INTRODUCTORY MATHEMATICAL STATISTICS

(STAT2001/6039)

LECTURE NOTES

OVERVIEW (or PREVIEW)

Introduction (Chapter 1 in textbook)

What is statistics? Basic concepts. Summarising (or characterising) a set of numbers, graphically and numerically.

Probability (Ch 2)

Eg1: A coin is tossed twice. What's the probability (pr) that exactly one head comes up? 1/2 (2/4 = 1/2, from HH, HT, TH, TT)

Eg2: A committee of 5 is to be randomly selected from 12 people. What's the pr it will contain the 2 oldest people? 5/33 (C(10,3)/C(12,5) = 5/33)

Eg3: (The coin tossing problem)

John and Kate take turns tossing a coin, starting with John.

The first to get a head wins. What's the pr that John wins? 2/3

(1/2 + 1/8 + 1/32 + ... = 2/3)

Discrete random variables (Ch 3)

A formalisation of concepts in Ch 2.

Eg: A coin is tossed twice. Let Y be the number of heads that come up. Then Y is a discrete random variable (rv), one with possible values 0, 1, 2. Y has probability distribution (pr dsn) given by $P(Y = 0) = 1/4$, $P(Y = 1) = 1/2$, $P(Y = 2) = 1/4$.

Continuous random variables (Ch 4)

Eg: A stick is measured as 1.2 m long. Let Y be the *exact* length of the stick. Then Y is a continuous random variable (cts rv), one with possible values between 1.15 and 1.25 (a 'continuum' of values). How can we describe Y's pr. dsn?

Multivariate probability distributions (Ch 5)

Eg: A die is rolled twice. Let X be the no. of 5's that come up and Y the no. of 6's. X and Y are 2 rv's with a multivariate (or bivariate or joint) pr dsn. (Possible values for (X,Y) are (0,0), (0,1), (0,2), (1,0), (1,1), (2,0); eg, (2,1) is not possible.)

Functions of random variables (Ch 6)

Eg: A coin is tossed twice. Let Y = no. of H's. Then $X = 3Y$ is a function of Y. What's the pr dsn of X? (Possible values of X are 0, 3, 6.)

Sampling distributions and the Central Limit Theorem (CLT) (Ch 7)

Eg: A coin is tossed 100 times. What's the pr that at least 60 heads come up? This pr is hard to work out exactly, but can be well approximated using the CLT and the normal distribution (Ch 4).

$$Y \sim \text{Bin}(100,0.5) \overset{\cdot}{\sim} N(50,25) \Rightarrow P(Y \geq 60) \approx P(Z > (60 - 0.5 - 50)/\sqrt{25})$$

$$= P(Z > 1.9) = 0.0287. \text{ The exact probability is } 0.0284.$$

Point estimation and confidence intervals (CI's) (Ch 8)

Eg: p = pr of H on a single toss of a bent coin. We toss the coin 100 times and get 60 heads. So we estimate p by 60/100 = 0.6 (a point estimate).

A 95% CI for p is (0.504, 0.696). So we are 95% 'confident' that p lies between 0.504 and 0.696. What exactly this means will be made clear later.

$$(.6 \pm 1.96\sqrt{.6(.4)/100}) = (.6 \pm .096) = (.504, .696)$$

Methods of estimation (Ch 9)

It was reasonable to estimate p (above) by .6. In some situations it is not so clear how to proceed. We'll look at 2 general methods for estimating a quantity:

1. the method of moments (MOM)
2. the method of maximum likelihood (ML).

(In fact, .6 above is both the MOME and MLE (E = estimate) of p.)

Hypothesis testing (Ch 10)

Eg: We toss a coin 100 times and get 60 heads. Can we conclude that the coin is unfair? (Ie, should we *accept* or *reject* the statement "$p = 1/2$"?)

INTRODUCTION (Chapter 1)

Statistics = descriptive statistics + inferential statistics (or "inference").

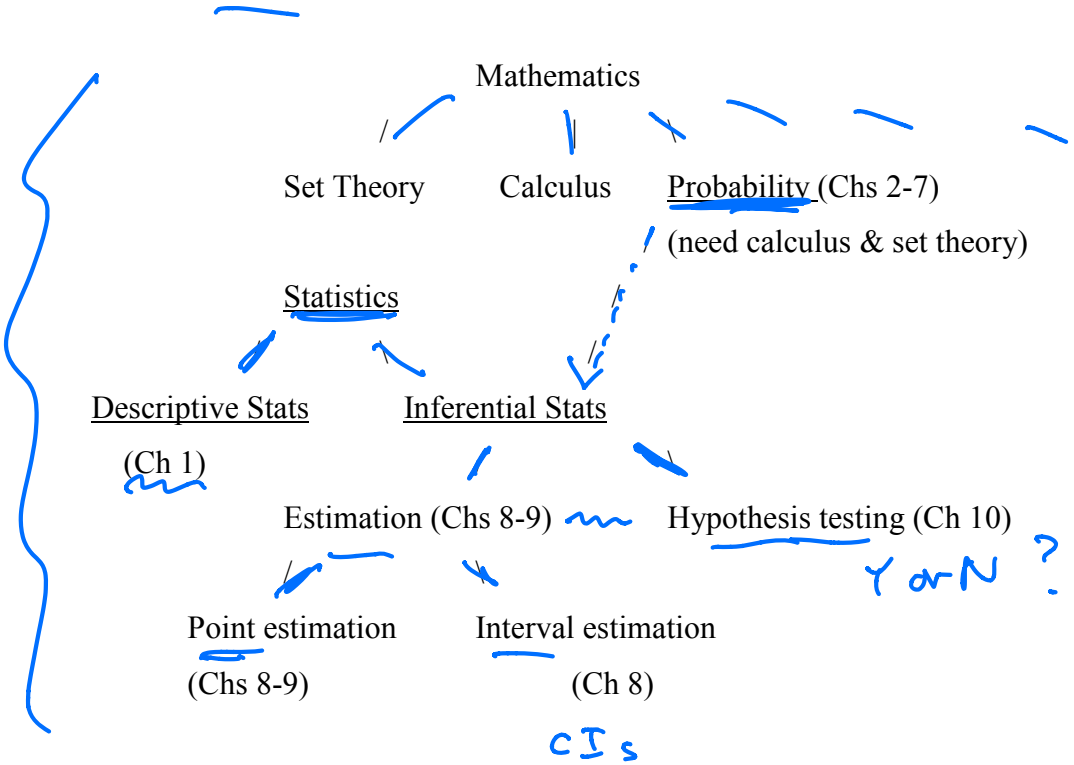
Descriptive statistics involves tables, graphs, averages and other simple summaries of data (not much maths is required). This is the focus of Chapter 1, covered here.

Inferential statistics has to do with using information from a *sample* to make statements about a *population*.

Eg: Sample 100 people from Canberra. On the basis of the sample, we estimate the average height of all Canberrans as 1.74 m, with 95% CI (1.72, 1.76).

To make inferential statements like this we need a good understanding of *probability*, which is a branch of *mathematics*. The term *mathematical statistics* refers to the combined fields of probability and inferential statistics (not *descriptive statistics*).

Chapters 2 through 7 deal with probability. Chapters 8 to 10 deal with inferential statistics. Field map:



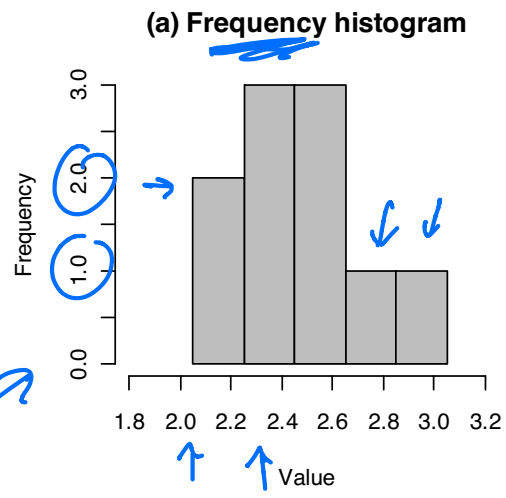
Descriptive Statistics

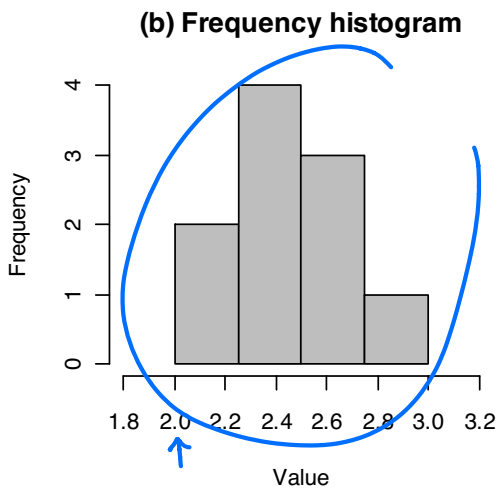
Consider a small dataset of $n = 10$ values (same as given on page 4 in the text):

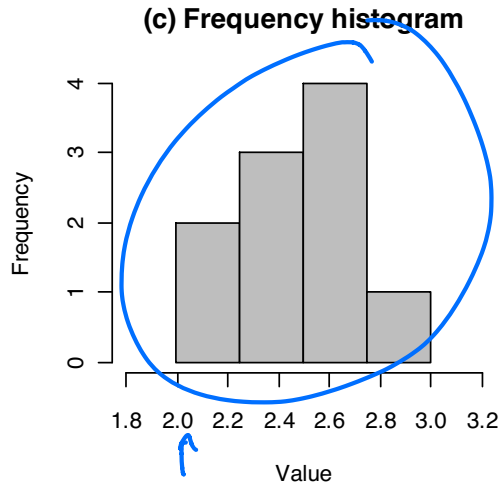
i	1	2	3	4	5	6	7	8	9	10
y_i	2.1	2.4	2.2	2.3	2.7	2.5	2.4	2.6	2.6	2.9

Here, for example, y_i represents the profit of a particular business on Day i (in units of \$1000). How can we *graphically summarise* these data?

One way is to create a number of *bins* that span the *range* of the data (2.1 to 2.9), and draw vertical *bars* which show the number of observations (frequencies) in each bin. This results in a *frequency histogram*, as shown in Figure (a) below. Here there are 5 bins, defined by breaks 2.05, 2.25, 2.45, 2.65, 2.85, 3.05. (Eg, Bin 1 has 2 values, namely 2.1, and 2.2, because these are the ones that lie between 2.05 and 2.25.)







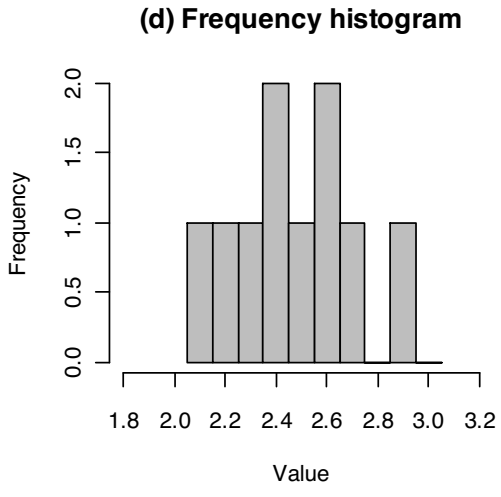


Figure (a) is a *frequency histogram* and is similar to Figure 1.1 in the text (page 4), which shows a *relative frequency histogram*. The difference is that Fig. 1.1 has 0 to 0.3 on the vertical axis, rather 0 to 3. (Eg, the relative frequency for Bin 1 is the proportion or fraction of values that lie between 2.05 and 2.25, namely 2/10 = 0.2.)

Note that none of the data values lies on a bin boundary. This is a nice feature because it allows for no ambiguity regarding how many values are in any particular range. Suppose we define the bins according to boundaries 2, 2.25, 5, etc. Then there are two ways this may be interpreted. These ways are as shown in Figure (b) above, where the bins are (2,2.5], (2.5,5],..., and in Figure (c), where the bins are [2,2.5), [2.5,5),....

(For example, there are three values in the interval (5,7.5], namely 2.7, 2.6, 2.6, whereas there are four values in the interval [5,7.5), namely 2.7, 2.5, 2.6, 2.6.)

Figure (d) shows yet another histogram of the same data, one where the bins are defined by boundaries 2.05, 2.15, 2.25,.... Often a good choice for the number of bins is 5 to 20, with a high number being appropriate when there are many data values.

R Code (non-assessable)

```
y = c(2.1, 2.4, 2.2, 2.3, 2.7, 2.5, 2.4, 2.6, 2.6, 2.9)
n = length(y); n # 10
help(hist)
par(mfrow=c(2,2))
hist(y,  breaks=seq(2.05,3.05,0.2),xlim=c(1.8,3.2), col='grey',probability=FALSE,
     main="(a) Frequency histogram", xlab="Value")
hist(y,  breaks=seq(2,3,0.25),xlim=c(1.8,3.2), col='grey',probability=FALSE,
     main="(b) Frequency histogram", xlab="Value")
hist(y,  breaks=seq(1.999,2.999,0.25),xlim=c(1.8,3.2), col='grey',probability=FALSE,
     main="(c) Frequency histogram", xlab="Value")
hist(y,  breaks=seq(2.05,3.05,0.1),xlim=c(1.8,3.2), col='grey',probability=FALSE,
     main="(d) Frequency histogram", xlab="Value")
```

Let's now consider how we might *numerically summarise* the small dataset.

Two basic types of descriptive measures are *measures of central tendency* and *measures of dispersion* (or *variation*).

The most common measure of central tendency is the *arithmetic mean* (or *average*), defined by

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}(y_1 + \ldots + y_n),$$

and in our example this works out as

$$\bar{y} = \frac{1}{10}(2.1 + 2.4 + \ldots + 2.9) = 2.47 \quad (\$2470).$$

If y_i represents the profit of a particular business on Day i (in units of \$1000), then $\bar{y}$ is an estimate of the average profit of the business *per day over a long period of time*, which may also be called the *population mean*, μ. (This is an example of statistical inference and will be discussed in greater detail in later chapters.)

A common measure of dispersion is the *sample variance*, defined by

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

which in our example works out as

$$s^2 = \frac{1}{10-1}\left\{(2.1-2.47)^2 + \ldots + (2.9-2.47)^2\right\} = 0.057889.$$

A problem with s^2 is that it is in *squared units*. Here, one unit is 1000 dollars, and so a squared unit is 1000000 square dollars. That is, $s^2 = 57{,}889\ \$^2$ (a bit awkward).

Another measure of dispersion is the *sample standard deviation*, s, which in our example is $\sqrt{0.057889} = 0.24060$. This is in the same units as $\bar{y}$, and so $s = \$240.60$.

The sample variance s^2 and sample standard deviation s are estimates of the *population variance* σ^2 and *population standard deviation* σ, respectively. Here, σ^2 may be thought of as the sample variance of a 'very large' number of y_i values. (More will be said about this later, and also about why s^2 involves "$n-1$".)

A useful principle which involves μ and σ is the **Empirical Rule**:

> About 68% of the values lie in the interval $\mu \pm \sigma$
> About 95% of the values lie in the interval $\mu \pm 2\sigma$
> Almost all (>99%) of the values lie in the interval $\mu \pm 3\sigma$

This Rule is most accurate when the n is *large* and histograms of the data are *bell-shaped*, but still has some validity otherwise. It also applies when μ and/or σ are changed to $\bar{y}$ and s. For example, consider our dataset with only $n = 10$ values. Here:

$\bar{y} \pm s = 2.47 \pm 0.24 = (2.23, 2.71)$, which contains 70% of the values ($\approx 68\%$)

$\bar{y} \pm 2s = 2.47 \pm 0.48 = (1.99, 2.95)$, which contains 100% of the values ($\approx 95\%$)

$\bar{y} \pm 3s = 2.47 \pm 0.72 = (1.75, 3.19)$, which contains 100% of the values ($> 99\%$)

R Code (non-assessable)

```
y = c(2.1, 2.4, 2.2, 2.3, 2.7, 2.5, 2.4, 2.6, 2.6, 2.9)
ybar = mean(y); ybar # 2.47
s2 = var(y); s2 # 0.05788889
s = sqrt(s2); s # 0.2406011

ybar + c(-1,1)*s # 2.229399 2.710601
ybar + c(-1,1)*2*s # 1.988798 2.951202
ybar + c(-1,1)*3*s # 1.748197 3.191803
```

Another empirical rule:

About 50% of the values lie above μ.

Also true with μ changed to $\bar{y}$.

NB: This can be very wrong for a highly skewed population or sample (of values)

Eg Suppose the IQs of 100 people in a room are $99, \underbrace{100, 100, \ldots, 100}_{99}$

Then the average IQ is $\frac{1}{100}(99 + 99 \times 100) = 99.99$

What proportion of these people have above-average IQ (intelligence)?

Answer: $\frac{99}{100} = 99\%$. $\begin{cases} \not\approx 50\% \\ \gg 50\% \\ \approx 100\% \end{cases}$