BIOSTATISTICS
TOPIC 2: DESCRIPTIVE STATISTICS

The amount of statistical information that is disseminated to the public and indeed, medical literature, for one reason or another is sometimes beyond comprehension, and what part of it is "good" statistics and what part of it is "bad" statistics is anybody's guess. Certainly, all of them can not be accepted uncritically. Sometimes, entirely erroneous conclusions are based on unsound data. Indeed, use of statistics has already been replaced by overuse and abuse. People are writing books and papers based on inappropriate application of statistics. Alvan Feinstein recently commented: "some of these authors are very popular because they are not afraid to provide solutions to problems that have not yet been solved." We, of course, do not want to go down to that path. We need to use statistics wisely.

In this topic we will deal with the use of some basic statistical indicators which are usually referred to as descriptive statistics. Specifically, we will be concerned with the summarising of continuous data. We will discuss four main themes:

Measures of central tendency
Measures of variability
Measures of shapes of distribution
Application of descriptive statistics.

I. MEASURES OF CENTRAL POSITION

1. THE **ARITHMETIC MEAN** of a set of observations $x_1, x_2, ..., x_n$ is defined by

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

When the data are in the form of frequency distribution, the mean takes into account the number of observations per category. Suppose that we have *k* categories, each with $n_1, n_2, \ldots, n_k$ number of observations (total sample size: $N = n_1 + n_2 + \ldots + n_k$) and associated with means $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$. Then the overall mean is given by:

$$\boxed{\bar{x} = \frac{1}{N}\sum_{i=1}^{k} n_i \bar{x}_i}$$

Example 1: The number of subjects and mean lumbar spine BMD for three genotypes are as follows:

Genotype	n	Mean
TT	40	1.25 g/cm^2
Tt	45	1.10 g/cm^2
tt	15	1.00 g/cm^2

We calculate the mean by using the above relation as follows:

$$\begin{aligned}\bar{x} &= \frac{1}{N}\sum_{i=1}^{k} n_i \bar{x}_i \\ &= \frac{1}{100}\left[(40\times 1.25)+(45\times 1.10)+(15\times 1.00)\right] \\ &= 1.145\ g/cm^2 \qquad //\end{aligned}$$

2. THE **GEOMETRIC MEAN** of a set of observations $x_1, x_2, \ldots., x_n$ is the antilogarithm of the arithmetic mean of the logarithms of the values, i.e.

$$\boxed{\log G = \frac{\log x_1 + \log x_2 + \ldots + \log x_n}{n}} = \frac{\log(x_1 \times x_2 \times \ldots \times x_n)}{n}$$

then the mean is $\bar{x} = anti\log(G) = (x_1 \times x_2 \times \ldots \times x_n)^{1/n}$.

The geometric mean is a useful measure of position for data involving ratios. As it can be seen from this relation, the geometric mean is undefined for a set of values with zeros or negative values.

Example 2: The percentage increase in osteocalcin in a group of 10 patients between visits was as follows:

Between visit 2 and 1:	5.4%
Between visit 3 and 2:	8.9%
Between visit 4 and 3:	9.6%
Between visit 5 and 4:	6.4%

To calculate the average percentage increase over the 5 visits, we need to (i) firstly convert the percentage data into ratio and (ii) apply the geometric formula.

The 4 percents could be written in ratio terms as: 1.054 1.089 1.096 1.064
Then the average log(ratio) is:

$$\ln(\mathrm{R}) = \frac{\ln\left(1.054\times 1.056\times 1.096\times 1.064\right)}{4}$$
$$= \frac{0.2608}{4} = 0.0652.$$

and the average ratio is $e^{0.0652} = 1.067$ or 6.7%//

3. THE **HARMONIC MEAN** of a set of observations $x_1, x_2, \ldots., x_n$ is the reciprocal of the arithmetic mean of the reciprocals of the values, i.e.

$$\boxed{\frac{1}{H} = \frac{\frac{1}{x_1}+\frac{1}{x_2}+\ldots+\frac{1}{x_n}}{n}}$$

So: $$\bar{x} = H = n / \sum_{i=1}^{n} \frac{1}{x_i}$$

When a data set contains values which represent rates of change, the harmonic mean is an useful measure of central tendency.

4. THE **MEDIAN** of a set of observations is the value of the middle term when all observations are arranged in order of magnitude. It is symbolised by *Md*.

Example 3: For the set of values 14, 17, -13, 41, 12. We can find the median as follows:

(i) firstly, rearrange the numbers:	-13	12	**14**	17	41
(ii) ranking them:	1	2	**3**	4	5

The median is obviously 14.

However for a set of values: -13 12 **14** **17** 41 66

The median is (14+17)/2 = 15.5.

5. The **MODE (*m*)** is another measure of central tendency which occurs at the most frequently observed value of the variable.

For example, for a set of data {4, 5, 3, 2, 4, 1, 7, 4, 2, 4}, the mode would be 4 since it is the most frequently occurred number.

II. MEASURES OF VARIABILITY

1. **VARIANCE**. The most commonly used measure of dispersion in statistical analysis is called the variance. It is a measure that takes into account all the values in a set of observations.

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

which is equivalent to: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \frac{n(\bar{x})^2}{n-1}$

For weighed data: $\boxed{s^2 = \frac{1}{n-1}\sum_{i=1}^{k} w_i(x_i - \bar{x})^2} = \frac{1}{n-1}\left[\sum_{i=1}^{k} w_i x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k} w_i x_i\right)^2\right]$

where $n = \sum_{i=1}^{k} w_i$

The wider the dispersion of the values around their mean, the greater the variance. If there is no dispersion (eg 5, 5, 5, 5) then all values are equal to the mean; it follows that the variance is 0.

Example 4: Consider the data set 5, 17, 12 and 10, whose mean is $\bar{x}$=11. We calculate the variance as follows:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{4-1}\left[(5-11)^2 + (17-11)^2 + (12-11)^2 + (10-11)^2\right]$$

$$= \frac{(-6)^2 + 6^2 + 1^2 + (-1)^2}{3}$$

$$= 24.67.$$

Example 1 (continued): For the data in Example 1, we can treat the number of subjects in each genotype as weights. The calculation of variance can be illustrated by the following table:

Genotype	n (w_i)	Mean (x_i)	$w_i x_i^2$	$w_i x_i$
TT	40	1.25	62.50	50.0
Tt	45	1.10	54.45	49.5
tt	15	1.00	15.00	15.0
Total	100		131.95	114.5

then
$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{k} w_i x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k} w_i x_i\right)^2\right]$$

$$= \frac{1}{99}\left[131.95 - \frac{(114.5)^2}{100}\right]$$

$$= 0.00856\ g^2 / cm^4$$

2. **Standard Deviation.** The positive square root of the variance is called the standard deviation and is denoted by s.

$$s = \sqrt{s^2}$$

The variance is expressed in units that are the square of the unit of measure of the variable under study. For instance the variance of BMD is measured as $(g/cm^2)^2$. However, the standard deviation is expressed in the original unit of measure of the variable e.g. g/cm^2.

In Example 4, the standard deviation is: $s = \sqrt{24.67} = 4.97$ g/cm^2.

If the data set has a large number of observations and approximately symmetrical, the standard deviation can be roughly approximated by using the maximum and minimum values as follows:

$$
\begin{aligned}
s &= (\text{max} - \text{min})/\sqrt{n} && \text{for } n < 12 \\
&= (\text{max - min}) / 4 && \text{for } 20 < n < 40 \\
&= (\text{max - min}) / 5 && \text{for } n \text{ about } 100 \\
&= (\text{max - min}) / 6 && \text{for } n > 400.
\end{aligned}
$$

3. **Standard Error** (SE) is the standard deviation of the means of samples of given size drawn from a particular parent population. If n is the sample size and N is the size of the parent population and σ is the standard deviation of the parent population, then the SE is defined by: $\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$. Therefore, for a large parent population or for sampling with replacement this equation may be simplify to: $\frac{\sigma}{\sqrt{n}}$. However, in a sample of data, SE is estimated by:

$$\text{SE} = \frac{s}{\sqrt{n}}$$

SE is a measure of a reasonable difference between a sample mean and the parent population mean and is used to test of whether a particular sample could have drawn from a given parent population. It is used to work out the confidence limit.

The SE for the data set in Example 4 is: $\text{SE} = \frac{s}{\sqrt{n}} = \frac{24.67}{\sqrt{4}} = 12.3$ g/cm^2.

4. **Coefficient of Variation**. The standard deviation is a measure of the absolute variability in a set of observation. For a number of problems, however, the relative variability is a more significant measure. The most commonly used measure of relative variability is the coefficient of variation:

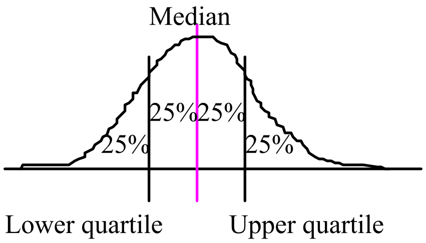
$$\text{CV} = \frac{s}{\bar{x}} \times 100$$

CV is used when *all* values of a variable are positive. When the values are both negative and positive the CV becomes rather meaningless.

The CV for the data set in Example 4 is estimated by: $\text{CV} = \frac{4.97}{11} \times 100 = 45.2\%$

5. **Percentile**. The pth percentile of a set of observations arranged in order of magnitude is the value that has at most p% of the measurements below it and at most (100 - p)% above it.

The following figure illustrates the 25th, 50th and 75th percentiles, often called the lower quartile, the middle quartile (median) and the upper quartile, respectively.



Example 5: Consider the following data set with 10 observations:

-15 -9 1 3 5 9 13 17 23 92

where the median can be estimated to be: (5+9)/2 = 7. So, the 50th percentile is 7. Similarly, the 25th percentile is 1 and the 75th percentile is 17, and so on.

III. Measures of Shapes

Skewness. One way to study the skewness of a frequency distribution is to compare the values of the mode, median (*Md*) and mean ($\bar{x}$). We know that the mode is the position on the scale that has the greatest concentration of observations; the median is the value where half of the observations lie below and above; and the mean tends to be pulled in the direction of the extreme values. Therefore, for a symmetrical and unimodal distribution, all the values of the mean, median and mode should be identical; otherwise, the distribution is not symmetrical and unimodal. The coefficient of skewness (S) is defined by:

$$\mathrm{S} = \frac{3\left(\bar{x} - Md\right)}{s} \qquad \text{or} \qquad S = \frac{\bar{x} - Mode}{s}$$

where *s* is the sample standard deviation.

If S is positive (mode < mean), the distribution is skewed toward the right side; if S is negative (mode > mean), the distribution is skewed toward to left side.

IV. Applications of Descriptive Statistics

1. Empirical Relations Between Mean, Median and Mode.

We have surveyed three main measures of central tendency. The question now is which measure is the most appropriate and reliable? The answer to this question depends on the distribution of the observed data. However, it can be stated that, like any physical measures, none of the above statistics is perfect in describing a central position of a distribution.

What can reasonably be stated is that from a theoretical point of view, the mean is the best measure of central tendency of a distribution. This is because it can be computed for numerical data, makes use of all the observations and is unique. Furthermore, it is readily understood by most people. While the mean is influenced by extreme values, the median does not. However, the median is not likely to be representative when number of observations is small because it is a positional average; it is also not unique. On the other hand, unless the number of observations is sufficiently large and the distribution of the data reveals a clear picture of central tendency, the mode has no significance.

If the distribution of a data set is symmetrical as in figure 1, the mean, the median and the mode are the same (or at least similar). If the distribution is skewed to the right (as in Figure 2), the mean is larger than the median. If the distribution is skewed to the left (Figure 3), the mean is smaller than the median.

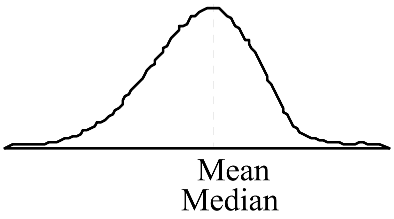


Figure 1

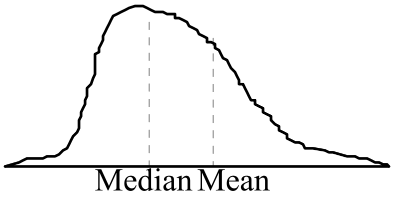


Figure 2

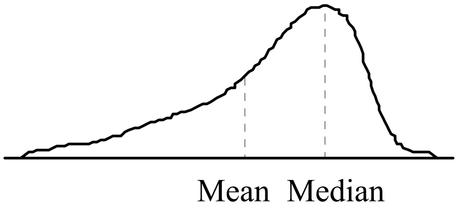


Figure 3.

For a reasonably large data set with approximately symmetrical, an empirical relation between mean, median and mode can be established:

$$\text{Mean} - \text{Mode} = 3(\text{Mean} - \text{Median});$$

That is, given a median and a mean, the value of the mode can be approximated by:

$$\text{Mode} = 3(\text{Median}) - 2(\text{Mean})$$

2. Chebyshev's Theorem and Estimating Ranges of Values and Confidence Interval.

It is important to emphasise here again that a set of data is a sample drawn from the population of all possible measurements. Thus, the sample mean $\bar{x}$, standard deviation *s,* etc. may not equal to the true population mean and standard deviation which are usually denoted by Greek characters such as μ and σ. The purpose of a parametric estimation is not just to get an estimate of the mean in the general population, but also to indicate its "uncertainty", i.e. how close or far off the estimate may be from the true value. Related to this estimation is the concept of confidence limit and is introduced here via the **Chebyshev's theorem**, one of the great theorem in probability which was named after the great Russian mathematician. The exact statement of this theorem is quite mathematically involved, however, it can be interpreted as follows:

(a) the interval $\bar{x}$-3s to $\bar{x}$+3s contains at least 89% of measurements;
(b) the interval $\bar{x}$-2s to $\bar{x}$+2s contains at least 75% of measurements;
(c) the interval $\bar{x}$-s to $\bar{x}$+s contains at least 0% of measurements.

In practice, this statement is rather conservative. For reasonably symmetrical and large data set, the **empirical rule** states that:

(a) 68% of measurements can lie between $\bar{x}$-s to $\bar{x}$+s;
(b) 95% of measurements can lie between $\bar{x}$-2s to $\bar{x}$+2s;
(c) 99.7% of measurements can lie between $\bar{x}$-3s to $\bar{x}$+3s.

Use of Standard Deviation. For any symmetrical data set with given mean ($\bar{x}$) and standard deviation (s), we could estimate the range of individual measurements with certain accuracy. For example, the mean and standard deviation of (natural) logarithmic osteocalcin of a sample of Sydney subjects are 2.86 and 0.45 respectively; it could be inferred that approximately 95% of subjects in this sample have their log(osteocalcin) between 2.86-2(0.45) to 2.86+2(0.45) (or 1.96 to 3.76).

Use of Standard Error. The standard error (SE) which we discussed earlier is often referred to as *standard deviation of the mean,* since it indicates the difference between a *sample* mean and the parent *population* mean. The latter is often unknown.

However, one can apply the Chebyshev's theorem to estimate the range of possible values of the population mean with certain confidence.

For example, the mean and standard error of femoral neck BMD among 20 fracture women from a community in Sydney was found to be 0.70 g/cm^2 and 0.02 g/cm^2, respectively. The true mean femoral neck BMD of *all* fracture subjects in Sydney was unknown. However, it could be stated that the true mean could lie between 0.70-2(0.02) = 0.66 g/cm^2 to 0.70+2(0.02) = 0.74g/cm^2. What it means here is that, if we keep sampling 20 fracture women from the Sydney population repeatedly (each time with different subjects) and each time the mean of 20 women was calculated, then we would expect that 95% of the times, the mean lies between 0.66 g/cm^2 to 0.74g/cm^2.

3. **Transformation:**

For a set of values $x_1, x_2, x_3, \ldots, x_n$, let the mean be $\bar{x}$ and the variance be s_x^2, then for any constants *a* and *b*, we have the following properties:

(a) Linear transformation: $y_i = a + bx_i$. The mean and variance of *Y* is defined as:

$$\bar{y} = a + b(\bar{x})$$

and $$s_y^2 = b^2\left(s_x^2\right)$$

For example, the mean and variance of a variable *X* was 10 and 8, respectively. If a new variable *Y* = 12 + 2*X*, then the mean and variance of *Y* are:

mean(*Y*) = 12 + 2.mean(*X*) = 12 + 2(10) = 32

and variance(*Y*) = 2^2.variance(*X*) = 4(8) = 32.

(b) Z-transformation: $z_i = \frac{x_i - \bar{x}}{s_x}$. The mean and variance of *Z* could be shown to be:

$$\bar{z} = 0$$

and $$s_z^2 = 1.$$

4. **Presentation of Descriptive Statistics:**

It is not uncommon nowadays in biomedical journals such presentation as $a \pm b$ is increasingly common. Some researchers indicate the two values as mean ± SE or mean ± SEM or mean ± SD; others do not care to mention what these numbers actually stand for.

In customary scientific usage, of course, the b of an $a\pm b$ expression refers to the accuracy of the measurement. Thus, if someone reports that a specimen weighs 27±2 mg, the idea is that its weight can be anywhere from 25 to 29 mg. In statistical usage, the ± usage has this same meaning if it refers to a confidence interval around a mean. A statement such as "the 95% confidence interval was 250 ± 10" implies that in a series of random samples taken from this same population, 95% of the means would lie between 240 and 260. But what is the value of the ± sign when it refers to the standard deviation or standard error. A reader who wants to use the information can not do directly. Perhaps a "mean (SD)" expression would be more helpful.

V. EXERCISES

1. Write down a list of 5 numbers satisfying both the following criteria:
(a) the median < the mean (b) the mode < the median.

2. Show that the sum of the deviations of a set of measurements, x_i, about their mean, $\bar{x}$, is zero, i.e. $\sum_{i=1}^{n}\left(x_i - \bar{x}\right) = 0$.

3. The hospitalised cost of fracture (in $AUS) for 29 patients in Dubbo is as follows:
5373, 15984, 7478, 3446, 11004, 9116, 3213, 5418, 16386
2857, 3656, 61876, 2972, 3057, 14449, 9400, 27518, 23278
23548, 3016, 12921, 4640, 4644, 23098, 2654, 7975, 10245
4045, 5018.
Construct a histogram of distribution of cost (you may use 5000-interval such as 5000-1000, 10001-15000, 15001-20000 etc.)
Calculate the mean, standard deviation, median, coefficient of skewness etc. and comment on the distribution of data.

4. What can be said about a set of measurements which has a standard deviation of zero?

5. A set of 10 numbers gave a mean of 13 and a standard deviation of 2. Later it was found that the number 12 in the set should have been 21. Find the corrected mean and standard deviation.

6. When hunting insects, bats send out high-frequency sounds and then listen for the echoes. One interest question is the distances (in cm) between the bat and its intended prey when the bat's echo-location system first detects the insect.
The following data comprise the bat-to-prey detection distances for 11 catches:
62 52 68 23 34 45 27 42 83 56 40
(a) Find the mean of the data set.
(b) Calculate the standard deviation of the data set, using: (i) the exact mean (calculated to 2 d.p) (ii) the rounded mean.
(c) Calculate 95% confidence interval (CI) for the measurements and 95% CI for the mean.
Comment on the difference between these results.

7. The osteocalcin of 5 subjects are as follows: 4, 3, 7, 11 and 10.

(a) Calculate the mean ($\bar{x}$), variance (s^2), standard deviation and standard error (SE) *manually*. Show your working fully.

(b) Transform the original observation by subtracting the mean from each observation (eg $(x_i - \bar{x})$). Show that the mean of $(x_i - \bar{x})$ is zero.

(c) Let $z_i = \frac{x_i - \bar{x}}{s}$. Show that the mean and variance of *Z* is 0 and 1, respectively.

8. A set of 340 scores exhibiting a bell-shaped relative frequency distribution has means $\bar{x}$ = 72 and standard deviation s = 8. How many of the scores would you expect to fall in the interval 64 to 80? 56 to 88?

9. The theoretical frequency and phenotype value of a 2-allele gene locus (A and a) with respective frequency *p* and *q*, are normally given by:

Genotype	**No. of subjects**	**Phenotype**
AA	p^2	$\mu + a$
Aa	$2pq$	$\mu + d$
aa	q^2	μ - a

Where q = 1-p. Express the overall mean and variance of the phenotype in terms of *μ, a, d, p* and *q*.

10. Data on lumbar spine BMD from 123 twins in Sydney stratified by VDR genotypes are as follows:

Genotype	*n*	lumbar spine BMD
TT	32	1.25 g/cm^2
Tt	61	1.17 g/cm^2
tt	30	1.07 g/cm^2

n: number of individuals in each genotype.

Find the mean and variance of lumbar spine BMD for these twins.

11. Given a set of observations $X = \{3,5,6,7,9\}$.

(a) Calculate the mean, standard deviation and median.

(b) Find the mean and variance of Y when

(i) $y_i = x_i - 8$ (ii) $y_i = 7x_i$ (iii) $y_i = \frac{x_i}{12}$ (iv) $y_i = \frac{x_i - 5}{7}$.

What relation can you deduce for each of the cases ?

12. Use the technique of transformation (page 9) to calculate the mean and variance (and hence SD) of the following samples: 997, 995, 998, 992 and 995, without using a calculator.

13. Let $X = \{4,3,7,10,11\}$. Transform the above observations by natural logarithm of x_i. Find the mean and variance of *X* and ln(*X*). Are these statistics similar between the two variables. Is the mean of ln(*X*) equal to the log of mean of *X* ? Why ?

14. Osteocalcin among a sample of 100 subjects from Denmark has the following characteristics:

Mean: 6.9 ng/ml
Standard deviation: 5.1 ng/ml
Median: 6.2 ng/ml.

Comments on the distribution of the data.

15. Some characteristics of bone mineral contents (BMC) for Black and White people are as follows:

	Mean	Median	SD
Black:	2872	2812	374
White:	2744	2805	250

Calculate the coefficient of skewness for each group and comment on the results.

16. The changes in the vitamin D 1,25 level for a patient in 4 consecutive days are as follows:

Day 1: 35; Day 2: 36; Day 3: 38; Day 4: 40

(a) Obtain the ratio of the change in one day to that in the preceding day for days 2, 3 and 4.

(b) Obtain the geometric mean of the three ratios. Show that the change in day 4 can be obtained from knowledge of the change in day 1 and the geometric mean.

17. Data on lumbar spine BMD from a sample of 10 subjects are as follows: 0.98, 1.05, 1.01, 0.97, 0.95, 0.87, 0.50, 0.89, 1.05 and 1.08. Notice that there is one subject with very low BMD. Would you exclude this subject from estimating the mean ?

18. In an experiment designed to answer the question "does environment affect the anatomy of the brain", rats from a genetically pure strain were randomly allocated to two groups: a treatment group and a control group. Those in the treatment group were placed in large cages with new toys every day. Those in the control group were isolated in separate cages with no toys. After a month, the cortex (grey matter of the brain) were weighed. The weights in mg were as follows:

    Treatment group: 707 740 745 652 649 676 699 696 712 708 749 690
    Control group: 669 650 651 627 656 642 698 648 676 657 692 621

    (a) Present the data in a graphical format so that it could be visualised easily.
    (b) Calculate the relevant statistics and discuss on their values.