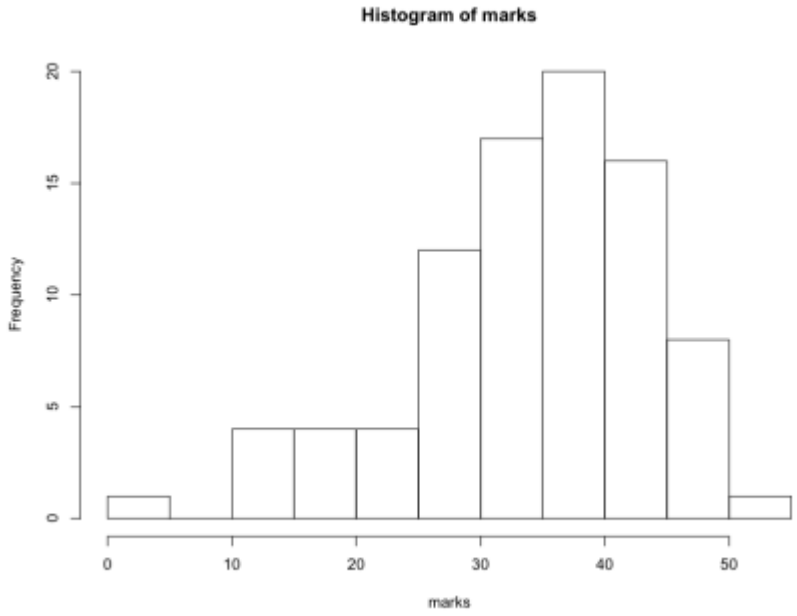
STATS 260 Class 3

Gavin Jaeger-Freeborn

1. Histograms



Modal

Unimodal	only one mode
Bimodal	2 modes
Multimodal	more then 2

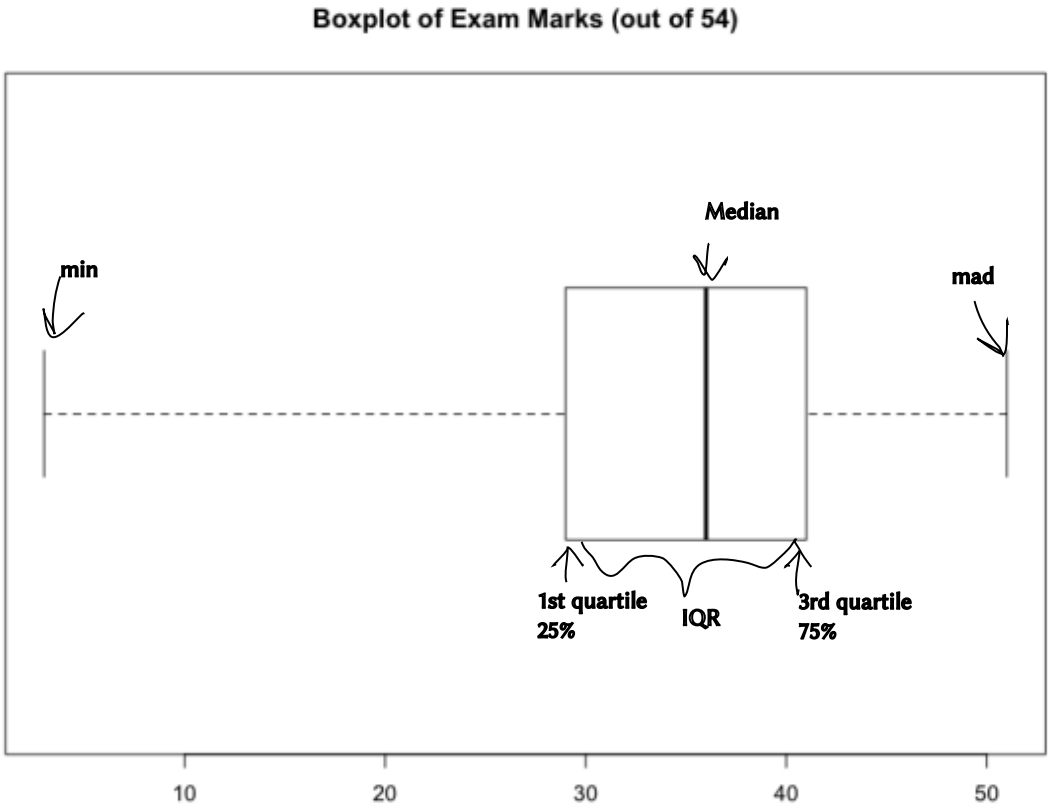
symmetric	even tail on both sides
asymmetric	uneven tail

In this example it is unimodal

Skewed

Where the data is mostly tailing in terms of the Mode(peak)
In this case it is negatively skewed

2. Boxplot



2.1. Interquartile

Range is the area between the 1st quartile and 3rd quartile

2.2. Outliers

Outside of the interval

$$[lowerquartile - 1.5\ IQR, upperquartile + 1.5\ IQR]$$

2.3. Example

Suppose I have the following sample data:

$1.7, 0.9, 3.8, 2.1, 1.9, 0.6, 0.5, 5.0, 2.4, 0.1, 5.0, 0.3, 8.8, 0.3, 0.3, 3.3, 4.8, 0.2, 2.2, 3.5$

I've used R to find that the lower quartile is **0.45**, the upper quartile is 3.575, and the IQR is $3.575 - 0.45 = 3.125$

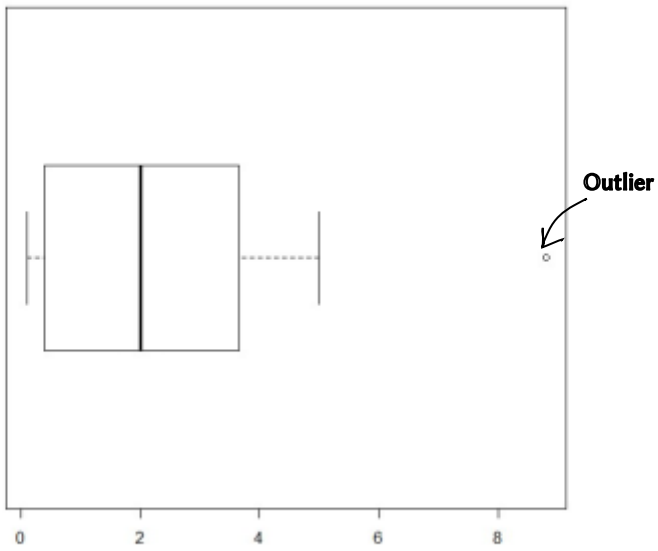
3.575, and the IQR is $3.575 - 0.45 = 3.125$

NOTE: Sometimes outliers are actually an error

Data outside $[0.45 - (1.5)(3.125), 3.575 - (1.5)(3.125)] = [-4.2375, 8.2625]$ Would be an **Outlier**.

8.8 which is outside that range. The right whisker ends at 5.0 (our largest non-outlier),

The Outlier 8.8 is indicated with a circle.



3. Bivariate data

Two variables. Set of pairs: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Common question - Whether or not there is a relationship between the two variables.

4. Scatterplot

A **Scatterplot** is used to visually depict **bivariate** data. The observations are plotted as a set of points on the plane.

Important

For a scatterplot to be appropriate, each pair of measure-

Example

I select 20 people, and for each person, I record **x**, their age, and **y**, their maximum heart rate.

Here, the data is **clearly bivariate** (one sample of size n = 20, with pairs of measurements being made);

A scatterplot would be appropriate.

Example

I select 20 people and put them on Diet A, and measure x, their blood pressure after two weeks. I select another 20 people and put them on Diet B, and measure y, their blood pressure after two weeks.

Here, we have two samples, of sizes $n_1 = 20$ and $n_2 = 20$. The data is **not bivariate**;

a scatterplot would be inappropriate.

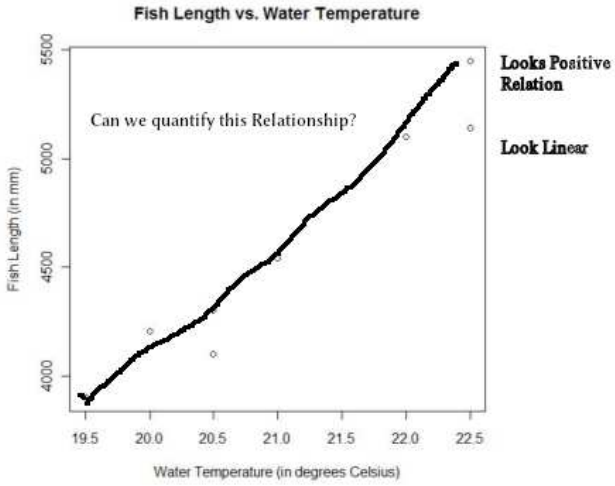
4.1. Scatterplot Example

Several of a particular species of fish are grown from eggs in tanks set at particular temperatures. After a fixed number of days, all fish are measured.

We wish to investigate the relationship between **y**, the length of the fish (in mm), and **x**, the temperature of the tank (in degrees Celsius).

y	3900	4205	4100	4300	4540	5100	5450	5140
x	19.5	20	20.5	20.5	21	22	22.5	22.5

n = 8



4.2. Sample Correlation Coefficient (r)

Used assess the **linearity** of **bivariate data.**

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{i})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Computation Form

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Denominator could be written in terms of s_x and s_y (the standard deviation of x and y, respectively).

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{(n-1)s_x s_y}$$

Example: For our fish data:

We find $\sum_{i=1}^{n} x_i y_i = (19.5)(3900) + \cdots + (22.5)(5140) = 778165.$

Then, we find $\bar{x}, \bar{y}, s_x, s_y$ using our calculator. We have r ≈ 0.973.

NOTE: This is fairly linear

Interpretation

r takes on values **between -1 and 1**. *no units*

- An r value of -1 indicates a perfect **decreasing linear** relationship.
- An r value of 1 indicates a perfect **increasing linear** relationship.
- An r value of 0 indicates a **non linear** relationship.

Warning An r value of 0 does not mean there is no relationship, only that the relationship is not linear.

5. Correlation Vs Causation

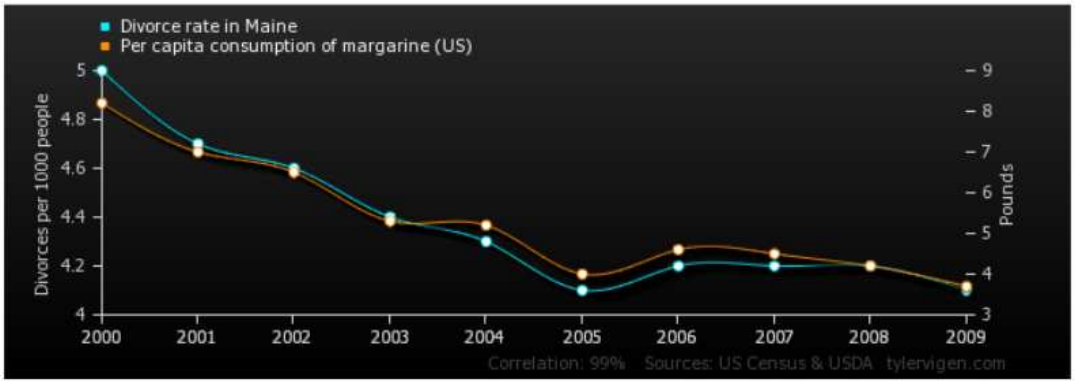
When we examine variables x and y and find there appears to be some correlation between them, there are many possible explanations:

- x causes y
- y causes x
- There is some other unexplored variable which relates to both x and y
- The correlation is spurious (there's no actual relationship; the correlation is just a coincidence)

NOTE. Spurious = no real correlation

Example

The image below shows that there appears to be a strong correlation between the divorce rate in Maine and the consumption of margarine. This is one of many examples of spurious correlation.



6. Introduction to Probability

Experiment

An activity we measure, or observe the results **Example** - Flipping a coin three times and noticing the sequence of heads and tails is an experiment.

Outcomes

The observations from our experiment.

Sample Space *S*

The set of all possible outcomes. The sample space may contain a finite or an infinite number of outcomes.

Sample Point

A single outcome in the sample space.

Event

Any subset of S (i.e. any collection of outcomes).

Simple event

An event consisting of one outcome.

Compound event

An event consisting of more than one outcomes.

6.1. Example

Consider the experiment where we flip a coin three times and note the sequence of heads and tails.

For this experiment the sample space is as follows:

$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$

Each of these eight elements of S are sample points. Some examples of events are:

$A = \{HHH, HHT, HTH, THH\} \leftarrow$ at least 2 heads

$B = \{HHT, HTT, THT, TTT\}$

$C = \{HHH, TTT\}$

Events are usually described in words. For example, B is the event that the third flip is tails

We say that an event **occurs** if one of its sample points is an observedwhen we carry out the experiment when we carry out the experiment.

7. Set Theory

A and B	The **intersection** of A and B is $A \bigcap B$
A or B	The **union** of A and B is $A \bigcup B$
not A	The **complement** of A is $\bar{A}$ or A'

Example

Suppose we select an integer from 1 to 10 at random. Let A be the event that an even number is selected. Let B be the event that a number 7 or larger is selected.

Find $A \bigcap B$, $A \bigcup B$, and$\bar{B}$

$$S = \{1, 2, \cdots, 10\},\ A = \{2, 4, 6, 8, 10\},\ B = \{7, 8, 9, 10\}$$

$$1 \rightarrow \text{outcome sample point}$$

$$A \bigcap B = \{8, 10\}$$

$$A \bigcup B = \{2, 4, 6, 8, 10, 7, 9\}$$

$$B' = \{1, 2, 3, 4, 5, 6\}$$