1 Lecture 1

1.1 Introduction, Sampling, Histograms

Statistics is a mathematical science pertaining to the collection, analysis interpretation and presentation of data. Statistical methods can be used to summarize or describe a collection of data, this is called descriptive statistics. Patterns in the data may be modeled in a way that accounts for randomness and uncertainty in the observation and are then used to draw inferences about the process or population.

The term population is often used in statistics. By a population we mean a collection of items. This maybe an actual population of person, animals, etc or a population in the extended sense, such as a collection of plants etc. Each member of the population has certain characteristics, which we would like to measure. There are two issues: A. it is not feasible to measure the characteristic of each member of the population (the size of the total population is too big), B. we would like to be able to say something about future population with the same characteristic attached to them.

What we do in such situations is known as sampling. That is, we choose randomly a number of representatives out of the population. The selection of a particular member of the population is not supposed to affect the choice or not of any other member. This means that each member of the population is chosen (or not) independently of each other member. The chosen set is referred as sample. The next step is to measure the characteristic of interest on each member of the sample, collect the data, organise them and finally extract some information on the characteristic of the population, as a whole. This means that we will try to determine the distribution of the characteristic. Often, it will also be the case that we know (or pretend to know) the distribution of the characteristic, up to certain parameter. In this case our effort concentrates on determining the parameters out this sampling procedure. A crucial assumption on such a procedure is that of independence. The characteristic of each member of the population is assumed to be independent of that of the another. This assumption might not always be true (especially in financial data), but later on we will see why such an assumption is important.

The term random variable is used to describe a variable that can take a range of values and applies when taking a sample of items. For instance with a population of people, 'height' is a random variable.
Let us start describing some methods to organise the data we collect.

**Example 1** Suppose we want to make a study regarding the incomes of the salespeople in Coventry. A statistics student makes an inquiry into the incomes of 20 sales people that he/she picks at random and comes up with the following table:

<table>
<thead>
<tr>
<th>Income (in pounds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>850</td>
</tr>
<tr>
<td>620</td>
</tr>
</tbody>
</table>

Such a process is called sampling.

Of course this table is of little value, unless he/she tells you what exactly this numbers are. Suppose, therefore, that these numbers are the incomes in pounds.

The next step following the collection of data is their summary or classification. That is the students will group the values together. Since the values collected are not that many the student decides to group them in 5 groups, grouping them to the nearest £500.

The next step would be to measure the frequency of each group, i.e. how many times each group appears in the above sampling and the relative frequency of each group, that is

\[
\text{Frequency of Group } i = \frac{\text{number of times group } i \text{ appears in the sampling}}{\text{sample size}}
\]

The results are shown in the following table:

<table>
<thead>
<tr>
<th>Class</th>
<th>250-749</th>
<th>750-1249</th>
<th>1250-1749</th>
<th>1750-2249</th>
<th>2250-2749</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>6</td>
<td>7</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Rel. Frequency</td>
<td>.30</td>
<td>.35</td>
<td>.25</td>
<td>.05</td>
<td>.05</td>
</tr>
</tbody>
</table>
A histogram can reveal most of the important features of the distribution of the data for a quantitative variable such as

1. What is the typical, average value?

2. How much variability is there around the typical value? Are the data values all close to the same point, or do they spread out widely?

3. What is the general shape of the data? Are the data values distributed symmetrically around the middle or are they skewed, with a long tail in one direction or the other?

4. Are there outliers, i.e. wild values that are far from the bulk of the data?

5. Does the distribution have a single peak or is it clearly bimodal, with two separate peaks separated by a pronounced valley?

In the above example the average value appears to be slightly above 1000, but there is substantial variability with many values around 500 and a few around 2500. The data are skewed, with a long tail to the right. There don’t seem to be any wild values, nor separate peaks.

1.2 Stem-and-leaf diagram

A stem-and-leaf diagram is a clever way to produce a histogram like picture from the data. We introduce this by an example

Example 2 We record the grades of 40 applicants on an aptitude test. The results are as follows

| 42 | 21 | 46 | 69 | 87 | 29 | 34 | 59 | 81 | 97 |
| 64 | 60 | 87 | 81 | 69 | 77 | 75 | 47 | 73 | 82 |
| 91 | 74 | 70 | 65 | 86 | 87 | 67 | 69 | 49 | 57 |
| 55 | 68 | 74 | 66 | 81 | 90 | 75 | 82 | 37 | 94 |

The scores range from 21 to 97. The first digits, 2 through 9, are placed in a column - the stem- on the left of the diagram. The respective second digits are recorded in the appropriate rows- the leaves.

We can, therefore, obtain the following stem-and-leaf diagram

```
2 | 1 9
3 | 4 7
4 | 2 6 7 9
5 | 5 7 9
6 | 0 4 5 6 7 8 9 9 9
7 | 0 3 4 4 5 5 7
8 | 1 1 1 2 2 6 7 7 7
9 | 0 1 4 7
```
Often it is necessary to put more than one digits in the first (stem) column. If, for example, the data ranged from 101 – 1000 we wouldn’t like to have 900 rows, but we would rather prefer to group the data more efficiently.

We may also need to split the values. For example, if we had data that ranged from 20 – 43 we wouldn’t like just a stem with three columns consisting of the numbers 2, 3, 4. We would rather split the values of 20s in low 20s, i.e. 20 – 24, high 20s, i.e. 25 – 29 etc. If there were even more values we could split further the value like : 20, 21-22, 23-24, etc.

The resulting picture look like a histogram turned sideways. The advantage of the stem-and-leaf diagram is that it not only reflects the frequencies, but also contains the first digit(s) of the actual values.

1.3 Some First Statistical Notions.

Histograms and stem-and-leaf diagrams give a general sense of the pattern or distribution of values in a data set, but they really indicate a typical value value explicitly. Many management decisions are based on what is the typical value. For example the choice of an investment adviser for a pension fund will be based largely on comparing the typical performance over time of each of several candidates.

Let’s define some standard measures of the typical value.

**Definition 1** The **mode** of a variable is the value of category with the highest frequency in the data.

**Definition 2** The *(sample)* median of a set of data is the middle value, when the data are arranged from lowest to highest (it is meaningful only if there is a natural ordering of the values from lowest to highest). If the sample size $n$ is odd, then the median is the $(n + 1)/2$-th value. If $n$ is even then the median is the average of the $n/2$-th and the $(n + 2)/2$-th value.

**Example 3** Suppose we have the following data for the performance of a stock

\[
25\ 16\ 61\ 12\ 18\ 15\ 20\ 24\ 17\ 19\ 28
\]

Arranged in an increasing order the values are

\[
12\ 15\ 16\ 17\ 18\ 19\ 20\ 24\ 25\ 28\ 61
\]

The median value of the sample is 19.

**Definition 3** The *(sample)* mean or average of a variable is the sum of the measurements taken on that variable divided by the number of measurements.
Suppose that \( y_1, y_2, \ldots, y_n \) represent measurements on a sample of size \( n \) selected from a population. The sample mean is usually denoted by \( \bar{y} \):

\[
\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}.
\]

The above definitions deal with the notion of typical or average value. Since most likely data are not equal to the average, we would like a quantity that measures the deviation of our data from the typical value. The most commonly used quantities are the variance and standard deviation. These measure the deviation from the mean.

**Definition 4** Consider a sample data \( y_1, y_2, \ldots, y_n \) obtained from measurements of a sample of \( n \) members. Then the variance is defined as

\[
s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n - 1}
\]

and the standard deviation is

\[
s = \sqrt{s^2}
\]

Notice that \( (y_i - \bar{y}) \) is the deviation of the \( i^{th} \) measurement from the sample mean. So the standard deviation \( s \) should be thought as the average deviation of the sample from its mean. Someone might then naturally object that \( n - 1 \) should be replaced by \( n \). This is true at first sight, but it turns out that it is not generally preferred. If \( n \) was present then the standard deviation would be a biased estimator and what we want is an unbiased estimator. We will come to this point later on.

How should we interpret the standard deviation? There is an empirical rule, which, nevertheless, assumes that the measurements have roughly bell-shaped histogram, i.e. single peaked histogram and symmetric, that falls off smoothly from the peak towards the tails. The rule then is

**Empirical Rule**

- \( \bar{y} \pm 1s \) contains approximately 68\% of the measurements
- \( \bar{y} \pm 2s \) contains approximately 95\% of the measurements
- \( \bar{y} \pm 3s \) contains approximately all the measurements

**Definition 5** Consider sample data \( y_1, y_2, \ldots, y_n \) obtained from measurements on a sample of size \( n \). The skewness of the sample is defined as

\[
\gamma_3 = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i - \bar{y}}{s} \right)^3
\]
It is easy to see that if the sample data are symmetric around the sample mean
then the skewness is zero. So if the skewness is not zero the sample data have to
be asymmetric. Positive skewness means that the measurements have larger (often
called heavier) positive tails than negative tails. In accordance with the Empirical
Rule a positive tail should be thought of as the values of the data larger than $\mp 2s$
and a negative tail the values of the data less than $\bar{y} - 2s$.

Definition 6 Consider sample data $y_1, y_2, \ldots, y_n$. The sample kurtosis is defined
to be

$$\gamma_4 = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \bar{y}}{s}\right)^4$$

The kurtosis measures how much the data concentrate in the center and the tails
of the histogram, rather than the shoulders. We think of the tails as the values of
the data which lie above $\bar{y} + 2s$ or below $\bar{y} - 2s$. The center as the values in between
$\mu \pm \sigma$ and the rest as the shoulders.

1.4 Random Variables and Probability Distributions

Informally a random variable is a quantitative result from a random experiment.
For example each measurement performed during the sampling process can be con-
sidered as an experiment. Each measurement gives a different outcome, which is
unpredictable. In the situation of Example 1, we can define the random variable as
the income of a sales person in Coventry. Probability theory and statistics concen-
trate on what is the probability that a random variable will take a particular value,
or take values within a certain set. Statistics tries to infer this information from the
sample data. For example, in the case of Example 1, one is tempted to say that the
probability that the income of a sales person in Coventry is 750-1249 is .35. The
histogram, then, is to be interpreted as the distribution of these relative frequencies
and thus the histogram tends to represent the shape of the probability distribution
of the random variable. A random variable can take a values in a discrete set, like
the income of the sales persons. It may also take values in a continuoum set, e.g.
the exact noon temperature at Coventry. In the first case we talk about a discrete
random variable (or better discrete distibution) and in the second case about a
continuous random variable (or better continuous distribution). It is customary
to denote the random variables with capital letters $X, Y, Z, \ldots$.

In the case of a discrete random variable $Y$, we denote the probability distribu-
tion by $P_Y(y)$. This should be interpreted as $\text{Prob}(Y=y)$. Let us concentrate in
discrete random variables that take values in the real numbers.
Definition 7 (Properties of a Discrete Probability Distribution)

1. The probability $P_Y(y)$ associated with each value of $Y$ must lie in the interval $[0, 1]$
   \[ 0 \leq P_Y(y) \leq 1 \]

2. The sum of the probabilities of all values of $Y$ equals 1
   \[ \sum_{y} P_Y(y) = 1 \]

3. For $a \neq b$ we have
   \[ \text{Prob}(Y = a \text{ or } Y = b) = P_Y(a) + P_Y(b) \]

Modelled as in Example 1, let us consider the random variable $Y$ with probability distribution

<table>
<thead>
<tr>
<th>$y$</th>
<th>500</th>
<th>1000</th>
<th>1500</th>
<th>2000</th>
<th>2500</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_Y(y)$</td>
<td>.30</td>
<td>.35</td>
<td>.25</td>
<td>.05</td>
<td>.05</td>
</tr>
</tbody>
</table>

The graph of the probability distribution in this case coincides with the histogram of Example 1.

Suppose, next, that we consider the data gathered from Barclays stock (look at the end of these lecture notes) and suppose that we consider a very fine resolution for the groups. Then the histogram looks very similar to a continuous curve. This will be an approximation of a continuous probability distribution. In the case of a continuous distribution, formally, it does not make sense to ask what is the probability that the random variable is equal to, say, $y$, since this is formally zero. In this case, the graph that we obtained above, represents empirically the probability density function of the distribution. Formally, in the case of continuous random variables, it makes more sense to ask what is the probability $F_Y(y) = \text{Prob}(Y \leq y)$. In this case the function $F_Y(y)$ is called the cumulative distribution function of the random variable $Y$. The probability density function, introduced before, is usually denoted by $f_Y(y)$ and satisfies the relation

\[ f_Y(y) = \frac{d}{dy} F_Y(y). \]

The cumulative distribution function has the properties (again we assume that the random variable takes values in the real numbers)

Definition 8 (Properties of Cumulative Distribution)

1. $F_Y(\cdot)$ is nondecreasing.
2. $\lim_{y \to -\infty} F_Y(y) = 0$ and $\lim_{y \to +\infty} F_Y(y) = 1$.

3. For $-\infty < b < a < \infty$ we have $\text{Prob}(a < Y < b) = F_Y(b) - F_Y(a) = \int_a^b f_Y(y) \, dy$.

In the sequel we will define several fundamental quantities that describe the properties of random variables and their distributions. These definitions should be compared with those of sample data in the previous section.

**Definition 9**

A. The **expected valued** of a discrete random variable $Y$ is defined as

$$E[Y] = \sum_y y P_Y(y)$$

B. The **expected value** or **mean** of a continuous random variable $Y$ is defined as

$$E[Y] = \int_{-\infty}^{\infty} y F_Y(y) \, dy$$

Often the expected value of $Y$ is denoted by $\mu_Y$.

**Definition 10**

A. The **variance** of a discrete random variable $Y$ is defined as

$$\text{Var}(Y) = \sum_y (y - \mu_Y)^2$$

B. The variance of a continuous random variable $Y$ is defined as

$$\text{Var}(Y) = \int_{-\infty}^{\infty} (y - \mu_Y)^2 f_Y(y) \, dy$$

The variance of a random variable $Y$ is often denoted by $\sigma_Y^2$. The **standard deviation** of a random variable $Y$ is $\sigma_Y = \sqrt{\text{Var}(Y)}$.

Often we need to consider more than one random variables, say, $X, Y$ and we would like to ask about the probability $\text{Prob}(X = x, Y = y)$. In this case, we are asking about the **joint probability distribution**. Empirically this corresponds to measurements on two characteristics of each member of the sample, e.g. height and weight.

**Definition 11** Consider two discrete random variable $X, Y$. The joint probability distribution is denoted by $p_{XY}(x, y) = \text{Prob}(X = x, Y = y)$.
The marginal of $X$ is defined as

$$p_X(x) = \sum_y p_{XY}(x, y).$$

The conditional probability distribution of $X$ given that $Y = y$ is defined as

$$p_{X|Y}(x|y) = \frac{p_{XY}(x, y)}{p_Y(y)}.$$

Similar definitions hold for continuous variables, as well.

**Definition 12** If $X, Y$ are discrete random variables, with expected values $\mu_X, \mu_Y$, standard deviations $\sigma_X, \sigma_Y$ and with joint probability distribution $p_{XY}(x, y)$, then the covariance of $X, Y$ is defined as

$$\text{Cov}(X, Y) = \sum_{x,y} (x - \mu_X)(y - \mu_Y)p_{XY}(x, y) = E[(X - \mu_X)(Y - \mu_Y)].$$

The correlation of $X, Y$ is defined as

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Similar definition hold for continuous random variables (can you give them?).

### 1.5 Probability Plots

Probability plots are a useful graphical tool for qualitatively assessing the fit of empirical data to a theoretical probability distribution. Consider sample data $y_1, y_2, \ldots, y_n$ of size $n$ from a uniform distribution on $[0, 1]$. Then order the sample data to an increasing order, as $y_{(1)} < y_{(2)} < \cdots < y_{(n)}$. The values of a sample data, ordered in the above fashion are called the order statistics of the sample. It can be shown (see exercises) that

$$E[y_{(j)}] = \frac{j}{n + 1}$$

Therefore, if we plot the observation $y_{(1)} < y_{(2)} < \cdots < y_{(n)}$ against the points $1/(n + 1), 2/(n + 1), \ldots, n/(n + 1)$ and assuming that the underlying distribution is uniform, we expect the plotted data to look roughly linear.

We could extend this technique to other continuous distributions. The theoretical approach is the following. Suppose that we have a continuous random variable with strictly increasing cumulative distribution function $F_Y$. Consider the random
variable $X = F_Y(Y)$. Then it follows that the distribution of $X$ is the uniform on $[0, 1]$ (check this!). We can therefore follow the previous procedure and plot the data $F(Y_{(k)})$ vs $k/(n + 1)$, or equivalently the data $Y_{(k)}$ vs. $F^{-1}(k/(n+1))$. Again, if the sample data follow the cumulative distribution $F_Y$, we should observe an almost linear plot.

In the next figure we see the probability plot of a sample of 100 random data generated from a Normal distribution. The probability plot (often called in this case normal plot) is, indeed, almost linear.

In the next figure we drew the probability plot of a sample of 100 random data generated from an exponential distribution, i.e. a distribution with cumulative distribution function $1 - e^{-x}$, versus a normal distribution.
The (normal) probability plot that we obtain is clearly non-linear. In particular we see that both the right and left tails of the plot are superlinear (notice that in the middle it is sort of linear). The reason for this is that the exponential distribution has heavier tails than the normal distribution. For example the probability to get a very large value is much larger if the data follow an exponential distribution, than a normal distribution. This is because for large $x$ we have $e^{-x} \gg e^{-x^2/2}$. There the frequency of large value in a sample that follows an exponential distribution would be much larger, than a sample of normal distribution. The same happens near zero. The sample data of the exponential distribution are superlinear, since the probability density of an exponential near zero is almost 1, while for a normal it is almost $1/\sqrt{2\pi}$.

### 1.6 Some more graphs

We have collected the data from the stock value of the Barclays Bank, from 16 June 2006 until 10 October 2008. Below we have drawn the histograms of the returns $R(t) = (P(t) - P(t-1))/P(t-1)$, where $P(t)$ is the value of the stock at time $t$. We exhibit histograms with different binning size. You see that as the size of the bin get small the histogram get finer and resembles more a continuous distribution. In the third figure we tried to fit a Normal distribution and in the last figure we drew the normal probability plot.
1.7 Exercises

1. Show that if a distribution is symmetric around its mean, then its skewness is zero.

2. Suppose someone groups sample data $y_1, y_2, \ldots, y_n$ into $k$ groups $G_1, G_2, \ldots, G_k$. Suppose that the sample mean of the data in group $i$ is $\bar{y}_i$. Find the relation between the sample mean $\bar{y} = (y_1 + \cdots + y_n)/n$ and the sample means $\bar{y}_i$'s.

3. Use Minitab to generate 100 random numbers following a uniform distribution on $[0, 1]$. Draw the normal probability plot and explain it. Does it deviate from linear? Why is this? Are there any negative values in the plot? Why is this?

4. Consider a sample $Y_1, Y_2, \ldots, Y_n$ of a uniform distribution and consider the order statistics $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$. Compute $E[Y_{(k)}]$.

5. Consider a random variable with strictly increasing cumulative distribution function $F_Y$. Consider, now, the random variable $X = F_Y(Y)$ and compute its distribution.

6. Consider two random variables $X, Y$, with correlation $\rho_{XY}$. What is the meaning of A. $\rho_{XY} = 0$, B. $\rho_{XY} > 0$, C. $\rho_{XY} < 0$?