CHAPTER THREE

Variability

Overview

Variation is the fluctuation of scores about a measure of central tendency. To describe a set of measurements accurately we need to know both the central tendency and the measure of the variation. Dispersion is a concept synonymous with variability.

The **range** is the simplest and most straightforward measure of variability; it is the difference of the lowest score from the highest score. It does not however, tell us anything about the pattern of the distribution of data.

The **variance** is the average sum of square deviation of a set of data from the mean and is often denoted by the symbol, $\mathbf{s}^2$ or $\boldsymbol{s}^2$.

The **standard deviation** is the square root of the variance, or the average sum of square deviation of a set of data from the mean. It is the most widely used measure of variability and is often reported in most research statistical summaries. It is often denoted by the symbol, **s** or ***s***.

The variability of a distribution can be represented by a single value (point estimator) or by graphical display from a frequency distribution of a Box-plot display.

Variability is a quantitative measure of the degree or extant to which scores in a distribution are spread out or clustered together.

3.1 Variability

Variability describes a distribution; it tells how close or far apart scores are from each other. Variability describes the relative distance one score is from another. It also can be used to describe how well individual scores represent the entire distribution. Common point estimates or measures of variability are range, semi-interquartile range for ordinal data, and variance and standard deviation. The variability of a distribution can also be represented by a display called the box-plot which combines measures of both central tendency and variability.

Range

The range is the simplest measure of variability to compute. It is also provides the least amount of information about the variability of a dataset. Simply, the range is the difference between the highest score and the smallest score of a distribution. Often when computing the range, it is best to order the scores from smallest to largest or largest to smallest in a frequency distribution.

The **range** is the difference between the upper real limit (maximum score) of the largest score and the lower real limit (minimum score) smallest score.

range = $X_{max} - X_{min}$

How would you find the range for the following data: {12, 10, 8, 20, 15, 22, 14}?

With the scores arranged in order, {8, 10, 12, 14, 15, 20, 22}, the range = 22 – 8 = **14**

Find the range from the frequency distribution below:

X	*f*
34 – 36	4
31 – 33	5
28 – 30	10
25 – 27	6
22 – 24	3

The range = **max***(real upper limit)* – **min***(real lower limit)* = 36 – 22 = **14**

Semi-Interquartile Range (Q)

Another measure of variability is called the semi-interquartile range, *Q*. The semi-interquartile range is commonly used with ordinal scales, where values represent some sort of ranking or ordering of scores. As with range, the larger the value of *Q*, the more dispersed the distribution. The semi-interquartile range, *Q* is computed from the formula:

$$Q = \frac{Q_3 - Q_1}{2}$$

where

Q_3 = the third quartile, or the score where 75% of the scores fall below

Q_1 = the first quartile, or the score where 25% of the scores fall below

Q_2 = the second quartile or the median (score where 50% of scores fall below)

The difference between the range and the semi-interquartile range (*Q*) is the semi-interquartile range represents a difference between percentile points, while the range represents a difference between the high and low scores of a distribution. The difference between Q_3 and Q_1 is called the *interquartile range* (*IQR*). Figure 3.1.1 shows an illustration of the interquartile range relative to a symmetrical distribution.

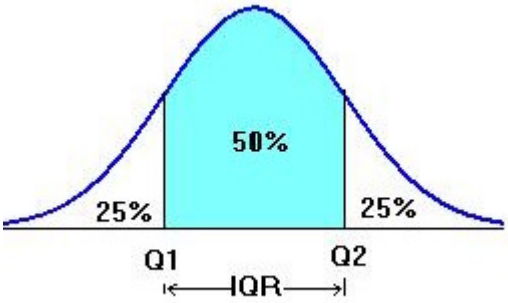

Figure 3.1.1 Illustration of Interquartile Range

When computing the semi-interquartile range, we first compute the 75 percentile, P_{75} and the 25 percentile, P_{25}. We then compute the difference and then divide by 2. Recall from chapter 2.1, we learned how to compute the percentile from a cumulative frequency distribution. Below are guides on how the compute the semi-interquartile range from ungrouped frequency distribution (or listed distribution, scores with frequency equal 1) and grouped frequency distribution. Often the only information given is the scores or group intervals and their corresponding frequencies, f_i. The formula for computing the P percentile is:

$$P_{\%} = LL_i + [(\frac{n_p - C_f}{f_i}) \cdot I]$$

where
$P_{\%}$ = any specified percentile point
LL_i = the *exact lower limit* or *real lower limit* of the group interval containing Percentile point, $P_{\%}$
n_p = number of scores or cases comprising the specified percentage for n.
$n_p = (p)(n)$, where p is decimal equivalent to % and n is total frequency
C_f = cumulative frequency up to but not including the percentile interval
f_i = frequency within the percentile interval
I = group interval size or class size

Construction Guide: Computing the Semi-Interquartile Range (Q) for Ungrouped Data

$$Q = \frac{Q_3 - Q_1}{2} = \frac{P_{75} - P_{25}}{2}$$

X_i	f_i	C_f	C_P	$CP(\%)$
15	**3**	**25**	0.12	100
14	**2**	22	0.08	88
12	**4**	20	0.16	80
10	**1**	16	0.04	64
8	**6**	15	0.24	60
6	**1**	9	0.04	36
4	**3**	8	0.12	32
2	**5**	5	0.2	20

Step 1. Compute 75 percentile, P_{75}

$$P_{75} = LL_i + [(\frac{n_p - C_f}{f_i}) \cdot I] = 11.5 + [(\frac{18.75 - 16}{4}) \cdot 1] = 12.1875$$

Step 2. Compute 25 percentile, P_{25}

$$P_{25} = LL_i + [(\frac{n_p - C_f}{f_i}) \cdot I] = 3.5 + [(\frac{6.25 - 5}{3}) \cdot 1] = 3.9167$$

Step 3. Compute Q

$$Q = \frac{Q_3 - Q_1}{2} = \frac{P_{75} - P_{25}}{2} = \frac{12.1875 - 3.9167}{2} = 4.1354$$

Note that this result will be slightly different from that calculated by just calculating the median of the lower and upper 50% of the ordered scores of the distribution for Q_1 and Q_2 respectively.

Construction Guide: Computing the Semi-Interquartile Range (*Q*) for Grouped Data

$$Q = \frac{Q_3 - Q_1}{2} = \frac{P_{75} - P_{25}}{2}$$

X_i	f_i	C_f	C_P	$CP(\%)$
75-79	2	**28**	0.07	100.00
70-74	3	26	0.11	92.86
65-69	1	23	0.04	82.14
60-64	4	22	0.14	78.57
55-59	6	18	0.21	64.29
50-54	2	12	0.07	42.86
45-49	3	10	0.11	35.71
40-44	3	7	0.11	25.00
35-39	4	4	0.14	14.29

Step 1. Compute 75 percentile, P_{75}

$$P_{75} = LL_i + [(\frac{n_p - C_f}{f_i}) \cdot I] = 59.5 + [(\frac{21 - 18}{4}) \cdot 5] = 61.75$$

Step 2. Compute 25 percentile, P_{25}

$$P_{25} = LL_i + [(\frac{n_p - C_f}{f_i}) \cdot I] = 39.5 + [(\frac{7 - 4}{3}) \cdot 5] = 44.5$$

Step 3. Compute Q

$$Q = \frac{Q_3 - Q_1}{2} = \frac{P_{75} - P_{25}}{2} = \frac{61.75 - 44.5}{2} = 8.625$$

The range and semi-interquartile range, though they provide us with a simple measure of variability, is not frequently reported or used in research studies. They do not lend themselves to further statistical manipulations and have little use in higher level statistical methods, such as inferential statistics. Instead, the variance and standard deviation measurement of variability provides insight about the deviations of the scores of a distribution form its center or mean.

Variance and Standard Deviations

Before we explore the concepts of variance and standard deviations, let us develop the concept of deviation and sums of deviations. We define deviation has the distance of a score from the mean of the distribution.

Deviation is the distance of a score from the mean of the distribution.

If the mean of the distribution is 45, then the deviation of $X = 34$ is $34 - 45 =$ ***-11.*** If the score $X = 50$, then the deviation is $50 - 45 =$ ***5***. An important property of deviation of scores is that the sum of all deviations of scores from the mean equal zero.

$\sum(X - M) = 0$ (sum of deviation = 0)

We can take the absolute values of these deviations or square them to avoid getting zero when they are summed. When we square and divide the deviations by the total number of scores we get a measure of the variability, know as the variance.

The **variance** is the mean of the square deviation of scores (the mean square deviation).

The **sum of square deviation** is denoted as *SS*:

$SS = \Sigma(X - M)^2$ (definition formula)

$$\text{population variance} = \frac{\text{sum of square deviation}}{\text{population sise}} = \frac{\Sigma(X - M)^2}{N} = \frac{SS}{N}$$

The square root of the variance is called the standard deviation, and it is often used in inferential statistics as a measure of variability.

The **standard deviation** is the square root of the variance:

standard deviation = $\sqrt{\text{variance}}$

When we calculate the variance of a population we divide the sum of square deviation by the population size, *N*. However, when we calculate the variance of a sample we divide the sum of square deviation by the sample size minus 1 or (*n – 1*). Most statistical packages compute the sample variance, using (n – 1) as the divider for the sum of square deviation. So the sample variance is given by the formula:

$$\text{sample variance} = \frac{\text{sum of square deviation}}{\text{sample sise - 1}} = \frac{\Sigma(X-M)^2}{n-1} = \frac{SS}{n-1}$$

The standard deviation is what is reported for statistical measurement of variability instead of the variance. Figure 3.1.2 illustrate how the standards deviation is used with the mean (measurement of central tendency) to describe a distribution. If the data is symmetrically distributed, then we can say that 34% of scores are between the mean and the score that is one standard deviation from the mean. This concept will be developed further in chapter 4 when we discuss the normal curve.

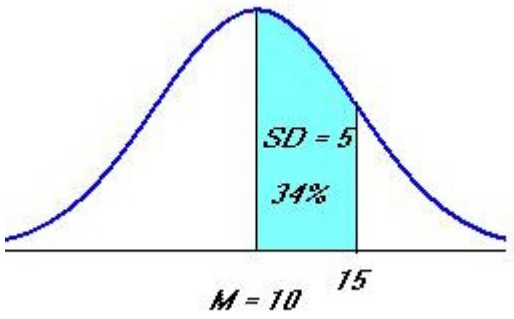

Figure 3.1.2 Graphical Illustration of Sample with Mean = 10 and Standard deviation = 5

When computing the variance of a sample we divide by $(n - 1)$; however, there are two possible formulas that we might use: the definition formula as above or the computational formula, which is often easier to use. Both of these are illustrated below.

The computational formula for the sum of square, *SS* is:

$$SS = \Sigma X^2 - \frac{(\Sigma X)^2}{n} \quad \text{(Computational formula)}$$

Construction Guide: Computing Sample Variance and Standard Deviation (definition)

$$\text{sample variance} = \frac{\text{sum of square deviation}}{\text{sample sise - 1}} = \frac{\Sigma(X - M)^2}{n - 1} = \frac{SS}{n - 1}$$

X_i	$X - M$	$(X - M)^2$
85	21.4	457.96
74	10.4	108.16
65	1.4	1.96
56	-7.6	57.76
76	12.4	153.76
45	-18.6	345.96
35	-28.6	817.96
65	1.4	1.96
55	-8.6	73.96
80	16.4	268.96
$M = 63.6$	$\sum(X - M) = \mathbf{0}$	$\sum(X - M)^2 = \mathbf{2288.4}$

Step 1. Compute the mean, *M* of *X*

$$M = \frac{\Sigma X}{n} = \frac{636}{10} = 63.6$$

Step 2. Compute deviations $(X - M)$ and square them $\boldsymbol{(X - M)^2}$.

Step 3. Sum the square deviations and divide by $n - 1$ to get the variance.

$$\text{variance} = s^2 = \frac{\Sigma(X - M)^2}{n - 1} = \frac{SS}{n - 1} = \frac{2288.4}{9} = 254.27$$

Step 4. Compute the standard deviation by taking the square root of result of step 3

$$\text{standard deviation} = \sqrt{\text{variance}} = \sqrt{254.27} = 15.95$$

Construction Guide: Computing Variance and Standard Deviation (computational)

$$\text{sample variance} = \frac{SS}{n-1} = \frac{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}{n-1}$$

X_i	X^2
85	7225
74	5476
65	4225
56	3136
76	5776
45	2025
35	1225
65	4225
55	3025
80	6400
$\sum X =$ ***636***	$\sum(X^2) =$ ***42738***
$(\sum X)^2 =$ ***404496***	

Step 1. Square each X

Step 2. Compute the following summations: $\sum X =$ ***636,*** $\sum(X^2) =$ ***42738***

Step 3. Square $(\sum X)^2 =$ ***404496***

Step 4. Use the computational formula to compute the sample variance

$$\text{sample variance} = s^2 = \frac{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}{n-1} = \frac{42738 - \frac{404496}{10}}{9} = 254.27$$

Step 5. Compute the standard deviation by taking the square root of result of step 4

$$\text{standard deviation} = \sqrt{\text{variance}} = \sqrt{254.27} = 15.95$$

The variance and standard deviations can also be computed from a frequency distribution. The formula is a modification of the variance computational formula. The following guide shows how to compute the variance from a frequency distribution.

The formula for computing the variance from a frequency distribution is:

$$\text{sample variance} = s^2 = \frac{\sum fX^2 - \frac{(\sum fX)^2}{n}}{n-1}$$

where

f is the frequency of each score or the midpoint of each grouped interval

Construction Guide: Computing Variance and Standard Deviation (Grouped Distribution)
(For ungrouped frequency distributions or listed data, use the raw score, *X* instead of the midpoint)

$$\text{sample variance} = s^2 = \frac{\sum fX^2 - \frac{(\sum fX)^2}{n}}{n-1}, \; MP = X$$

X_i	f_i	Midpoint (MP)	MP^2	f(MP)	$f(MP^2)$
75-79	2	77	5929	154	11858
70-74	3	72	5184	216	15552
65-69	1	67	4489	67	4489
60-64	4	62	3844	248	15376
55-59	6	57	3249	342	19494
50-54	2	52	2704	104	5408
45-49	3	47	2209	141	6627
40-44	3	42	1764	126	5292
35-39	4	37	1369	148	5476
30-34	3	32	1024	96	3072
				$\sum f(MP)$ = ***1642***	$\sum f(MP^2)$ = ***92644***

Step 1. Square the *midpoint*, *MP* of each interval, MP^2

Step 2. Compute the following products and: $f(MP)$ and $f(MP^2)$ sum them $\sum f(MP)$; $\sum f(MP^2)$

Step 3. Square $\sum f(MP)$ and compute the variance from formula, $(\sum f(MP)) = (1642)^2 =$ **2696164**

$$\text{sample variance} = s^2 = \frac{\sum fX^2 - \frac{(\sum fX)^2}{n}}{n-1} = \frac{92644 - \frac{2696164}{31}}{30} = 189.03$$

Step 4. Compute the standard deviation by taking the square root of the result from step 3

$$\text{standard deviation} = \sqrt{\text{variance}} = \sqrt{189.03} = 13.75$$

Factors Affecting the Variance and Standard Deviation

When we add of subtract a constant to every score of a distribution it affects the mean (central tendency) in a similar manner, but have no effect on the variance or standard deviation. For, example, given a dataset with *n* values of a variable we shall call *X* with a mean $M = 20$, variance, $s^2 = 16$ and standard deviation, $s = 4$; if 3 is added to all the scores in *X*, then the new $M = 23$. Similarly, if we subtract 5 from all *X*, then the new $M = 15$. In both these cases, the variance and standard deviations would remain the same.

When we multiply or divide all the scores or values of the variable *X* above by a constant, the mean would change correspondingly; however, the variance would change in the following manner: for *X* multiplied by a constant, the new variance would become the original variance multiplied by the *square of the constant*. For *X* divided by a constant, the new variance would become the original divided by the *square of the constant*. For example, for 3*X*, the new variance would change from 16 to $16(3^2) = 144$, while the new standard deviation would become 12 (the square root of 144). Similarly, if we divide *X* by 2, *X*/2, the new variance would become, $16/(2^2) = 4$.

Table 3.1.1 shows the SPSS output for the variability measures for the *pass9th* variable from the ODE data table. Figure 3.1.3 shows the SPSS procedure for computing variability measurements.

Table 3.1.1 Measures of Dispersion for pass9th Variable (ODE)

Statistics	**Values**
Std. Deviation	13.61
Variance	185.22
Range	72.00

N is 94

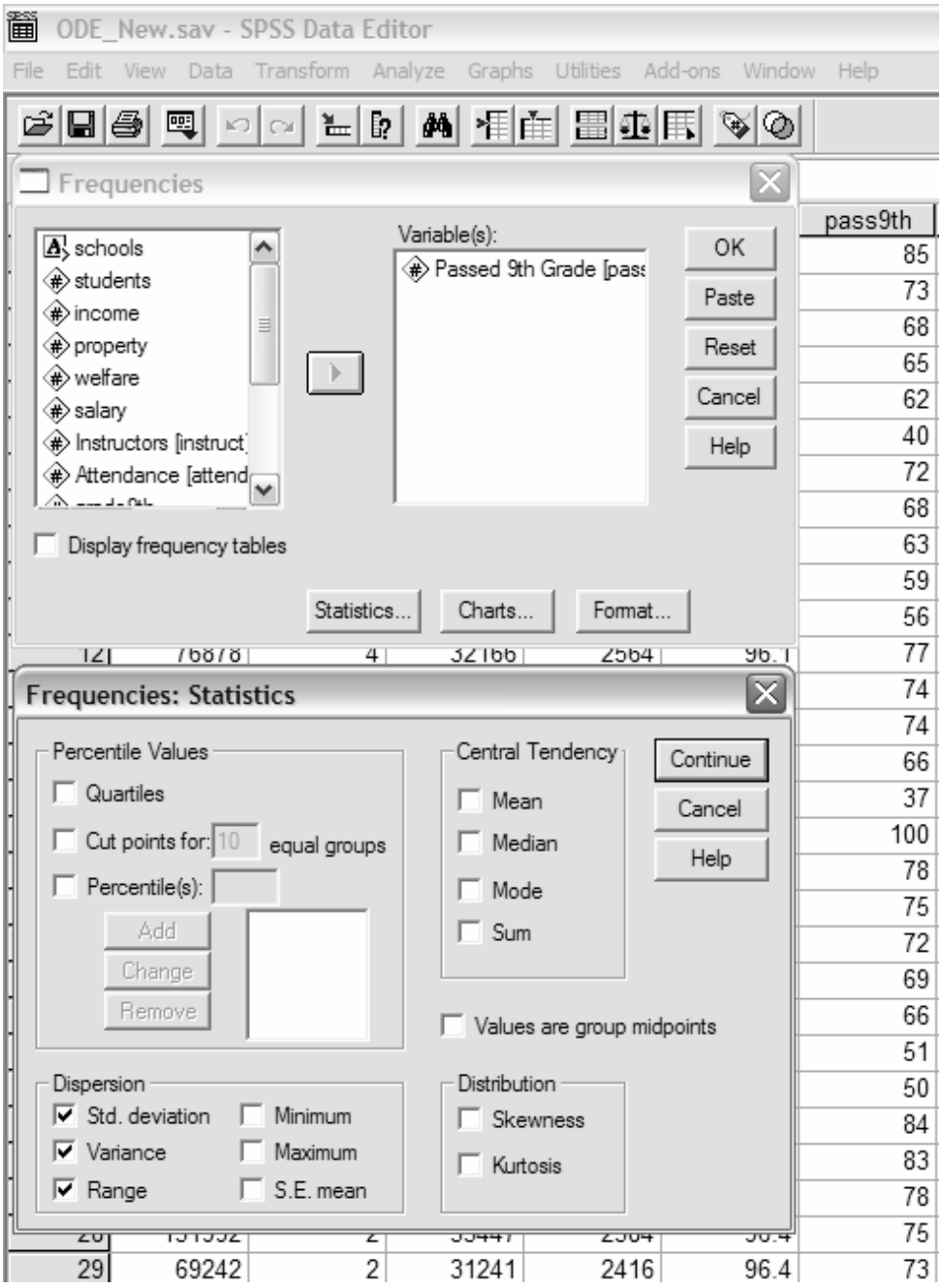

Figure 3.1. 3 SPSS Variability Procedure