MEASURES OF DISPERSION

In addition to knowing the measures of central tendency, it is often important to know how widely dispersed certain measurements are. For instance, San Diego ($64.2^\circ F$) and Las Vegas ($67.1^\circ F$) have roughly the same average temperatures. However, the difference between the highest and lowest average monthly temperatures in San Diego is $13.6^\circ F$ ($71.0^\circ - 57.4^\circ$), but the difference for Las Vegas is $45.6^\circ F$ ($91.1^\circ - 45.5^\circ$). Even with similar average temperatures, you need to dress differently in the cities.

There are various types of measures of dispersion, just as there are different measures of central tendency. We will only concern ourselves with three:

RANGE –

The range for a set of data is merely the highest value minus the lowest. An example of this would be the all-time temperature range in Farmington, New Mexico. Since weather records have been kept in Farmington, the all-time high has been $115^\circ F$, and the all-time low has been $-23^\circ$ for a range of $138^\circ F$. Note that we used the range to compare San Diego and Las Vegas dispersion in temperatures.

AVERAGE DEVIATION –

The average deviation of a set of numbers is merely the mean distance from the mean of the data set (huh?). In other words, the average deviation is a measure of how far the typical entry is away from the mean. To find the average deviation, the first thing to do is to find the mean of the set of numbers. Then the distance of each number from the mean ($\mu$) is found. Their absolute values are added. This sum is then divided by the number of data points. Mathematically, the formula is:

$$\text{Average deviation} = \frac{\sum |x_i - \mu|}{n}$$

The reason for the absolute values being required is that if all of the deviations were added directly, the sum of the negatives would equal the sum of the positives, and the average deviation would be zero.
STANDARD DEVIATION –

One problem with the range is that sometimes there may be one number which is totally out of whack. For instance, it would be possible, if the temperatures in Farmington were written by hand, for someone to accidentally put a negative sign before a temperature, or possibly a cataclysmic event might happen such as a forest fire heating the temperature gauge up to 150° as the building next door burns down. Of course, it is also possible for the one data entry to legitimately be outlandishly high or low. We call this type of number an “outlier”.

Also, one data set might show almost all data points right near the “average”, with just a few towards the high or low end, while another data set with the same mean and range might have a tendency for the data points to be near one end with few or none in the center. An example of this difference might be the grades in two classes, both of which have almost an exact “C” (2.0) average. In the first class, all the students seemed to do about the same, and all but a few got a “C”, with someone getting an “A” and someone an “F”. In the other class, half the class quit and received an “F”, but almost everyone who finished the class got an “A”. These are obviously very different classes, even though the “average” and the range were the same.

Average deviation assumes that if one value is twice as far from the mean as another, it’s deviation is twice as important. In actuality, twice the distance from the mean might be much more than twice as important. For instance, assume you travel to a city with an expected temperature of 70°F. If the temperature is within 10 degrees of that, you may be comfortable if you assumed 70°F. However, if the temperature gets 30°F away, most of us would probably say it’s more than 3 times as bad that we didn’t pack warmer or cooler clothes.

Statisticians have come up with a measure of dispersion which takes account of how close most of the observations are to the average (mean), still overweighting the ones far away. This measure of dispersion is called the standard deviation (σ), and it’s formula is:  

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

where $\sigma^2$ is the variance, and is defined by:  

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

Note that in the standard deviation, the absolute value signs are not required: the squaring of the terms makes absolute value signs redundant.
EXAMPLE

Let’s assume we have a population with ten data points:

2, 2, 2, 4, 4, 5, 6, 8, 8, 9

We will find all the measures of central tendency and the measures of dispersion for these numbers:

MEASURES OF CENTRAL TENDENCY

The **Mean** is \((2 + 2 + 2 + 4 + 4 + 5 + 6 + 8 + 8 + 9) / 10 = 5\)

The **Median** is half-way between 4 and 5 (the two middle values) = 4.5

The **Mid-range** is the average of 2 and 9 (the highest and lowest) = 5.5

The **Mode** occurs the most and = 2

MEASURES OF DISPERSION

The **Range** is the largest minus the smallest = 9 – 2 = 7

**Average deviation** =

\[
\frac{|2-5|+|2-5|+|2-5|+|4-5|+|4-5|+|5-5|+|6-5|+|8-5|+|8-5|+|9-5|}{10} = 22 / 10 = 2.2
\]

**Variance** = \(\frac{(2-5)^2+(2-5)^2+(2-5)^2+(4-5)^2+(4-5)^2+(5-5)^2+(6-5)^2+(8-5)^2+(8-5)^2+(9-5)^2}{10}\) / 10 = 64 / 10 = 6.4

**Standard deviation** = \(\sqrt{6.4} \approx 2.53\)

Note: Technically, there are two different types of standard deviation (sample and population), which are slightly different. For the purposes of this course, we will assume we are using the population standard deviation. Standard deviation has many good mathematical properties which are not discussed in these notes.
NORMAL DISTRIBUTION

When observations are made for various studies, such as weights of schnauzers or test scores on a national test, they tend to follow a “normal distribution.” This is an extremely important concept in Statistics.

In a normal distribution, most of the observations tend to congregate within a relatively short distance from the mean, with very few a much longer distance away. In a theoretical normal distribution, the mean should be the same as the median and the mode. In addition, you can expect the same number to be about two units above the mean as two units below the mean, or as many .5 units above as .5 units below the mean. The standard deviation becomes very important in this distribution. In theory, about 68% of the observations should be within one standard deviation of the mean, about 95% within two standard deviations, and about 97.7% with 3 standard deviations.

Notice on the 10 data point example just given: the mean, median, and mode are not equal. That is because these numbers were manipulated to make the Arithmetic easy. Therefore, this is not a normal distribution. Of course, with such a small sample, these could have been randomly selected from an approximately normal distribution.

Interestingly, the Central Limit Theorem states that no matter what distribution we have, if we randomly pick equal-sized samples from that distribution, the samples themselves will tend to be normally distributed when compared to other samples, tending more towards the normal distribution as the size of the samples get larger. However, for many of the projects normally dealt with, the data itself will tend to be normally distributed, and this is especially true in the fields of biology, psychology and sociology.

STANDARD NORMAL DISTRIBUTION

To find out how many standard deviations an observation is from the mean, take the difference between the number of units in the observation and the number of units in the mean, and divide that by the standard deviation: i.e. \( \frac{(x_i - \mu)}{\sigma} \). That will give us the “z-score”, which will in turn give us an indication of how far the observation is from the mean compared to others. We can use this z-score to determine the percentage of the population between two points, or what points will give a certain percentage (also known as a probability). Note the bell-curve shape of the graph, extending infinitely in both directions.
### Standard Normal Distribution

![Standard Normal Distribution Graph]

<table>
<thead>
<tr>
<th>Z - Score</th>
<th>% of Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>3.98</td>
</tr>
<tr>
<td>0.2</td>
<td>7.93</td>
</tr>
<tr>
<td>0.3</td>
<td>11.79</td>
</tr>
<tr>
<td>0.4</td>
<td>15.54</td>
</tr>
<tr>
<td>0.5</td>
<td>19.15</td>
</tr>
<tr>
<td>0.6</td>
<td>22.57</td>
</tr>
<tr>
<td>0.7</td>
<td>25.8</td>
</tr>
<tr>
<td>0.8</td>
<td>28.81</td>
</tr>
<tr>
<td>0.9</td>
<td>31.59</td>
</tr>
<tr>
<td>1.0</td>
<td>34.13</td>
</tr>
<tr>
<td>1.1</td>
<td>36.43</td>
</tr>
<tr>
<td>1.2</td>
<td>38.49</td>
</tr>
<tr>
<td>1.3</td>
<td>40.32</td>
</tr>
<tr>
<td>1.4</td>
<td>41.92</td>
</tr>
<tr>
<td>1.5</td>
<td>43.32</td>
</tr>
</tbody>
</table>
EXAMPLES  (For all of these, assume a normal distribution) Note in doing these we are assuming all values are possible, and not just integers. The actual probability that a particular number comes up approaches zero in theory, but that’s not true in practical terms. Therefore, the answers we get are approximate. It is strongly suggested that you sketch a normal curve for each problem, filling in the areas affected.

1) On a certain IQ test, the mean is 100 with a standard deviation of 15. Find the percentage of those who took the test with an IQ above 120.
   The Z-score is \((120 - 100) / 15 = 1.333\) Looking at the Z-score of 1.3 equating to 40.3% and a Z-score of 1.4 equating to 41.9%, the percentage should be VERY ROUGHLY 1/3 of the way from 40.3% to 41.9%, or about 40.8%. Looking at the diagram for the standard normal distribution, 50% are below 100, and 40.8% are between 100 and 120. Therefore, about 90.8% are 120 or less, which leaves about 9.2% above 120.

2) Let’s say a certain brand of light-bulb is supposed to last 1,000 hours with a standard deviation of 200 hours.
   a) Find the probability that the bulb will last more than 700 hours.
      The Z-score is \((700 - 1,000) / 200 = -1.5\) The Z-score of 1.5 equates to 43.32 percent. However, since the Z-score is -1.5, approximately 43.32% of the bulbs should last between 700 and 1,000 hours, which means that approximately 93.32% of the bulbs should last longer than 700 hours.
   b) Find the probability that the bulb will last less than 700 hours.
      The probability that the bulb will last less than 700 hours is 100% minus the probability that it will last more than 700 hours, or about 100% - 93.32% = 6.68% .
   c) Find the probability that the bulb will last less than 1300 hours.
      This is the converse of a), with a Z-score \((1,000 + 300) / 200 = 1.5\) . That is 43.32% added to 50% = 93.32%. Note the symmetry with part a).
   d) Find the probability that the bulb will last more than 1300 hours.
      From part c), 100% - 93.32% = 6.68% . Again, note how parts c) and d) are just the converse parts of a) and b).
3) With a distribution having a mean of 14 and a standard deviation of 2, find the probabilities that:

a) The observation is between 12 and 14
   12 is 1 standard deviation (\( \sigma \)) below the mean, so since a Z-score of 1.0 has a probability of 34.13 %, approximately 34.13 % of the observations are between 12 and the mean, and the likelihood of any particular observation being between 12 and 14 = 34.13 %.
   (Obviously a Z-score of the mean = 0.0, and represents 0% of the observations.)

b) The observation is less than 12
   With 34.13 % of the observations being between 12 and the mean, with 50% being less than the mean, the probability of a particular observation being less than 12 = 50.00 % - 34.13 % = 15.87 %.

c) The observation is less than 9.5
   9.5 is 2.25 \( \sigma \) below the mean. A Z-score of 2.25 should have a percentage approximately half-way between the probabilities for 2.2 and 2.3 (i.e half-way between 48.61 % and 48.93 %, or 48.77 %.) Therefore, approximately 50.00 % - 48.77 % = 1.13 % of the observations should be below 9.5. Note: A Z-score of 2.2 should give a probability approximately two-tenths of the way between 48.61 % and 48.93 %, or 48.67 %.

d) The observation is greater than 20
   20 is 3 \( \sigma \) above the mean for a probability of 49.87 %. Therefore, the probability that the observation is greater than 20 is about 50.00 % - 49.87 % = .13 % (i.e. very small)

e) The observation is greater than 25
   Since the probability is so small that an observation is more than 3.1 standard deviations from the mean, we assume it is zero.

f) The observation is less than 25
   Since the total probability = 100 % and the chance of being more than 25 =0 %, the chance of being less than 25 must be about 100%.

g) The observation is below 17
   17 is 1.5 \( \sigma \) above the mean, so the probability of being less than 17 is approximately 50 % + 43.32 % = 93.32 % (note that the probability of being less than the mean is 50 %).
4) With a distribution having a mean of 100 and a standard deviation of 10, find the probabilities that:

a) The observation equals 104
   The probability of any specific number theoretically = zero, since there are an infinite number of possibilities right around it.

b) An observation is between 80 and 105
   80 is 2 standard deviations (σ) below the mean, and 105 is ½ σ above, so this is a two-part problem. With a z-score of 2, the probability of being between 80 and 100 is about 47.72 %. The chance of being between 100 and 105 (1/2 σ ) is about 19.15 %. Therefore the probability of being between 80 and 105 is about 47.42 % + 19.15 % = 66.57 %.

c) The observation is less than 80 or more than 105
   Looking up at part b), this must be the rest of the curve, or about 100 % - 66.57 % = 33.43 %.

d) The observation is within 1 of the mean
   The probability for a Z-score of .1 (because that’s 1/σ ) is about 3.98 %. ∴ the probability of being less than .1 σ from the mean is about 2 (3.98 % ) = 7.96 %

e) The observation is more than 1 away from the mean
   From part d), the probability should be 100 % - 7.96 % = 92.04 %
5) A particular distribution has a mean of 18 and a variance of 9.

a) 75% of the observations are below what number?
   First of all, if the variance = 9, the standard deviation (σ) = 3.
   Looking at the graph, we note that we need 25% above the mean.
   That is between $Z = .6 \ (22.57\%)$ and $Z = .7 \ (25.80\%)$ standard deviations above the mean. To be somewhat more accurate, the difference between 25.80% and 22.57% = 3.23%. We need to be at 25%, which is 2.43% above $Z = .6$. Therefore, we need to go about $(2.43 / 3.23)$ of the way from .6 to .7, which gives about $Z = .675$. Therefore, we are about .675 standard deviations above the mean, or $18 + .675 \ (3) = 20.025$. Therefore, about 75% of the observations should be below 20.025, and I’d round this to the nearest tenth, or 20.0. Note this is approximate.

b) 75% of the observations are above what number?
   Since 75% are below 20.0 (that’s $18 + 2.0$), I would expect 75% to be above 16.0 (that’s $18 - 2.0$). Of course, from part a), we should note that 25% should be above 20.0 and 25% should be below 16.0.

c) 5% of the observations are above what number?
   Looking at our bell-curve and the table again, we are looking for the Z-score for 45%. Note the Z-score should be between 1.6 (that’s 44.52%) and 1.7 (that’s 45.54%). Similar to part a), 45.54% - 44.52% = 1.02%, and we should be about $.48 / 1.02 = .47$ above 1.6. Therefore, our Z-score should be 1.647, and 5% of the observations should be above $18 + 1.647 \ (3) = 22.94$, or about 22.9.

d) 5% of the observations are below what number?
   5% of the observations should be below $18 - 1.647 \ (3) = 13.06$, or about 13.1. Note that this means about 90% of the observations should be between 13.1 and 22.9.

e) 50% of the observations are below what number?
   50% of the observations should be below the mean (18).
EXERCISES

1) Find the mean, median, mode, mid-range, range, average deviation and standard deviation for the following observations:
   1, 3, 3, 5, 7, 9, 10, 11, 12, 14

2) Find the mean, median, mode, mid-range, range, average deviation and standard deviation for the following observations: Does any appear to be an outlier?
   2, 2, 2, 2, 2, 4, 4, 4, 4, 4, 25

3) Given a normal distribution with a mean of 17’ and a standard deviation of 2’,
   a) Find the median and mode
   b) Find the probability that a particular observation is equal to 15’
   c) Find the probability that a particular observation is less than 15’
   d) Find the probability that a particular observation is more than 15’
   e) Find the probability that a particular observation between 15’ and 19’
   f) Find the percentage of observations between 15’ and 19’
   g) Find the probability that a particular observation is more than 4’ from the mean
   h) Find the probability that a particular observation is between 11’ and 16’
   i) Find the probability that a particular observation is within 1’ of the mean
   j) Find the probability that a particular observation is between 14’ and 18’
   k) Find the probability that a particular observation is not between 12’ and 20’
   l) Find the probability that a particular observation is below 10
   m) Find what measurement below which 10% of the observations fall
   n) Find what measurement above which 90% of the observations fall
   o) Find what measurement below which 90% of the observations fall
   p) Find what measurement below which 50% of the observations fall
   q) Find what measurement below which 80% of the observations fall
ANSWERS TO EXERCISES

1) Mean: 7.5, median: 8, mode: 3, mid-range: 7.5, range: 13, average deviation: 3.7, standard deviation: 4.15

The observation is definitely an outlier.

3)
   a) Median: 17’, mode: 17’
   b) Zero
   c) 15.87 %
   d) 84.13 %
   e) 68.26 %
   f) 68.26 %
   g) 4.56 %
   h) 30.72 %
   i) 38.30 %
   j) 62.47 %
   k) 7.3 %
   l) Zero
   m) About 14.43
   n) About 14.43
   o) About 19.57
   p) 17
   q) About 18.69