Descriptive Statistics

Types of variables / data

Data can be obtained from either **qualitative** (**categorical**) or **quantitative (numeric)** variables. On the basis of qualitative data study units can be categorized and counted. When the different categories can be ordered in some logical way the data are termed **ordinal data** and when no ordering of the categories is possible the data are termed **nominal data**.

Quantitative data are obtained by either making measurements or by counting (or quantifying) qualitative variables. Quantitative data can be either discrete or continuous. **Discrete variable** means that the variable can take only some particular values within a given range. A **continuous variable** can take any value within a given range (the possible values are infinite but in practice the values obtained are limited by the measuring instrument).

Table 1 gives some examples of each of the 4 types of variables found in the biomedical literature.

Table 1 - Examples of the 4 types of variables/data

Qualitative or categorical data		Quantitative or numeric data	
Nominal	Ordinal	Discrete	Continuous
Religion Blood group Sex	Social class Stage of disease Exam results	Number of siblings Parity Age at last birthday	Height Weight Haemoglobin level

Collecting data

Data are usually collected by using questionnaires or by making observation or measurements. Sometimes we also make use of available data or secondary data – that is we make use of data collected for some other purpose.

When collecting data we have to ensure that it is of good quality. This is achieved by paying attention to the instruments used, training of data collectors and ensuring uniform/standard conditions for data collection. These measures will reduce variance and bias.

Reliability, Repeatability & variance

A measurement can be considered as a combination of the true value and the error associated with measurement. When we repeat the measurement of height of a person (either by the same or a different observer or by using a different instrument) and the results are very close we could say that we were able to achieve low variance in measuring the height. In such situation we could say that the measurement of height has good reliability or repeatability.

Bias & Validity

If we use an old stretched tape measure to measure height, the measurement obtained will be lower than the person's actual height. That is we have introduced bias into our measurement. We could also say that the validity of the data is low or questionable.

Measures of Central Tendency/Average

We collect data regarding many continuous variables (i.e. birth weight, Haemoglobin etc.) in our day-to-day work. If you spend some time looking at these data you could observe a pattern in their distribution. If you take the birth weight of babies born at term as an example you could see that most of them would be near 2900gm. The number of babies with birth weight lower or higher than 2900gm will decrease as we go further and further away from 2900gm.

We use this property of majority of values tending to fall near a particular value (which is called the central tendency) to describe or summarize a quantitative variable. This value is also called the average. When talking about the average the term that comes to our mind is the **mean** (technically the more accurate term for this is arithmetic mean). This is calculated by summing all the values and dividing the total by the number of observations.

The mean is a good measure of average because -

- it is what most people consider as *the* average
- it makes use of all the observations
- it is easy to calculate

The drawbacks of the mean are -

- could be unduly influenced by a few or a single extreme value/s
- may not correspond to an actual data value

The mean is not the only measure of average. The median and the mode are two other measures of average.

The **median** is the middle value when the data are arranged in ascending order. When the number of observations is an even number the median is the mean of the middle two values.

The median is a good measure of average because -

- it often corresponds to an exact data value
- it is not affected by extreme values
- it can be identified from the first 50% of survival data

The drawbacks of the median are -

- it only uses the central value, or values, in the data set
- it is not what most people think of when they speak of an *average*

The **mode** is the most frequently occurring value.

The mode is sometimes used as a measure of average as -

- it always corresponds to an exact data value
- it is not affected by extreme values
- it is simple to identify

Disadvantages of the mode as a measure of average-

- Not what most people think of when they speak of an *average*
- Not used in statistical analysis.

Is the average adequate to describe or summarize the data?

If we get back to the birth weight example we could see that there are many situations in which we could get an average birth weight of 2900gm.
When all babies weigh exactly 2900gm at birth
When the vast majority weigh 2900gm and a few weigh a little less or more than 2900gm
When some weigh 2900gm and the other weights vary widely from 2900gm

The possibilities of variation of individual birth weights with a mean birth weight of 2900gm are endless. Therefore if we want to describe the data adequately we should indicate the amount of variation or dispersion in the data in addition to indicating the average.

Measures of dispersion/variation

Range - The difference between the highest value and the lowest value is called the Range. This is easy to calculate but has the drawback of being influenced by extreme values and it tends to increase as the number of observations increase.

Quantiles - Quantiles are divisions of a distribution into ordered subgroups of equal size (Altman & Bland, 1994). **Centiles** are hundredths; **quartiles**, quarters; **quintiles**, fifths; **deciles**, tenths. When the distribution is divided into centiles the nth centile will have n% of observations below it.
Is it possible to calculate a measure of variation/ dispersion that indicates how much the observations vary from the average or mean?

Standard deviation - If we represent the individual observations by x and the mean by $\bar{x}$ (pronounced x bar) and the number of observations by n then the expression $(x - \bar{x})$ indicates how much each individual observation deviates from the mean. The expression $\sum(x - \bar{x})/n$ (where $\sum$ capital sigma is for summation) will give the mean deviation but unfortunately this will always be zero as the sum of positive deviations is equal to the some of negative deviations. This problem can be overcome by squaring the deviations and making all of them positive quantities. When these squared deviations are summed and divided by (n-1) we get the quantity called the **variance** (the reasons for the denominator being (n-1) rather than n are beyond the scope of this note). The variance is a useful measure of variation. The positive square root of variance is called the **standard deviation** and it is the most widely used measure of dispersion.

Data presentation

Data can be summarized and presented in the form of text, tables and graphs/charts. The same data need not/ should not be presented in more than one form.

Graphs/charts are very effective when used as visual aids during an oral presentation or as part of a poster presentation. But these can also be used in a report and tables can be used during an oral presentation or in a poster. In some instances the use of text (a single sentence or a short paragraph) may be the most efficient use of space (and ink).

Tables

Tables should be numbered.
The title of the table should be informative enough.
The other important parts of a table are the column and row headings.
Sometimes it is necessary to include footnotes.

A well-constructed table should be self-explanatory. There should be no need to refer to the accompanying text to make sense of a table

Rounding and Significant Figures

Avoid misleading precision in summary statistics (Altman & Bland, 1996).

In summary statistics use only one decimal place more than the original data. (How about decimals in percentages when the total sample size is less than 100?)

Numbers with more than three significant figures are rarely needed.

Normal Distribution

The distribution of many biological variables (e.g. height, weight) follow a pattern. An important feature of this pattern is that the majority of the observations are concentrated around the mean and the frequency of observations decrease as one gets further away from the mean in either direction. The frequency distributions of these continuous variables take the shape of a bell.

Features / Properties of the Normal distribution.

1. Data are concentrated around the mean (unimodal)

2. Values on either side of the mean occur in approximately equal frequency (symmetrical; Mean = Median = Mode)

3. Frequency of values falls as they get away from the mean.

The shape of the Normal distribution is rather like a bell. It is also called the Gaussian curve after the 19^{th} century German mathematician C. F. Gauss.

Most values (≈ 68%) are within one standard deviation on either side of the mean. ($\mu \pm \sigma$)

Large majority (≈ 95%) are within two standard deviations on either side of the mean. ($\mu \pm 2\sigma$)

Almost all (≈ 99.7%) are within three standard. deviations on either side of the mean. ($\mu \pm 3\sigma$)

Normal Range or 95% Reference Interval

We make use of this property of the Normal distribution to calculate 95% reference interval or normal range. This is the range of values within which 95% of values of healthy people will lie. When the variable follows the Normal distribution this is given by

Mean – 2 SD to Mean + 2 SD

When the distribution of the variable is not Normal the 95% reference range is given by the 2.5^{th} centile to 97.5^{th} centile.

Another technique used when the distribution of the variable is not Normal is to try and make it Normal by a transformation, calculate the 95% reference interval for the transformed data using the mean and SD of the transformed data and then express this reference range in the original units. When the distribution has a long right tail i.e. positive skew it may be made a Normal distribution by taking its logarithm.

Z score

When the difference between the mean and a particular value (or observation) is expressed in terms of the standard deviation it is called the Z score. Z scores make it possible to compare values obtained from different distributions i.e. different means and standard deviations.

Standard Normal Distribution

This is a hypothetical distribution with a mean of zero (0) and standard deviation of one (1). Any series of observations that form a Normal distribution can be converted to standard normal distribution. This is done by first calculating the mean and the standard deviation. The next step is to deduct the mean from each of the observations and dividing the results by the standard deviation.

We have earlier noted that when the data are from a Normal distribution, within 1, 2 or 3 standard deviations from the mean a certain (69, 95 & 99.7) percent of observations could be found. Statisticians have worked out the percentage of observations that can be found within any multiple of standard deviation (not just the numbers 1, 2 & 3 but any value from zero upwards). These are expressed as probability values for given Z value. These can be found in standard textbooks and statistical computer programs.

Sampling

In many instances when we need to determine a characteristic of a population (mean height of adult males in Sri Lanka) we don't make observations on the total population. We select a representative sample and collect information from the sample and estimate the population value - mean height (parameter).

If we take one large sample will the sample mean be exactly equal to the population mean?

If we take two large samples of equal size will the two sample means be exactly the same?

If we take many large samples and plot their mean values the sample means will form a Normal distribution with a mean equal to the population mean. The standard deviation of the sample means is called **standard error of mean**. We make use of this property of sample means to estimate the population mean from sample means.

The term **Population** (or **Universe**) refers to the whole collection of **Units** or **Elements** (Persons, Records, Institutions or Events) under investigation.

The term **Sample** refers to a selected subset of the **Population,** which is used to gather information about the population.

Sampling is the process of selecting a number of units from the population.

Benefits of sampling when compared to complete coverage are-

Reduction in time
Reduction in labour
Reduction in cost
Improved quality of data.

These advantages are of value only if the sample is representative of the population.

Representative samples are selected by various forms of **Random Sampling**. The essential feature of Random Sampling is that each unit has a known probability of being selected.

Sampling Techniques
Simple Random Sample – Requires a complete sampling frame (a listing of all units) and random numbers.

Systematic Sample – Random numbers are not required. Every n^{th} person from the sampling frame is selected.

Cluster sampling – The units are not selected in to the sample individually but as groups or clusters. This may be done when a complete sampling frame is not available but it is known that the units belong to some larger groupings.

Stratified sampling – When it is known that the characteristic under investigation is related to another variable the population under study could be stratified according to the second variable. The sample is selected separately from within the different strata.

Multistage sampling – As the name implies this procedure involves selecting the sample in more than one stage. Multistage sampling enables the investigator to concentrate/restrict the fieldwork to selected areas.

It is important to remember that these above terms are technical terms with exact meanings and there are other named sampling methods. In many instances sampling involves the use of combinations of these methods and some modifications as well. The objective of sampling is not to adopt one of these methods rigidly but to ensure that the sample obtained is representative of the population. So when reporting the sampling of a particular study it is important to describe in detail what was actually done.

Estimation

Let us assume that we are interested in estimating the mean height of a population. We could do this by taking a large random sample from this population and measuring the heights of the selected individuals and calculating the mean height of the sample. The sample mean is a point estimate of the population mean. We have to be extremely lucky to obtain a sample that gives a mean value exactly equal to the population mean.

In this situation we can be reasonably certain that the value of the population mean will be close to the sample mean. We would like to refine this and give a range for the population figure. The techniques of estimation based on the Normal distribution allow us to do this.

The calculated range is called the confidence interval. This also allows us to give a numerical value to the term reasonably certain.

95% Confidence Interval: When we use this technique to calculate the interval that is likely to contain the population mean (or another parameter) our prediction will be correct on 95 out of 100 occasions.

We make use of the standard error of the mean to do these calculations. The standard error of the mean is derived by dividing the population standard deviation by the square root of the sample size. But in real life we don't know the population standard deviation we have only got the sample standard deviation. When the sample size is large (preferably more than 60 but at least more than 30) the sample standard deviation is a very good approximation of the population standard deviation. Therefore we could make use of the sample standard deviation to calculate the standard error.

SE – Standard Error of Mean, SD – Standard Deviation, n – Sample size. Read this paper to learn more about SD & SE (Streiner, 1996).

Standard errors and 95% confidence intervals can be calculated for not only the mean but also for proportions (or percentages), Relative Risks, Odds Ratios, Regression Coefficients etc. Further standard errors and 95% confidence intervals can be calculated for difference between two means and difference between two percentages.

$$SE = \frac{\sqrt{p \times q}}{n}$$

SE – Standard Error of Percentage, p – Percentage of interest, q = 100 – p, n – Sample size

Standard error of difference

Let's assume that the mean height of Sri Lankan women is 160 cm. If we take two samples of women it is very unlikely that the two sample means would be identical, either 160 cm or some other value. Intuitively we could say that the larger the size of the samples the closer their means would be to each other as well as the population mean.

If we repeat this process of taking two samples at a time, on some occasions the mean of the first sample will be greater than the second sample and on other occasions the mean of the second sample will be greater. Differences closer to zero are more likely than differences away from zero. In fact if we plot the differences between two means they will form a Normal distribution with a mean of zero and a standard deviation known as the standard error of difference between means.

In real life we may be faced with a situation where we would have taken samples from two groups and find that there is a difference between the means of the two samples (Say we measured the height of two samples of women from urban and rural areas and found the mean and standard deviation to be 162 & 8 in a sample of 100 urban women and 158 & 7 in a sample of 100 rural women). What is the likely difference in the mean height of all women from urban areas and rural areas? We can be reasonably certain that it will be close to 4cm (162 – 158). It is known that the difference between two sample means form

a Normal distribution. The standard deviation of this distribution is called the standard error of difference.

This may have happened because the two groups have actually different mean values or it may be that the means of the two groups are the same but we found the difference due to the fact that we have taken samples – chance occurrence.

We make use of the fact that difference between two sample means form a Normal distribution to resolve this. We'll assume that there is no difference in the mean height of women from urban and rural areas – null hypothesis. We'll try and see how much support is there for this assumption in the data provided. That is how likely is it to see such difference in sample means when there is no difference in the population means of the two groups. We have just learnt that when the null hypothesis is true the difference between two means form a Normal distribution with a mean of zero and standard deviation called standard error of difference between two means. So in the above example we just have to look at the difference between the two means and see how far away is this from zero in terms of the standard error. If this is more than 1.96 we know that the probability of this happening is less than 0.05.

By convention when the probability is less than 0.05 ($P<0.05$) we say that it is unlikely that the two samples came from the same population. In other words we reject the null hypothesis. One could argue that there is no real difference between a probability value of 0.049 and 0.051 even though with 0.049 we reject the null hypothesis and with 0.051 we fail to reject the null hypothesis. This is the reason why it is better to report the exact P value rather than state either $P < 0.05$ or $P > 0.05$.

An alternative or supplementary way of reporting results obtained from samples is to use confidence intervals. Just as we calculate the 95% confidence interval for a mean or a percentage we could calculate 95% confidence interval for difference between two means or difference between two percentages.

Tests of Significance

Null Hypothesis - The statistical hypothesis that one variable has no association with another variable or set of variables, or that two or more population distributions do not differ from one another. In simple terms, the null hypothesis states that the results observed in a study, experiment, or test are no different from what might have occurred as a result of the operation of chance alone.

Type I Error - The error of rejecting a true null hypothesis i.e., declaring that a difference exists when it does not.

Type II Error - The error of failing to reject a false null hypothesis i.e., declaring that a difference does not exist when in fact it does.

Truth in the population versus the results in the sample: the four possibilities

		Truth in the population	
		The two variables are associated	No association between the two variables
Results in the sample	Reject null hypothesis	CORRECT	TYPE I ERROR
	Fail to reject null hypothesis	TYPE II ERROR	CORRECT

Types of data	Significance test
Analyzing single mean	
Large sample	Normal test (Z test)
Small sample (less than 30)	t test
Comparing two means	
Large samples	Normal test (Z test)
Small samples (less than 30)	
Independent samples	t test
Paired samples	Paired t test
Comparing two percentages	Normal test (Z test) or Chi square test
Comparing three or more percentages or Association between two qualitative variables	Chi square test
Association between two quantitative variables	Correlation

Steps in performing a significance test

State the null hypothesis
Calculate the test statistic
Refer the test statistic to a known distribution
Find the probability of a value of the test statistic arising which is as extreme or more extreme than that observed
Conclude that the data are consistent or inconsistent with the null hypothesis

Graphs/ Charts

Frequency distribution of numeric variable can be presented as a histogram, stem and leaf plot or box and whisker plot.

Histogram
Y axis – Frequency (either absolute or relative). The height of bars/lines proportional to frequency
X axis – Classes of a continuous variable (or a discrete variable with many discrete values) generally with equal class intervals.
By inspecting a histogram one could say whether the distribution is uni-modal or bi-modal. If it is uni-modal, whether it's symmetric or not. If it is asymmetric, whether it has a positive or negative skew.

Stem and leaf plot
This also shows the shape of the distribution; in addition to that it also displays the original data.

Box and whisker plot or Box plot
Constructed using summary statistics of a numeric variable. Box and whisker plots are useful in visually comparing the distribution of a numeric variable in more than one group.

Pie chart
Frequency distribution of a categorical variable can be displayed as a pie chart. The angle and the surface area of each segment are proportional to the relative frequency of that particular category. It is not easy to judge the angle or surface area of a segment by looking at the 'pie' obliquely. This is why three-dimensional pie charts are not used in academic publications.

Bar chart
Bar charts are used to show the relationship between two or more categorical variables.

Line graph
Line graphs are useful in demonstrating change or trend over time.

Scatter plots
Scatter plots are used to show the relationship between two numeric variables

Minimum information required to calculate sample size

Study to determine a population proportion / prevalence of a disease or risk factor
Required information: Estimate of the proportion – if not available assume it as 0.5 and
Desired precision or confidence interval – usually 95%

Study to determine population mean
Required information: Estimate of the standard deviation
Desired precision or confidence interval – usually 95%

Study to determine the association between an exposure and outcome
Required information:
Estimate of prevalence of exposure or outcome in the control group
Estimate of effect size or estimate of prevalence of exposure or outcome in the study group
Level of significance – usually 0.05
Power – usually 90% or 80%

For a more detailed account of sample size calculation please refer to one of these papers - (Campbell, Julious, & Altman, 1995), (Florey, 1993) or (Streiner, 1990).

Further reading

Books

Kirkwood, B. B., & Sterne, J. (2003). *Essential medical statistics*. *Malden, MA: Blackwell Science* (pp. 1–512). doi:10.1002/sim.1961

Papers

Altman, D. G., & Bland, J. M. (1994). Quartiles, quintiles, centiles, and other quantiles. *BMJ (Clinical Research Ed.)*, *309*(6960), 996.

Altman, D. G., & Bland, J. M. (1996). Presentation of numerical data. *BMJ (Clinical Research Ed.)*, *312*(7030), 572.

Campbell, M. J., Julious, S. A., & Altman, D. G. (1995). Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. *BMJ (Clinical Research Ed.)*, *311*(7013), 1145–8.

Florey, C. V. (1993). Sample size for beginners. *BMJ (Clinical Research Ed.)*, *306*(2), 1181–1184. doi:10.1016/B978-0-240-51910-4.50004-0

Streiner, D. L. (1990). Sample size and power in psychiatric research. *Canadian Journal of Psychiatry. Revue Canadienne de Psychiatrie*, *35*(7), 616–20.

Streiner, D. L. (1996). Maintaining standards: differences between the standard deviation and standard error, and when to use each. *Canadian Journal of Psychiatry. Revue Canadienne de Psychiatrie*, *41*(8), 498–502.