Statistics

Week 2: Summarizing and Exploring Data (Chapter 4)

ESD, SUTD

Term 5, 2017

SINGAPORE UNIVERSITY OF
TECHNOLOGY AND DESIGN

Established in collaboration with MIT

Update

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

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

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

Outline

1. Categorical data

2. Numerical data

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 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

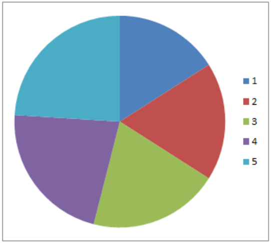
A pie chart uses the relative area, or angle, of the sectors represent the relative frequencies. An informative pie chart should label the frequencies.

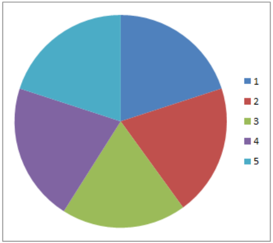
Many experts *do not recommend* the use of pie charts. There are several reasons for this: pie charts are very often misused; also, humans are not good at comparing angles.

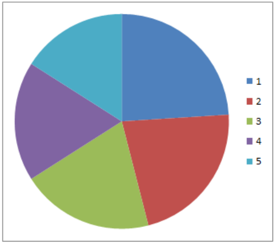
Do not use pie charts for too many or too few (such as 2) categories. Do not use 3D pie charts.

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?

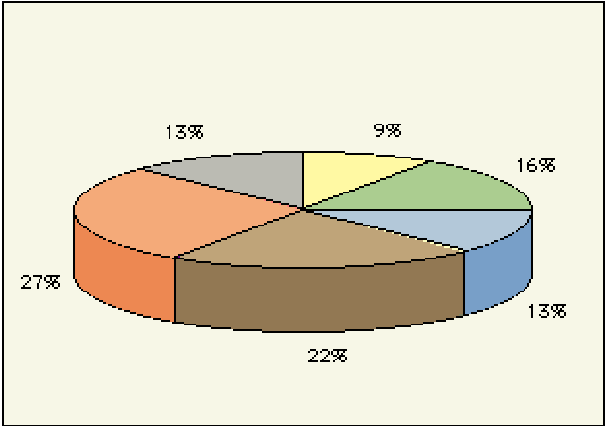






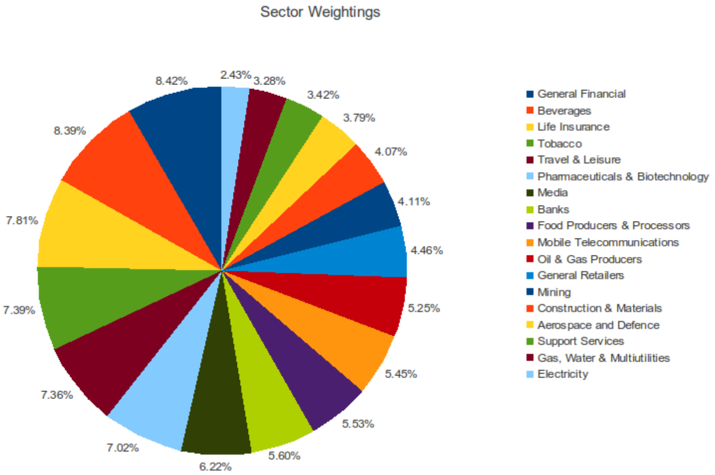
Misuse 2

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



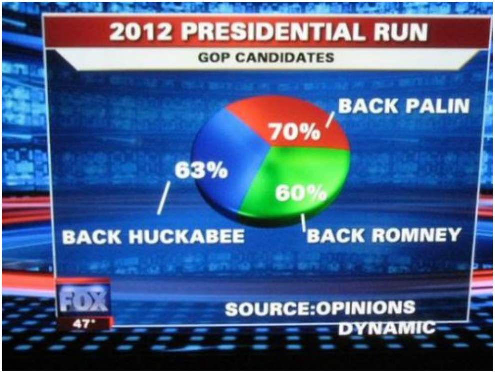
Misuse 3

Is this pie chart a good representation of the data?



Misuse 4

This chart was used on FOX News.



Outline

1. Categorical data
2. Numerical data

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

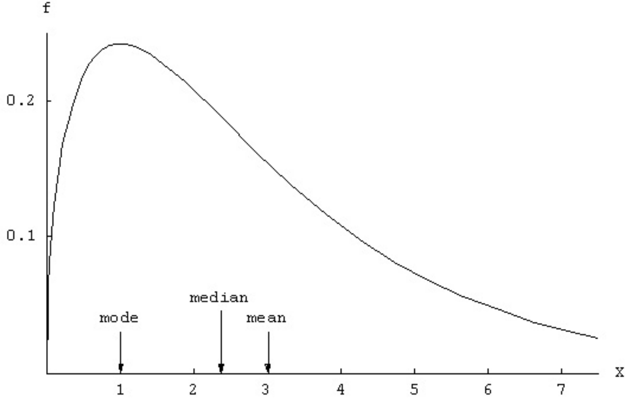
Given data values $x_1, x_2, \ldots, x_n$, order them as follows: $x_{\text{min}} = x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} = x_{\text{max}}$. The median is defined as

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

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 (or figure out) how to find the mean, median and mode for a continuous 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,
$|a| + |b| \geq |a + b|$.

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* common 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 one of range and IQR is 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

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

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

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.

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.