Basic Statistics Review – Part Two

Sampling Distribution of the Mean; Standard Error
(See Zar 4th ed. pages 65-76; or Zar 5th ed. pages 66-72; 87-91)

In our discussion of frequency distributions in Part One, we were discussing the frequency distribution of values of data. However, we are also concerned about frequency distributions of the values of statistics. [Note: The distinction between data and statistics is very important here.]

For example, suppose every person in this class went out to the quad, stopped 30 random students, and asked how tall they were. Now, if there are 24 people taking this class, and they each now had a sample of n=30 data points, that's 24x30 = 720 data points. We could prepare a frequency distribution of the values of the 720 data points. It would have to be frequencies of data intervals, i.e. how many students were between 5'0" and 5'2"; how many between 5'2" and 5'4"; and so on – but it's still a frequency distribution of the data.

But we could also have each person in the class calculate the mean of their data points. If there are 24 people in class, we now have 24 means. We could prepare a frequency distribution of the values of the mean, i.e. how often did different intervals of the mean occur. Since the mean is a statistic, this would be a **frequency distribution of the values of a sample statistic**. Since "frequency distribution of the values of a sample statistic" is a lot of words, we have a shorter term: **sampling distribution**.

A sampling distribution is a frequency distribution of the values of a sample statistic.

Next, we could calculate the standard deviation of our 24 means. We would use the formula from above, except that our mean would be the "mean of the means", and each X_i value would be one of the 24 means. We would symbolize the standard deviation of the means as: $s_{\bar{X}}$. Notice that the subscript tells us that we are describing the standard deviation of a set of means. We can write our formula as:

$$s_{\bar{X}} = \sqrt{\frac{\sum(\bar{X}_i - \bar{\bar{X}})^2}{n-1}}$$ Note that n=24.

Since the mean is a statistic, we have calculated the standard deviation of the values of a statistic.

The standard deviation of the values of a statistic is called a **standard error.**

In the formula above, we have calculated the standard deviation of the values of the mean, so $s_{\bar{X}}$ is called the **standard error of the mean**.

In our example, we were able to calculate the standard error of the mean because we had 24 samples, and therefore 24 means. But what if we had only one sample, and therefore only one mean? You couldn't calculate the standard error of the mean with the above formula, because you divide by n-1. If n is 1, then you'd be dividing by zero – which is not allowed!

However, you can estimate the standard error of the mean when you have only one sample of data. The formula is:

$s_{\bar{X}} = \frac{s}{\sqrt{n}}$ where s is the standard deviation of your data, and n is the sample size.

This formula comes from an important mathematical theorem called the **Central Limit Theorem**, which will be discussed in lecture.

Normal Distribution; Proportions of the Normal; Z Scores
(See Zar 4th ed. pages 76-79; or Zar 5th ed. pages 72-74)

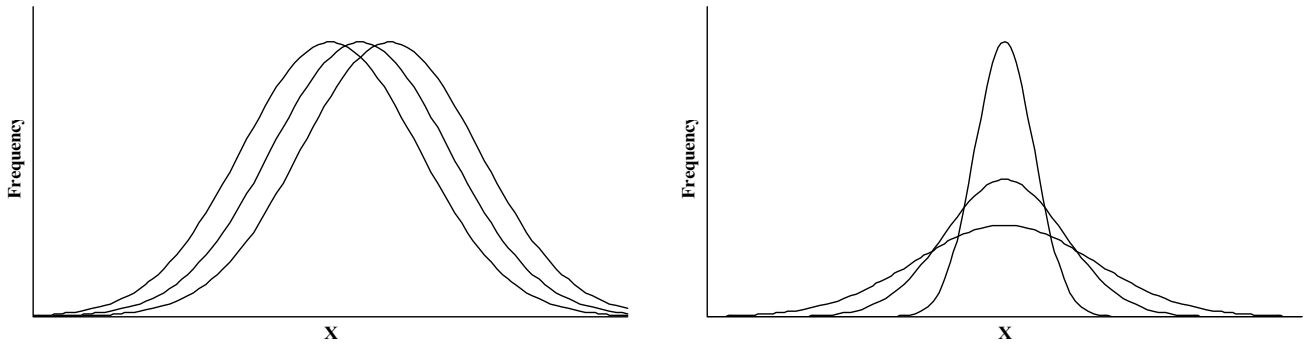
A very important frequency distribution in biology, science, and life in general is a frequency distribution called the **Normal Distribution**, which is also called the **Gaussian Distribution**. The normal distribution is a **symmetrical, bell-shaped curve**, described by a very specific mathematical equation:

$$Y_i = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(\frac{-(X_i-\mu)^2}{2\sigma^2}\right)}$$

This equation is a type of function known as a **probability density function**. A density function allows mathematicians to determine probabilities. Do you recall from Part One why frequency distributions were important? Because they allow us to determine probabilities!

Don't attempt to memorize this equation, or worry where it comes from – what you really need to see is that in order to specify a unique normal distribution, you must have a value for the mean (μ) and the standard deviation (σ). Thus, there isn't just one normal distribution; there is an infinite number of normal distributions – all with different means and/or standard deviations. Mathematicians call this a "family" of distributions.

Below on the left you can see three normal distributions with different means, but the same standard deviation. On the right are three normal distributions with the same mean, but different standard deviations.



Many biological variables are normally distributed. Therefore, the normal distribution is very important in biology, as well as in the other natural sciences.

Properties of the Normal Distribution

The shape is completely determined by only two parameters: The mean (μ) and the standard deviation (σ).

The curve is symmetrical about the mean and the mean = the median = the mode. That is, in a normal distribution, the mean, median, and mode are all the same value.

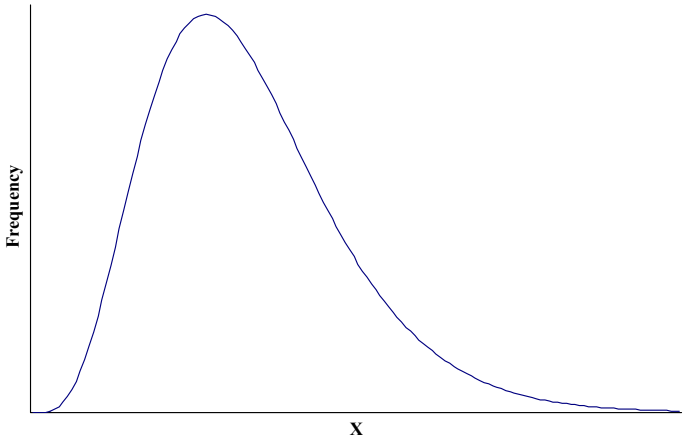
68.27% of the observations (measurements) are within 1 standard deviation of the mean.
95.44% of the observations are within 2 standard deviations of the mean.
99.73% of the observations are within 3 standard deviations of the mean.
95% of the observations are within 1.96 standard deviations of the mean.

Departures from the Normal Distribution

Common departures are **skewness** and **kurtosis**.

Skewness

This refers to the symmetry of the distribution. If the distribution is asymmetrical, with the tail on the right "pulled out", the distribution is "skewed to the right".

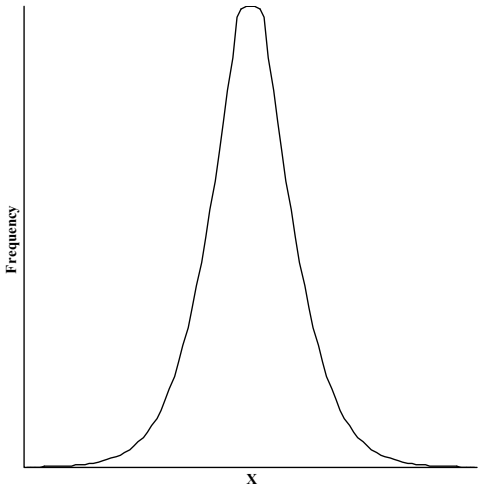


The distribution in this picture is skewed to the right. (Note: the distribution pictured is another important distribution in science and mathematics, the **Gamma** distribution.)

If the left tail is "pulled out", then the distribution is "skewed to the left" (not pictured).

Kurtosis

Kurtosis can be characterized as "peakedness". If the distribution is not normal because the peak is too high, and both tails are "pulled out", that's called **leptokurtic**.



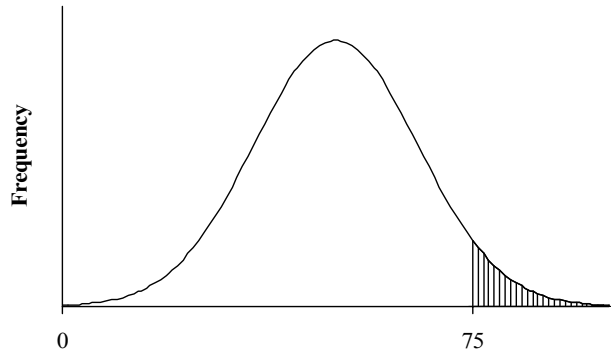
The distribution in this picture is **leptokurtic**. (Note: an important distribution that differs from the normal distribution by being leptokurtic is the **Student's *t* distribution**. We will encounter the *t* distribution several times in this course.)

A distribution that is not normal because the peak is too low, and the tails pulled in, is called **platykurtic** (not pictured).

Standard Scores and the Standard Normal Distribution

Because we have the exact mathematical equation describing the normal distribution, we can determine probabilities (or "proportions of the normal distribution") rather easily. What we need is a value for X, the mean and standard deviation for the normal distribution, and a table of "Proportions of the Normal Curve" (see Zar 4th ed., Table B.2, page App 17; or Zar 5th ed., Table B.2, page 676).

In Part One, we saw this graph, and asked what is the probability of finding a value greater than or equal to 75? This distribution is normal, has a mean of 50, and a standard deviation of 15. For this example, we are going to ask: "What is the probability of finding a value greater than 75?" We've changed the question slightly just to be consistent with the format of the table in Zar – it is not a matter of great importance.



Since there is an infinite family of normal distributions, we need a way to find the probability without having to have a different table for each possible normal distribution. That would require an infinite number of tables, and that would be a very large and expensive book!

So what we do is convert our data point (75) to what's called a **Z Score** or a **Standard Score**. To convert a data point (X_i) to a Z score (Z_i), use this formula:

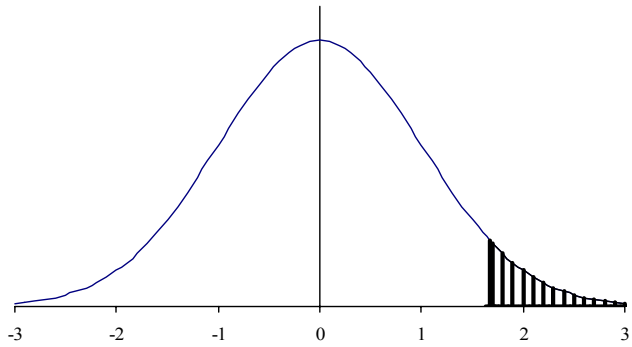
$$Z_i = \frac{X_i - \overline{X}}{s}$$

One reason this is called a "Standard Score" is because it has no units. The units "cancel out" in the calculation. Recall that the standard deviation is in the original units of the data, so both the numerator and denominator of the formula have the same units.

In our example, $X_i = 75$, the mean = 50, and the standard deviation = 15. Our Z_i is:

$$Z_i = \frac{X_i - \overline{X}}{s} = \frac{75 - 50}{15} = \frac{25}{15} = 1.67$$

The graph at left now depicts the standard normal distribution of our example. The shaded area is the area greater than 1.67, which is the Z score for our value of 75.



To determine the probability of obtaining a Z score greater than 1.67, we use Table B.2 (4th ed. page App 17; or 5th ed. page 676) in Zar. This table, as are normal distribution tables in all books, is based on the **Standard Normal Distribution**, that is, a normal distribution of Z scores.

To understand how to use and interpret Table B.2 in Zar, carefully read the instructions at the top of the table. First go down the column in the table labeled "Z" to 1.6, and then follow this row over to the column under "7".

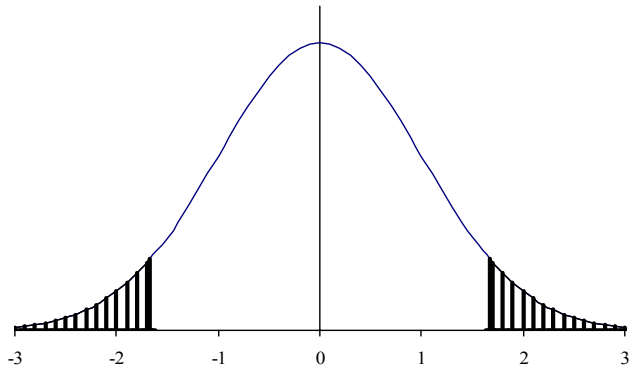
This is the proportion of the normal distribution that is more extreme than an absolute value of Z = 1.67. The table value is 0.0475. The proportion of the normal distribution greater than 1.67 (and therefore greater than 75) is 0.0475.

"More extreme than…" means "going away from zero". So the proportion greater than 1.67 is 0.0475; but this also means that the proportion less than -1.67 is also 0.0475.

If the proportion less than -1.67 is 0.0475; and the proportion greater than 1.67 is 0.0475; then what is the proportion from -1.67 to 1.67? This is easy if you remember that the total proportion always has to be one (1). The answer is

$$1 - (0.0475+0.0475) = 1 - 0.095 = 0.905$$

The graph below may help you understand these calculations:



The shaded portion in each "tail" is 0.0475, so the two tails together are 0.0475 + 0.0475 = 0.095.

The area between the two tails (not shaded) is 1 – 0.095 = 0.905.

Finally, there's one additional property of Z scores that is important. If you convert a set of data to their Z scores (using the mean and standard deviation of the data), and then calculate the mean and standard deviation of the Z scores, you will find that the mean of the Z scores is always 0, and the standard deviation of the Z scores is always 1. Consider this example, done on 10 data points:

	X_i	Z_i
	20	1.33
	12	-0.72
	17	0.56
	12	-0.72
	10	-1.23
	20	1.33
	19	1.08
	11	-0.97
	14	-0.21
	13	-0.46
n	10	10
Mean	14.8	0
Standard Deviation	3.9	1

Z scores are sometimes referred to as "standardizing to zero mean and unit variance" – which simply means that any set of Z scores has a mean of 0, and a standard deviation (and therefore a variance) of 1. You might go back to the graph on the previous page to see what this looks like.

If you do this example on your calculator, your means and standard deviations may differ slightly due to rounding error.

Practice Problems

1. Your point on a standard normal distribution is at Z = 0.91.

   What percentage of the normal distribution is greater than your point?

   What percentage of the normal distribution is less than your point?

2. Your point on a standard normal distribution is at Z = -1.98.

   What percentage of the normal distribution is less than your point?

   What percentage of the normal distribution is greater than your point?

3. Calculate the Z-Score for each value of X_i:
   $X_1 = 16$
   $X_2 = 15$
   $X_3 = 25$
   $X_4 = 18$
   $X_5 = 29$

   What is the mean of your five Z-Scores? Even though you know what the mean should be, please calculate it anyway. This will give you valuable practice and show if you made any mistakes in calculating the Z scores or the mean.

   What is the standard deviation of your five Z-Scores? Again, calculate the standard deviation to see if your value agrees with what it should be.