1. STATISTICS

In statistics we are faced with data, which could be measurements in an experiment, responses in a survey, etc. There will be some randomness, which may be inherent in the problem or due to errors in measurement. The problem in statistics is to make various kinds of inferences about the underlying distribution, from realizations of the random variables. We shall consider a few basic types of problems encountered in statistics. We mostly deal with examples, but sufficiently many so that the general ideas should become clear too. It has become accepted in today's world that in order to learn about something, you must first collect data. Statistics is the art of learning from data. It is concerned with the collection of data, its subsequent description, and its analysis, which often leads to drawing of conclusions.

2. Descriptive Statistics

Univariate (real-valued) data: $X_1, \ldots, X_n$ in $\mathbb{R}$

Data type:

- discrete (say, outcomes of the tosses of a coin),
- continuous (say, heights of students in MSO201A).

2.1. **Data visualization.** We can plot univariate data!
For discrete data, one can plot a bar graph or a pie chart.
For continuous data, we can do the following:

- frequency table of the data, and draw a bar graph/frequency polygon,
- divide the data into groups, and draw a histogram.

Use the following links (with R codes) to explore further:
Discrete data: http://www.r-tutor.com/elementary-statistics/qualitative-data.
Continuous data: http://www.r-tutor.com/elementary-statistics/quantitative-data.
Also, check Chapter 2 of Sheldon M. Ross's book for detailed description of the various types of plots stated above.

Additional links:
http://www.r-tutor.com/elementary-statistics, and
https://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf.

2.2. **Data summarization.** We will now look into the sample analogues (based on the data $X_1, \ldots, X_n$ in hand) of the population versions. Also, recall Module 2 of Note 5.

Central tendency: The sample mean and sample median are both useful statistics for describing the central tendency of a data set. The sample mean makes use of all the data values and is affected by extreme values that are much larger, or smaller than the others; the sample median makes use of only one, or two of the middle values and is thus not affected by extreme values. A measure of central tendency, or location (also called an average) gives us idea about the central value of the probability distribution around which values of the random variable are clustered. Three commonly used measures of central tendency are mean, median and mode.

Sample Mean: $\bar{X}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$ (*recall the population mean*).

Recall that $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ are the order statistics of $X_1, \ldots, X_n$.
Sample Median: M_n is $X_{((n+1)/2)}$ if n is odd, and any number in $[X_{(n/2)}, X_{((n+2)/2)}]$ if n is even (*recall the population median*).

Dispersion: We have presented statistics that describe the central tendencies of a data set, we are also interested in ones that describe the spread, or variability of the data values. A statistic that could be used for this purpose would be one that measures the average value of the squares of the distances between the data values and the sample mean. This is accomplished by the sample variance.

Measures of central tendency give us the idea about the location of only central part of the distribution. Other measures are often needed to describe a probability distribution. The values assumed by a random variable X usually differ from each other. The usefulness of mean (or, median) as an average is very much dependent on the variability (or, dispersion) of values of X around the mean (or, median). A probability distribution (or, the corresponding random variable X) is said to have a high dispersion if its support contains many values that are significantly higher, or lower than the mean (or, median) value. Some of the commonly used measures of dispersion are standard deviation, quartile deviation (or, semi-inter-quartile range) and coefficient of variation.

Sample Standard Deviation: $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$ (*recall the population standard deviation*).

Sample Quartile Deviation: $(X_{([3n/4])} - X_{([n/4])})/2$ (*recall the population quartile deviation*). Here, $X_{(r)}$ is the rth order statistic.

Sample Mean Absolute Deviation: $\frac{1}{n}\sum_{i=1}^{n}|X_i - \bar{X}_n|$, or $\frac{1}{n}\sum_{i=1}^{n}|X_i - M_n|$ (*recall the population mean absolute deviation*)

Skewness and kurtosis: Skewness is a measure of asymmetry of the probability distribution about its mean. The skewness value can be positive, zero or negative. A negative skew commonly indicates that the tail is on the left side of the distribution, while positive skew indicates that the tail is on the right. Zero value indicates that the tails on both sides of the mean balance out.

Sample Skewness: $\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^3}{\{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\}^{3/2}}$ (*recall the population skewness*).

Exercise 1. For a symmetric distribution, prove that the population skewness is 0.

Like skewness, kurtosis describes the shape of a probability distribution. The standard measure of kurtosis is a scaled version of the fourth moment of the distribution, and it measures the peakedness and thickness of the tail of a probability distribution.

Sample Kurtosis: $\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^4}{\{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\}^2}$ (*recall the population kurtosis*)

Exercise 2. The population kurtosis of any univariate normal distribution is 3.

It is common to compare the kurtosis of a distribution to the value 3. Distributions with kurtosis less than 3 are said to be platykurtic. Example of a platykurtic distribution is the uniform distribution. Distributions with kurtosis greater than 3 are said to be leptokurtic. Example of a leptokurtic distribution is the Laplace distribution. It is also common practice to use an adjusted version of Pearson's kurtosis, the excess kurtosis, which is the kurtosis minus 3 (to provide the comparison to the standard normal distribution).

Use the link below (with R codes) to explore further:
http://www.r-tutor.com/elementary-statistics/numerical-measures.

Let us now look into one the following plot:

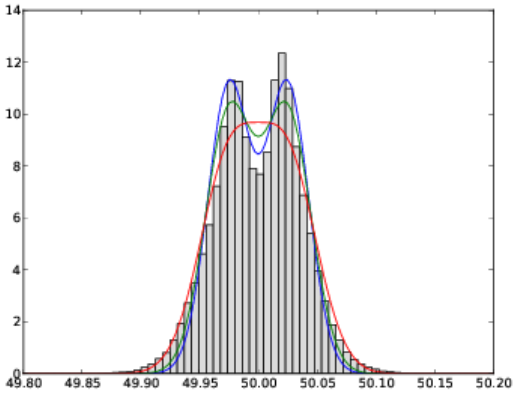

FIGURE 1. A typical histogram

This figure suggests that we should fit an *appropriate* pdf f to the data $X_1, \ldots, X_n$ in hand.

3. STRUCTURE

General setting: We have data (measurements perhaps), usually of the form $X_1, \ldots, X_n$ which we assume are realizations of independent random variables from a common distribution. The underlying distribution is *not* known. In the problems we consider, typically the distribution is known, except for the values of a few parameters. Thus, we may write the data as $X_1, \ldots, X_n$ i.i.d. f_θ, where f_θ is a pdf (or, pmf) for each value of the parameter(s) θ.
For example, the density could be of $N(\mu, \sigma^2)$ (two unknown parameters μ and σ^2), or of Ber(p) (one unknown parameter p).

(1) Estimation: Here, the question is to guess the value of the unknown θ from the sample $X_1, \ldots, X_n$. For example, if X_i are i.i.d. from Ber(p) distribution (p is unknown), then a reasonable guess for p would be the sample mean $\bar{X}_n$ (an *estimator*). Is this the only one? Is it the "best" one? Such questions are addressed in estimation. An estimator is *always* a function of the data points $X_1, \ldots, X_n$.

(2) Confidence intervals: Here, again the problem is of estimating the value of a parameter, but instead of giving one value as a guess, we instead give an interval and quantify how sure we are that the interval will contain the unknown parameter. For example, a coin with unknown probability p of turning up head, is tossed n times. Then, a confidence interval for p could be of the form

$$\left[\bar{X}_n - \frac{3}{\sqrt{n}}\sqrt{\bar{X}_n(1-\bar{X}_n)}, \bar{X}_n + \frac{3}{\sqrt{n}}\sqrt{\bar{X}_n(1-\bar{X}_n)}\right],$$

where $\bar{X}_n$ is the proportion of heads in n tosses. The reason for such an interval will come later. It turns out that if n is large, one can say that with probability 0.99 ("confidence level"), this interval will contain the true value of the parameter.

(3) Hypothesis testing: In this type of problems, we are required to decide between two competing choices ("hypotheses"). For example, it is claimed that one batch of students is better than a second batch of students in mathematics. One way to check this is to give the same exam to students in both batches, and record the scores. Based on the scores, we have to decide whether the first batch is better than the second (one hypothesis), or whether there is not much difference between the two (the other hypothesis). One can imagine that this can be done by comparing the sample means etc., but that will come later.

A good analogy for testing problems is from law, where the judge has to decide whether an accused is guilty, or not guilty. Evidence presented by lawyers take the role of data (but of course, one does not really compute any probabilities quantitatively here!).