Lecture-17
Algorithmic Mathematics(CSC545)

Prepared by Asst. Prof. Bal Krishna Subedi

CDCSIT, TU

1 Measures of Central Tendency

1.1 The Mean, Median and Mode

When given a set of raw data one of the most useful ways of summarising that data is to find an average of that set of data. An average is a measure of the centre of the data set. There are three common ways of describing the centre of a set of numbers. They are the mean, the median and the mode and are calculated as follows.

The mean	–	add up all the numbers and divide by how many numbers there are.
The median	–	is the middle number. It is found by putting the numbers in order and taking the actual middle number if there is one, or the average of the two middle numbers if not.
The mode	–	is the most commonly occurring number.

Let's illustrate these by calculating the mean, median and mode for the following data.

Weight of luggage presented by airline passengers at the check-in (measured to the nearest kg).

$$18 \quad 23 \quad 20 \quad 21 \quad 24 \quad 23 \quad 20 \quad 20 \quad 15 \quad 19 \quad 24$$

$$\text{Mean} = \frac{18+23+20+21+24+23+20+20+15+19+24}{11} = 20.64.$$

Median = 20.

$$15 \quad 18 \quad 19 \quad 20 \quad 20 \quad \underset{\substack{\uparrow \\ \text{middle value}}}{20} \quad 21 \quad 23 \quad 23 \quad 24 \quad 24$$

Mode = 20. The number 20 occurs here 3 times.

Here the mean, median and mode are all appropriate measures of central tendency.

For example, suppose a class of 20 students own among them a total of 17 pets as shown in the following table. Which measure of central tendency should we use here?

Type of Pet	Number
Cat	5
Dog	4
Goldfish	3
Rabbit	1
Bird	4

If our focus of interest were on the *type* of pet owned, we would use the mode as our average. 'Cat' would be described as the 'modal category', as this is the category that occurs most often.

If, on the other hand, we were not interested in the type of pet kept but the average *number* of pets owned then the mean would be an appropriate measure of central tendency. Here the mean is $\frac{17}{20} = 0.85$.

Also, if we are interested in the average number of pets per student then our data might be presented quite differently as in the table below.

Number of Pets	Tally	Frequency
0	~~||||~~ ~~||||~~ |	11
1	||||	4
2	|||	3
3	|	1
4	|	1

Now we are concerned only with a quantity variable and the average used most commonly with quantity variables is the mean. Here, again, the mean is 0.85.

$$\text{Mean} = \frac{(11 \times 0) + (4 \times 1) + (3 \times 2) + (1 \times 3) + (1 \times 4)}{20} = 0.85.$$

Note that (4×1) is really $1 + 1 + 1 + 1$, since 4 students have 1 pet each, and (3×2) is really $2 + 2 + 2$, since 3 students have 2 pets each. Since there are 20 scores the median score will occur between the tenth and the eleventh score. The median is 0, since the tenth and the eleventh scores are both 0, and the mode is 0.

Let's look again at our pets example and suppose that one of the students kept 18 goldfish.

Number of Pets	Tally	Frequency
0	$\cancel{||||}$ $\cancel{||||}$ |	11
1	||||	4
2	||	2
3	|	1
4	|	1
18	|	1

The mean is now 1.8, but the median and the mode are still 0. The effect of the outlier was to significantly increase the mean and now the median is a more accurate measure of the centre of the distribution.

With the exception of cases where there are obvious extreme values, the mean is the value usually used to indicate the centre of a distribution. We can also think of the mean as the balance point of a distribution.

2 Measures of Dispersion

The mean is the value usually used to indicate the centre of a distribution. If we are dealing with quantity variables our description of the data will not be complete without a measure of the extent to which the observed values are spread out from the average.

We will consider several measures of dispersion and discuss the merits and pitfalls of each.

2.1 The Range

One very simple measure of dispersion is the range. Lets consider the two distributions given in Figures 3 and 4. They represent the marks of a group of thirty students on two tests.

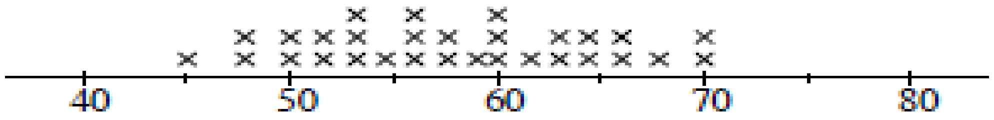

Figure 3: Marks on test A.

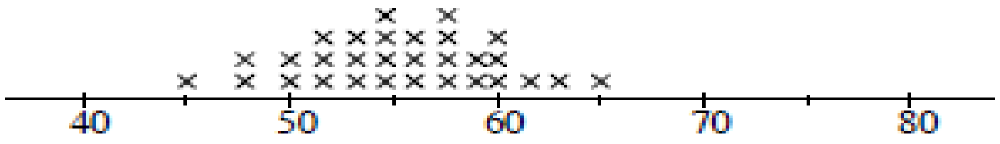

Figure 4: Marks on test B.

Here it is clear that the marks on test A are more spread out than the marks on test B, and we need a measure of dispersion that will accurately indicate this.

On test A, the range of marks is $70 - 45 = 25$.

On test B, the range of marks is $65 - 45 = 20$.

Here the range gives us an accurate picture of the dispersion of the two distributions.

However, as a measure of dispersion the range is severely limited. Since it depends only on two observations, the lowest and the highest, we will get a misleading idea of dispersion if these values are outliers. This is illustrated very well if the students' marks are distributed as in Figures 5 and 6.

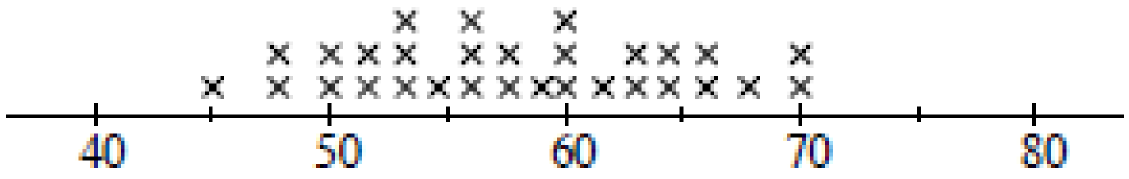

Figure 5: Marks on test A.

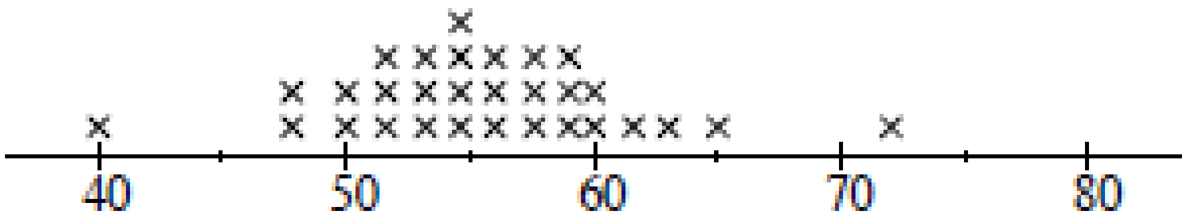

Figure 6: Marks on test B.

On test A, the range is still $70 - 45 = 25$.

On test B, the range is now $72 - 40 = 32$, but apart from the outliers, the distribution of marks on test B is clearly less spread out than that of A.

We want a measure of dispersion that will accurately give a measure of the variability of the observations. We will concentrate now on the measure of dispersion most commonly used, the standard deviation.

2.2 Standard Deviation

Suppose we have a set of data where there is no variability in the observed values. Each observation would have the same value, say 3, 3, 3, 3 and the mean would be that same value, 3. Each observation would not be different or *deviate* from the mean.

Now suppose we have a set of observations where there is variability. The observed values would deviate from the mean by varying amounts.

The standard deviation is a kind of average of these deviations from the mean.

This is best explained by considering the following example.

Take, for example, the following grades of 6 students:

$$56 \quad 48 \quad 63 \quad 60 \quad 51 \quad 52.$$

Mean $= 55$.

To find how much our observed values deviate from the mean, we subtract the mean from each.

Observed values	56	48	63	60	51	52
Deviations from Mean	$+1$	-7	$+8$	$+5$	-4	-3

We cannot, at this stage, simply take the average of the deviations as their sum is zero.

$$(+1) + (-7) + (+8) + (+5) + (-4) + (-3) = 0$$

We get around this difficulty by taking the square of the deviations. This gets rid of the minus signs. (Remember $(-7) \times (-7) = 49$.)

Deviations	$+1$	-7	$+8$	$+5$	-4	-3
Squared deviations	1	49	64	25	16	9

We can now take the mean of these squared deviations. This is called the variance.

$$\text{Variance} = \frac{1 + 49 + 64 + 25 + 16 + 9}{6} = 27.33.$$

The variance is a very useful measure of dispersion for statistical inference, but for our purposes it has a major disadvantage. Because we squared the deviations, we now have a quantity in square units. So to get the measure of dispersion back into the same units as the observed values, we define standard deviation as the square root of the variance.

$$\text{Standard Deviation} = \sqrt{\text{Variance}} = \sqrt{27.33} = 5.228.$$

The standard deviation may be thought of as the 'give or take' number. That is, on average, the student's grade will be 55, give or take 5 marks. The standard deviation is a very good measure of dispersion and is the one to use when the mean is used as the measure of central tendency.

Example: Calculate the mean and standard deviation of the following set of data.

Birthweight of ten babies (in kilograms)

2.977 3.155 3.920 3.412 4.236 2.593 3.270 3.813 4.042 3.3

Solution:

Birthweight in kilograms	Deviations from Mean score − mean	Squared Deviations $(\text{score} - \text{mean})^2$
2.977	−0.5035	0.2535
3.155	−0.3255	0.1060
3.920	0.4395	0.1932
3.412	−0.0685	0.0047
4.236	0.7555	0.5708
2.593	−0.8875	0.7877
3.270	−0.2105	0.0443
3.813	0.3325	0.1106
4.042	0.5615	0.3153
3.387	−0.0935	0.0087
Sum = 34.805	Sum = 0	Sum = 2.3948

$$\text{Mean} = \frac{\text{sum of observations}}{\text{number of observations}} = \frac{34.805}{10} = 3.4805 = \mu.$$

$$\text{Variance} = \frac{\text{sum of squared deviations}}{\text{number of observations}} = \frac{2.3948}{10} = 0.2395 = \sigma^2.$$

$$\text{Standard Deviation} = \sqrt{\text{Variance}} = \sqrt{0.2395} = 0.4894 = \sigma.$$

2.3 The Interquartile Range

The interquartile range is another useful measure of dispersion or spread. It is used when the median is used as the measure of central tendency. It gives the range in which the middle 50% of the distribution lies. In order to describe this in detail, we first need to discuss what we mean by quartiles.

2.3.1 Quartiles

Suppose we start with a large set of data, say the heights of all adult males in Sydney. We can represent these data in a graph, which if smoothed out a bit, may look like Figure 9.

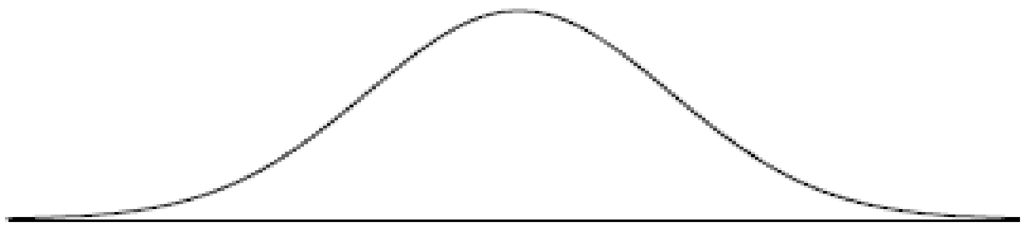

Figure 9: Graph representing heights of adult males.

As the name 'quartile' suggests, we want to divide the data into four equal parts. In the above example, we want to divide the area under our curve into four equal areas.

The second quartile or median

It is easy to see how to divide the area in Figure 9 into two equal parts, since the graph is symmetric. The point which gives us 50% of the area to the left of it and 50% to the right of it is called the second quartile or median. This is illustrated in Figure 10.

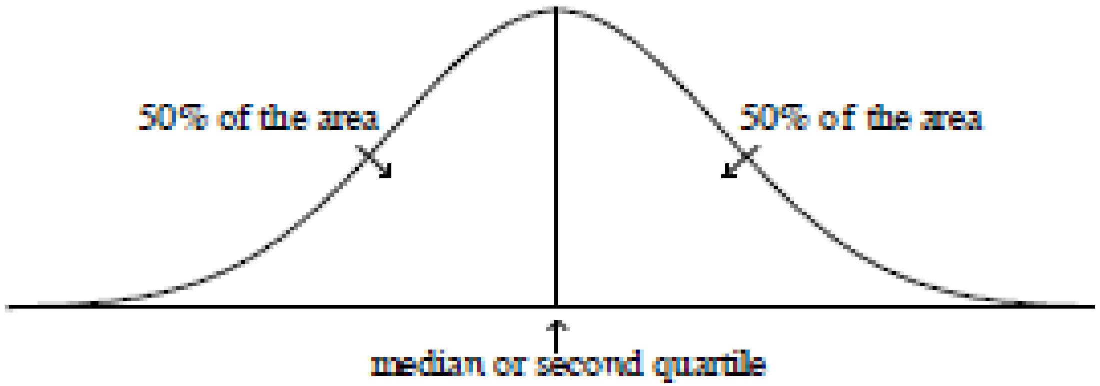

Figure 10: Graph showing the median or second quartile.

This exactly corresponds to our previous idea of median as the middle value.

The first quartile

The first quartile is the point which gives us 25% of the area to the left of it and 75% to the right of it. This means that 25% of the observations are less than or equal to the first quartile and 75% of the observations greater than or equal to the first quartile. The first quartile is also called the 25th percentile. This is illustrated in Figure 11.

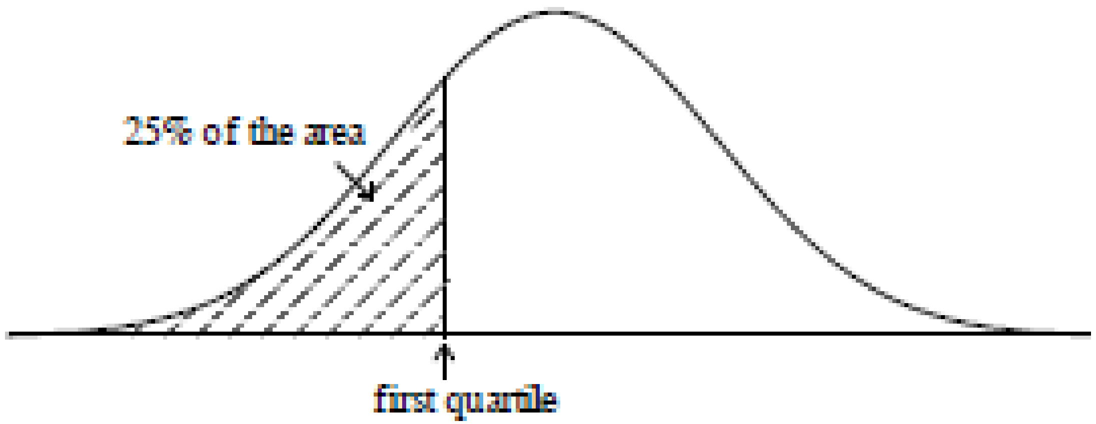

Figure 11: Graph showing the first quartile.

The third quartile

The third quartile is the point which gives us 75% of the area to the left of it and 25% of the area to the right of it. This means that 75% of the observations are less than or equal to the third quartile and 25% of the observation are greater than or equal to the third quartile. The third quartile is also called the 75th percentile. This is illustrated in Figure 12.

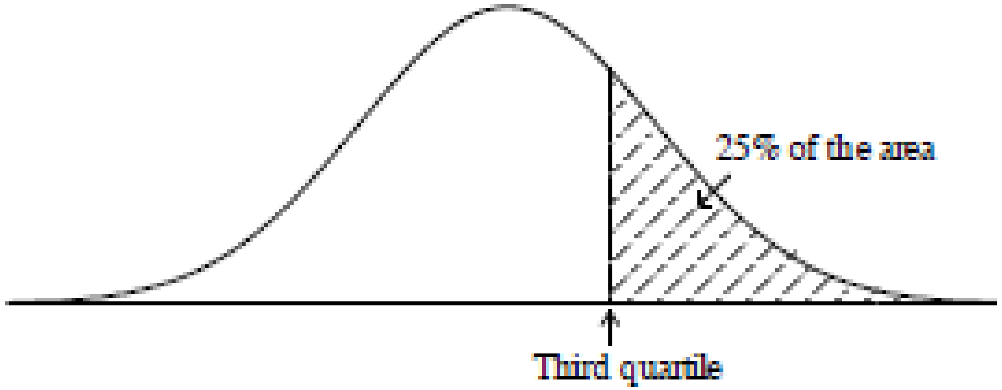

Figure 12: Graph showing the third quartile.

Summary

The first (Q_1), second (Q_2) and third (Q_3) quartiles divide the distribution into four equal parts. This is illustrated in Figure 13.

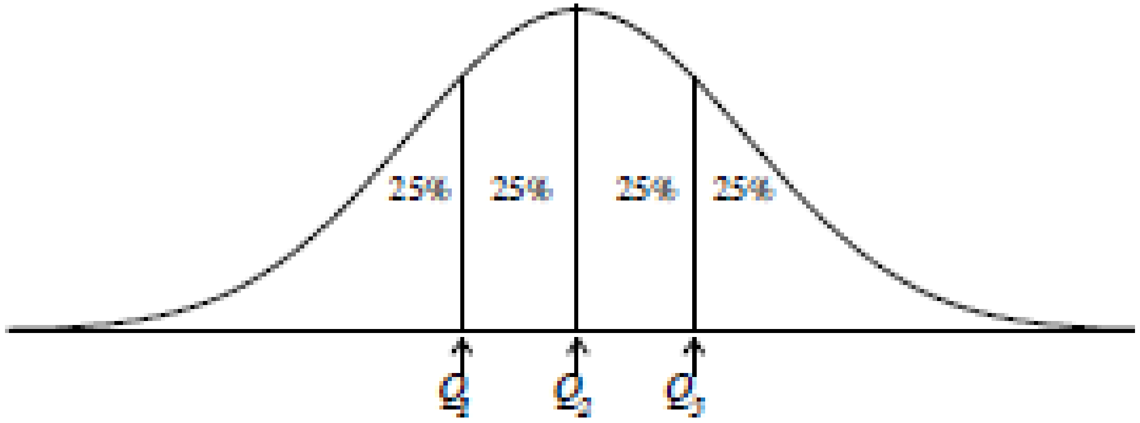

Figure 13: Graph showing all quartiles.

2.3.2 Quartiles for small data sets

Suppose now we have a small data set of twelve observations which we can write in ascending order as follows. (A data set, where the number of observations is a multiple of four, has been chosen to avoid some technical difficulties.)

$$15 \quad 18 \quad 19 \quad 20 \quad 20 \quad 20 \quad 21 \quad 23 \quad 23 \quad 24 \quad 24 \quad 25$$

In this case, we want to divide the data into four equal sets, so that there are 25% of the observations in each.

First, we find the median just as we did in Chapter 1.

$$15 \quad 18 \quad 19 \quad 20 \quad 20 \quad 20 \underset{\text{median}}{\uparrow} 21 \quad 23 \quad 23 \quad 24 \quad 24 \quad 25$$

The **median** is 20.5 (half way between the 6th and 7th observations), and divides the data into two equal sets with exactly 50% of the observations in each: the 1st to the 6th observations in the first set and the 7th to 12th observations in the other.

To find the first quartile we consider the observations less than the median.

$$15 \quad 18 \quad 19 \quad 20 \quad 20 \quad 20$$

The **first quartile** is the median of these data. In this case, the first quartile is half way between the 3rd and 4th observations and is equal to 19.5.

Now, we consider the observations which are greater than the median.

$$21 \quad 23 \quad 23 \quad 24 \quad 24 \quad 25$$

The **third quartile** is the median of these data and is equal to 23.5.

So, for our small data set of 12 observations, the quartiles divide the set into four subsets.

$$15 \quad 18 \quad 19 \underset{Q_1}{\uparrow} 20 \quad 20 \quad 20 \underset{Q_2}{\uparrow} 21 \quad 23 \quad 23 \underset{Q_3}{\uparrow} 24 \quad 24 \quad 25$$

2.3.3 The interquartile range

The interquartile range quantifies the difference between the third and first quartiles. If we were to remove the median (Q_2) from Figure 13 we would have a graph like that in Figure 14.

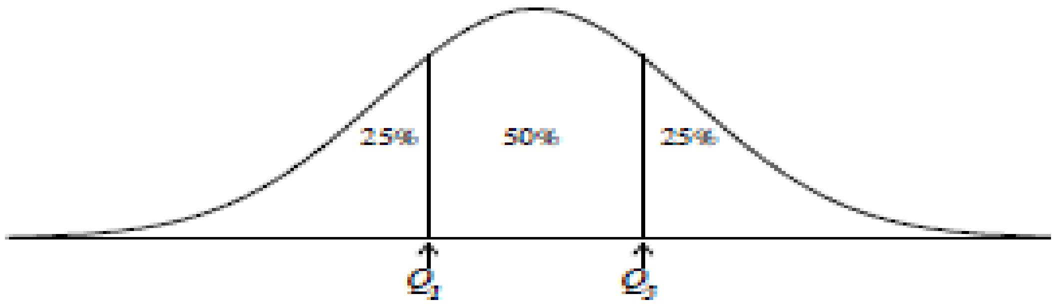

Figure 14: Graph showing the first and third quartiles.

From Figure 14, we see that 50% of the area is between the first and third quartiles. This means that 50% of the observations lie between the first and third quartiles.

We define the interquartile range as:

$$\text{The interquartile range} = \text{Third quartile} - \text{First quartile}.$$

For our small data set, the first quartile was 19.5 and our third quartile was 23.5. So, the interquartile range is $23.5 - 19.5 = 4$.

3 Formulae for the Mean and Standard Deviation

So far we have avoided giving the formulae for mean or standard deviation but no discussion would be complete without them. If you are not familiar with sigma notation, do not attempt this section. An explanation of sigma notation can be found in the Mathematics Learning Centre booklet: *Introductory Algebra for Social Scientists.*

3.1 Formulae for Mean and Standard Deviation of a Population

The formula for the mean (average) of N observations is given by:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

where x_1 is the value of the first observation, x_2 is the value of the second observation, etc.

Example: The weights of five children in a family are:

$$x_1 = 3.5\text{kg} \qquad x_2 = 12.3\text{kg} \qquad x_3 = 17.7\text{kg} \qquad x_4 = 20.9\text{kg} \qquad x_5 = 23.1\text{kg}.$$

Find the mean and standard deviation of the weights of these children.

Solution:

$$\begin{aligned}
\mu = \frac{1}{N}\sum_{i=1}^{N} x_i &= \frac{1}{N}(x_1 + x_2 + x_3 + x_4 + x_5) \\
&= \frac{1}{5}(3.5 + 12.3 + 17.7 + 20.9 + 23.1) \\
&= \frac{1}{5}(77.5) \\
&= 15.5.
\end{aligned}$$

A measure of how spread out the scores are, called the variance, has the following formula:

$$\begin{aligned}
\sigma^2 &= \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 \\
&= \frac{1}{5}((-12)^2+(-3.2)^2+(2.2)^2+(5.4)^2+(7.6)^2) \\
&= \frac{1}{5}(246) \\
&= 49.2.
\end{aligned}$$

The standard deviation is the square root of the variance so,

$$\begin{aligned}
\sigma &= \sqrt{\sigma^2} \\
&= \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}} \\
&= 7.0 \qquad \text{to one decimal place.}
\end{aligned}$$

3.2 Estimates of the Mean and Variance

We have, so far, concerned ourselves with the mean, variance, and standard deviation of a population. These have been written using the Greek letters μ, σ^2, and σ respectively.

However, in statistics we are mainly concerned with analysing data from a sample taken from a population, in order to make inferences about that population. Our data sets are usually random samples drawn from the population.
When we have a random sample of size n, we use the sample information to estimate the population mean and population variance in the following way.

The mean of a sample of size n is written as $\overline{x}$ (read x bar).

To find the sample mean we add up all the sample scores and divide by the number of sample scores. This can be written using sigma notation as:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The sample mean is used to estimate the population mean. If we took many samples of size n from the population, calculated their sample means, and then averaged them, we would get a value very close to the population mean. We say that the sample mean is an unbiased estimator of the population mean.

An estimate of the population variance of a sample of size n is given by s^2 where

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Notice that we are dividing by $n-1$ instead of n as we did to find the population variance. We need to do this because the value obtained if we divide by n, tends to underestimate the population variance. Calculated in this way, s^2 is an unbiased estimator of population variance. In fact, s^2 can be described as the *estimated population variance*. (It is sometimes called the 'sample variance' but this is strictly speaking not accurate.)

4 Presenting Data Using Histograms and Bar Graphs

We can now calculate two very important characteristics of a distribution, namely its 'average value' and a measure of its spread. In this section we will discuss one way of organising our data to give a visual representation of our data set. One of the most effective ways of presenting data is by a histogram.

Before we discuss histograms, we need to revise some facts about area.

4.1 Areas

Do you remember how to find the area of a rectangle? A rectangle with length l and breadth b is given in Figure 16.

Figure 16: Rectangle of length l and breadth b.

The area of a restangle = length × breadth.

For example, a rectangle of length 4 units and breadth 2 units has an area of 8 square units. This is illustrated in Figure 17.

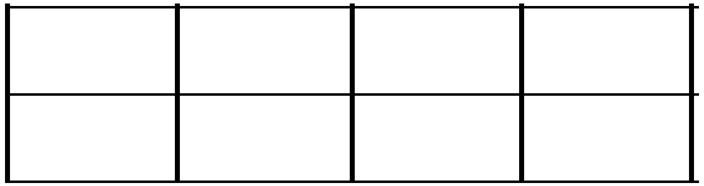

Figure 17: Rectangle of length 4 units and breadth 2 units.

4.2 Histograms

In statistics, data is often represented using a histogram. A histogram is constructed by dividing the data into a number of *classes* and then number in each class or *frequency* is represented by a vertical rectangle. The area of the rectangle represents the frequency of each class.

The table below gives the marks of 80 students on an exam. The data has already been grouped for us into 10 classes. The exam scores are given in whole marks.

Range of marks	Frequency
1–10	2
11–20	2
21–30	4
31–40	6
41–50	7
51–60	8
61–70	15
71–80	22
81–90	10
91–100	4
Total	80

A histogram of these data has been drawn in Figure 19.

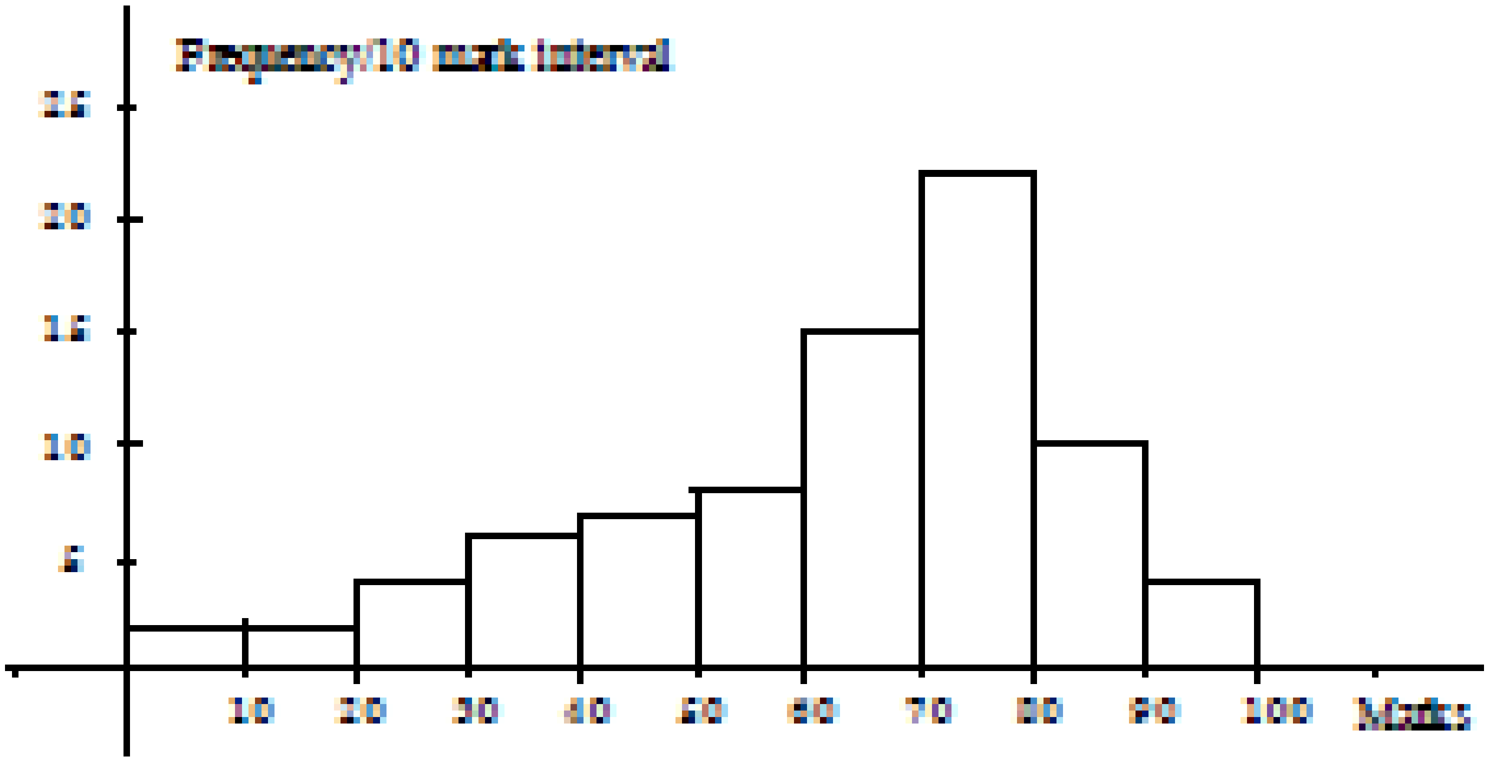

Figure 19: Histogram of students' exam marks.

Probability basics

Begin with a set Ω - *the sample space*

e.g., 6 possible rolls of a die.

$\omega \varepsilon \Omega$ is a *sample point/possible world/atomic event*

A *probability space* or *probability model* is a sample space with an assignment $P(\omega)$ for every $\omega \varepsilon \Omega$ s.t.

$0 \leq P(\omega) \geq 1$

$\sum P(\omega) = 1$

e.g., $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$.

An event A is any subset of Ω

$P(A) = \sum_{\{\omega \varepsilon A\}} P(\omega)$

E.g., $P(\text{die roll} < 4) = 1/6 + 1/6 + 1/6 = 1/2$

Prior probability

The prior or unconditional probability associated with a proposition is the degree of belief accorded to it in the absence of any other information.

Example:

$P(Weather = sunny) = 0.72$

$P(Cavity = true) = 0.1$ or $P(cavity) = 0.1$

Probability distribution gives values for all possible assignments:

$P(Weather) = (0.72, 0.1, 0.08, 0.1)$

Joint probability distribution for a set of r.v.s gives the probability of every atomic event on those r.v.s (i.e., every sample point)

P(Weather, Cavity) = a 4 × 2 matrix of values.

$Weather =$	$sunny$	$rain$	$cloudy$	$snow$
$Cavity = true$	0.144	0.02	0.016	0.02
$Cavity = false$	0.576	0.08	0.064	0.08

Every question about a domain can be answered by the joint distribution because every event is a sum of sample points

Conditional probability

The conditional probability "P(a|b)" is the probability of "a" given that all we know is "b".
e.g., $P(cavity|toothache) = 0.8$ means if a patient have toothache and no other information is yet available, then the probability of patient's having the cavity is 0.8

Definition of conditional probability:

$$P(a|b) = P(a \wedge b)/P(b) \text{ if } P(b) \neq 0$$

Product rule gives an alternative formulation:

$$P(a \wedge b) = P(a|b)P(b) = p(b|a)P(a)$$

Chain rule is derived by successive application of product rule:

$$P(X_1,\ldots,X_n) = P(X_1,\ldots,X_{n-1})\, P(X_n|X_1,\ldots,X_{n-1})$$

$$\ldots\ldots$$

$$= \prod^{n}_{i=1} P(X_i|X_1,\ldots,X_{i-1})$$

The axioms of probability

1. All probabilities are between 0 and 1:

$$0 \leq P(a) \leq 1$$

2. Necessarily true propositions have probability 1 and necessarily false propositions have probability 0:

$$P(true) = 1 \qquad P(false) = 0$$

3. The probability of a disjunction is given by:

$$P(a \cup b) = P(a) + P(b) - P(a \cap b)$$

Inference using full joint probability distribution

We use the full joint distribution as the knowledge base from which answers to all questions may be derived.

The probability of a proposition is equal to the sum of the probabilities of the atomic events in which it holds.

$$P(a) = \Sigma\, P(e_i)$$

Therefore, given a full joint distribution that specifies the probabilities of all the atomic events, one can compute the probability of any proposition.

Full joint probability distribution: example

We consider the following domain consisting of three Boolean variables: Toothache, Cavity, and Catch (the dentist's nasty steel probe catches in my tooth).

The full joint distribution is the following 2x2x2 table:

	toothache		¬toothache	
	catch	¬catch	catch	¬catch
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

The probabilities in the joint distribution must sum to 1.

The probability of any proposition can be computed from the probabilities in the table.

Each cell represents an atomic event and these are all the possible atomic events

Full joint probability distribution: example

	toothache		¬toothache	
	catch	¬catch	catch	¬catch
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

We simply identify those atomic events in which the proposition is true and add up their probabilities

P(cavity or toothache) = P(cavity, toothache, catch) + P(cavity, toothache, ¬catch) + P(cavity, ¬toothache, catch) + P(cavity, ¬toothache, ¬catch) + P(¬cavity, toothache, catch) + P(¬cavity, toothache, ¬catch)

$= 0.108+0.012+0.072+0.008+0.016+0.064=0.28$

Marginalization or summing out

	toothache		¬toothache	
	catch	¬catch	catch	¬catch
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

P(cavity) → cavity row

P(¬toothache) → ¬toothache columns

P(cavity, ¬toothache) → cavity row of ¬toothache columns

That is, a distribution over **Y** can be obtained by summing out all the other variables from any joint distribution containing **Y**.

$$\mathbf{P(Y)} = \Sigma\, \mathbf{P(Y}, z)$$

$P(cavity) = 0.108+0.012+0.072+0.008 = 0.2$

$P(\neg toothache) = 0.072+0.008+0.144+0.576 = 0.8$

$P(cavity, \neg toothache) = 0.072+0.008 = 0.08$

Conditioning

P(**Y**) = Σ P(**Y**, z)
P(**Y**, z) = P(**Y**|z)P(z)
Therefore, for any set of variables **Y** and **Z**:
P(**Y**) = Σ P(**Y**|z)P(z) - This is the conditioning rule

	toothache		¬toothache	
	catch	¬catch	catch	¬catch
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

$$P(\neg cavity|toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)}$$
$$= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

Bayes' Rule

Product rule:

$$P(a \wedge b) = P(a|b)P(b)$$
$$P(a \wedge b) = P(b|a)P(a)$$

Bayes rule:

$$P(b|a) = \frac{P(a|b) * P(b)}{P(a)}$$

Why is the Bayes' rule is useful in practice ?

Bayes' rule is useful in practice because there are many cases where we have good probability estimates for three of the four probabilities involved, and therefore can compute the fourth one.

Useful for assessing diagnostic probability from causal probability:

$$P(Cause|Effect) = \frac{P(Effect|Cause)P(Cause)}{P(Effect)}$$

Diagnostic knowledge is often more fragile than causal knowledge.

Applying Bayes' Rule

Example:

A doctor knows that the disease meningitis causes the patient to have a stiff neck 50% of the time. The doctor also knows that the probability that a patient has meningitis is 1/50,000, and the probability that any patient has a stiff neck is 1/20.

Find the probability that a patient with a stiff neck has meningitis.

$p(s|m) = 0.5$

$p(m) = 1/50000$

$p(s) = 1/20$

$P(m|s) = P(s|m)P(m)/P(s) = (0.5*1/50000)/(1/20) = 0.0002$

Thank You

Any Query??