Statistics

Week 2: Summarizing and Exploring Data (Chapter 4)

ESD, SUTD

Term 5, 2017

Update

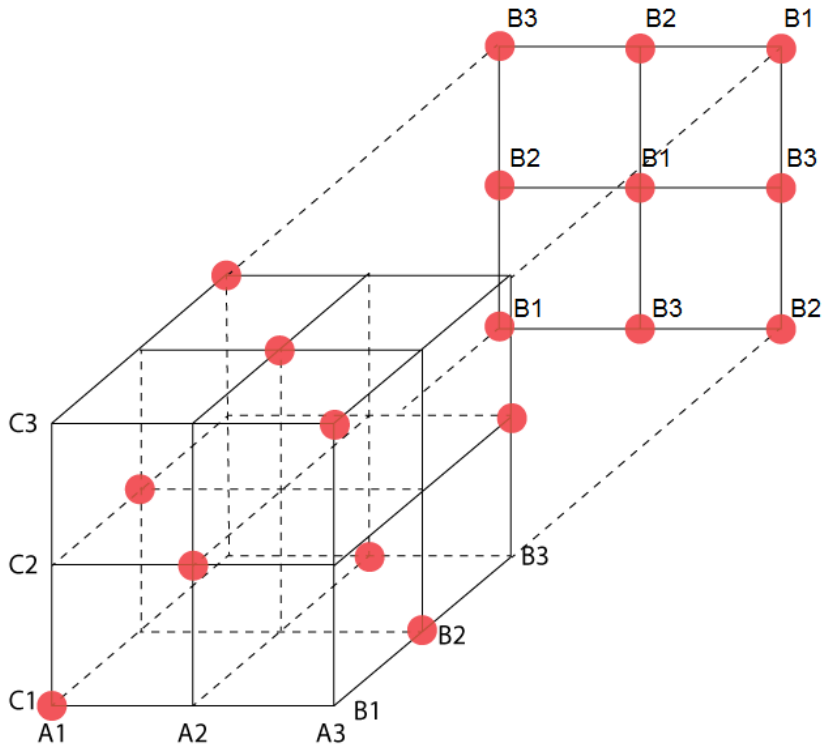
Wednesday 1 Feb, 11am – 1pm: makeup class, TT21.

Thursday 2 Feb, 1pm – 2pm: recitation, TT21.

Thursday 2 Feb, 2pm – 3pm: project recitation, **TT22**.

Recap

Last week, factorial design:



The projection is a Latin square!

Outline

Types of data

Data can be classified into two types: **categorical** (qualitative) and **numerical** (quantitative).

Categorical data can be either *nominal* (distinct labels, such as red, blue, yellow) or *ordinal* (ranked, such as disagree, neutral, agree).

Numerical data can be either *discrete* (results of counting, such as number of people) or *continuous* (results of measurement, such as distance).

Frequency table

A frequency table can be used to show the number of occurrences for each category. The relative frequency (the proportion in each category) can also be given.

Example: a survey of 100 people is conducted on the top 2 reasons why they are late to class or work.

Reason	Frequency	Relative frequency (%)
Bad weather	24	12
Overslept	58	29
Alarm failure	36	18
Family issue	6	3
Traffic	68	34
Other	8	4
Total	200	100

Bar chart

A bar chart uses rectangular bars to denote the frequencies.

Example: use *Excel*'s column chart or bar chart option to construct a bar chart for the previous table.

Note: the bars in a bar chart should not touch, in order to separate the different categories.

Pie chart

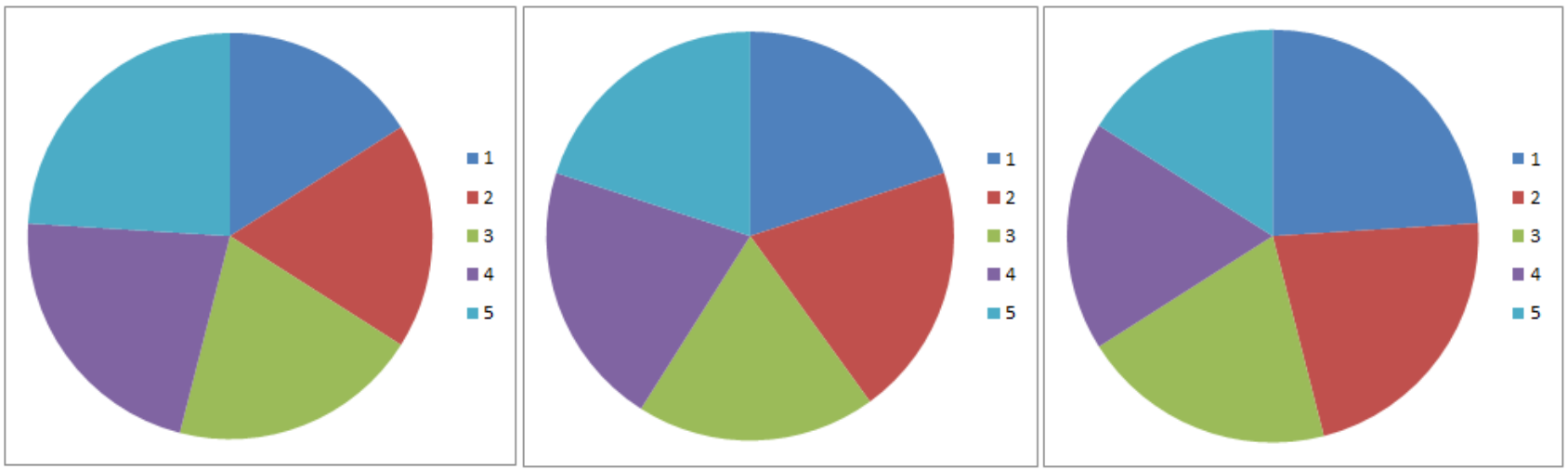
A pie chart uses the relative area, or angle, of the sectors represent the relative frequencies. An informative pie chart should label the frequencies.

Many experts *do not recommend* the use of pie charts. There are several reasons for this: pie charts are very often misused; also, humans are not good at comparing angles.

Do not use pie charts for too many or too few (such as 2) categories. Do not use 3D pie charts.

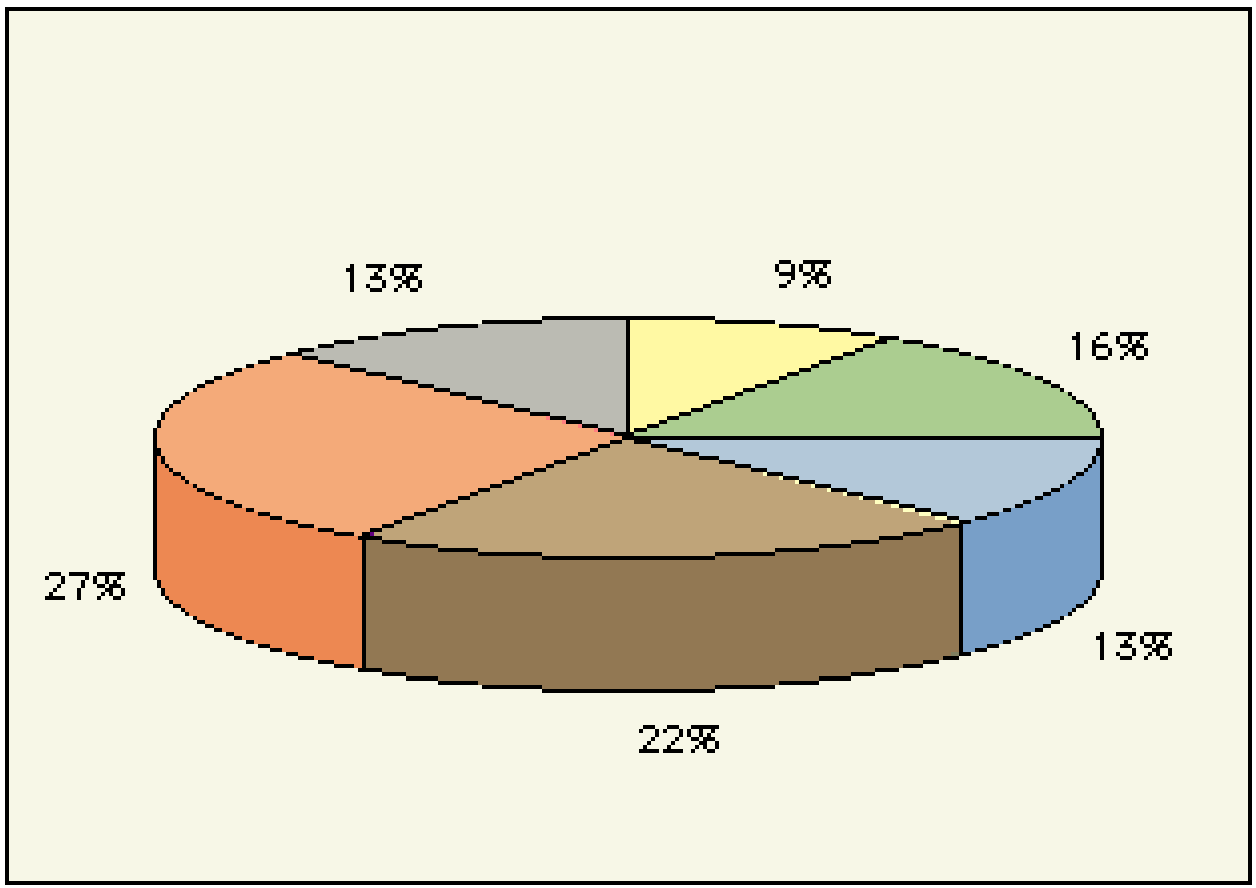
Misuse 1

The charts below represent the results from a local election with five candidates at three different locations. What are some of the problems with using pie charts here?



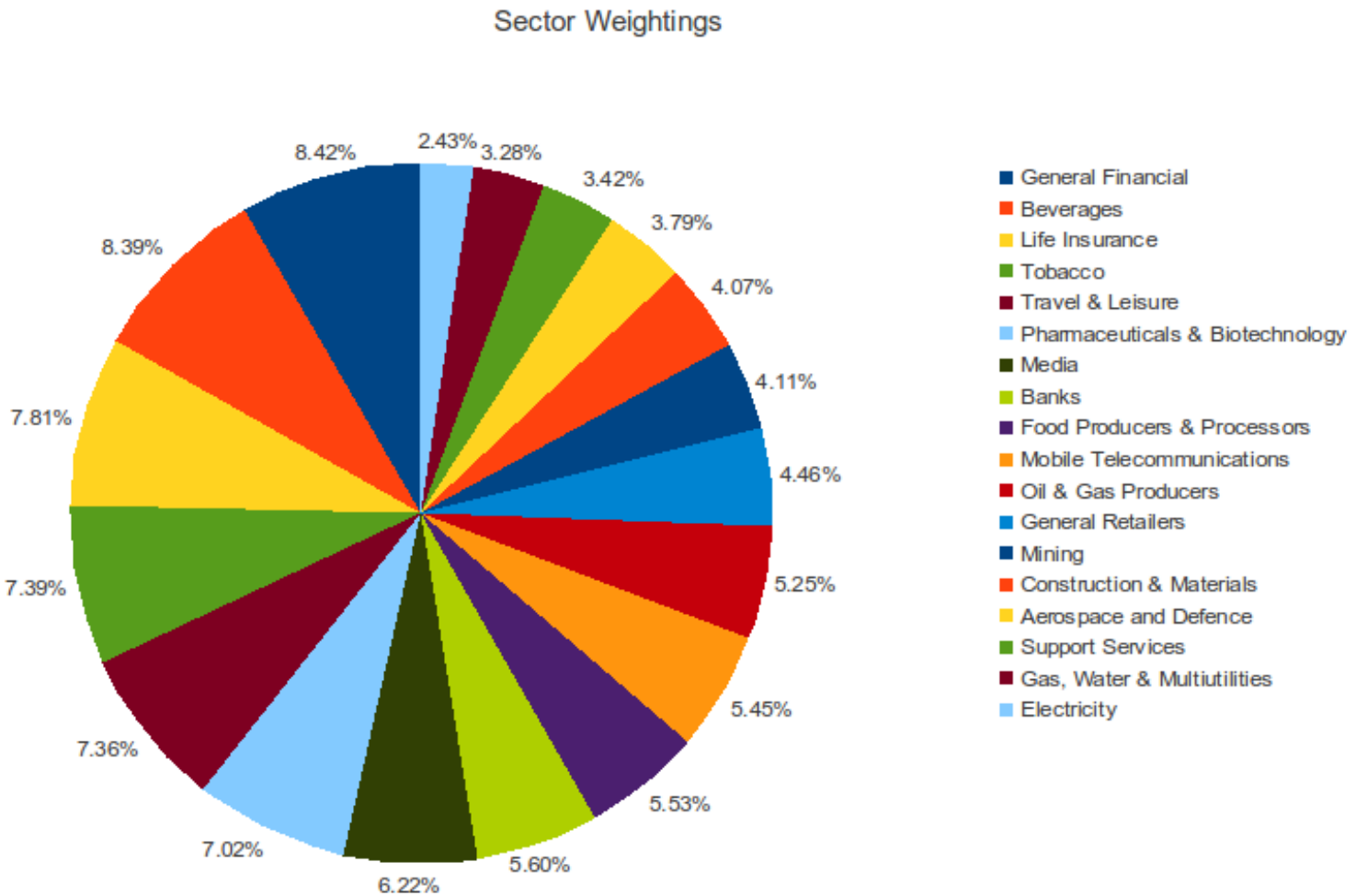
Misuse 2

Is this pie chart is a good representation of the data?



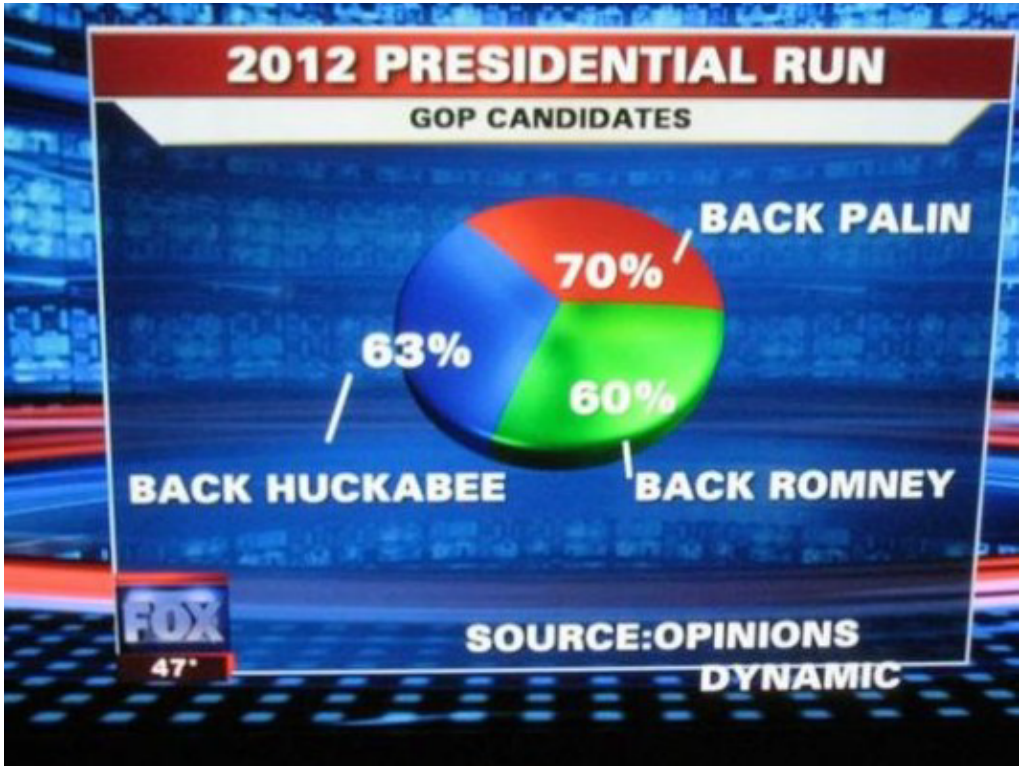
Misuse 3

Is this pie chart a good representation of the data?



Misuse 4

This chart was used on FOX News.



Outline

Summary statistics – measures of centre

Mean (average)

Given data values $x_1, x_2, \ldots, x_n$, the (sample) mean is defined as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

The sample mean is used to estimate the population mean μ (more on this later).

Median

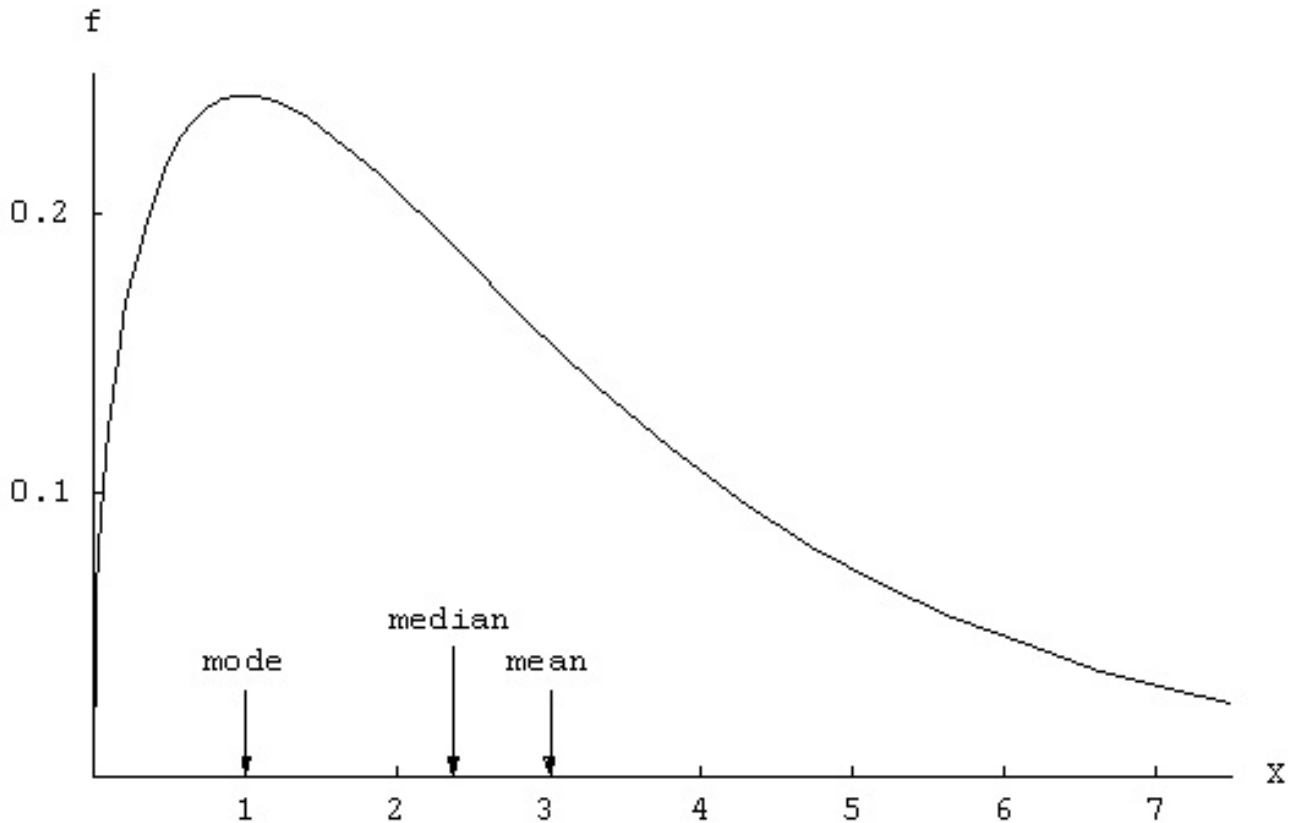
Given data values $x_1, x_2, \ldots, x_n$, order them as follows: $x_{\text{min}} = x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} = x_{\text{max}}$. The median is defined as

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right] & \text{if } n \text{ is even.} \end{cases}$$

Mean, median and mode

The *mode* is a value that appears most often in a data set. It is not always a good measure of center, nor is it always unique.

You should know how to find the (population) mean, median and mode for a continuous or discrete distribution.



Optimization

Exercise: (1) Show that $\bar{x}$ is the optimal solution to

$$\min_x \sum_{i=1}^{n} (x_i - x)^2.$$

(2) Show that $\tilde{x}$ is an optimal solution to

$$\min_x \sum_{i=1}^{n} |x_i - x|.$$

Hint: draw a picture, and use the triangle inequality, $|a| + |b| \geq |a + b|$.

Answer to (2)

Order the data values as $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. The proof here assumes that n is even; the odd case is very similar.

We have

$$\begin{aligned}\sum_{i=1}^{n} |x_i - x| &= \Big(|x_{(1)} - x| + |x_{(n)} - x|\Big) + \Big(|x_{(2)} - x| + |x_{(n-1)} - x|\Big) \\ &\quad + \cdots + \Big(|x_{(n/2)} - x| + |x_{(n/2+1)} - x|\Big) \\ &\geq |x_{(1)} - x_{(n)}| + |x_{(2)} - x_{(n-1)}| + \cdots + |x_{(n/2)} - x_{(n/2+1)}|,\end{aligned}$$

where we have applied the triangle inequality $n/2$ times.

Note that the last line depends only on the data values, and not on x, so it is fixed; let us call this fixed value X.

Answer to (2)

So we have

$$\sum_{i=1}^{n} |x_i - x| \geq X,$$

and equality is achieved if and only if x lies between $x_{(1)}$ and $x_{(n)}$, and between $x_{(2)}$ and $x_{(n-1)}$, ..., and between $x_{(n/2)}$ and $x_{(n/2+1)}$.

In other words, equality is achieved if and only if x is between $x_{(n/2)}$ and $x_{(n/2+1)}$ (since the data values are ordered).

Therefore $\sum_{i=1}^{n} |x_i - x|$ achieves its minimum value when x is between $x_{(n/2)}$ and $x_{(n/2+1)}$, which occurs, for instance, when x equals the median $\tilde{x}$.

Use of the mean

Example 1: SUTD jelly bean contest.

The entries were:

404, 225, 1228, 1119, 1117, 1234, 1125, 5566, 1234, 920, 1695, 987, 1400, 1467, 1425, 1650, 1545, 1600, 1250, 1272, 1350, 1783, 1199, 2359, 777, 777, 1500, 908, 1317, 1445, 1876, 888, 1370, 1560, 1000, 688, 1360, 1275, 1700, 2215, 1234, 911, 1028, 1524, 888, 945, 159, 1212, 1518, 999, 1456, 1200, 1313, 1086, 2359, 1763, 1800, 1452, 1500, 857, 1239

If we take out the 'obviously' too high guess of 5566, and the 'obviously' too low guesses of 159 and 225, then the sample mean of the remaining guesses is 1315.57, which is closer than the winning entry of 1317.

Use and misuse of the mean

Example 2: a stopped clock shows the correct time twice a day, while a clock that is one minute too slow never shows the correct time.

Naïvely, it may seem that the stopped clock is more accurate. However, if we compare the *mean* error of each clock, then the slow clock is far more accurate.

Example 3:

"The average annual income of leading research mathematicians (those, say, with at least three articles in the Annals of Mathematics) is about 10,000,000 USD." – James Simons

Outlier and robustness

Outlier

An outlier is a data point that is 'distant' from the main body of data points. Outliers can occur by chance in any distribution, but they are often indicative either of measurement *error* or that the population has a *heavy-tailed* distribution.

There is no hard and fast rule for detecting outliers.

The median is a *robust* statistic, meaning that it is resistant to outliers, while the mean is not robust.

If an outlier is a measurement error, then we should discard it or use a robust statistic. If an outlier results from a heavy-tailed distribution, then we should be cautious in using tools that assume a normal distribution.

Summary statistics – measures of spread

Sample variance

Given data values $x_1, x_2, \ldots, x_n$, the sample variance is defined as

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

The sample standard deviation is s.

The population variance is denoted by σ^2.

We will explain the '$n-1$' later.

Range, interquartile range

The *range* is define as $x_{\max} - x_{\min}$.

The *interquartile range* IQR is defined as $Q_3 - Q_1$. *One* definition of Q_1 (resp. Q_3) is the median of the lower (resp. upper) half of the ordered data. If there are an odd number of data points, do not include the median in either half.

The five number summary is $\{x_{\min}, Q_1, \tilde{x}, Q_3, x_{\max}\}$. They can be used to construct a **box plot**, which is useful for comparing different data sets.

For a box plot, data values that are more than 1.5 IQR below Q_1 or above Q_3 are usually considered outliers.

Box plots are poorly supported in *Excel*, but can be done with *R*.

Quantiles

Exercise: which of standard deviation, range, and IQR are robust?

The pth quantile $\tilde{x}_p$ is a value such that fraction p of the data are less than or equal to it.

There are different definitions! The textbook uses

$$\tilde{x}_p = x_{(m)} + \big[p(n+1) - m\big]\,\big(x_{(m+1)} - x_{(m)}\big),$$

where m denotes the integer part of $p(n+1)$. (For some values of p, $\tilde{x}_p$ is not defined.)

You may check that $\tilde{x} = \tilde{x}_{0.5}$. On the other hand, another definition for the quartiles is $Q_1 = \tilde{x}_{0.25}$ and $Q_3 = \tilde{x}_{0.75}$, but this is *different* from the one given on the last slide. The *Excel* command is `quartile.exc`.

Histogram

To construct a histogram, first divide up the range of values into intervals (*bins*), then counts how many values fall into each bin.

For each bin, a rectangle is drawn with height proportional to the count and width equal to the bin size. The rectangles should touch each other.

A histogram can be used to estimate the probability distribution of a continuous variable.

It is often a good idea to use at least 2 different plots to explore a data set.

Exercise: see *Excel* file on speed of light data.

- Find the mean, standard deviation, the five number summary, and any outliers.
- Make a box plot (by hand).
- Make a histogram.