Describing Data 1

Measures of Centre and Variation

This worksheet focuses on describing data through measuring its centre and variation. These measurements will give us an idea of what our data set looks like overall.

CENTRE

There are several ways to describe the **centre** of a data set: *mean*, *median*, *mode*, and *midrange.* All of these figures represent a typical value of a data set.

Mean is the technical term for what most people call an average. In statistics, "average" is too vague, so you aren't likely to see it.

To find the mean, we first take the sum of all numbers in the data set. The symbol for taking the sum of a set of numbers is the capital Greek letter sigma, **Σ**, so "Σ x" means "take the sum of all values of x". Then we divide by **n**, the number of observations in the data set. You will see the notation for *mean* represented two ways: $\bar{\mathbf{x}}$ (pronounced "x bar") is used when we are finding the average of a **sample** of data, and **μ** (pronounced "mew", the Greek letter mu) represents the mean of a **population** of data. The population is *all* the members of the group of interest (e.g., all ducks) while a sample is a smaller group, or subset, of the population (e.g., 250 of the ducks at Trout Lake).

Example 1: Find the mean of the following sample: {3, 5, 4, 9, 8, 5, 7, 8, 9, 12}

Solution: We take the sum of all numbers, and then we divide by n, which is 10:

$$\Sigma x = 3 + 5 + 4 + 9 + 8 + 5 + 7 + 8 + 9 + 12 = 70$$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{70}{10} = 7$$

The mean of our data set is 7. The formula above can be used to calculate the mean of any data set.

The **median** is the middle value of an ordered data set. If there is an odd number of observations, then the median is the middle number. If there is an even number of observations, we take the mean of the two middle numbers. We find the position of the central observation using the formula:

$$L = \frac{n + 1}{2}$$

Here the L stands for "locator", and it represents *the position within the list* where the median can be found. The median is a more useful measure of central tendency if the data is **skewed**, meaning that the data favours high numbers over low numbers, or vice versa. In graphical form, a skewed curve appears asymmetrical, with one longer tail leading off in the direction of skew. (So data that is right skewed has a long tail to the right.) Any data set that's limited on one side (for example, when negative values are impossible, but low values are likely) will exhibit skewness.

Example 2: Find the median of the data set in Example 1.

Solution: The first step is to put the data set in order from smallest to largest:

{3, 4, 5, 5, 7, 8, 8, 9, 9, 12}

Next we have to determine which position within the data set is the middle position. Our data set has 10 observations, so n = 10. The middle observation is located at number (10+1)/2 = #5.5, meaning it's halfway between #5 and #6. We must take the mean of the observations in the 5^{th} and 6^{th} positions. Those observations are 7 and 8, so the median of our data set is 7.5. (Note: the mean is *not* 5.5! That's a position number and has nothing to do with the observations themselves.)

The **mode** of a data set is the value that occurs most frequently. It is possible to have more than one mode in a data set. In the data set {3, 4, 5, 5, 7}, the number 5 occurs twice so it is the mode. In the data set {2, 4, 2, 6, 7, 7, 7, 8, 2}, both the numbers 2 and 7 occur three times each. This would be a **bimodal** data set.

Example 3: Identify the mode(s) in the data set from Exercise 1 if any exist.

Solution: There are three modes in this data set: 5, 8, and 9 (each value occurs twice). This is called a **multimodal** data set.

The **midrange** of a data set is the mean of the highest and lowest observations in the set. The midrange of that same data set would be (3 + 12) ÷ 2 = 7.5.

MEANS IN SPECIAL CIRCUMSTANCES

Sometimes we don't get to see the raw data (the individual observations), and instead we see a summary of the data. Without all the original raw data, it's impossible to know what the actual mean is, but we can estimate it.

The frequency distribution shows a sample of observations of people's ages, but sorted into bins as though we were going to make a histogram from the data. We know three of the subjects are under 10, but we don't know their precise ages, and so on. We can estimate the mean of this data set by assuming all the members of each bin have the central age for the bin, and we do that by calculating the midrange.

Age	*Frequency*
0–9	3
10–19	6
20–29	5
30–39	9
40–49	7

Example 4: Estimate the mean for the data in the table at the right.

Solution: The midrange of the first bin is (0 + 9) ÷ 2 = 4.5. We can get the other midranges quickly by noticing the bins are 10 apart, so the other midranges should be 10 apart: 14.5, 24.5, 34.5, 44.5. So, our best guess as to what the original data looks like is:

{4.5, 4.5, 4.5, 14.5,…[6 times], 24.5,…[5 times], 34.5,…[9 times], 44.5,…[7 times]}

We can now take the mean as normal. There's even a short cut, since our data consists of a lot of repeated entries. We can find Σ x by multiplying each midrange by its frequency, and we can find n by adding the frequencies:

$$\bar{x} = \frac{\Sigma (f \cdot x)}{\Sigma f} = \frac{(3 \cdot 4.5) + (6 \cdot 14.5) + (5 \cdot 24.5) + (9 \cdot 34.5) + (7 \cdot 44.5)}{3 + 6 + 5 + 9 + 7} = \frac{845}{30} = 28.2$$

We may also need to calculate a **weighted mean**, a mean of a data set where various observations might be assigned a **weight** or point value. The American "grade point average" is a common example of this. In this system, a grade of A is awarded 4 points, B is 3 points, and so on down to an F being 0 points. Each course has a weight of a certain number of credits, and a student's GPA is found by adding the points from each course times that course's weight, and dividing by the total of all the weights.

Example 5: Find the GPA of a student with the following grades: B (4 cr), C (4 cr), A (3 cr), B (3 cr), D (2 cr).

Solution: The total number of credits is 4 + 4 + 3 + 3 + 2 = 16 credits. The total points are (3 × 4) + (2 × 4) + (4 × 3) + (3 × 3) + (2 × 2) = 45. The student's GPA is 45 ÷ 16 = 2.8.

VARIATION

Measures of centre are only half the story. It's also useful to know how far away from the centre an observation can be and still be a reasonable observation. **Variation** (or sometimes spread or variability) tells us how different observations are likely to be.

The **range** is the difference between the highest (maximum) and lowest (minimum) value in the data set, but it doesn't tell us much. If a data set has 50 observations, of which one is 20, another is 100, and the rest are all 60, the range is 100 − 20 = 80, but that's not a true measure of how much the data varies.

The most common measure of variation of a data set is **standard deviation**. It reflects the deviations, or differences, of all values in the data set from the mean. A larger standard deviation would indicate greater variation for a data set.

If you calculated the mean mark on a class midterm to be 65, that only tells you the typical mark. Did the marks in the class look like {66, 64, 67, 66, 62, 70, …} or like {48, 97, 83, 57, 62, 81, …}? The first set of marks has low standard deviation—most of the marks are quite close to the mean. The second set has a higher standard deviation as there is a greater spread of values from the mean. The notation for standard deviation of a <u>population</u> is σ (sigma again, but this is the lower case form of the Greek letter). The notation for standard deviation of a <u>sample</u> is s.

Many modern calculators will let you input a data set and calculate standard deviation for you — *this is worth it!* The calculation is long and tedious, and just wastes time on a test. Become familiar with the Stats functions of your calculator if it has them! Be careful, though: calculators can find either the sample or the population standard deviation. Be sure to know how your calculator differentiates between them.

To calculate standard deviation we use the following formulas:

Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{n}}$$

Sample standard deviation

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\frac{n(\sum x^2) - (\sum x)^2}{n(n-1)}}$$

The rightmost formula for sample standard deviation is the easiest one to use for calculating s by hand.

Variance is another related measure of variation that is simply the square of the standard deviation (σ^2 or s^2). Variance is less useful as a measure of variation, since its scale often doesn't match the spread of the data. (Data that varies by no more than 10 might have a variance of 20. If so, the 20 is meaningless in the context of the problem. The standard deviation, on the other hand, would be about $\sqrt{20} \approx 4.5$, which would mean you might expect data to vary from the mean by 4.5 on average.) Variation will turn out to have its uses, which you'll see later in the course.

Example 6: Calculate the standard deviation of the data set from Example 1.

Solution: We know n = 10 from Example 1. We also know $\Sigma x = 70$ from Example 1. The only term left to figure out is $\Sigma(x^2)$. $\Sigma(x^2)$ is the sum of the square of all data values:

$$\Sigma(x^2) = 3^2 + 5^2 + 4^2 + 9^2 + 8^2 + 5^2 + 7^2 + 8^2 + 9^2 + 12^2 = 558$$

Now we plug into the formula:

$$s = \sqrt{\frac{10 \cdot 558 - 70^2}{10 \cdot (10 - 1)}} = \sqrt{7.555\ldots} \approx 2.7$$

If the actual standard deviation is not necessary, we can estimate it by using the **range rule of thumb**: dividing the range by 4 will often provide a decent estimate of s. For the data set that we've been using through this worksheet, this estimate would be $(12 - 3) \div 4 = 2.25$, which isn't a bad estimate.

EXERCISES

A. For the following sets of data, calculate (a) sample mean, (b) median, (c) mode, (d) midrange, (e) range, (f) standard deviation, (g) estimate of s.d. using the range rule of thumb. Round to one extra decimal place than the source data.

1) {8, 24, 9, 6, 10, 18, 7, 14, 16, 21, 13, 24}

2) {3, 6, 5, 4, 6, 5, 9, 10, 11, 7, 9}

3) {41, 39, 38, 42, 43, 39, 40, 43, 26, 42, 42, 41, 41, 42, 27, 55, 60}

B. Estimate the mean mass of the observations in this table of processed data:

Mass (kg)	50–74	75–99	100–124	125–149	150–174
Frequency	13	25	29	22	11

C. Calculate the GPA of a student who earned these grades: A (4 cr), C (4 cr), F (3 cr), C (3 cr), B (3 cr).

SOLUTIONS

A. 1) (a) 14.2 (b) 13.5 (position #6.5) (c) 24 (d) 15 (e) 18 (f) 6.5 (g) 3.8
2) (a) 6.8 (b) 6 (position #6) (c) 5, 6 and 9 (d) 7 (e) 8 (f) 2.6 (g) 2
3) (a) 41.2 (b) 41 (position #9) (c) 42 (d) 41 (e) 34 (f) 7.9 (g) 8.5

B. Midranges: 62, 87, 112, 137, 162. n = 100. x-bar = 110.3

C. 17 credits total, 16 + 8 + 0 + 6 + 9 = points, GPA = 2.3