CHAPTER 24

ERRORS IN HYPOTHESIS TESTING AND POWER

First of all, welcome to the last Chapter! Much of this text has concentrated on making inferences from sample data to the target populations of interest. In many of these inference situations, the inference being made was in the form of testing some hypothesis about the population mean, or population variance, or population correlation coefficient. We also discussed other hypothesis tests. What we need to do for the closing section is to briefly examine the consequences of the decisions we make when testing hypotheses. Since we are using sample data to make inferences about the population parameters, and in this context we retain or reject some null hypothesis, we can never be sure that the decision we make is the correct one. Since we only have sample data and not the population data, our inferences are only our best estimates of the truth.

Types of Decisions: Correct and Incorrect

In the typical hypothesis testing situation, we come to the final decision where we either retain or reject the null hypothesis. Thus, a decision is made in the study. But, juxtaposed to this decision is the real truth. The null hypothesis is either true or false. What this yields is a two by two table that indicates the various combinations of decisions you make compared to the actual truth. Consider the following table.

	H(0) is True	H(0) is False
Retain H(0)	Correct Decision	Incorrect Decision (**Type II or Beta Error**)
Reject H(0)	Incorrect Decision (**Type I or Alpha Error**)	Correct Decision (**Power**)

At the side, notice that we have the two possible decisions we can make: either retain the H(0) hypothesis or reject the H(0) hypothesis. At the top, we have the two possible states of truth: either the H(0) hypothesis is true or it is false. This sets up 4 possible different combinations of decision and truth. For purposes of discussion, assume for a moment that the null hypothesis is the population mean IQ score is 100. If the population mean is really 100 and we retain the null hypothesis, then we are in the upper left cell and have made a correct decision. To retain the null hypothesis when the null hypothesis is true is the right thing to do. On the other hand, if the true mean is 100 and you reject the null hypothesis, then you have made the wrong decision and you are in the lower left cell. You have made what is called a **TYPE I OR ALPHA ERROR**.

A TYPE I OR ALPHA ERROR IS WHEN YOU REJECT THE NULL HYPOTHESIS WHEN THE NULL HYPOTHESIS IS TRUE.

If the null hypothesis is false and you retain the null hypothesis (upper right cell), then again you have made a wrong decision. You have retained the null hypothesis when the null hypothesis is not correct. Since this is the opposite of a Type I error, this is called a **TYPE II OR BETA ERROR**.

A TYPE II OR BETA ERROR IS WHEN YOU RETAIN THE NULL HYPOTHESIS WHEN THE NULL HYPOTHESIS IS NOT TRUE.

Finally, if you reject the null hypothesis when the null hypothesis is false, then you have made a correct decision. If the true population mean IQ score is not 100 and you reject the null hypothesis <u>of</u> 100, then you are saying that you do not think that the mean is 100 and you would be doing the right thing by rejecting the null hypothesis.

Let's put the above diagram into another context, a jury trial. The H(0) hypothesis in a trial, when we give the benefit of doubt to the person on trial, is that the person is not guilty and did not commit the crime. Look at the following diagram.

	H(0) is True: Innocent	H(0) is False Guilty
Retain H(0) (Acquit)	Correct Decision (**Let an innocent person go free**)	Incorrect Decision (**Let a guilty person go free**)
Reject H(0) (Convict)	Incorrect Decision (**Convict an innocent person**)	Correct Decision (**Convict a guilty person**)

If the person is really not guilty and the jury votes "not guilty" and acquits the person, then they have made the correct decision. The jury has set free an innocent person. If the person is actually not guilty but the jury convicts the person anyway, then the jury has made a TYPE I or alpha error. Convicting an innocent person is an error of the "first kind". On the other hand, if the person is really guilty and the jury votes to acquit the person, then they again have made a mistake and this type of mistake in decision is called a TYPE II or beta error. This is an error of the "second kind". Finally, if the jury votes to convict the person and the person really did commit the crime, then the jury has made the correct decision. In the literature, the decision of rejecting a false null hypothesis [H(0)] is called **POWER**. Although you may not have thought about it this way, the main purpose of a trial is to<u>convict guilty persons</u>. Certainly, we want the jury to arrive at the correct decision and if the person is not guilty, then we hope the jury will acquit the person. However, the defense side of the case (to defend the person on trial) is not the side to bring the case to trial. It is the prosecution side; they are the ones to instigate the entire trial process. In this context therefore, the hopes of the trial system (in addition to finding the truth) is to convict persons of the crime they were charged with especially if they really committed the crime.

Thus, what we see here is the possibility of 4 different consequences of our decisions: two good and correct ones but also two bad and incorrect ones. Of course, whether it be in a jury trial situation or an experiment where we retain or reject the null hypothesis, our goal is to **MAXIMIZE THE CHANCES OF MAKING A CORRECT DECISION** and **MINIMIZE THE CHANCES OF MAKING AN INCORRECT DECISION**. To summarize, we will do one the following.

Make **Correct Decision** by:

Retaining the null hypothesis when the null hypothesis is **true**, or

Rejecting the null hypothesis when the null hypothesis is **false**.

Make **Incorrect Decision** by:

Retaining the null hypothesis when the null hypothesis is **false** (TYPE II or beta error), or

Rejecting the null hypothesis when the null hypothesis is **true** (TYPE 1 or alpha error).

Finally, you will recall that on several occasions, I stated that you could look at the p value from the output from t tests, ANOVA, etc. and use that as a guide to decide whether the retain or reject the null hypothesis. If the p value was = or less than .05, we would reject. If the p value was > .05, then we would retain. **THIS p VALUE REPRESENTS THE PROBABILITY OF MAKING A TYPE I ERROR AND IS SOMETIMES CALLED THE LEVEL OF SIGNIFICANCE OR ALPHA** á. Generally speaking, researchers are not willing to reject the H(0) hypothesis if the chance of rejecting it (concluding that your treatment made a difference, that there is a non-0 correlation, etc.) wrongly is greater than about 5%. Thus, the p value is important because it gives you some estimate of what the chances are of making a Type I error. Remember, we usually want to reject the null hypothesis but we only want to do this if the chance that we do it incorrectly is rather low. Unfortunately, setting some criterion about how much Type I or alpha error rate we are willing to tolerate does little to regulate or establish what level of Type II error or beta we are willing to tolerate. The literature seems to be fixated only on the rate of Type I errors without much attention to Type II error.

However, consider for a moment two situations where it is possible that Type I error rate may be more important in one case and Type II error may be more important in the other case. What if we had a new drug that has tremendous potential for positive effects. In the one case, there may be no reason to suspect that there will be any adverse side effects. It either works or does nothing. In this case, you would be willing to reject the null hypothesis almost all of the time in hopes of finding evidence that the drug would be beneficial. What is there to lose? In this case, what we do not want to make is a Type II error and fail to conclude that the drug is beneficial when it actually is. Thus, in this case, we will want to make sure that Type II error is very low and we do not really care how high Type I error rate may be.

On the other hand, what about the opposite situation where this drug has tremendous potential for positive benefit only if it really works but and the fact is, if it does not work, we strongly suspect that the side effects could be very harmful. In this situation, we are still very interested in rejecting the null hypothesis (of no positive benefit) but are only willing to do it if we are very sure that the drug works. In this case, we do not want to reject the null hypothesis incorrectly because of the potentially harmful side effects. Here we want the Type I error rate to be very low and we aren't as concerned with a higher beta or Type II error rate. We would rather say that the drug does no good and keep it from the market (even if it really does work) than to risk letting it flow into the marketplace and produce disastrous effects.

As you can see, Type I and Type II errors are both important and different circumstances demand that one be controlled more than the other. Ideally, we would like to make sure that both have very low rates. But, in most studies, one has to both compromise and concentrate on controlling one of these a little more than controlling the other. Which one receives greater attention in a particular investigation depends on what that investigation is all about. However, to give you a better feel for how the Type II error rate is controlled, the discussion below focuses on **POWER**. As you will see, power is the "key" concept in what we try to accomplish in research investigations. We would like to have our studies be as "powerful" as possible. In our discussion of power, you will see that the Type II or beta error is actually the "flipside" of power. If power is high, then Type II error rate is low, and vice versa. The same factors that impact on power also impact on Type II error rate and by controlling factors designed to improve power, we are therefore also controlling Type II or beta error.

Power of Statistical Tests

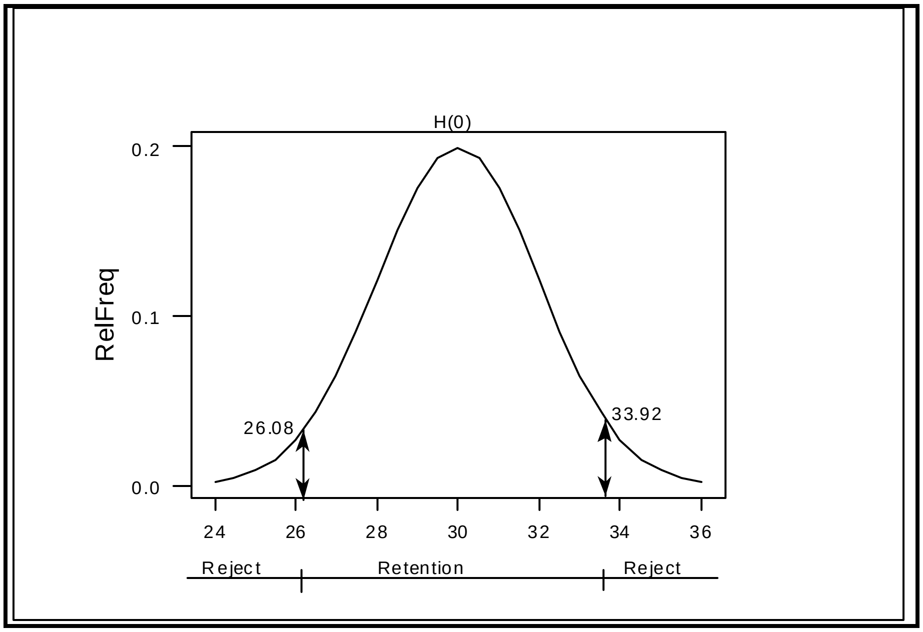
Now we come to one of the most important concepts in statistical inference and particularly hypothesis testing, and that concept is called power. Unfortunately, power is a complex topic and we are only able to scratch the surface. However, it is important to at least indicate what power is, show a simple example of how one could estimate power, and to indicate some of the factors that influence power.

First of all, what is power? In our two by two table of the different consequences of our decisions, recall that the lower right cell represented the instance when you rejected the null hypothesis when the null hypothesis is false. This represented one of the 2 correct decision possibilities. This particular cell in the model is highly meaningful in statistical hypothesis testing since, most of the time, that is the cell you would like to be in. For example, assume we are using the example of the null hypothesis that the population mean IQ score is 100, according to the test publishers. However, if one did a study in this area, it would not likely be to show support for the null hypothesis. Perhaps you believe that ability is going "downhill", thus you think the test publishers are incorrect when they say that the mean IQ score is 100 today. In fact, your directional prediction is that the true mean IQ is now less than 100. So, while we may give the benefit of the doubt to the test publishers in that we will use their view for the null hypothesis, our goal is to reject the null hypothesis. Of course, we should not want to reject the null if the null is really true. But, if our "hunch" is correct and the population mean IQ is really not 100, then, in this case, we want to make as sure as we can that our decision in our study will be to reject the null hypothesis. Power represents the chance that we will be able to reject the null hypothesis of 100 if in fact the true mean is not 100.

In a nutshell, power boils down to the following. Will your study be designed in such a way as to be sensitive enough to find what you are really looking for if what you are predicting and looking for is really the truth? Will your study have enough power? For example, if a new method of instruction is really superior to the old method, will your experiment be able to demonstrate that and reject the null hypothesis of no difference between the two methods? If there really is a non-0 correlation between two measures out in the population between X and Y, and you test the null hypothesis of 0 correlation, will your sample study allow you to reject the null hypothesis? What we see is that generally we would like to reject the null hypothesis. In fact, I said much earlier that it is called the null hypothesis because you hope that your sample data will be strong enough to nullify that hypothesis in favor of your own prediction. Thus, that lower right cell in the two by two model of decisions is very important.

How Power is Calculated

Inorder to better understand power, I will "walk" you through the steps of how we can calculate power and show you the things that are necessary to do so. First, we need to have a real H(0) hypothesis, hopefully to reject later. So, for discussion sake, let's assume that the null hypothesis is that the population mean is 30. If the population mean is really 30, then samples drawn at random from that population will exhibit sampling error. That is, sample means will vary around the null value of 30. How much are they likely to vary? That depends on the size of the sample and how much variability there is in the population. If we knew what the standard error of the mean was, then we would have a good idea of how much the means would vary around the null value of 30. Therefore, assume for the moment that not only is the H(0) value = 30, but the standard error of the mean (S sub X bar) is about 2. What would the sampling distribution look like under these conditions? See the following diagram.



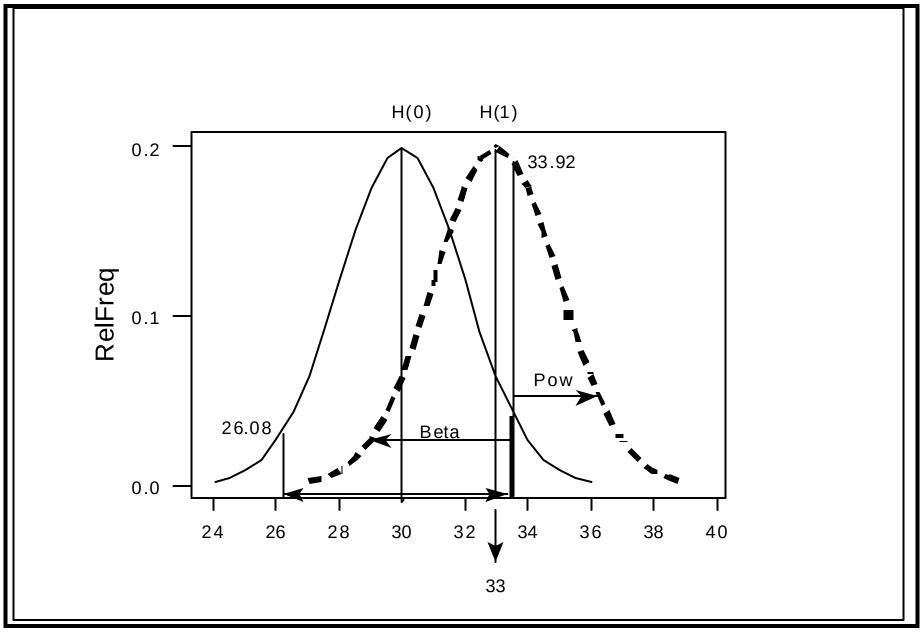
Remember that sampling distributions are similar to normal distributions. In this case, the center of this distribution would be 30. In addition, since this is a sampling distribution of means, the sample means would vary around 30 according to the standard error of 2. That means that sample means would vary from about 3 units of 2 (= 6) below 30 to about 3 units of 2 (= 6) above 30. However, when we do a hypothesis test, we usually "draw the line" somewhere away from the null value so that if our sample value (sample mean in our one study) falls at or beyond that point, then we would reject the null hypothesis. Thus, on this sampling distribution of H(0), we mark off critical points below the null value of 30 and above the null value of 30.

Recall that when we did hypothesis testing with means, we usually used the t value for a 95% confidence interval so that we would have 2 1/2 percent below the low end critical point and 2 1/2 percent above the high end critical point. Thus, in our case here, we would establish the critical points approximately 2 units of error below and 2 units of error above the value of 30. Technically, if this distribution was exactly normal, we would go 1.96 units rather than 2. The diagrams here are based on going 1.96 units, not 2. This is like a z value in the normal distribution. So, on the H(0) sampling distribution, we set up critical points that are about 1.96 units of 2 (error) below 30 and 1.96 units of 2 (error) above 30. 2 times 1.96 = 3.92 and if you subtract and add 3.92 to 30 you get the critical points of 26.08 and 33.92.

By setting the critical points of 26.08 at the low side and 33.92 at the high side, we have sectioned the H(0) distribution into 2 regions. Since any sample mean that is larger than 26.08 or smaller than 33.92 would cause you to retain the null hypothesis, the middle region between (but not including) 26.08 and 33.92 would be called the retention region. If the null hypothesis is true, we will be retaining the null about 95% of the time. However, any sample mean that is 26.08 or smaller, or 33.92 or larger would cause you to reject the null hypothesis. Therefore, the 2 end regions (combined together) would be called the rejection region. If the null hypothesis is really true, you would be rejecting the null (wrongly and making a Type I error) about 5% of the time. Thus, **on the H(0) distribution, you have the retention and rejection regions**.

As a next step in the process of examining power, we need to assume that the true population mean is some specific value. Yes, we have formulated the null hypothesis and established some critical points based on that. But, what is the real truth? Technically, to be able to calculate power, we need to know the true value for the parameter. This should give you some idea that power is not something we can know for sure (just like we cannot know whether our decision in hypothesis testing is correct) since we will never exactly know the value of the parameter. So, to continue our example, let's assume that the true population mean is 33, or 3 points above the H(0) value of 30. Thus, by knowing what the truth really is (or making an assumption about the population value), and knowing what the sampling distribution of means would look like if the null hypothesis had been true, we can compare the true value to the null sampling distribution and see where it fits. If the true value for the population mean is really 33, then the true sampling distribution will be around 33, not 30. And, if the true sampling distribution is around 33 and not 30, we would expect it to have the same type of sampling error as it would have if the parameter had been 3 points lower on the scale. Thus, the diagram at the top of the next page shows the overlapping nature of the true sampling distribution (highlighted by a dark dashed line) around the true population mean of 33 compared to the hypothesized sampling distribution around the null value of 30 (solid line), if it had been the truth. Note that I have deliberately kept on the H(0) distribution the critical points of 26.08 and 33.92.

Now, here comes the difficult part. If we had done a study where we had formulated a null hypothesis of 30 and had set the critical points at 26.08 and 33.92, any sample mean we obtained based on our data that had been larger than 26.08 and smaller than 33.92 (inside the retention region) would cause us to retain the null hypothesis. Note that I put a horizontal line with arrows at each end to indicate the width of the retention region. However, any sample mean we obtained based on our data that was 26.08 or smaller, or 33.92 or larger (in the rejection region), would cause us to reject the null hypothesis. In the present case, since the true population mean is 33 not 30, we should be rejecting the null hypothesis all the time since if we retain it, we will be making a Type II error. Unfortunately though, while the sampling distribution around the value of 33 is the true sampling distribution, some of it overlaps (takes up the same space!) with the H(0) sampling distribution around the null value of 30. In fact, in this example (because of the way the two distributions overlap) any sample mean value less than 33.92 that is truly in the H(1) sampling distribution around 33, which seems to be more than 50 percent of them (because the critical point of 33.92 is somewhat above the center value of 33), will result in a decision to retain the null hypothesis, which is a Type II or beta error. Note that I have drawn another horizontal line with an arrow at the left side to indicate the beta area. Power in this case is the area indicated by the horizontal line from the critical point of 33.92 to the right towards the arrow and as mentioned above appears to be less than 50% in this case. For this example therefore, the chance of making a Type II error is greater than correctly rejecting the null hypothesis.



If we wanted to find the **exact value of power** in this example, we would have to find the exact area in the true sampling distribution (dashed line distribution) above 33.92. Since the standard error is 2 and this is the standard deviation of this sampling distribution, the problem reduces to finding the percentile rank for a value of 33.92 in a normal distribution where the mean is 33 and the standard deviation is 2. Minitab could help us with this. See below.

```
 MTB > cdf 33.92;
 SUBC> norm 33 2.
    33.92   .6772
```

Approximately 67 percent of the area in the true sampling distribution around 33 is below the value of 33.92 and that is the Type II error rate or beta. Therefore, a value of 100% - 67% = 33% is the area above the value of 33.92 and that is our power value. In this problem, power is approximately 33%. In fact, if the conditions above were true, we would have only about a 1 in 3 chance of rejecting the null hypothesis and making the correct decision. Type II errors seem to have the upper hand here!

Factors Influencing Power

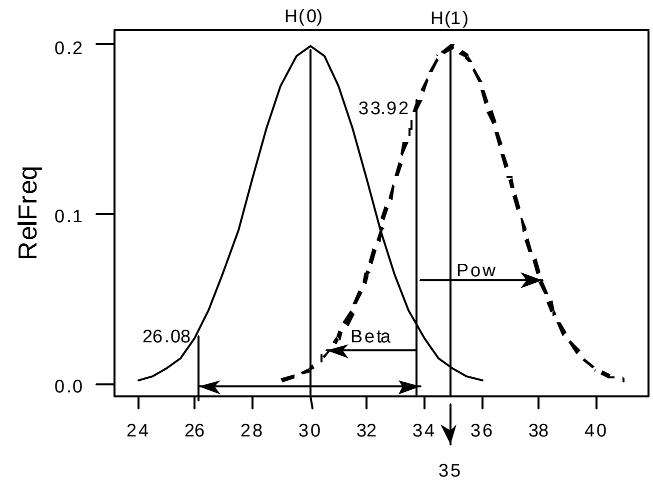
The example above showed a power value of approximately 33% and yielded not very good odds of making the correct decision by rejecting the null hypothesis. What are some of the factors that influence power, and hence allow us to regulate or build into our study a better power value? We now discuss these. An excellent software program by **Jack Barnette of the University of Alabama** (my buddy!) is also available to illustrate concepts below. Contact the present author for more information.

Distance between H(0) and H(1)

The basic example above on page 324 estimated power when the null hypothesis value was 30 but the true population mean value was 33. In that case, given the level of sampling error, our power value or probability of <u>correctly</u> rejecting the null hypothesis, was only about .33. This is what we found when the difference between the null value and the true value was 3 points. What would happen to this power value or probability of correctly rejecting the null hypothesis if the distance between the H(0) value of 30 and the true value changed; for example, got larger? The key to understanding this is to see whether the conditions produce <u>more or less overlap</u> of the two distributions. If there is less overlap, then there will be more area beyond the critical point in the true H(1) sampling distribution. If that happens, power will be larger and Type II or beta error will be smaller. Just the reverse is true if this distance becomes smaller because there will now be more overlap between the two distributions. Look at another example where the null is still 30 but, the true value has been made to be 35, a value more distant from the null. The graph is at the top of the next page.

In this graph, note that again the retention region on H(0) has been indicated by a horizontal line with arrows at each end. Any sample mean that has a value between 26.08 and 33.92 will cause us to retain the null hypothesis. However, any value equal to or less than 26.08 or equal to or greater than 33.92 will cause us to reject the null. Now, in this case, since the true sampling distribution of means around the true value of 35 would not likely produce a sample mean down in the 26 region, the lower critical point plays really no role in this power/beta calculation. The dashed line distribution is the true sampling distribution with 35 in the middle. But, lower than the middle at 33.92 is the upper critical point. So, any of the means in H(1) or the true distribution that are 33.92 or larger, will lead us to reject the null, and doing so is the correct or power decision. How much of the true H(1) distribution is to the right of 33.92? Well, it is clearly more than 50% in this case. To find out exactly, we could let Minitab find that area. Actually, what we need to find out is how much area is above the value of 33.92 in the true distribution where the mean is 35 and the standard error is 2. See below.

```
MTB > cdf 33.92;
SUBC> norm 35 2.
   33.9200    0.2946    <--- Power is 1 - .29 = .71
```

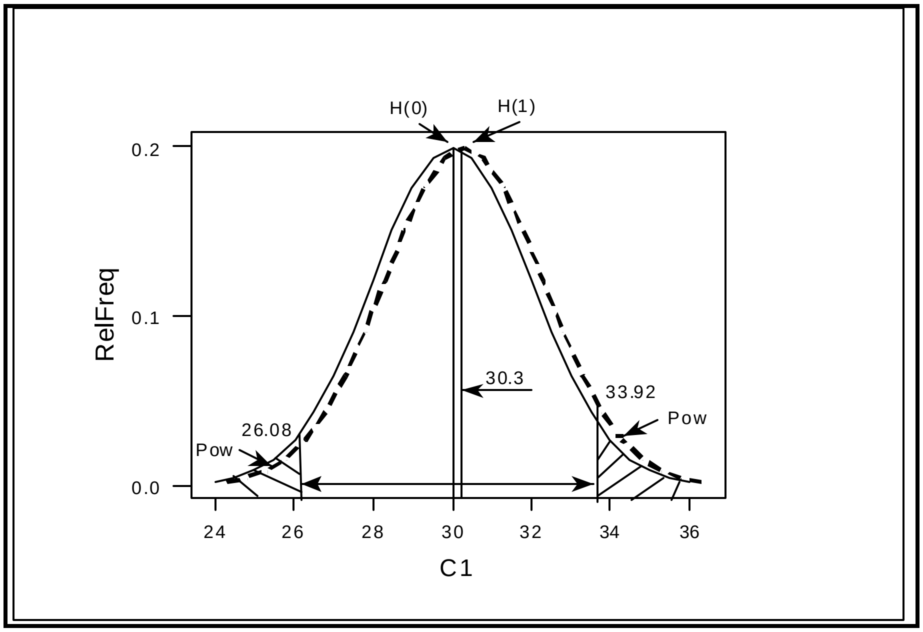


If you compare the power value of .71 in the above example where the true mean was 35, to the previous example where the power value was .33 when the true mean was only at 33, you should be able to see that the further out along the baseline the true value is compared to the null value, the less overlap there will be between the two distributions and hence, more power and less Type II error. The basic principle here is that: **POWER WILL BE LARGER AND BETA WILL BE SMALLER THE MORE DISTANT THE TRUE H(1) MEAN IS FROM THE NULL HYPOTHESIS H(0) VALUE.** The issue here is one of overlap and, the less overlap there is, the more we will be rejecting the null.

Based on this, what do you think the maximum value of power could be? Theoretically, the maximum power would be when the two distributions overlap the least amount possible. What if we had a situation where the two distributions, for all intents and purposes, did not overlap at all. For example, in the above case, what if the true mean was at a point of 50. Well, if you go 3 units of error of 2 to the left of 50, the lowest sample mean would be about 44. But, we saw in the H(0) distribution, the highest value was about 30 + 3 units of error of 2 or an upper point of 36. This would be a case when there seemed to be a clear separation between the critical point and the bottom of the true H(1) distribution. If that were the case, then all of the H(1) distribution would be beyond the critical point and therefore power would be approximately 1.00 or 100%. While this is theoretically possible, don't hold your breath waiting for a power value this high!

What is the Minimum Value of Power?

Above, we saw that the maximum value of power is about 1.00 or 100% in the case when the distance between the center points of the H(0) and the H(1) distributions is so far apart that the distributions do not overlap. But, when will power be at its minimum? If the maximum value is 1.00, it would be easy to think that the minimum value would be the opposite or 0. But, we will see that is not the case. Keep in mind two things: (1) the more the overlap, the less the power, and (2) power is always the area beyond the critical points on the H(1) distribution. What if I generated an H(1) distribution that has a mean that is very nearly the same value as the null hypothesis value of 30? For example, what if the true H(1) value was 30.3? Look at the diagram below.



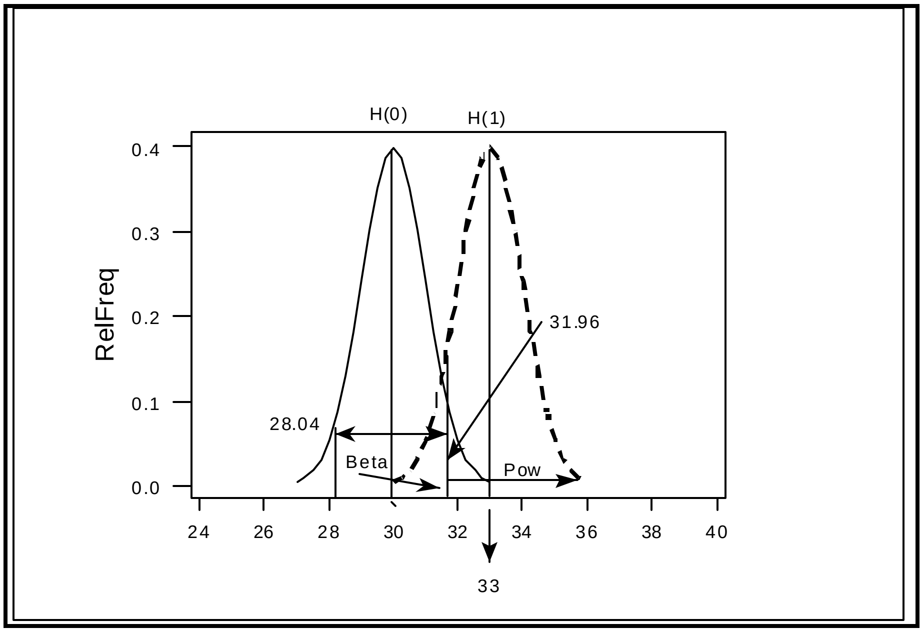
Again, I have drawn in the values of 30, 26.08 (lower critical point) and 33.92 (upper critical point). I also have put the true H(1) distribution in with the dashed line, and the center of it is (vertical line) 30.3. The little horizontal line with left arrow below 30.3 shows where the center of the true sampling distribution falls. Notice that almost all of the area on the true H(1) sampling distribution is now within the region between the crtical points (again, marked by the long horizontal line with arrows at the ends. This leaves very little room beyond the critical points on H(1). We could find the exact areas by using the cdf commands but I have spared you of that for the moment. The power part on the right side is the section topped by the dashed line over to the critical point of 33.92 where the arrow points down from POW. The left side part to power again is the segment topped by the dashed line over to the lower critical point of 26.08. The other short line ended by an arrow next to POW is that part. As you can see, the two sections that are power (right side part and left side part) are almost identical to the tail ends of the H(0) distribution beyond the critical points of 26.08 and 33.92. Those sections are the Type I error sections IF the null hypothesis had really been true. But, since H(0) is not true, and we should be rejecting the null all the time, we find that we will be rejecting the null only about the same number of times as defines the end sections, or alpha section, on H(0).

Thus, power in this case is almost the same as the Type I area as sectioned on the H(0) distribution. The point is, almost all of the H(1) distribution falls between the two critical points and therefore power would be very small and Type II error would be very large. However, could this power area ever get to be 0? No! The most the overlap could be would be when the true H(1) value is so close to the null hypothesis value that it is essentially the same as the null hypothesis value. For example, if the true H(1) value is 30.1 or 30.01 or 30.001; technically it is not the same as H(0) but very close. In this case, if you project the critical points of 26.08 and 33.92 down onto the H(1) distribution, there will always be at least an amount of area beyond the critical points that is <u>equal to the area beyond the critical points</u> on H(0)! Since this was our Type I or alpha error that was used in defining the critical points, <u>power can never be less than the Type I error rate</u>. Thus, although power could theoretically be 1.00 or 100%, it can never be less than the Type I error rate that you establish for your investigation.

How does Sample Size Impact on Power?

Another factor that has an important impact on power is the size of the sample used in the investigation. When we looked at the impacts of sample size on sampling error, we observed an inverse relationship in that as the size of the random sample increased (for example), the amount of sampling error decreased. This impact was evidenced by the fact that the <u>widths</u> of the sampling distributions became narrower. Keeping that in mind, what if we assumed that in the examples above, instead of having a standard error of about 2, that sample size had been increased sufficiently to cut that error factor in half. [Note: If you examine the formula for the standard error of the mean, it would take an increase in the sample size of 4 fold to reduce the error to half the original size.] Therefore, instead of the sampling error being 2, it is 1. We could then do our original calculations and see what power would be in that context. Remember, the first power problem we did on page 324 had a power value of approximately .33 or 33%. What I will now do is to create another similar graph where, the center points of H(0) and H(1) stay at 30 and 33 respectively but, where the widths of each of these sampling distributions, because of increased sample size, have been reduced. Look at the diagram at the top of the next page.

The first thing you should notice, compared to the first power simulation on page 324, is that the <u>widths</u> of the sampling distributions are narrower. Since the widths are narrower, and the center points on each and distance between the two are the same (30 and 33), this means that there should be less overlap between the two distributions. If there is less overlap, then power should be greater.



To find power, we need the area on H(1) that is beyond the critical point. Before, that upper critical point was 33.92. The value of 33.92 was determined by going 1.96 units of an error value of 2 around the H(0) value of 30 (1.96 times 2 = 3.92 added to and subtracted from 30 = 33.92 at the top and 26.08 at the bottom). However, if the unit of error is smaller, then we will be multiplying 1.96 units times a smaller error factor, in this case, 1 rather than 2. Thus, the critical points will be 1.96 units of an error of 1 above and below 30 which is 1.96 times 1 or adding and subtracting 1.96 to 30. The new critical points in this case, when sampling error has been cut in half, will be 28.04 and 31.96. Only the upper critical point is relevant to the present problem. So, instead of determining how much area is beyond the value of 33.92 in the H(1) true distribution, we have to find the area beyond 31.96 in the H(1) true distribution. Not only do the sampling distributions shrink in width, the critical points also back up closer to the null value.

To find the new power value, we could again let Minitab do it. See below.

```
MTB > cdf 31.96;
SUBC> norm 33 1.
   31.9600    0.1492    <--- Power is 1 - .15 = .85
```

Thus, increasing the sample size not only has the effect of narrowing the sampling distributions but also moving the critical point closer to the H(0) value. In any case, this produces more power and less Type II or beta error. It must be realized however that descreasing the sample size has the opposite impact. Manipulating sample size, particularly when increasing it, is one of the most direct ways in which to influence power. If there is a concern about whether the investigation will be sensitive enough to detect the truth, especially if the truth is different than the null hypothesis, then one of the best ways to increase the chance of rejecting the null hypothesis would be to increase the sample size. Unfortunately, increasing the sample size also has the negative effect of taking more time and costing more money and, in some cases, may not be practical. But, if one can control sample size, one should if you want better power. If you have already invested considerable time and resources into conducting an investigation, doesn't it make sense to go the "extra mile" to give yourself the best chance of finding what you are looking for?

Type I Error Rate and Power

In the examples above, we used .05 as the Type I or alpha rate of error. Using this in a two-tailed fashion, we split the .05 in half and put 2 1/2 of it at the low end of the H(0) sampling distribution and the other 2 1/2 at the upper end of the H(0) sampling distribution. While this is very typical, there is no hard and fast rule that says one must use .05. Within reasonable limits, some researchers may adopt a .01 rule or, in some cases in pilot investigations, some may adopt a .10 rule. What impact, if any, will changing this Type I error rate have on power? Let's see if I can "intuit" you through this.

Think again about the original power problem on page 324 where, based on defining alpha as .05 and splitting half of it at the lower and upper ends, we came up with critical points of 26.08 and 33.92. But, what if we had used 10% as our Type I error rate (for example), we would have been splitting 10% in half and putting 5% at the low end and 5% at the high end. Instead of needing the 2 1/2 and 97 1/2 percentile rank points, we would need the 5 and 95 percentile rank points. Again, we only need one and that information can be also used for the other. However, the lower critical point does not really come into play in this problem. See below.

```
MTB > invcdf .95;
SUBC> norm 30 2.
   0.9500   33.2897   <------ Upper critical point is about
                              33.29 if upper tail is 5%
```

Using the previous upper critical point (2 1/2 % at upper end), we saw that the area beyond that point in the H(1) distribution was about .33, which was the power in that case in the diagram on page 324. However, since putting 5% at the upper end means that the critical point moves backwards a small amount from 33.92 to 33.29, the area beyond the 33.29 point on the true H(1) sampling distribution will be somwehat greater. To see exactly how much more this is, we could use the cdf command from Minitab as below.

```
MTB > cdf 33.29;
SUBC> norm 33 2.
  33.2900    0.5576   <--- Power is 1 - .56 = .44 or 44 %
```

Thus, in this case, increasing the amount of Type I error or alpha error rate you are willing to tolerate means that power also increases (and beta decreases). Of course, if you had changed alpha in the opposite direction, that is made it smaller than 5% overall (say 1%), then the critical points would have moved further out and the area beyond the critical points on the H(1) sampling distribution would have been less: power would have been less.

Because of the connection between Type I error rate and power, you might conclude that we should increase the risk of making a Type I error since it means a greater amount of power. But, keep in mind that the null hypothesis might be true. In that circumstance, rejecting the null hypothesis is not the correct thing to be doing. Actually, you can only gain power if the null hypothesis is not true and you should be rejecting the null hypothesis. Thus, an apparent gain in power only really means a "potential" gain in power and therefore the possibility of gaining more power may be offset by the reality that you are merely making more Type I errors.

Building-in a Certain Level of Power in a Study

The simple reason why many experiments or investigations fail, in the sense that they are not able to support legitimate research hypotheses, is that the power that exsists is too small to give the experiment a favorable chance of rejecting the null hypothesis. Think about all the problems that can "interfere" with finding good results: sample sizes are way too small, treatments do not get implemented properly, instruments have very poor reliabilities, and the list goes on. One of the major sources of this "lack of sufficient power" is that there is too much sampling error present. And, as has been stated above, perhaps the best way to lower the amount of sampling error is to increase the size of the sample. How much power could we potentially gain if we increased the size of the sample by X amount? Or, stated differently, what size of a random sample would need to be taken from the target population (assuming that the experiment is done correctly) to achieve some acceptable and predetermined level of power? I have hinted above that a power of .50 is not really very good in that there is about as much chance of not finding significant results (by not rejecting the null hypothesis) in your study as finding significant results. Power values like .6 or .75 would seem to make better sense. However, one needs to keep in mind that if you set the power value too high, it might call for a sample that is larger than one can realistically obtain. But, let's explore one example of how you might consider, ahead of time, the type of sample that would be required to "build-in" a certain amount of power.

Before one can attempt to estimate what sample size is required to achieve a certain level of power, various pieces of information are needed. These are listed below.

A. You need to decide on how much of a difference, from the null hypothesis value, is meaningful. For example, if the average IQ score in the general population is 100 and you think you have discovered a "treatment" that will raise the average IQ, "how much" is an important rise in IQ? That is, if this treatment really works, is the amount of increase large enough to spend lots of money to implement the new treatment all across the country? I should point out that this is not a statistical decision but rather a rational decision.

B. What level of Type I error are you willing to tolerate? In many cases, it is arbitrarily set at 5% but, as I have said before, there is nothing "set in stone" on this value. See next item.

C. What level of power would you like to achieve? 75%? 60%? Given that some level of power is selected, keep in mind that 1 - power = beta or Type II error rate. So, if you decide on 60% for power, that means you are necessarily willing to tolerate 40% Type II errors. Since Type I has been set at 5%, and Type II would be 40% if power is 60%, that means that relatively speaking, you are 8 times more willing to take a risk of making a Type II error than a Type I error. This is not the place to argue about which error is more important and when, but you can see how this logic follows.

D. Recall in the standard error of the mean formula, there is both an estimate of the population standard deviation value in the numerator and a sample size (or n) factor in the denominator. Since we will be solving for n, the only other factor we need to know (or have a good idea of) is an estimate of the population standard deviation.

With all the information from points A to D, you can then set up a simple equation and solve for an estimate of n. Let's see how that is done. First we need to make some assumptions about points A to D above.

A. Let's assume that the new treatment for improving IQ would only be considered to be worthwhile if you could demonstrate at least a 5 point gain in IQ from 100 to 105. Thus, the meaningful difference value here is 5 points on the IQ scale.

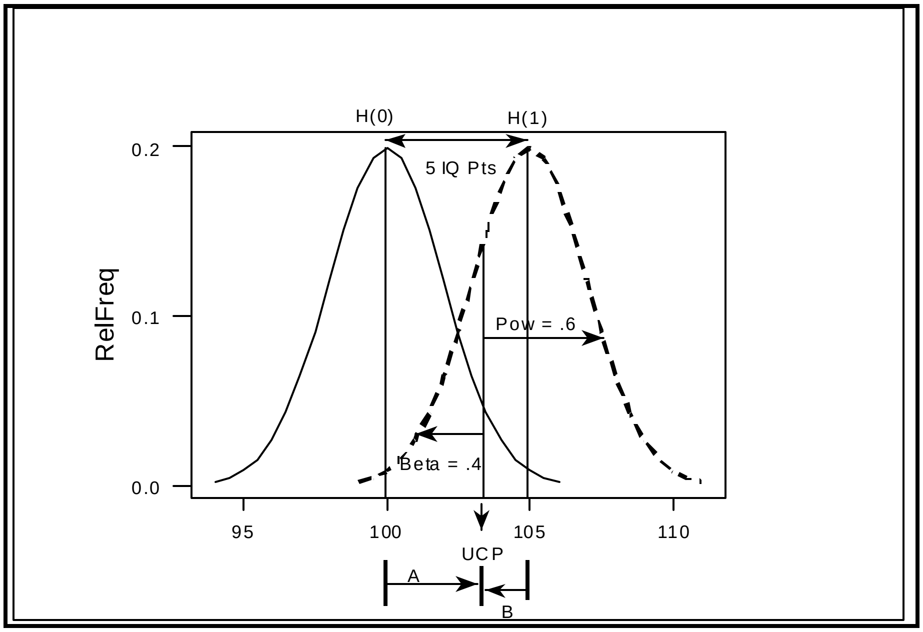
B. Let's establish the Type I error rate as 5%; rather traditional but good enough for our purposes.

C. Let's say that we want power to be about 60% and that means that Type II error rate would therefore be about 40%.

D. IQ scores are said to have a standard deviation of about 15 so let's make that assumption about the population standard deviation.

Given all of that, I have generated two sampling distributions where H(0) on the top has a mean of 100 (average IQ without treatment) and the bottom sampling distribution is for H(1).

We are assuming that the true mean IQ would be at least 105 if our treatment worked so that is why the center of H(1) is given as a value of 105.



Since we have specified Type I error rate as 5%, that means we would set up critical points on H(0) where the LCP has 2 1/2 percent below it (lower critical point) and the UCP has 2 1/2 percent above it (upper critical point). Since the LCP does not really overlap down onto H(1), I did not place it in the above diagram and will not refer to it any more in this discussion. The distance labeled as "A" represents the distance from the null hypothesis value of 100 up to the upper critical point. Since we have defined power as .6, then the beta area has to be .4, and is shown with the horizontal line with left side arrow. That means of course that .6 of the H(1) distribution or power is the section indicated with the horizontal line with right side arrow. Also, since we would travel 1.96 units of error (standard error of the mean) out to arrive at that critical point, we could label the point or UCP at the end of distance "A" as follows.

$$UCP = 100 + 1.96\ S_{\bar{X}}$$

That is, once we add to the null value of 100 the amount equal to 1.96 units of error, we have arrived at the position of the UCP! But, that upper critical point or UCP can be arrived at from another direction: it is at a position that is a certain distance below the H(1) value of 105. But, how far from 105 is it? Remember that we decided to achieve a power value of 60%, and that means that the area beyond and above the UCP on the H(1) distribution is equal to 60%. Since the area from 105 to the right equals 50%, that must mean that the area from 105 down to the critical point must be equal to 10% so that the total of the area on H(1) beyond the UCP is 60%. Therefore, how many standard deviation units below 105 would you have to go to include an area equal to 10%? If you look this value up in a normal curve table, you will see that you will need to travel about .25 standard deviation or error units out from the value of 105 (downward) to include about 10 percent of the area. Therefore, the UCP could also be arrived at by travelling distance "B" from 105 to the left, and that distance could be labelled as follows.

$$\text{UCP} = 105 - .25\, S_{\bar{X}}$$

Thus, what we have are two ways to arrive at the same point: move to the right from the null value of 100 or move to the left from the H(1) value of 105. Since these paths lead us to the same point, we can set them equal to one another.

From 100 to UCP $\longrightarrow \longleftarrow$ From 105 to UCP

$$100 + 1.96\, S_{\bar{X}} \;=\; 105 - .25\, S_{\bar{X}}$$

Keep in mind that the standard error of the mean has S, or the standard deviation, in the numerator part and the square root of sample size or n in the denominator part. So, we could further expand the "equality" as follows.

$$100 + 1.96\,(S\,/\text{Sqrt } n) = 105 - .25\,(S\,/\text{Sqrt } n)$$

Since we assumed above that the population standard deviation of IQ scores is about 15, we could further substitute into the "equality" above. See the following.

$$100 + 1.96\,(15/\text{Sqrt } n) = 105 - .25\,(15\,/\text{Sqrt } n)$$

Now what remains to be done is to rearrange the sides and solve for n. Without going through all the steps, we would have:

$$\frac{29.4}{\text{Sqrt } n} + \frac{3.75}{\text{Sqrt } n} = 105 - 100$$

$$\frac{33.15}{\text{Sqrt } n} = 5$$

$$33.15 = 5 * \text{Sqrt } n$$

Now square both sides and divide by 25.

$$1098.92 = 25 * n$$

$$n = 1098.92$$

$$n = \text{about } 44$$

Therefore, assuming that we are interested in detecting a true impact of the treatment on IQ scores if that impact is **at least** 5 IQ points, we would need a random sample of about 44 persons drawn from this population. If the treatment really does have an impact of at least 5 points to the better, that would give us a power of about .6 (that is, we would have about a 60% chance of rejecting the null hypothesis), and the chance of making a Type I error would be about 5% and Type II error would be about 40%. Random samples of less than about 44 will probably provide us with less than a 60% chance of rejecting the null hypothesis (or power).

In showing you this single example of trying to estimate how large a sample would have to be to achieve a certain level of power, the main purpose is to indicate that the answer to a question like: how big must my sample be... is not that easy to answer. To attempt to answer that question, one must consider several factors such as what type of power would you like, what levels of Type I and II error would you be willing to live with, and perhaps most important, what is a meaningful and important size of the impact of the treatment that you will want to observe before you are willing to change all the old methods and completely implement the new methods. Even if there is a real impact, if that impact is so small, it may be the case that it is not worth re-tooling the entire system for such a small effect.

Useful Minitab Commands

There are no new ones from this Chapter

Practice Problems

1. Assume that we are using 30 as the H(0) value, 33 as the assumed to be true H(1) value, alpha of 5% for two tails, and a standard error of the mean as 1.5. What is the value for power and Type II or beta error?

2. The H(0) value for the average IQ score is 100, the assumed to be true H(1) value is 97, the two tailed alpha is 5%, and the standard error of the mean is 3. What is the value for power and beta?

3. A researcher does a simple experiment where two types of reinforcement methods are used. In reality, there really is no difference in the effectiveness of the two methods. However, in the study, the null hypothesis was rejected. What has happened?

4. Someone formulates a null hypothesis that there is no correlation between X and Y in the population. In reality, that is true. In the investigation, the investigator retains the null hypothesis. What has happened?

5. When doing a t test on the difference between two means, there is an assumption of equal population variances. That would be the null hypothesis. Before doing the t test, the researcher checks on the validity of the equal variance assumption. In reality, there is a difference between the two population variances. The null hypothesis is actually rejected in the study. What has happened?

6. In a trial, the defendant acutally committed the crime. However, due to shaky evidence, the jury finds the defendant innocent. What has happened?

EPILOGUE

This text has presented a wide variety of methods for organizing and summarizing data from simple techniques for describing a single set of data (a classroom test for example) to fairly sophisticated techniques for testing inferential hypotheses about several population means (are 3 methods of teaching statistics the same or different?). Of course, this book is about data analysis so it makes sense to see many data analysis techniques discussed. You need to keep in mind however that there are dozens of other procedures not discussed here that are appropriate to other situations. No book can cover everything! For example, we examined using the t test to test a hypothesis about the equality of two population means where independent samples (different groups) were measured. However, not discussed was the situation where the same group of students was measured at two different times, such as once before a course and then again after the course was finished. While this appears to be a candidate for a t test for the difference between two means, and it is, you would need to alter somewhat the procedure that is used to implement the t test analysis. This same situation would exist if we wanted to compare the variances, for pretest and posttest, on the same group of people, or the correlation between X and Y for one group of people where X and Y were measured once at time A and then again at time B. In this text, we did not cover situations called dependent samples that involves the same people measured twice. Does that mean that these situations are not important to discuss? Absolutely not! The decision not to include many topics was a very practical one; there is a limit that every author has to impose on where to end his or her book!

You should also know that while the focus of this book is on data analysis techniques, albeit merely a sample of all possible procedures, it is important to point out that analytical techniques play only a partial role in the entire process of conducting investigations and interpreting results. My own personal opinion is that this role is a rather small one at that. Far more important in my view are the "components" of knowing the relevant literature that forms the base for doing an investigation, the use of proper and relevant instruments, and the careful implementation of the steps in the collection of the data which means careful selection of the proper sample on which to collect the data. In general, one cannot "mess up" too badly when it comes to the analysis stage (but please take that with a grain of salt!) but one can sure mess up by not conducting the study in the proper way. For example, I have stressed over and over again how important it is to draw and use the "correct" sample. If the sample is not representative of the population to which you want to generalize your findings, what is the purpose of the investigation in the first place? Thus, the point is that the statistical analysis part needs to be put in its proper perspective and one should not ascribe an "overimportance" to the "statistics" that they do not deserve. I am not arguing that the statistical part is not important; surely it is. However, there are other more important parts of conducting and interpreting research. There are no statistical methods that will "save" a bad study. A bad study simply gives you results that are uninterpretable. So, if you are interested in doing research or being a better consumer of research, spending all your "course" time taking statistics courses is not the way to go. Courses in measurement, survey sampling, research design, and writing are all important: equally so to statistics courses. Of course, this is just my opinion.

Also, while I have scattered the use of Minitab throughout this book as one software package that can be very helpful to you in your analysis needs, you should keep in mind that Minitab will not do everything. No package will! Minitab is a very useful general purpose package that will perhaps do 80% - 90% of what you need done, and I have provided some additional examples of using Minitab in the Appendix. However, I would encourage you to explore other packages to see what they offer. The more familiar you are with more than one "tool", the better off you will be. Other programs that are well worth your perusal include Systat, SPSS, SAS, Statgraphics, and the list goes on. Compare them to Minitab in terms of what they all will do and how they work. You may find that another package is more to your liking for doing certain types of analyses. In addition, you should know that special analysis needs in the areas of measurement (for example) are best handled by programs specifically designed for those purposes. The programs I have mentioned are primarily "good" for general statistical analysis.

Finally, while I have tried to include many general purpose data analysis procedures in this book, there will be many times when you are analyzing your own data that you will not be able to "go to THIS book" and find the answer. So, as we close our journey through some basic data analysis techniques, remember the following: **DON'T HESITATE TO SEEK OUT OTHER BOOKS AND SOURCES AND PEOPLE when you find that the current book is not adequate.** Believe me, it won't bother me a bit! Good luck!

Dennis M. Roberts
Educational Psychology
208 Cedar Building
University Park PA 16802
AC 814-863-2401
Email at: DMR at PSU.EDU