7.6.2 Measures of dispersion

(a) Sum of squared deviations

One of the important aspects in statistical analysis is to analyze the deviation of the dataset from the mean. We generally do this using the sum of squared deviations from the mean (SS). This is given as:

$$SS = \Sigma(X - \bar{X})^2 \qquad (7.40)$$

Let's take up a dataset representing the size of a particular species of fish as follows: 10.1, 11.2, 11.4, 11.7, 12.2, 13.4 and calculate the mean deviation of the dataset from the mean.

X	$\bar{X}$	$X - \bar{X}$	$(X - \bar{X})^2$
10.1	11.66	- 1.56	2.43
11.2		- 0.46	0.21
11.4		- 0.26	0.07
11.7		+0.04	0.0016
12.2		+0.54	0.29
13.4		+1.74	3.03
$\Sigma X = 70.0$			$\Sigma(X - \bar{X})^2 = 6.03$

(b) Variance

The measures of central tendency that is, the mean, median, and mode does not give us any idea of the dispersion or the deviation of the data from the mean. A measure of dispersion or variability is the variance (s^2). This is given as:

$$s^2 = \frac{SS}{DF} \tag{7.41}$$

where SS = sum of the squared deviations from the mean = $\Sigma X^2 - [(\Sigma X)^2/n]$ and DF = the degrees of freedom this is defined as (n - 1). The calculation of variance has been shown in box 7.10.

(c) Standard deviation

The universally used measure of dispersion is the standard deviation or the root mean square deviation represented as

$$\text{Standard deviation } \sigma = \sqrt{s^2} \tag{7.42}$$

An example has been provided in box 7.11.

Box 7.10 Problem: Calculate the variance from the dataset of fish length 10.1, 11.2, 11.4, 11.7, 12.2, 13.4

Solution: $\Sigma X = 10.1 + 11.2 + 11.4 + 11.7 + 12.2 + 13.4 = 70$

$\bar{X} = 70/6 = 11.67$

$\Sigma X^2 = (10.1)^2 + (11.2)^2 + (11.4)^2 + (11.7)^2 + (12.2)^2 + (13.4)^2$

$= 102.01 + 125.44 + 129.96 + 136.89 + 148.84 + 179.56$

$= 822.7$

$SS = 822.7 - [(70)^2/6] = 822.7 - 816.7 = 6$

$s^2 = SS/DF = 6 / (6 - 1) = 6 / 5 = 1.2$

Box 7.11 Problem: Calculate the standard deviation of the following set of plankton measurement 23, 24, 25, 26, 28, 30.

Solution:

X	X^2
23	529
24	576
25	625
26	676
28	784
30	900
156	4090

$$SS = \sum X^2 - \frac{(\sum X)^2}{n}$$

So $\sum X = 156$ and $\sum X^2 = 4090$

$$SS = 4090 - \frac{(156)^2}{6} = 34$$

$$s^2 = \frac{SS}{DF} = \frac{34}{6-1} \quad \text{and} \quad \sigma = \sqrt{s^2} = \sqrt{\frac{34}{5}} = 2.608$$

$$s_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \qquad (7.43$$

Box 7.12 Problem: Given below (X and Y) are the datasets for diurnal oxygen concentration in two polluted water bodies. Find out which dataset shows greater dispersion?

Solution:

X mg l^{-1}	X^2	Y mg l^{-1}	Y^2
4.6	21.16	5.4	29.16
4.4	19.36	8.0	64.00
4.8	23.04	8.8	77.44
4.0	16.00	3.6	12.96
7.0	49.00	4.0	16.00
4.8	23.04	5.6	31.36
9.6	92.16	3.2	10.24
4.2	17.64	7.2	51.84
5.4	29.16	7.2	51.84
5.4	29.16	5.6	31.36
4.8	23.04	4.0	16.00
7.2	51.84	4.0	16.00
5.8	33.64	4.0	16.00
5.6	31.36	3.6	12.96
$\Sigma X = 77.6$	$\Sigma X^2 = 459.6$	$\Sigma Y = 74.2$	$\Sigma Y^2 = 437.16$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{77.6}{14} = 5.543 \quad \text{and} \quad \bar{Y} = \frac{\Sigma Y}{n} = \frac{74.2}{14} = 5.3$$

$$SS_X = \Sigma X^2 - \frac{(\Sigma X)^2}{n} = 459.6 - \frac{(77.6)^2}{14} = 459.6 - \frac{6021.76}{14}$$

$$= 459.6 - 430.12 = 29.48$$

$$s^2 = \frac{SS}{DF} = \frac{29.48}{14-1} = 2.27 \quad \sigma_X = \sqrt{s^2} = \sqrt{2.27} = 1.506$$

$$CV_X = \frac{\sigma}{\bar{X}} = \frac{1.506}{5.543} \times 100 = 27.2\%$$

$$SS_Y = \sum Y^2 - \frac{(\sum Y)^2}{n} = 437.16 - \frac{(74.2)^2}{14} = 437.16 - \frac{5504.64}{14}$$

$$= 437.16 - 393.26 = 43.9$$

$$s^2 = \frac{SS}{DF} = \frac{43.9}{14-1} = 3.377 \quad \sigma_Y = \sqrt{s^2} = \sqrt{3.377} = 1.838$$

$$CV_Y = \frac{\sigma}{\bar{Y}} = \frac{1.838}{5.3} \times 100 = 34.68\%$$

So we can say that the dataset Y has a greater dispersion

Thus, the standard error in our previous example of plankton measurements is

$$s_{\bar{X}} = \frac{2.608}{\sqrt{6}} = \frac{2.608}{2.449} = 1.065$$

If a series of sub samples are taken from the population sample and their means are determined then two – thirds of the mean will deviate about the true mean (μ- mean of the population sample) by one standard error ($s_{\bar{X}}$).

(e) Coefficient of variation

To obtain a relative measure of dispersion for the purpose of comparing two distributions we generally use the Karl Pearsons coefficient of variation. This is obtained by dividing the standard deviation by the mean from which it computed:

$$CV = \frac{\sigma}{\bar{X}} \times 100 \qquad (7.44)$$

(f) Significance of standard deviation

We have seen how to calculate the standard deviation. Lets observe as to what use it is to us in any interpretation. In any distribution which is reasonably symmetrical and unimodal (single peak) we can assume that two-thirds of the distribution lies less than one standard deviation from the mean, that 95 percent of the distribution lies less than two standard deviations from the mean, and that less than 1 percent of the distribution lies more three standard deviations away from the mean. Suppose we are given a dataset that has a mean value of 100 and a standard deviation of 15. Then we might assume the distribution as shown in figure (figure 7.6).

This interpretation of the standard deviation provides us with a rough check of the value of σ obtained in a particular case. In a unimodal distribution

substantially all the values should be within three standard deviations of the mean. This implies that by adding 3σ we should get the greatest value of the variable, and by subtracting 3σ we should get the least value of the variable. This is called the three standard deviation check.

(g) Confidence limits t - test

Calculation of the standard deviation for a set of data provides us with an indication of the precision inherent in a particular procedure or analysis. But, unless there is a large number of data, it does not by itself give any information about how close the experimentally determined mean might be to the true mean value. Statistical theory however, allows us to estimate the range within which the true value might fall. The range is called the confidence interval, an interval that, with a stated level of confidence, may be said to include the population mean μ. The limits of this range are called confidence limits. The confidence limit is given as:

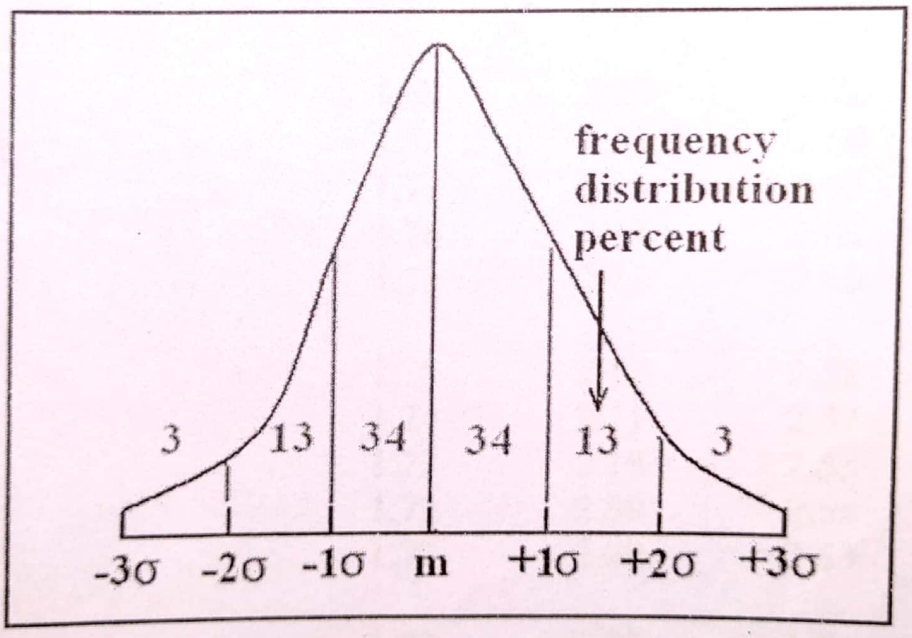

Figure 7.6 Normal distribution curve of a dataset with $\bar{x} = 100$ and $\sigma = 15$

$$\text{Confidence limit for } \mu = \bar{X} \pm ts_{\bar{X}} \qquad \text{Eq. 7.45}$$

Where μ = population mean
$\bar{X}$ = sample mean
$s_{\bar{X}}$ = sample standard error
and t is obtained from a statistical distribution known as student's t (table 7.1).

Table 7.1 Critical values of student's t (from Zar, 1996)

DF	Confidence level	90%	95%	99%	99.5%
	Significance level α =	0.01	0.05	0.02	0.01
1		6.31	12.71	31.82	63.66
2		2.92	4.31	6.96	9.92
3		2.35	3.18	4.54	5.84
4		2.13	2.78	3.75	4.60
5		2.01	2.57	3.36	4.03
6		1.94	2.45	3.14	3.71
7		1.89	2.36	3.00	3.50
8		1.86	2.31	2.90	3.36
9		1.83	2.26	2.82	3.25
10		1.81	2.23	2.76	3.17
11		1.80	2.20	2.72	3.11
12		1.78	2.18	2.68	3.06
13		1.77	2.16	2.65	3.01
14		1.76	2.14	2.62	3.00
15		1.75	2.13	2.60	2.95
16		1.75	2.12	2.58	2.92
17		1.74	2.11	2.57	2.90
18		1.73	2.10	2.55	2.88
19		1.73	2.09	2.54	2.86
20		1.72	2.09	2.53	2.85
22		1.72	2.07	2.51	2.82
24		1.71	2.06	2.49	2.80
26		1.71	2.06	2.48	2.78
28		1.70	2.05	2.47	2.76
30		1.70	2.04	2.46	2.75

DF = degrees of freedom (n – 1)

A significance level of 5% or a confidence level of 95% is most frequently used in biological research. For instance in the problem given in box 7.13 the mean length of fishes is 11.67 then

95% confidence interval for $\mu = 11.67 \pm 2.57 \times 0.45 = 11.67 \pm 1.16$ cm

Thus we can say that the mean of the entire population from which our sample was drawn is 11.67± 1.16. This assertion can be made with 95 percent confidence or we can assume that if 100 random samples were taken from this population, 95 of them would lie between 10.51 cm and 12.83 cm.

Box 7.13 Problem: The height of a group of 10 trees is as follows:
14.8m, 12.8m, 14.7m, 12.7m, 15.0m, 14.5m, 12.4m, 14.0m, 14.2m, and 14.9m.
Let's make the hypothesis that the average height of the group is 13, test whether it is true or not.

Solution:

X	X^2
14.8	219.04
12.8	163.84
14.7	216.09
12.7	161.29
15.0	225.00
14.5	210.25
12.4	153.76
14.0	196.00
14.2	201.64
14.9	222.01
140	1968.92

$SS = 8.92 \quad s^2 = 8.92/9 = 0.991 \quad s_{\bar{x}} = \sqrt{\frac{s^2}{n}} = 0.315 \quad t = \frac{\bar{X} - \mu}{s_{\bar{x}}}$

$= (14 - 13)/0.315 = 1/0.315 = 3.17$, for 9 degrees of freedom the tabulated value of t at 95 percent confidence level is 2.26. As the calculated value of t is higher than the tabulated value, we can conclude that the sample was not drawn from the population having an average height of 13 m.