Mathematics 2

Lecture 2: Describing Data (continued)

Danica Solina
School of Mathematical Sciences

Outline

- Reference: Devore §1.2, 1.4, 12.1

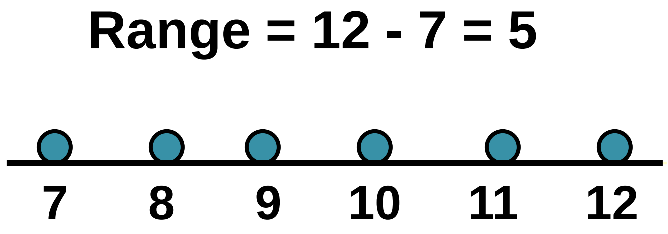
- This week, we continue to look at how to describe data.
- There are two main ways of describing data
  - Numerically
  - Graphically
- We consider these in turn.

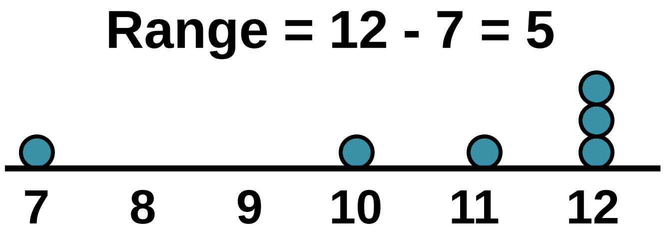
Measures of Variability - Range

- Difference between the largest and the smallest observations:

$$\boxed{\text{Range} = X_{\text{Largest}} - X_{\text{Smallest}}}$$

- Ignores the distribution of the data.





Measures of Variability - Variance

- Average distance of an observation from the mean.
- Variances of independent measurements **can be added together** (standard deviations cannot be added).
  - Sample Variance:

$$\boxed{S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n-1}}$$

  - Population Variance:

$$\boxed{\sigma^2 = \frac{\sum_{i=1}^{N} \left(x_i - \mu\right)^2}{N}}$$

Notice that sample variance has a denominator of n - 1, but population variance has a denominator of N

Measures of Variability - Standard Deviation

- The positive square root of the variance.
- It is easier to interpret as it has the **same units** as the data itself.
  - Sample Standard Deviation:

$$\boxed{s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}}$$

  - Population Standard Deviation:

$$\boxed{\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N}}}$$

Measures of Variability - Variance

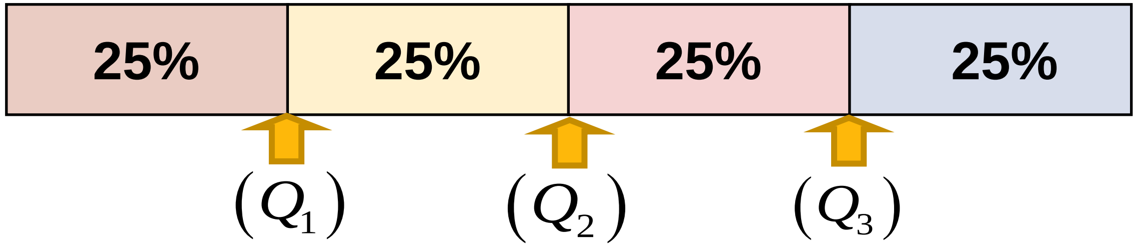
Examples

- Population of size $N = 6$:
  (9, 10, 10.5, 11.5, 12, 14)

- Sample of size $n = 4$:
  (9, 10.5, 11.5, 14)

Measures of Variability - Quartiles

- The third measure of variation is the **interquartile range**. To calculate this, we first must look at **quartiles**.
- The three quartiles (Q1,Q2, and Q3) split the data into four groups with the same number of observations
  - The second quartile is the median.



- We can express the position of i^{th} quartile as

$$(Q_i) = \frac{i(n+1)}{4}$$

Measures of Variability - Quartiles

- (Q_i) will not always be an integer
  - If (Q_i) is $x + 0.25$ or $x + 0.75$ (where x is a whole number), then **round down or up** (respectively)
    - If $(Q_3) = 5.75$ then let the third quartile be the 6^{th} observation.
    - If $(Q_3) = 5.25$ then let the third quartile be the 5^{th} observation.

  - If (Q_i) is $x + 0.5$ (where x is a whole number), then **take the mean** of the x^{th} and $(x + 1)^{st}$ observations
    - If $(Q_3) = 5.5$ then let the third quartile be the mean of the 5^{th} and 6^{th} observations.

Measures of Variability - IQR

- The **interquartile range** is a measure of variation
  - The range of the middle 50%
- Difference between the First and Third Quartiles

$$\boxed{\text{IQR (Interquartile Range)} = Q_3 - Q_1}$$

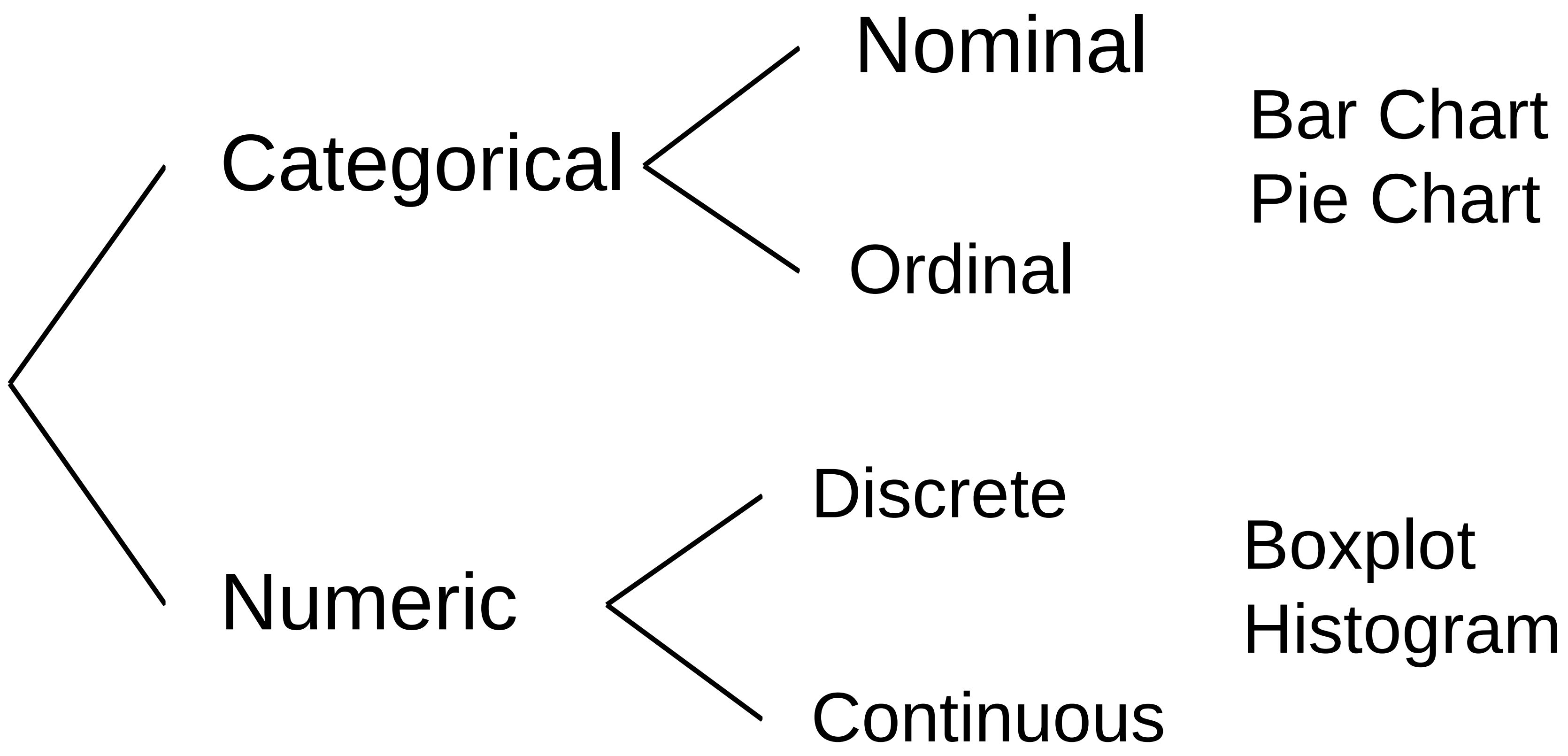
- The IQR is not affected by outliers as much as the standard deviation or the range, and is more sensitive to the distribution of the data than the range.

Measures of Variability - IQR

- 15 students with part time jobs were randomly selected and the number of hours worked last week was recorded.

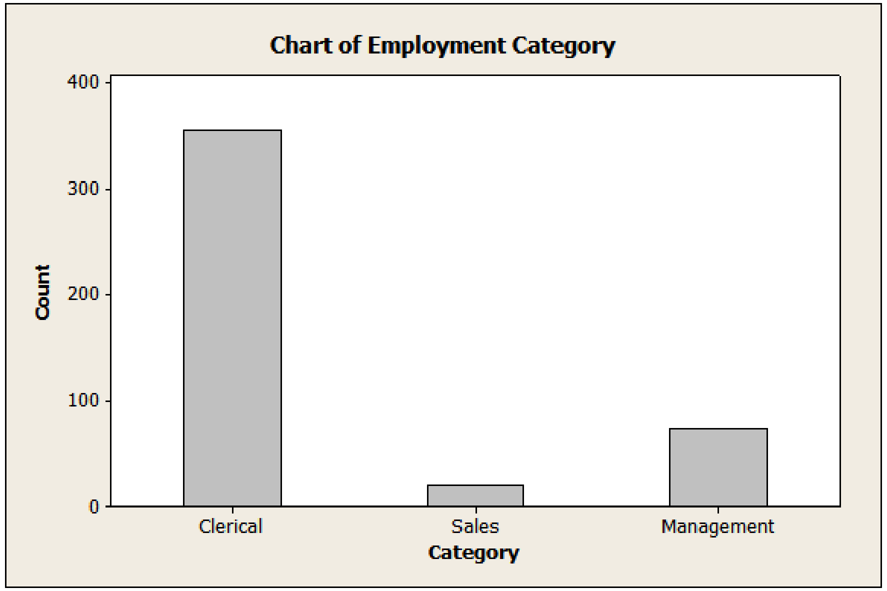
19, 12, 14, 10, 12, 10, 25, 9, 8, 4, 2, 10, 7, 11, 15

Types of Data – Graphical Displays



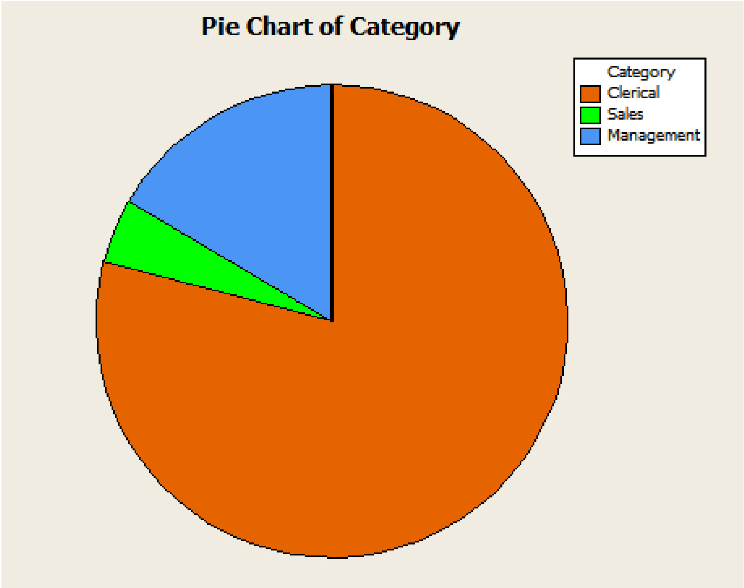
Displaying Categorical Data

Bar Graph



Good for comparing the number of Individuals in different groups

Pie Graph



Good for looking at parts of a whole
- what proportion of people…

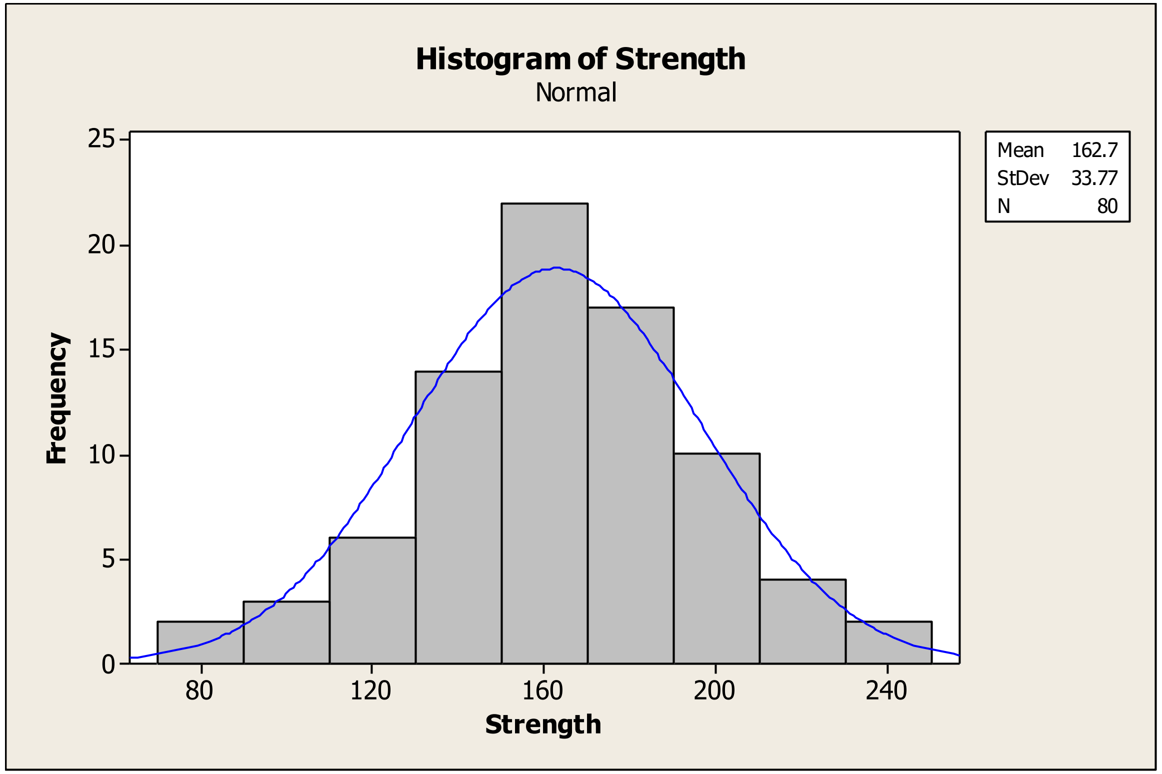
Displaying Quantitative Data: Histogram

- Histograms are used to get an overall idea of the shape of the data.

- Histograms are **not** the same as bar graphs
  - The *x*-axis on a histogram is a **continuous scale**.
  - The x-axis on a bar chart is a **set of categories**.

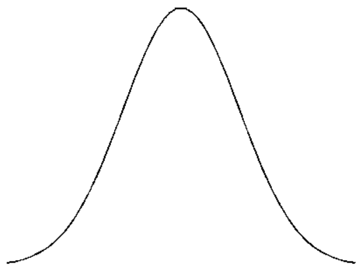
- We need to choose the widths of the columns carefully so we do not lose too much information.

Displaying Quantitative Data: Histogram

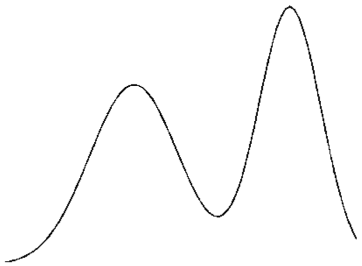


Note: The presence of a normal fit does not necessarily imply that the data are normally distributed. If the histogram (rectangles) and the curve have a similar shape (as they do here) then we can say that the data are normally distributed.

Illustrated Distribution Shapes

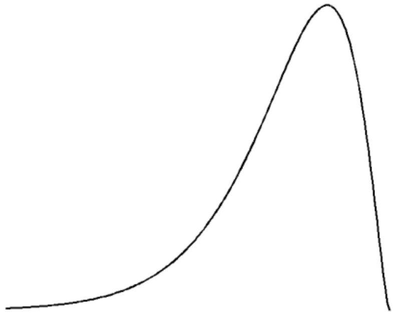


Unimodal

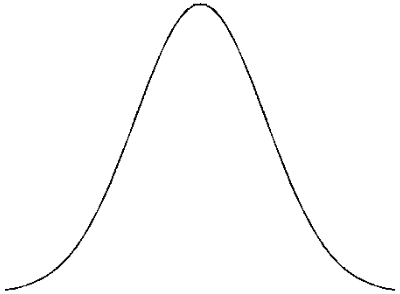


Bimodal

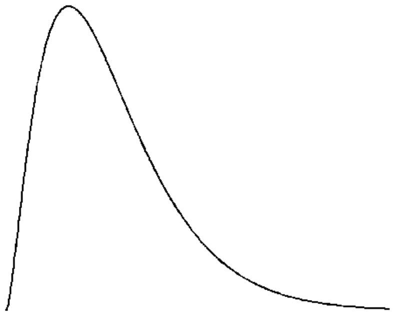
Multimodal



Left-Skewed

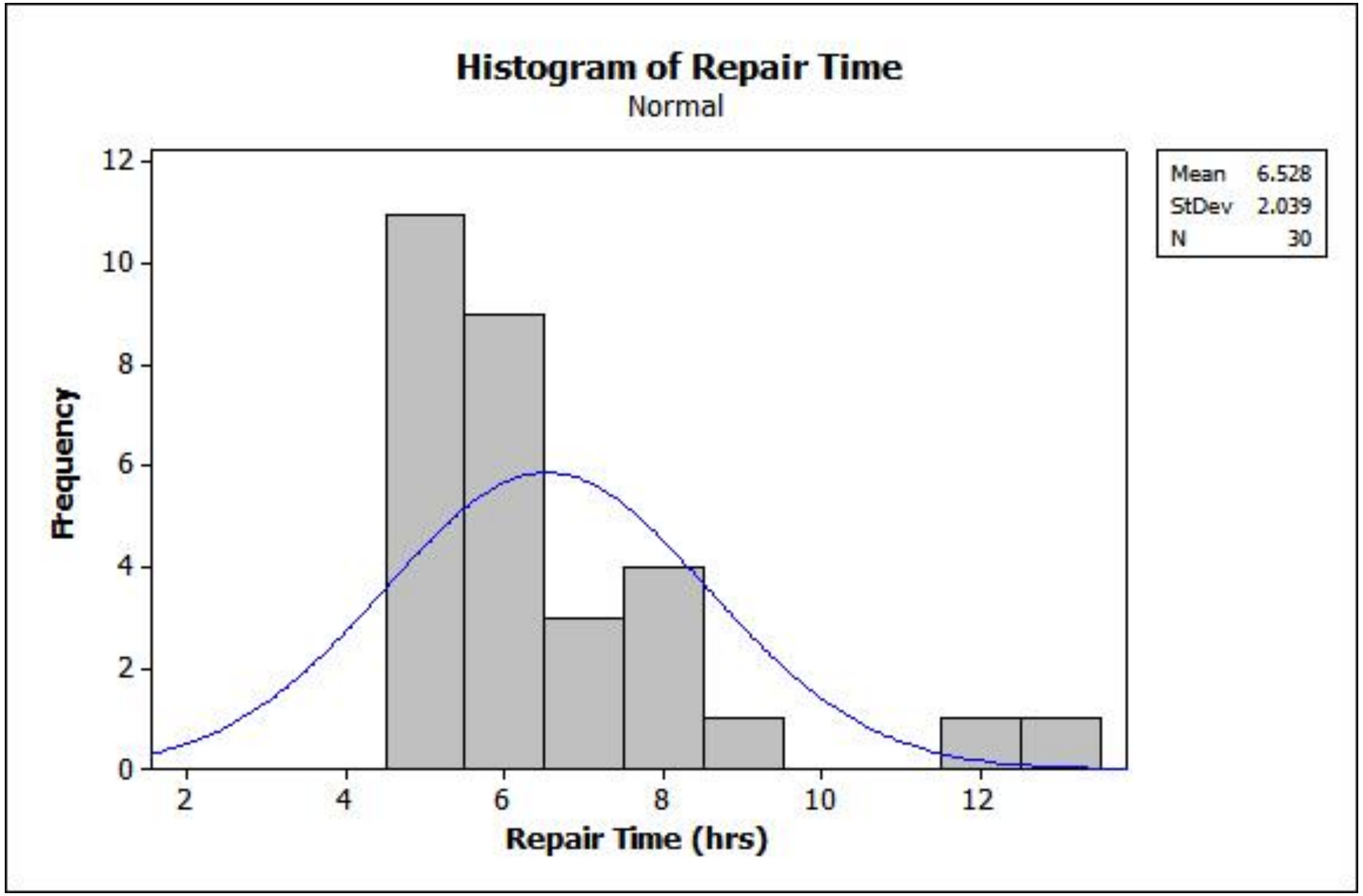


Symmetric

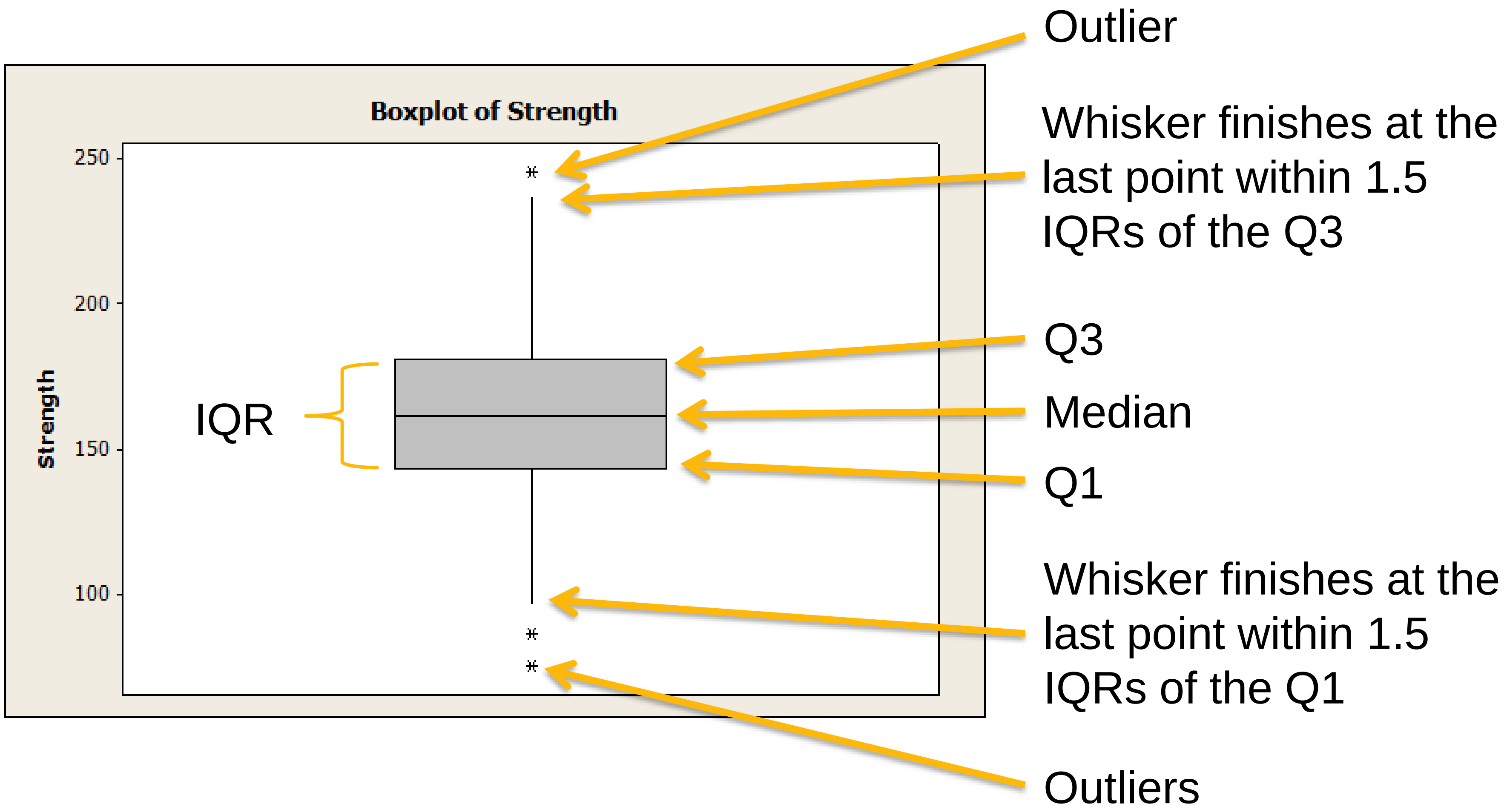


Right-Skewed

Exercise: Describing Distributions



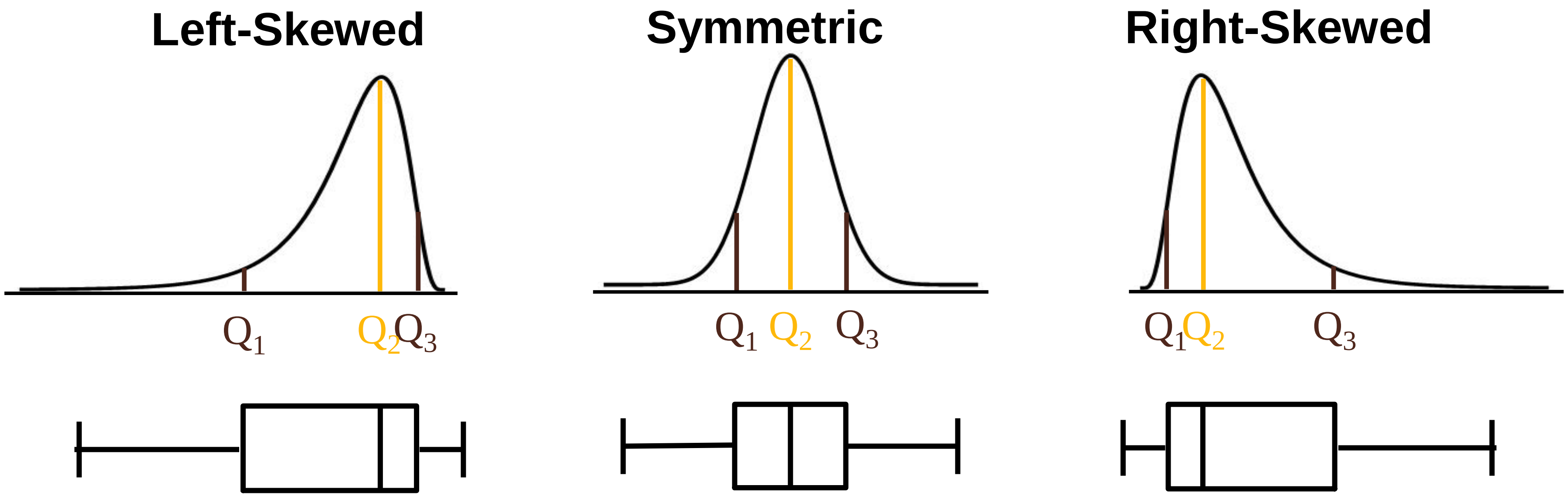
Displaying Quantitative Data: Boxplot



Distribution Shape & Box Plots

- Box plots also shows useful information about the symmetry of the data
  - If
    - the median is much closer to one quartile then the other **and**
    - the tail on the side closest to the median is much shorter than the other

    then the distribution is **skewed**.



From page 41 of Devore : observations on the time until failure (1000s of hours) for a sample of turbochargers from one type of engine (from “The Beta Generalized Weibull Distribution: Properties and Applications,” Reliability Engr. and System Safety, 2012: 5–15).

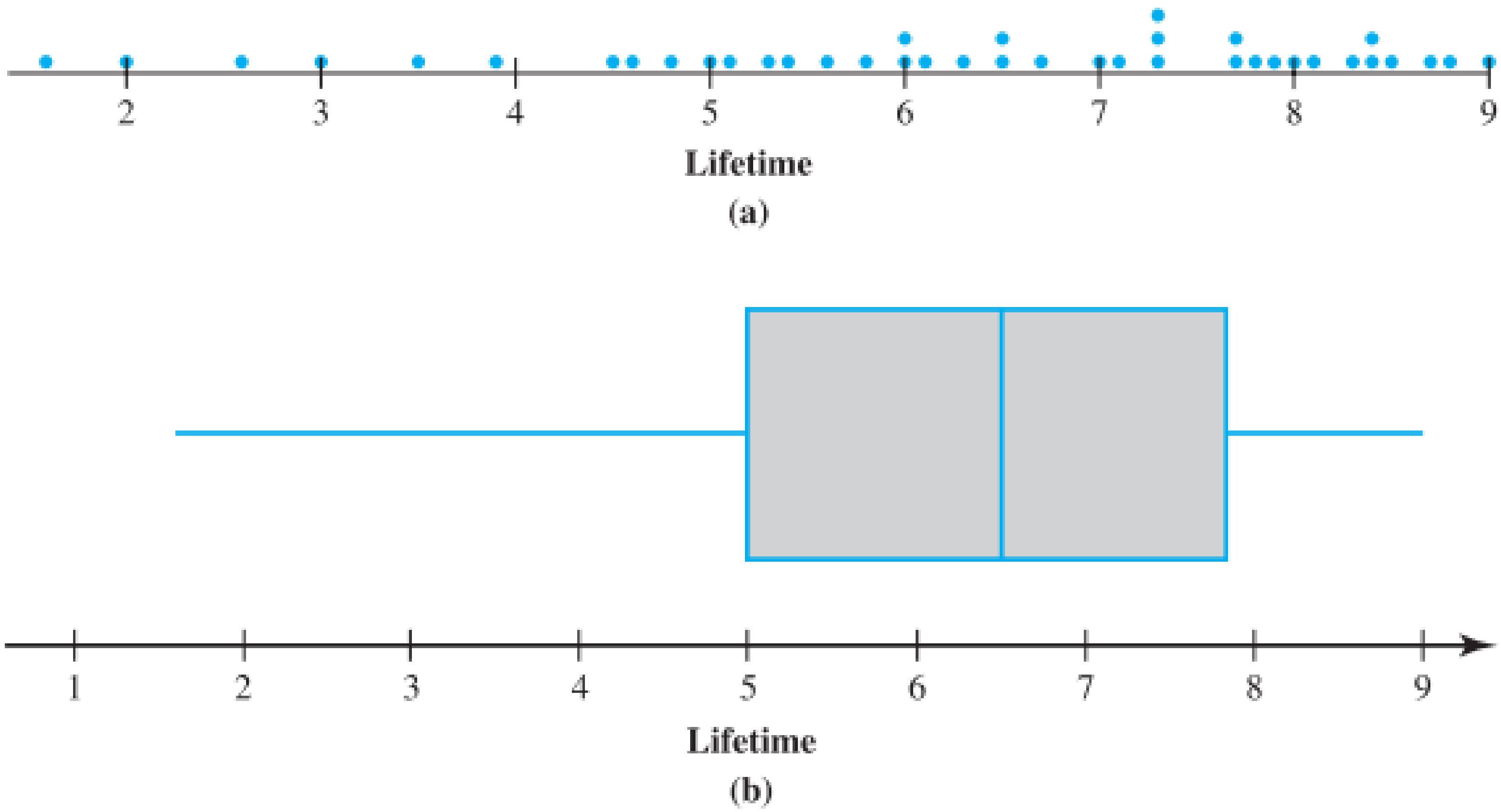
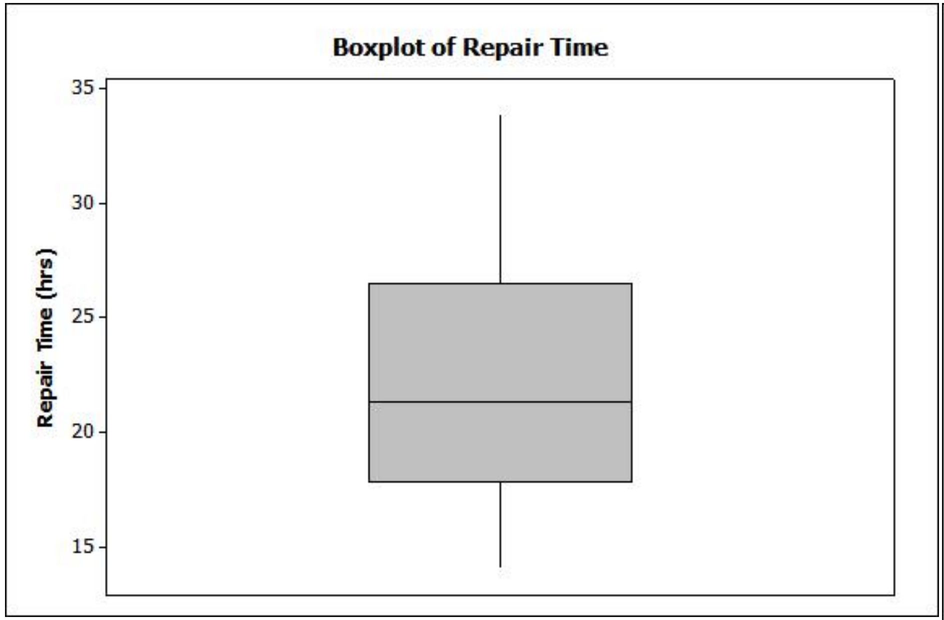
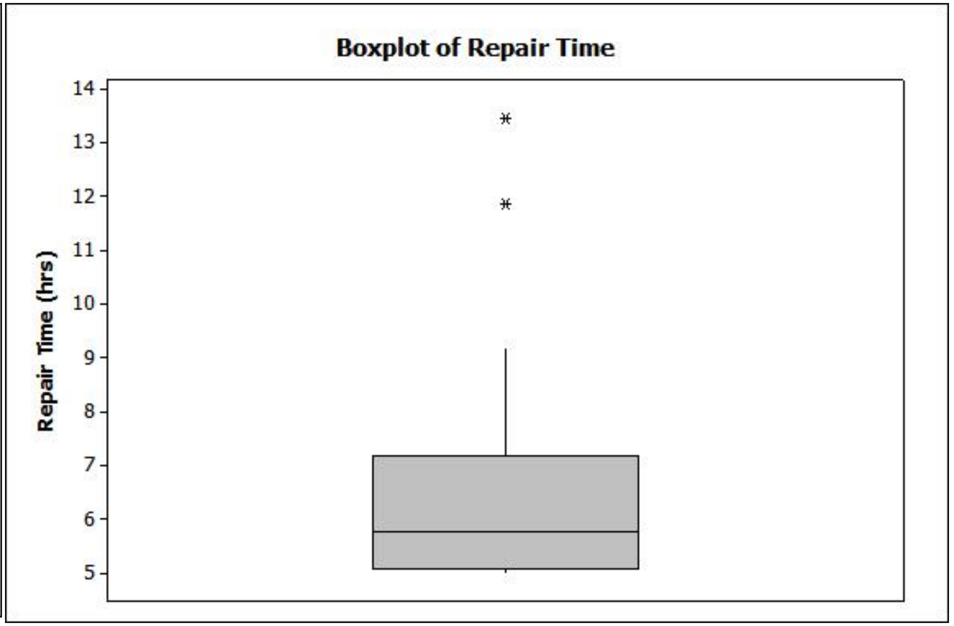


Figure 1.20 (a) Dotplot and (b) Boxplot for the lifetime data

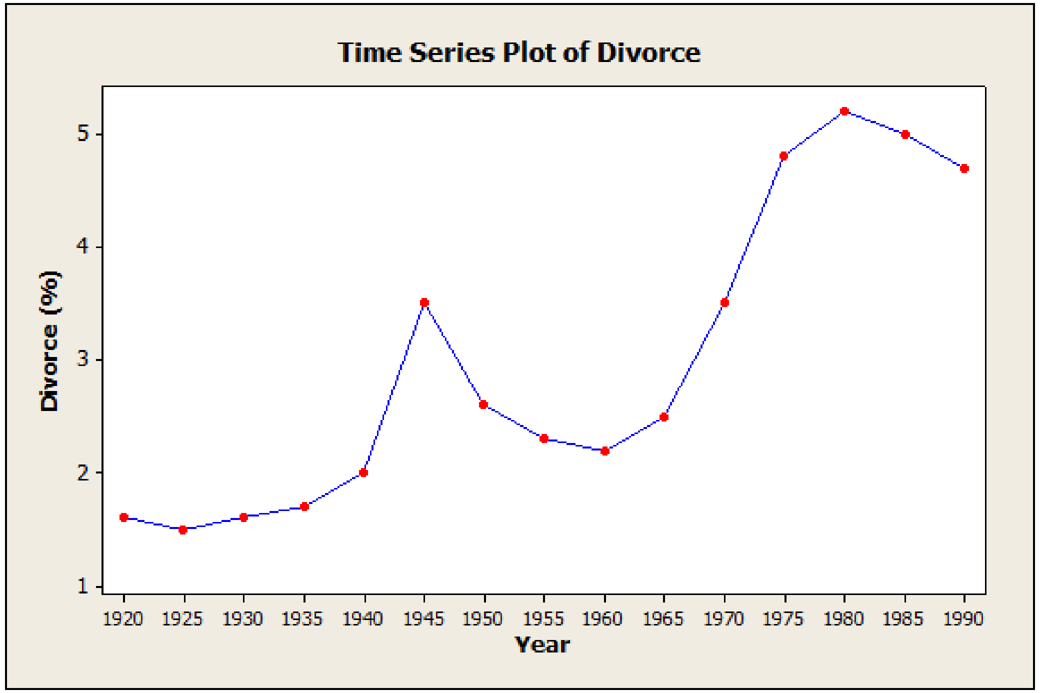
Exercise: Describing Distributions



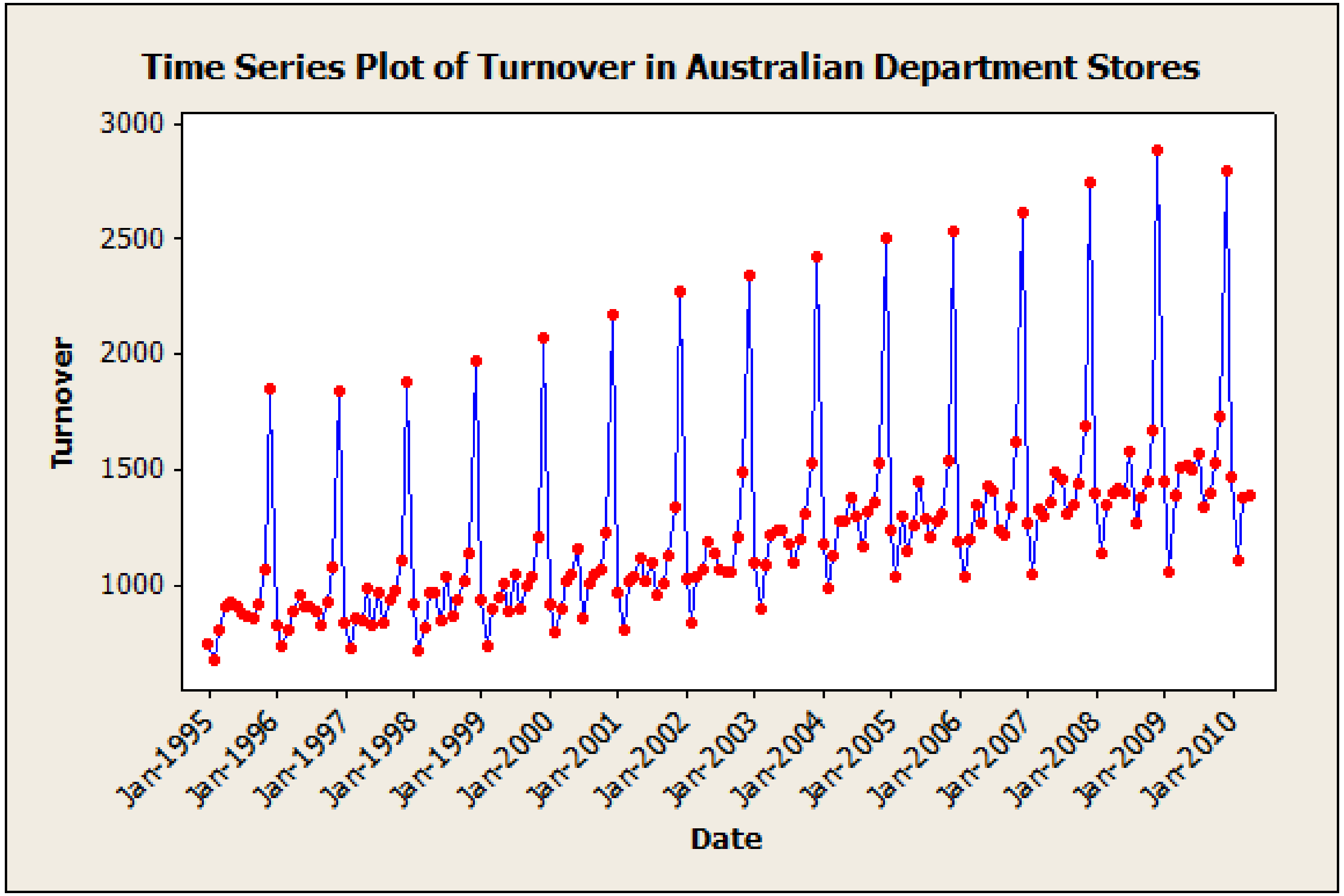


Time-Series Data

- Time Series are used to look at a phenomenon over time
  - Is there a trend? Can we see a cycle?



Time-Series Data: Exercise



Bivariate Relationships

- While graphs are useful for looking at the properties of a single variable, sometimes we are more interested in the relationship between two variables.
  - This is called a **bivariate relationship**.
- The method that we use to display bivariate relationships depends on the type of data.

Two Categorical Variables

- In some cases, we would like to compare two categorical variables to see whether there is a relationship between them.
  - We use a **contingency table** to do this.
- **Example:** A survey of 150 households was conducted 2 weeks after a nuclear mishap on Three Mile Island. This survey asked for the respondents distance from the island, and whether or not they believed that there should have been a full evacuation of the area. The results are as follows:

		Distance from Three Mile Island (miles)			
		1-6	***7-12***	***13+***	***Total***
Evacuation	***Yes***	18	15	33	66
	No	20	19	45	84
Total		38	34	78	150

Two Categorical Variables

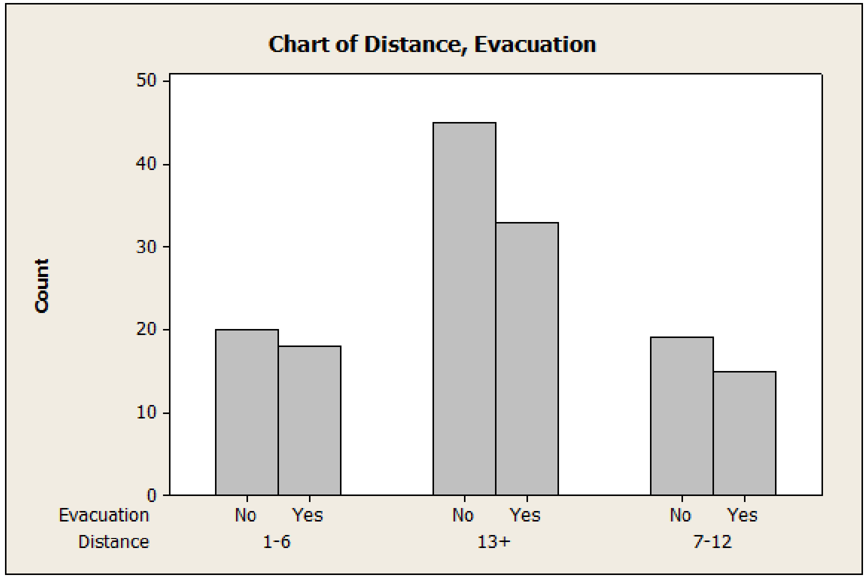
Tabulated statistics: Evacuation, Distance

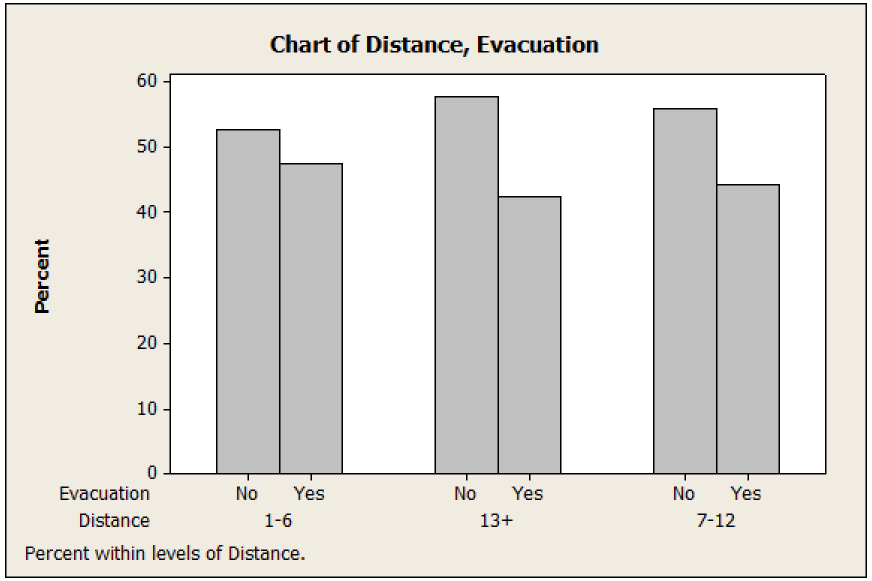
```
Using frequencies in Frequency


Rows: Evacuation   Columns: Distance

            1-6     7-12      13+      All

Yes          18       15       33       66
          27.27    22.73    50.00   100.00
          47.37    44.12    42.31    44.00

No           20       19       45       84
          23.81    22.62    53.57   100.00
          52.63    55.88    57.69    56.00

All          38       34       78      150
          25.33    22.67    52.00   100.00
         100.00   100.00   100.00   100.00

Cell Contents:      Count
                    % of Row
                    % of Column
```

1. What percentage thought that there should have been an evacuation?
2. What percentage of all respondents lived within 6 miles of Three Mile Is?
3. What percentage of respondents who live within 6 miles of Three Mile Is thought that there should have been an evacuation?
4. What percentage of respondents who thought that there should have been an evacuation live within 6 miles of Three Mile Is?

Two Categorical Variables

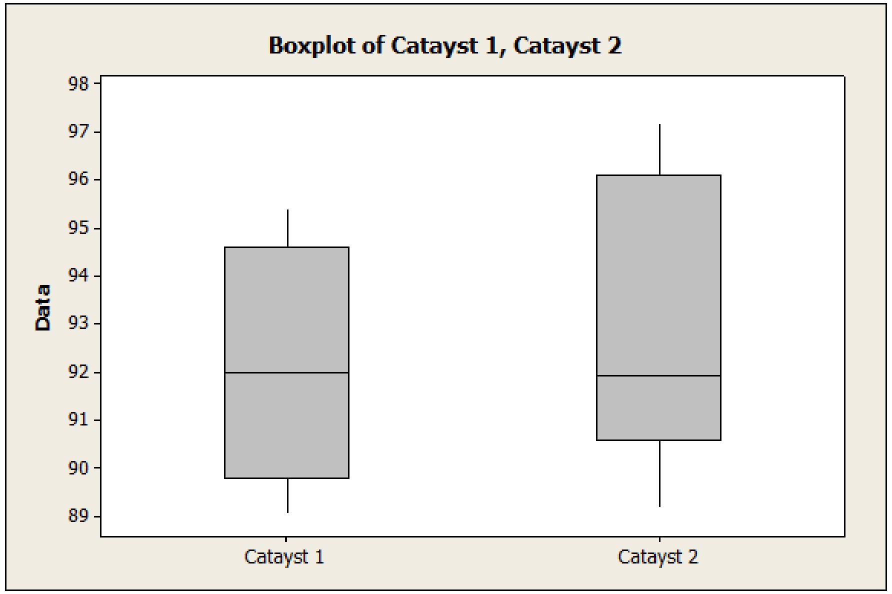
- We can also display this data as a **multiple bar chart**.



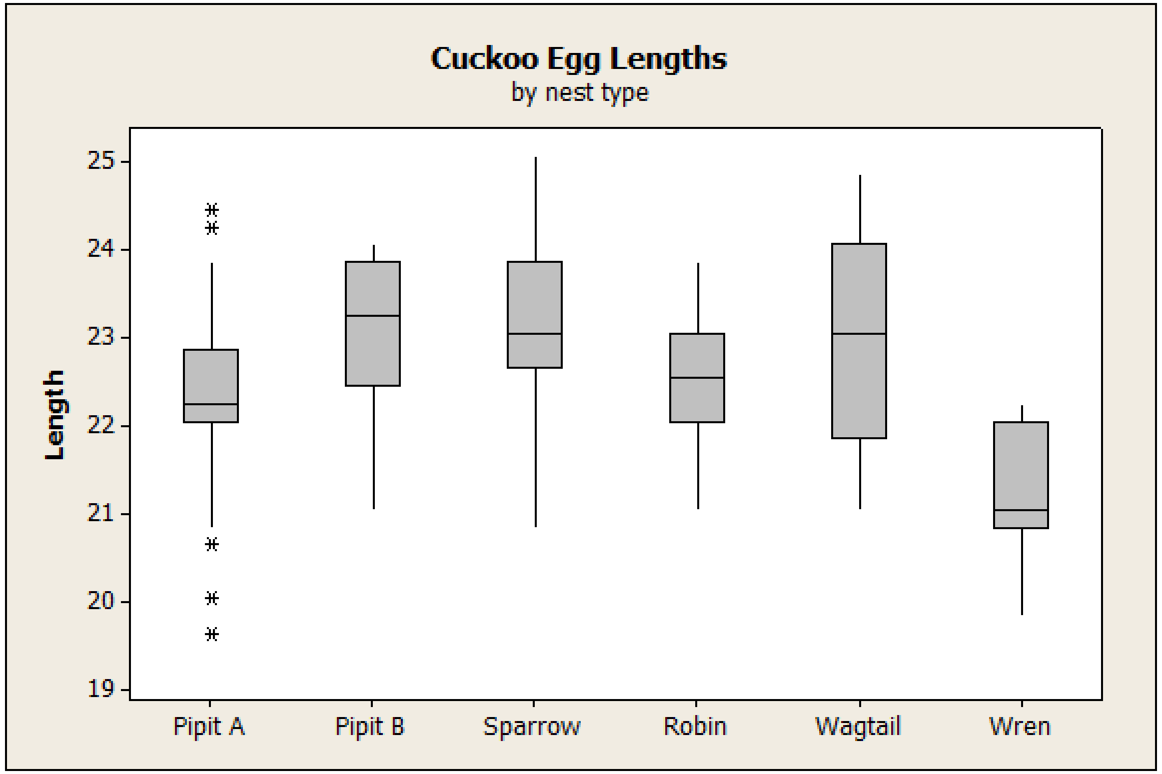


One Categorical, One Numeric

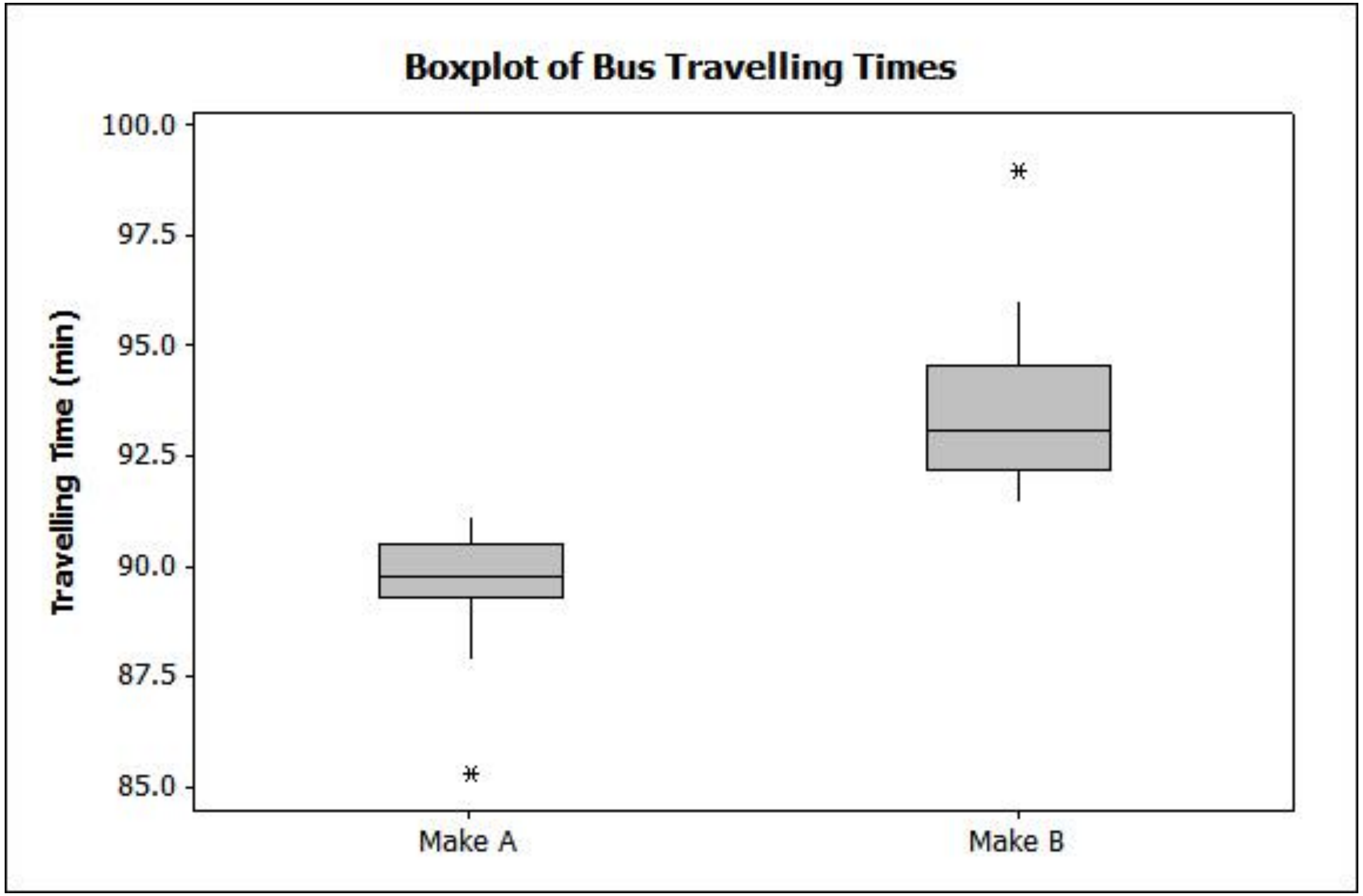
- We can use **boxplots** to determine differences in an numerical variable between groups (a categorical variable)
  - If one box is entirely above another then we can be sure that they have different means
  - If the boxplots overlap, then we cannot necessarily make any conclusions
  - We should also compare the variability of the observations in each group.



Multiple Boxplots: Exercise

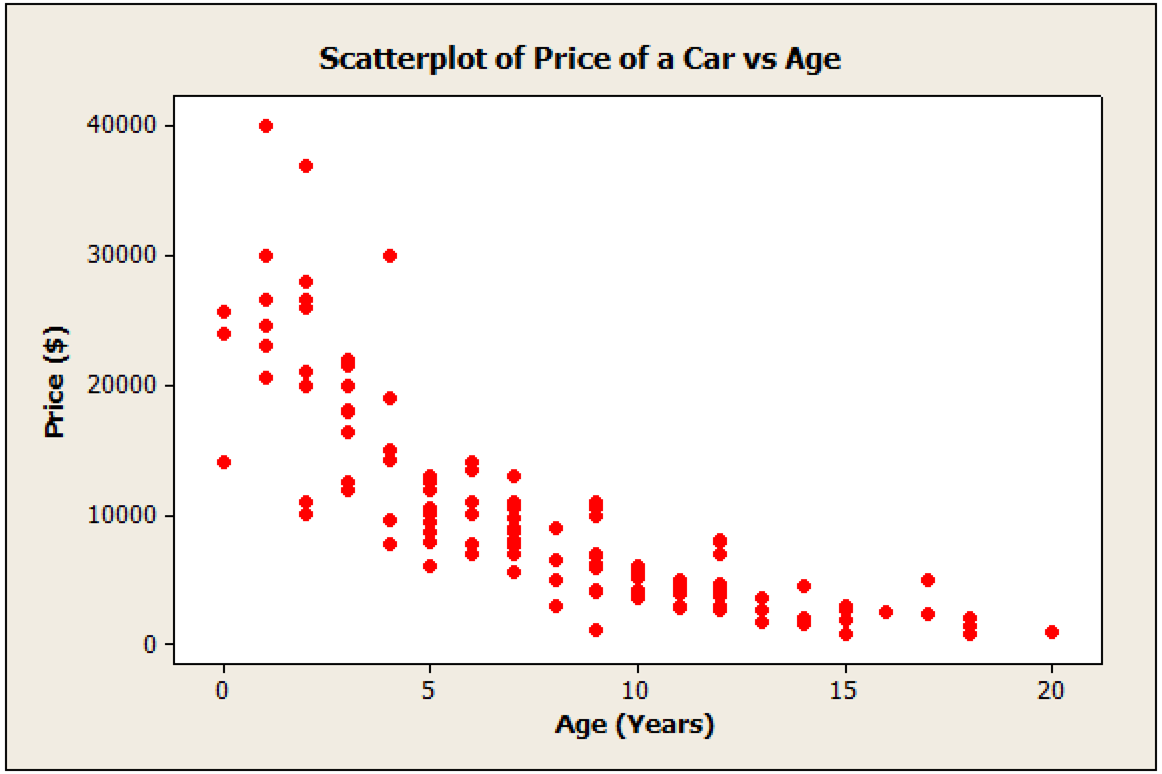


Multiple Boxplots: Exercise

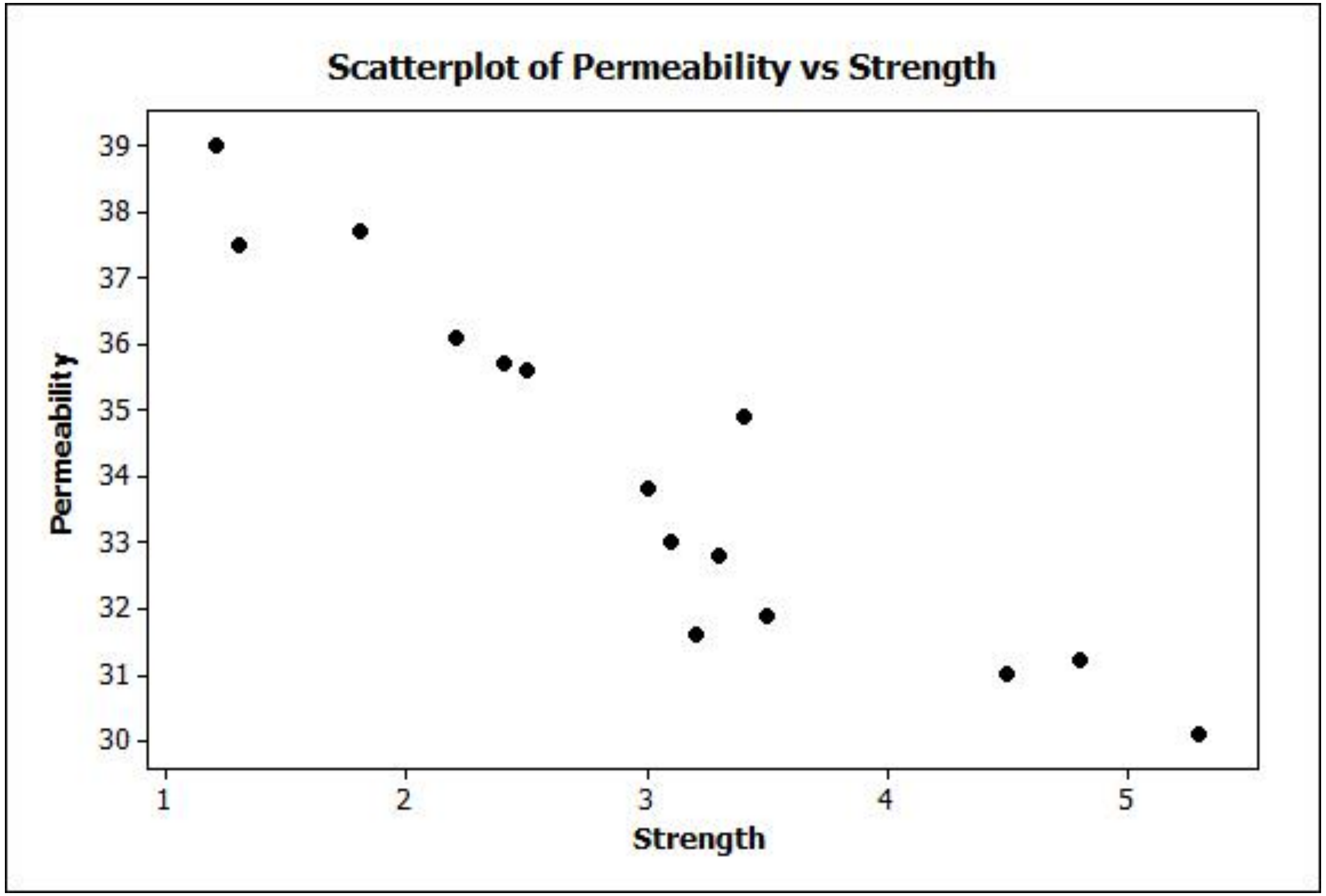


Two Numeric Variables

- When we have two numeric variables, we use a **scatterplot** to display their relationship.

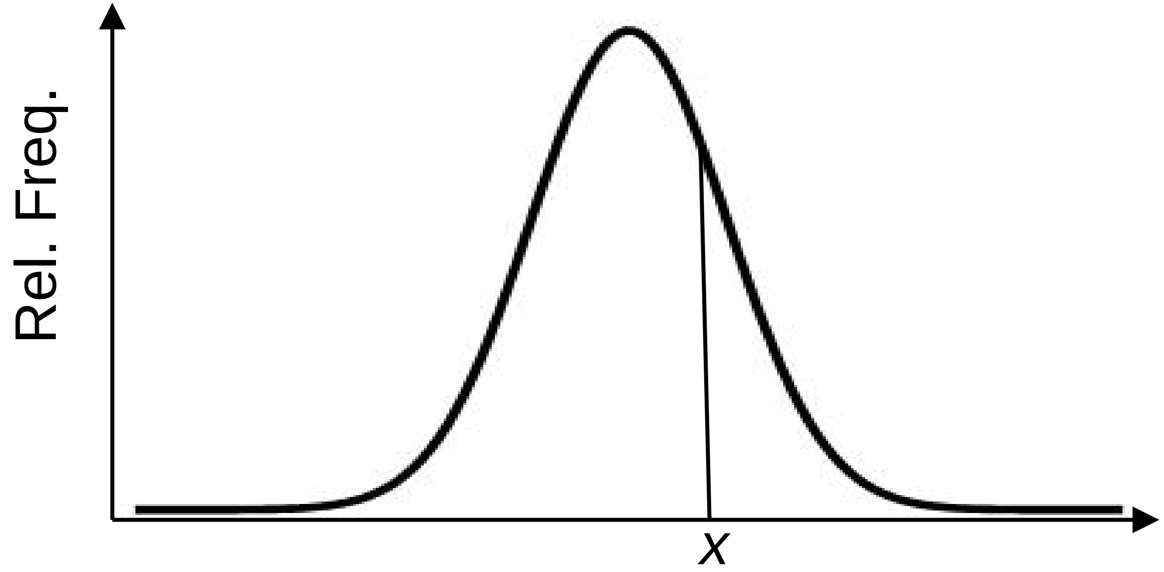


Scatterplot: Exercise



Measures of Relative Standing - Percentiles

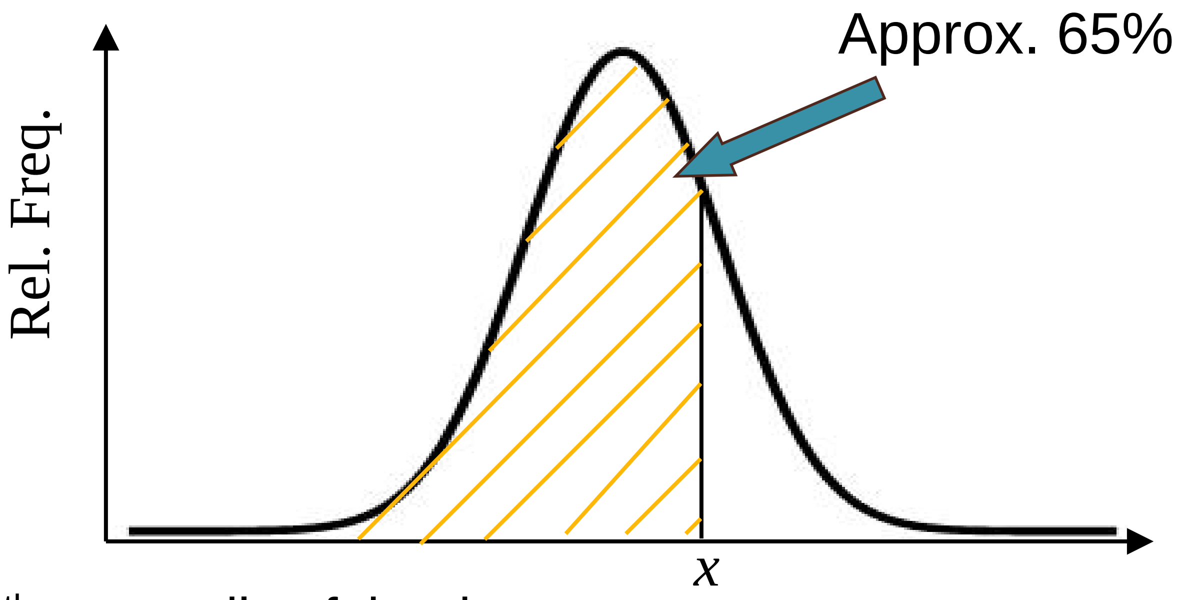
- Sometimes we are interested in the relative position of an observation amongst the set of data.



- We want to measure the **position** of an observation x within the data.

Measures of Relative Standing - Percentiles

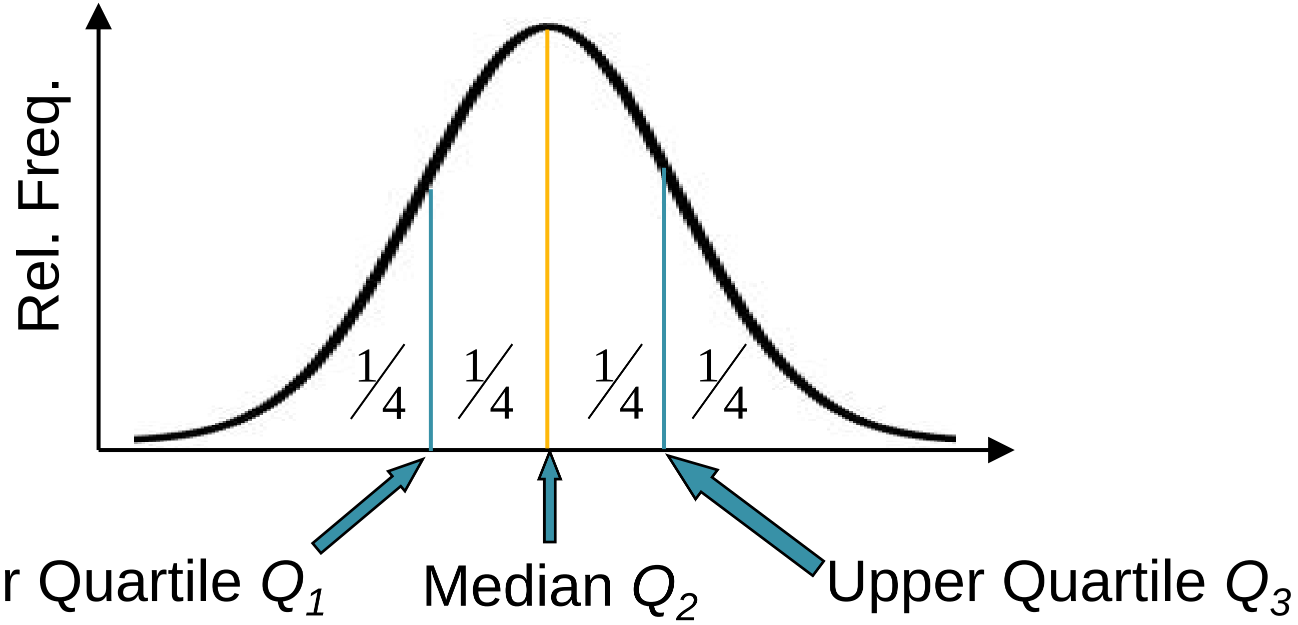
- The **percentile** of an observation *x* is the proportion of the data that is less than *x.*
  - E.g. University entrance scores (ATAR/UAI)



- Example: *x* is the 65th percentile of the data

Measures of Relative Standing - Percentiles

- We can describe quartiles as percentiles
  - lower quartile Q1 25th percentile
  - median Q2 50th percentile
  - upper quartile Q3 75th percentile



Measures of Relative Standing - Z Scores

- The **z-score** corresponding to a particular observation in the data set is

$$\boxed{z\text{–score} = \frac{\text{Observed} \; - \; \text{Mean}}{\text{Std Deviation}}}$$

  - The z score is **how many standard deviations the observation is from the mean**.
  - A positive z score indicates the observation is larger than the mean and a negative z score indicates the observation is smaller than the mean.
- When we use sample data, this formula becomes

$$\boxed{z\text{–score} = \frac{X_i - \bar{X}}{s}}$$

EXCEL

- Let's use EXCEL for some of these things.

Lecture 2 Revision Exercises

Use software to construct graphs

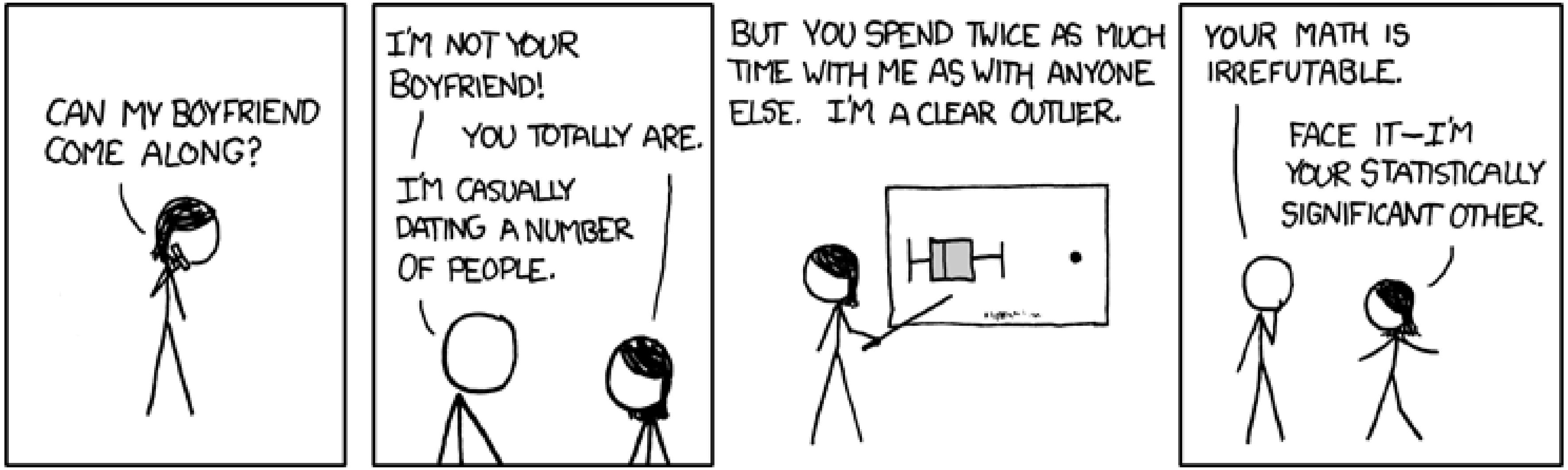
- Devore §1.2 Q 11, 13b, 15, 21, 29
- Devore §1.4 Q 44, 47, 51, 53(a-b), 54 (b-d), 56, 58, 59, 61
- Devore §12.1 Q 2, 3, 4

N.B. The “fourth spread” is the same as the interquartile range, and the fourths are the same as the quartiles.

Where the text asks for a stem and leaf display, construct a histogram instead. Use a multiple boxplot instead of a comparative stem and leaf display.

Lecture 2 Objectives

- Be able to find a sample standard deviation for a set of data by hand.
- Be able to calculate the lower quartile, median, upper quartile and inter-quartile range by hand.
- Produce and interpret measures of variation using statistical software.
- Be able to use graphical and numerical summaries to describe the distribution of the data.
- Use appropriate methods to investigate bivariate relationships.
- Be able to calculate and interpret a z-score.



http://xkcd.com/539/