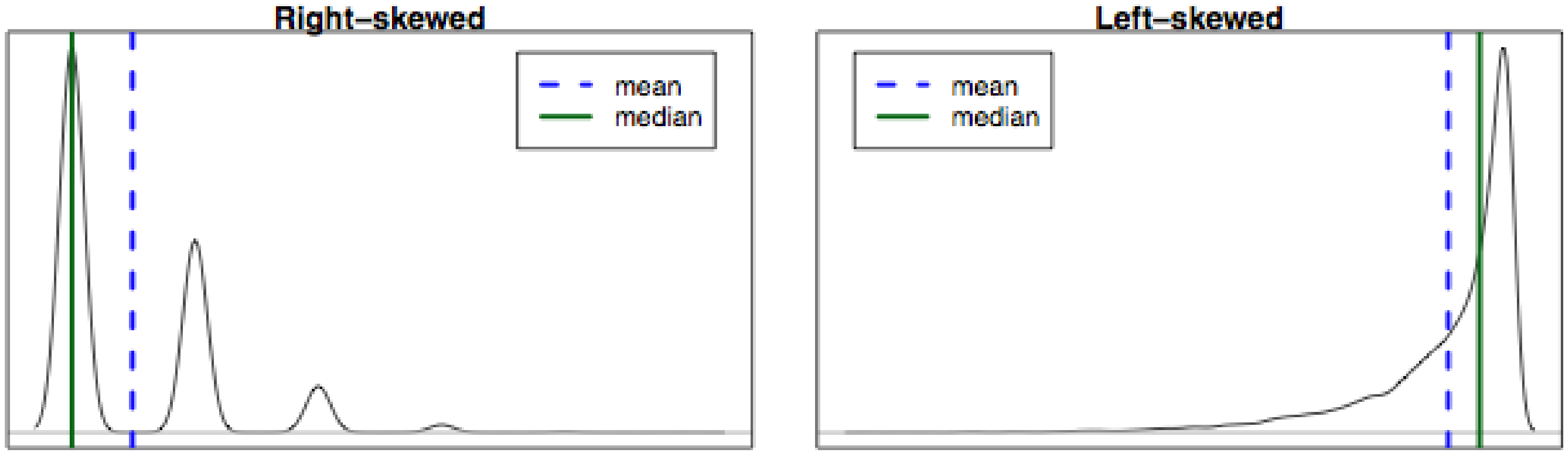
Robust Statistics

- In symmetric distributions, the mean is typically used to describe the center
  - mean ~ median



Robust Statistics

- In skewed distributions or where extreme outliers are present, the median is typically used to describe the center
  - Right skewed: mean > median
  - Left skewed: mean < median



Robust Statistics

- For symmetric distributions, use $\bar{x}$ and s to describe the center and spread
- For skewed distributions: use median and IQR to describe the center and spread