Week 1 Lecture 3

A second way of looking at data differences or similarities is to consider how likely a given outcome is. In looking at our data set, we could ask questions such as, what is the *probability* (likelihood) of a male or female salary exceeding 60K, what is the probability that a person's salary is within the range of 38K to 52K, etc. Probability questions about a data begin to help us look at distributions, a topic we will delve into in more detail in the upcoming weeks.

Probability is the likelihood that a specific outcome will occur; it is always positive and ranges from 0 (will never occur) to 1.00 (will always occur). Generally speaking, we have 3 kinds of probability – *empirical* (counting actual outcomes), *theoretical* (using theory/logic to determine what should occur) and *subjective* (our individual guesses and feelings). Obviously the theoretical and empirical are the best approaches for business research questions, but at times the best we can get is an expert's guess.

Theoretical probability is just as it sounds – the theory of what the probability should be. For example, if we flip a fair coin, our theory says we should get heads 50% of the time – one outcome out of the two possible. If we flip the coin a number of times, we will get the empirical probability – the number of actual heads divided by the number of flips. While this is generally close to .5, achieving this is usually the result of a lot of flips rather than just a few (even up to 100) (Lind, Marchel, & Wathen, 2008).

While many approaches to theoretical probability exist (binominal, hypergeometric, Poisson, etc.) (Lind, Marchel, & Wathen, 2008) exist, we will look at just two particular types – the *binominal* and the *normal curve* based probabilities. The binominal requires that we have only two outcomes, such as heads and tails when flipping a coin. This is not as restrictive as it might seem, as we can always create 2 groups out of what we have. For example, if we have a single die (one of a pair of dice), we could form several two group situations – evens versus odds, 1 – 3 versus 4 – 6, etc. We will use the binominal to discuss several basic probability rules.

Four general probability (P) concerns exist. Typically, we want to know one or more of the following probabilities:

- of something happening – called P(event),
- of two things happening together – called joint probability: P(A and B),
- of either one or the other but not both events occurring – P(A or B),
- of something occurring given that something else has occurred – conditional probability: P(A|B) (read as probability of A given B).
- Compliment rule: P(not A) = 1- p(A) (Lind, Marchel, & Wathen, 2008).

Two other issues are needed, the idea of *mutually exclusive* means that the elements of one data set do not belong to another – for example, males and pregnant are mutually exclusive data sets. The other term we frequently hear with probability is *collectively exhaustive* – this simply means that all members of the data set are listed (Lind, Marchel, & Wathen, 2008).

Some rules, which apply for both theoretical and empirical based probabilities, for dealing with these different probability situations include:

- P(event) = (number of success)/(number of attempts or possible outcomes)
- P(A and B) = P(A)*P(B) for independent events or P(A)*P(B|A) for dependent events (This last is called conditional probability the probability of B occurring given that A has occurred).
- P(A or B) = P(A) + P(B) – P(A and B); if A and B cannot occur together (such as the example of male and pregnant) then P(A and B) = 0
- P(A|B) = P(A and B)/P(B) (Lind, Marchel, & Wathen, 2008).

Binominal Probability

Binominal probabilities deal with dichotomous outcomes – those that have only 2 possible outcomes. A typical example is flipping a coin, the result can only be a head or tail. Another common example is gender, we are born as either male or female. The interesting element about binominal outcomes is that while every single trail (such as the flip of a coin) has the same probability, the outcome of a group of trails will not necessarily match that probability. For example, the probability of getting exactly 5 heads out of 10 flips of a fair coin is not .5, but rather 24.6%! This is due to the number of ways the 10 outcomes can be distributed (Lind, Marchel, & Wathen, 2008).

We can turn almost any outcome into a dichotomous outcome by creating groups. For example, we can say that when we toss a six-sided die (half of a pair of dice), we have two outcomes: getting a 1 or 2 versus getting anything else. Now we have two outcomes of interest instead of the original 6 possible outcomes.

Tables exist to determine the likelihood, but the easier way is to use the Excel functions found in the *fx* or Formulas lists. For example, Excel's *BINOM.DIST* functions can quickly provide us with the correct probability of getting a certain number of outcomes within a given number of attempts.

Research Question Example

Understanding the distribution of the data is an important element of understanding what the data is trying to tell us. Probabilities can give us a sense of the data set and allow us to compare results across groups.

Equal Pay Example. In thinking about equal pay, we might be interested in the probability that both males and females appear to be grouped in similar ways as the overall group. This would be an example of an empirical probability, as we would be counting how many of each group fall into each of the ranges we would set up.

We noted that the overall salary mean was 45, with the female mean equaling 38 and the male mean equaling 52. This suggests one group to look at – what is the probability of someone having a salary between 38 and 52 in each group – overall, females, and males?

Translating this into "probability" terms, we want to know:

- What is p(38 <= salary <=52)? What is the probability that salaries are between 38 and 52 inclusive?
- What is p(38 <= salary <=52|Female)? What is the probability that salaries are between 38 and 52 inclusive, given a female salary? Or, if a female, what is the probability that salaries are between 38 and 52 inclusive?
- What is p(38 <= salary <=52|Male)? What is the probability that salaries are between 38 and 52 inclusive, given a Male salary? Or, if a Male, what is the probability that salaries are between 38 and 52 inclusive?

We know a couple of things right off. First the entire sample has 50 members, and we have 25 males and females. These become the denominators in the respective probabilities. Since you do not have the exact data set we are working with, the counts for salaries in these ranges are: Overall: 8, Females: 3, and Males: 5.

So, we have:

P(38 <= salary <=52) = 8/50 = .16.

P(38 <= salary <=52|Female) = 3/25 = .12.

P(38 <= salary <=52|Male) = 5/25 = .20.

We can see if gender influences being within this range by seeing if the formula for independent events is true. Above we had stated P(A and B) = P(A)*P(B) for independent events. In this case, the P(within salary range AND Female) is the same as P(within salary range|Female); (this is not always the case). So, since P(within salary range) = .16, and P(Female) = .5, we would have:

P(within salary range and Female) = P(within salary range) * P(Female)

Replacing these with the associated values, we would have:

.12 = .16 * .5 (= .08). An expression that is clearly not correct or true.

Since the two sides of the equation are clearly not the same; we can say that gender and being within this salary ranger are not independent elements. Doing this for other ranges produces similar results, so we have a clue that gender and salary interact in some ways that suggest males and females are not paid equally.

What we still do not know yet, is how to consider equal work in our examination.

The Normal Curve

The normal curve is a data distribution that is often called the bell curve, as when you plot the likelihood of outcomes occurring, the resulting graph looks like a bell – the most outcomes in the middle (where the mean = median = mode), and then smoothly decreasing on each side.

As a probability distribution, the normal has some interesting characteristics. First, the probability of any outcome equals the area under the curve for that range of outcomes. (Tables and Excel give us these values.) Second, the curve technically extends from - infinity to + infinity (although this range is rarely actually used). Third, since the normal curve is continuous data, the probability of any single outcome (for example, getting a 76 on a test) is 0, so to overcome this we develop a range of values – the 76 score outcome would be the area from 75.5 to 76.5 – the adding of +/- half a unit to a value allows us to translate discrete data into a continuous range (Lind, Marchel, & Wathen, 2008).

The normal curve is important due to its wide spread appearance in everyday situations. Some examples of data that follow the normal curve are height, weight, IQ, standardized test scores such as the college boards, many manufacturing measures (above and below the average result), etc.

To make working with different normal curves (having different means and standard deviations), we can convert them all into the *Standard Normal Curve*, which has a mean of 0 and a standard deviation of 1.0.

We do this using a z-score – subtract the mean from the data value, and divide the result by the standard deviation. Doing this for every value in a data set would change the mean of the new distribution to 0 (due to the subtraction), while the division changes the standard deviation to 1. The resulting data values are now z-scores, and the area between z-scores is the probability of an outcome within that range of values. One characteristic of the z-score is that it tells us, in standard deviations, how close or far from the mean any individual score is; so in some ways this is another location measure but one that focuses on individual values.

Here is an example using the normal distribution and related Excel functions found in the *fx* list. (See the Excel Week 1 lecture for guidance on using this function if you are unclear about it.) To find the probability of an outcome between a z-score of 1.63 and 2.0, we would need to find the area between these two scores. To do this, we would subtract the area under the curve up to the z score of 1.63 from the area under curve up to the area of the z-score for 2.0. In excel we use the *fx* function NORM.S.DIST (z, cumulative) this way"

=NORM.S.DIST(2.0,1)-NORM.S.DIST(1.63,1) = 0.0288.

This tells us that the probability of finding a sample value within this range is about 2.9%.

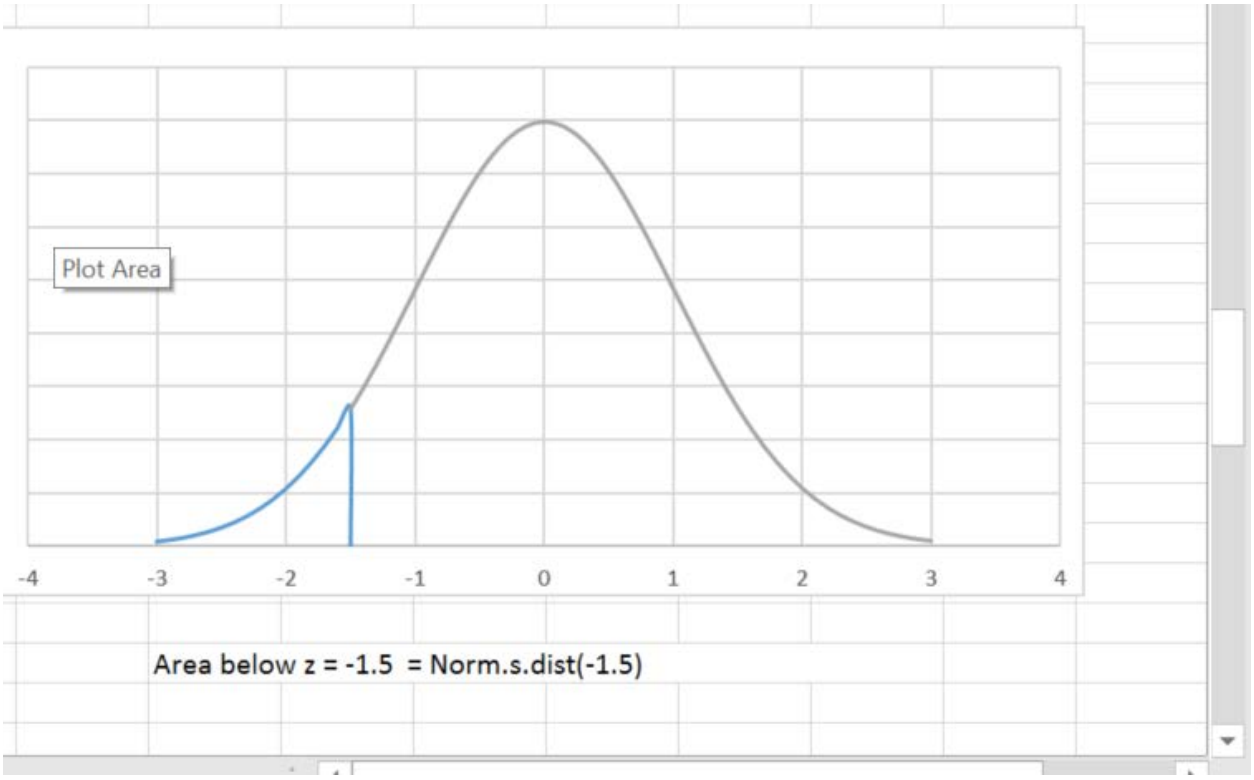
A second example with values that are above and below the mean would be done the same way. Looking to find the probability that we would find a sample value between the z scores of -1.63 and 2.0 would be:

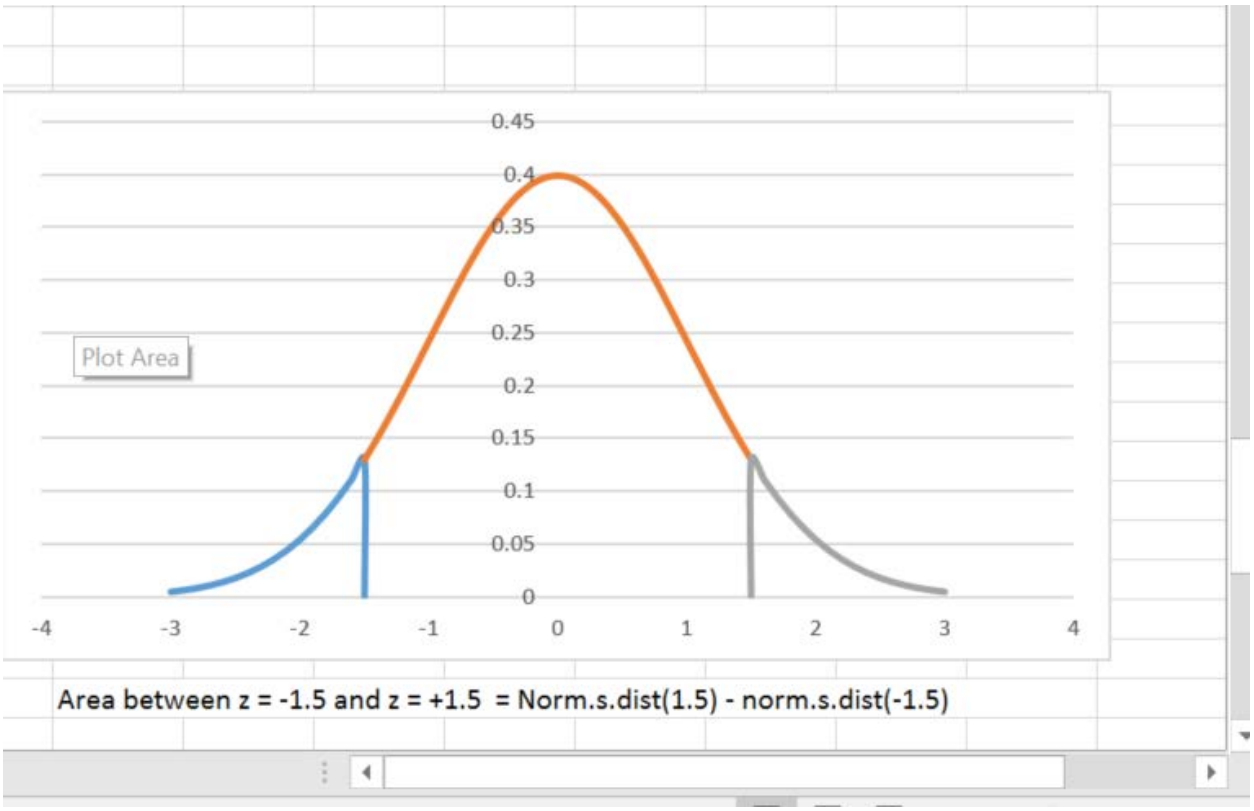
=NORM.S.DIST(2.0,1)-NORM.S.DIST(-1.63,1) = 0.9257 or 92.6%.

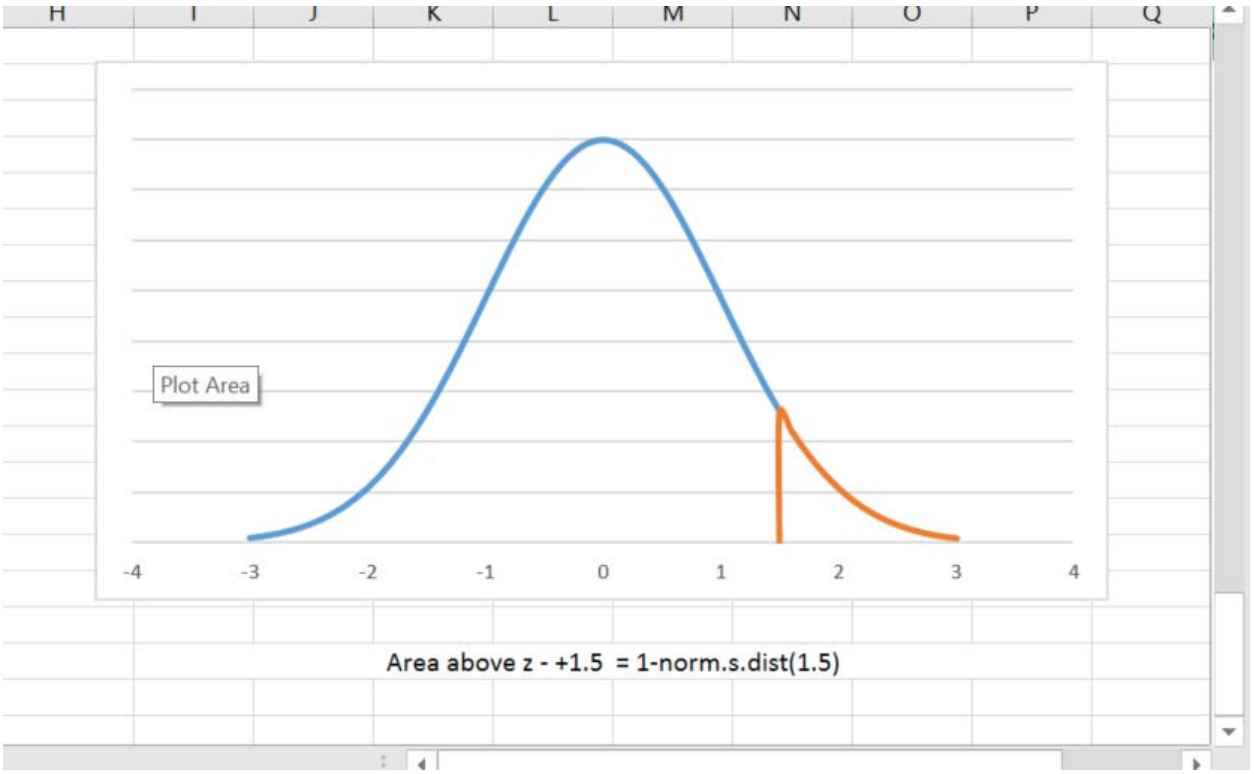
The final example of finding a normal curve based probability is determining the probability of being greater than some value; for example, what is the probability of exceeding a z score of 2.0? =1 - NORM.S.DIST(2.0,1) = 0.02275 or 2.3%

Finding the area below a negative z score is found by simply using the NORM.S.DIST function. = NORM.S.DIST(-2.0,1) = 0.02275 or 2.3%

A hint on doing these kinds of problems is to draw a picture of the normal curve, and draw a vertical line at each of the z score values you are working with. Then shade the area you are interested in. There are three cases – the area below a certain value, the area above a certain value, and the area between two values. This visual guide helps in determining what we subtract from what.







Side note – the probability of exceeding a particular outcome by pure chance alone is called the **p-value**. We will start using this idea next week.

Research Question Example

We will be assuming that the variables we are using for our equal pay question come from a normally distributed population. This allows us to use normal curve based probabilities and statistical tests to examine the data we are using to answer our question.

Equal Pay Example. Earlier we found that the likelihood of males and females having a salary between 38 and 52K were not the same, suggesting that gender and salary interacted in some way. Let us ask if the probability of having a salary greater than the overall mean of 45 is the same for both genders. Since we are assuming that salary is normally distributed as a whole and for each gender, we can use a normal curve probability to examine this.

The first step is to find the z scores for each gender for the data value of 45. We found earlier that the female mean is 38, and the sample standard deviation is 18.3, and the male mean and standard deviation is 52 and 17.8 respectively. This gives us the information needed to determine our z scores:

Female z = (45-38)/18.3 = (7)/18.3 = 0.38.

Male z = (45-52)/17.8 = (-)7/17.8 = -0.39.

The second step is to find the probability of exceeding each of these values using the NORM.S.DIST function.

Female: =1-NORM.S.DIST(0.38,1) = 0.352

Males = 1-NORM.S.DIST(-0.39,1) = 0.652

So, it again appears that males and females have a different salary distribution as males are almost twice as likely to be above the overall average of 45 as females are. Again, we have not yet considered the equal work element.

References

Lind, D. A., Marchel, W. G., & Wathen, S. A. (2008). *Statistical Techniques in Business & Finance*. (13th Ed.) Boston: McGraw-Hill Irwin.