Frequencies and percentages

FREQUENCIES, also known as tallies, are used to count up the number of things or people in different categories. For instance, let's say the respondents to the *Language Teaching/Learning Beliefs Questionnaire* turned out to be seventeen people and you wanted to analyze their gender. You could count up the frequencies in two categories: male and female. If it turned out that there were six males and eleven females, six and eleven would be your frequencies for males and females. Such results can be expressed as RAW FREQUENCIES as in the previous sentence, or they can be converted to PERCENTAGES. This can often be a clearer way of expressing the information. As you know, percentages are calculated by dividing the total number in one category by the total number in all categories and multiplying the result by 100. For instance, using the example above, if you want to know what percentage of the total group is represented by the males, you would divide the number in the male category (6) by the number in all categories (males + females = 6 + 11 = 17), which would be calculated as follows: (6 / 17) x 100 = .353 x 100 = 35.3% or about 35%

Percentages are easier for many people to understand than raw frequencies. Telling people that your study included 35% males and 65% females may be clearer for them than saying that you studied six males and eleven females. Other people may find the actual raw frequencies clearer than the percentages, so you may find it best to report both the raw frequencies and percentages.

Such frequency and percentage analysis could equally well be applied to the answers that participants give to the various items. For instance, in analyzing the participants' answers to item number one in the survey, '1. Some people have a special aptitude for learning foreign languages', it might turn out that nine of the participants out of seventeen selected 'strongly agree', four selected 'agree', three selected 'disagree', and one selected 'strongly disagree'. Presenting those frequencies in prose, as we just did, or in a table would be helpful in analyzing the results. It might also be useful to calculate the percentage of participants who selected each (by dividing the number who selected each by the total of seventeen and moving the decimal point two places to the right). The result for the item one example above would be as follows: 52.9% of the participants selected 'strongly agree', 23.5% selected 'agree', 17.6% selected 'disagree', and 5.9% selected 'strongly disagree'. You might have noticed that these percentages add up to 99.9% rather than 100%, a difference that you would have to attribute to what is called ROUNDING ERROR.

Central tendency

Another convenient way of summarizing data is to find a single statistic, called the central tendency, which represents an entire set of numbers. CENTRAL TENDENCY can be defined as the propensity of a set of numbers to cluster around a particular value. Three statistics are often used to find central tendency: the mean, the mode, and the median.

Mean The most widely used measure of central tendency is the MEAN, which is more commonly called the AVERAGE. The mean is the sum of all the values in a distribution divided by the number of values. The formula is written as follows:

$$M = \frac{\sum X}{N}$$

where: M = mean
$\sum$ = sum of (or add up)
X = values
N = number of values

Look at these ages for yet another group of respondents to the questionnaire:

24 26 26 27 28 29 29 29 31 32

Using the formula $M = \sum X / N$, the mean for these ages is

$(24+26+26+27+28+29+29+29+31+32)/10 = 281/10 = 28.1$

Mode The MODE is that value in a set of numbers that occurs most frequently. In a way, the mode is the simplest of the three central tendency statistics discussed here because it requires no computation. You simply need to examine the number of occurrences of each value.

Look at the respondents' ages again: 24 26 26 27 28 29 29 29 31 32

In this case, the mode is 29 because it is the most frequent age.

Note that there can also be more than one mode in a distribution. For instance, in the set of ages that we have been working with, if we were to add one additional 26-year-old, there would be two modes (26 and 29) as follows:

24 26 26 26 27 28 29 29 29 31 32

When there are two modes, the distribution is referred to as BIMODAL. If there are three modes, it is TRIMODAL, and so on.

Median The MEDIAN is the point in the distribution below which 50% of the values lie and above which 50% lie. To find the median, you should first place the age values in order from low to high. Then, examine the age above and below which 50% of the ages lie.

Dispersion

Knowing about the central tendency of a set of numbers is very helpful way of characterizing the most typical behavior in a group. It doesn't, however, tell us anything about the way the numbers spread out around that central or typical behavior. Consider the following two sets of ages for groups answering the *Language Teaching/Learning Beliefs Questionnaire*:

Group A	Group B
65	54
61	53
60	51
54	50
50	50
50	50
47	49
41	47
22	46

Low-high The LOW-HIGH involves finding the lowest value and the highest value in a set of numbers. Look at these ages: 24 26 26 27 28 29 29 29 31 32

In this case, by putting the numbers in order from high to low, you can see immediately that the lowest age is 24 and the highest age is 32. Thus, the low-high is 24 to 32, or 24–32.

Range The RANGE is the highest value minus the lowest value plus one. You must add one because otherwise you will be excluding either the highest number or lowest number from the range. The formula is written as follows:

$$\textit{Range} = H - L + 1$$

where: H = high number
L = low number

Look at these ages: 24 26 26 27 28 29 29 29 31 32

$$\textit{Range} = 32 - 24 + 1 = 9$$

Thus the range is nine including both the 24 and 32, as you can see by counting up the nine possible numbers in that range: 24 25 26 27 28 29 30 31 32

Standard deviation The best overall indicator of dispersion is the STANDARD DEVIATION. Brown (1988: 69) defined it as 'a sort of average of the differences of all scores from the mean'. The standard deviation is 'a sort of average' because you are averaging some values by adding them up and dividing by the number of values, just as you did in calculating the mean. So the equation for the standard deviation starts with adding *something* up and dividing by the number of *somethings*. In calculating the standard deviation, the *something* you are adding up is the 'differences of all scores from the mean,' or $\Sigma(X-M)$. So, the equation for the standard deviation includes $\Sigma(X-M)$ (for summing up the differences from the mean of each student divided by the number of students):

$$\frac{\Sigma(X-M)}{N}$$

Because half the students will be above the mean and half below the mean, the differences will tend to cancel each other out when you sum them up. Hence they will add up to something close to zero, which would tell you nothing. To get around this problem, statisticians eliminate the minus signs (for those students below the mean when you subtract the mean from each student's value) by *squaring the result for all students* before adding them up (rather than simply adding up the absolute numbers). So, the equation for the standard deviation includes (for summing up the distances from the mean *squared* for each student divided by the number of students):

$$\frac{\Sigma(X-M)^2}{N}$$

Because the distances of the values from the mean are squared, you need to take the square root of the result (when you are finished summing and dividing) in order to bring it back down to the original scale. This is for the sake of ease of interpretation. (Note that the squaring and square root part of the equation are why Brown (1988) began his definition with '*a sort of* average' (emphasis ours).) So, the full equation for calculating the standard deviation requires you to subtract the mean from each value and square that result for each student, then add those squared values up, divide the result by the number of students, and take the square root of the result of that division.

$$\text{SD} = \sqrt{\frac{\Sigma(X-M)^2}{N}}$$

Where: SD = standard deviation
X = values
M = mean of the values
N = number values

Let's consider an example calculation of standard deviation. Look at these ages:

24 26 26 27 28 29 29 29 31 32

From calculations earlier, you already know that the mean of these ages is 28.1.

You should begin by lining the ages up vertically (as shown in column C1) with the mean next to each one (as in column C2). Next, calculate the difference between C1 and C2 for each student by subtracting C2 from C1 and put the result in column C3. Then square each of the values in C3 and put the result in column C4, and add up all the values in C4 at the bottom of that column.

C1		C2		C3	C4
ages	–	*mean*	=	*difference*	*squared difference*
24	–	28.1	=	–4.1	16.81
26	–	28.1	=	–2.1	4.41
26	–	28.1	=	–2.1	4.41
27	–	28.1	=	–1.1	1.21
28	–	28.1	=	–0.1	.01
29	–	28.1	=	0.9	.81
29	–	28.1	=	0.9	.81
29	–	28.1	=	0.9	.81
31	–	28.1	=	2.9	8.41
32	–	28.1	=	3.9	15.21

52.90 = $\Sigma\,(X - M)^2$, or the sum of squares
10 = N, or the number of participant ages

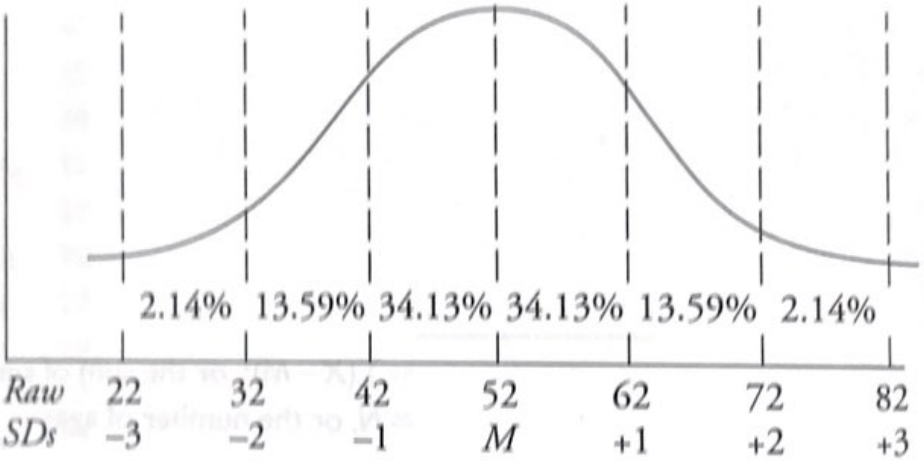
You now have everything you need to calculate the standard deviation. So, plug the sum of squares and number of ages into the equation for the standard deviation, then divide, and take the square root, and you have the standard deviation: 2.3.

$$SD = \sqrt{\frac{\Sigma(X-M)^2}{N}} = \sqrt{\frac{52.90}{10}} = \sqrt{5.29} = 2.3$$

This value indicates critical information about the spread or dispersion of the data, which we describe in more detail in the next section.

Normal distribution

In any set of values that form a NORMAL DISTRIBUTION (commonly referred to as a BELL CURVE), the standard deviation can become a useful way of describing the dispersion of the scores. Let's say we have a large set of ages that range from 22 to 82. We calculate the mean and standard deviation, and they turn out to be 52 and 10, respectively, each ten-year period being one standard deviation. In the following graph, we have shown how the mean can be marked off in the middle of the distribution, and how the 10 years of the standard deviation can be used to mark off standard deviation units across the bottom:



Notice that the mean of 52 is exactly in the middle. Notice also that one standard deviation above the mean is 62 (52 + 10 = 62); two above the mean is 72 (52 + 10 +10 = 72); and so forth. Next, notice that one standard deviation below the mean is 42 (52 – 10 = 42); two standard deviations below the mean is 32 (52 – 10 – 10 = 32); and so forth. Thus we can refer to a student as being three standard deviations below the mean, or being two standard deviations above the mean, or being right in the middle (that is, exactly on the mean).

Based on experience with large groups of data, we can expect certain percentages of students to fall within one standard above the mean (34.13%), between one standard deviation and two above the mean (13.59%), and between two standard deviations and three above the mean (2.14%). Since the distribution is symmetrical, the same percentages also apply below the mean –1, –2, and –3 standard deviations. Thus the percentage of participants who will fall within the area between one SD above and one SD below the mean is 68.26% (34.13 + 34.13 = 68.26).