Chapter 1 What is statistics?

Statistics is a branch of mathematics dealing with the collection, analysis, interpretation and presentation of numerical data

In many disciplines, the collection and analysis of data is the basis to create scientific knowlege. Statistical techniques play a fundamental role in medicine, economics, engineering, biology, psicology

Statistics is concerned with two different topics.

(1) Design of experiments. Design the ways to collect data more effectively and impartially

(2) statical inference (or Statistical Learning). Analyze the data to gain information

When comparing two different statistical methods we are interested in which methods give the most possible information under a limited cost.

Characterizing a set of measurements: graphical methods.

For one dimensional data, a common representation is a histogram.

A frequency table is constructed by dividing the real line into intervals and counting the number of data points in each interval. For each interval, the number of data points and/or the percentage of the values of the data in each interval are displayed. The relative frequency distribution is the percentage of the values of the data each interval.

Sometimes cumulative frequencies are displayed. The cumulative frequency of an interval is the percentage of the values of the data in that interval and all intervals smaller than that one.

A histogram is constructed from a frequency table. For each interval, a rectangle is represented. The height of the rectangle is the (relative) frequency of that interval.

The used intervals are called bins.
Usually, the bin width is the same for each bin.
So, a histogram is found taking

$a,\ a+h,\ \dots,\ a+Mh.$

and letting

f_i = number of observations in the interval $(a+(i-1)h,\ a+ih]$

When looking to a histogram, we must discern the shape of the histogram.
We must look whether the histogram is symmetric or skewed.
We must look whether the histogram has one or more bumps

one modal

bimodal.

Numerical methods to characterize a set of measurements

Assume that the data consists of $y_1 \cdots y_n$.

The mean of the data $\bar{y}$ is $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ = sample mean

$\bar{y}$ is the average of the values of the data.

$\bar{y}$ is a measure of location (or central tendency) of the data

The data $y_1 \cdots y_n$ is around $\bar{y}$

Suppose that we order the data $y_1 \cdots y_n$ into $y_{(1)} \cdots y_{(n)}$ from the smallest the largest value

The sample median m is

$$m = \begin{cases} y_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{y_{(n)} + y_{(n+1)}}{2} & \text{if } n \text{ is even} \end{cases}$$

The sample median is another location parameter

Roughly 50% of the observations are below the median and 50% of the observations are above the median

The sample variance is $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$

The sample standard deviation is $s = \sqrt{s^2}$

Empirical rule

For a distribution of measurements that is approximately (bell-shaped) it follows that the interval with endpoints

(1) $\mu \pm \sigma$ contains approximately 68% of the observations

(2) $\mu \pm 2\sigma$ contains approximately 95% of the observations

(3) $\mu \pm 3\sigma$ contains approximately all observations.