18

The Normal Curve Error Rates and Power

Type I and Type II Error Rates
Increasing Statistical Power
Statistical versus Practical Significance

Because decisions to reject or retain null hypotheses are based on probabilities (necessary since we work with sampling error and inferences), *our results are always subject to errors*. Such is science: problem, hypothesis, data gathering, analysis, and probabilities of findings. Our goal through proper research design and statistical analysis is to minimize errors and maximize "true findings." We complete our journey through the Normal Curve and hypothesis testing with considerations of error rates and power.

The chapter is divided into the following sections:

Error Rates. What are Type I and Type II Error Rates? How can we reduce the likelihood of committing Type I and Type II errors?

Power. What is statistical power? How do we increase the power of a statistical test?

Statistical versus Practical Significance. What is the difference between statistical significance and practical significance?

Type I and Type II Error Rates

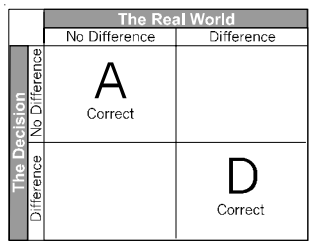
The dependence on "laws of probability" to make decisions means that we can be wrong when we retain or reject a null hypothesis. What's the probability that we're right? Or wrong? How can we improve our chances of making right decisions concerning our data? Before we can answer these questions, we must establish the elements of the problem, and how they relate to making correct decisions.

In hypothesis testing, there are two independent realities at work. The first reality is the "**real world**" itself. An anti-cancer drug either works or it doesn't. A prescribed teaching technique improves learning or it doesn't. A counseling procedure reduces anxiety in counselees or it doesn't.

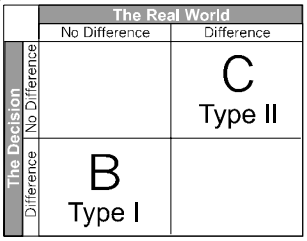
The second reality, at least within the context of our study, is "**our decision**" about the effectiveness of the treatment. Based on measurements, we will decide — on the basis of statistical analysis — whether our anti-cancer treatment succeeded or not, whether our teaching procedure was effective or not, whether our counseling ap-

proach reduced anxiety or not.

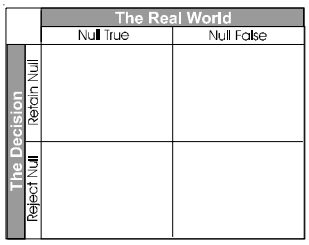
Decision Table Probabilities



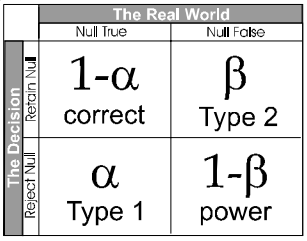
Correct



Error Types



Statistical Language



Probabilities

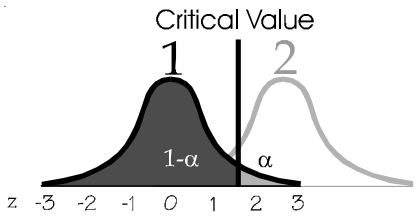
These two related, but independent, realities set up four possible outcomes for the analysis. These four outcomes are best considered in a 2x2 decision table, like the one at left. When there is *no real difference* in the world, and the researcher decides statistically that there is *no difference*, he makes a correct decision (box A). When there *is a real difference* in the world, and the researcher decides statistically that there *is a difference*, he also makes a correct decision (box D). In both cases the decision of the researcher matches the real world.

When there is *no real difference* in the world, and researchers decide statistically that there is a difference (B), they make a mistake, called a ***Type I error***. Where there *is a real difference* in the world, and researchers decide statistically that there isn't (C), they also make a mistake, called a ***Type II error***.

Now, let's translate the English labels above into statistical language. The term "null" refers to the stated "null hypothesis" for the study. If there is no difference in the real world, we say that the ***null is true***. If there is a difference in the world, we say that the ***null is false***. When we statistically decide that a difference does not exist, we ***retain the null***. When we statistically decide that a difference does exist, we ***reject the null*** and accept the research (alternative) hypothesis. The statistical labels mean exactly the same as the English labels above.

We add probability values (1-α, α, β, 1-β) to the boxes at left, and one new term, "power." These probabilities refer to the likelihood of a mean falling into the conditions of each particular box. When I set α to 0.05 (the probability of committing a Type 1 error), I automatically set the probability of making the correct statistical decision to retain the null to 1-α (0.95). Adding α to 1-α results in 1 (1-α+α=1), or 100% of all the scores in a single population.

When I set power (1–β) to 0.80 (the probability of declaring a real difference significant), I automatically set β to 0.20 (the probability of committing a Type 2 error). Adding β to 1–β results in 1 (1–β+β=1), or 100% of a second, different population. We will use the normal curve diagrams at left to identify exactly what these probability values refer to.



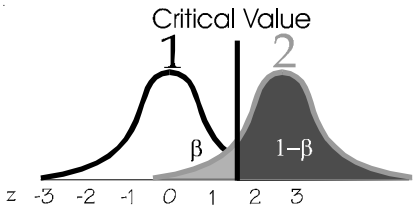
Normal Curve Areas

These two normal curves represent two distributions that differ on some variable. We'll call them "1" and "2." Normal Curve 1 is the population represented by the ***sampling distribution*** of our study, centered on the hypothesized mean. Normal Curve 2 represents a ***theoretical distribution*** of means centered on a different

population mean higher than our own.

The critical value line of Curve 1 cuts off the region of rejection (light gray area). This area is equal to α, the probability of committing a ***Type 1 error: rejecting a true null***. Any mean falling into this region is declared "significantly higher than μ_1." The dark gray area to the left of the critical value is the region of non-rejection and is equal to 1-α, the probability of retaining a true null. If we set α at 0.05, then 1–α equals 0.95. This means that when we declare a difference "significant," we are 95% sure of our decision.

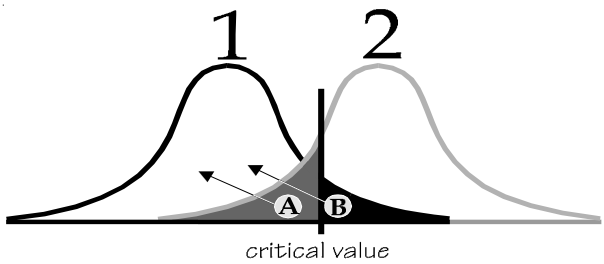
But notice that the critical value line also cuts off the lower part of Curve 2 (light gray). This area is symbolized by β, the probability of committing a ***Type 2 error: retaining a false null***. Any mean from distribution 2 (which should be declared different) falling in this area will be declared "not significantly higher than μ_1." The dark gray area to the right of the critical value line equals 1–β, the probability of rejecting a false null (power).



Now let's put the curves and boxes together with the labels A, B, C, and D. The diagram at right shows two sample means (A,B) which are **true nulls*** (arrows show they belong with population 1).

Mean A falls to the left of the critical value and is declared "not significantly different" from μ. This is a correct decision, and reflects **box A** (p=1-α).

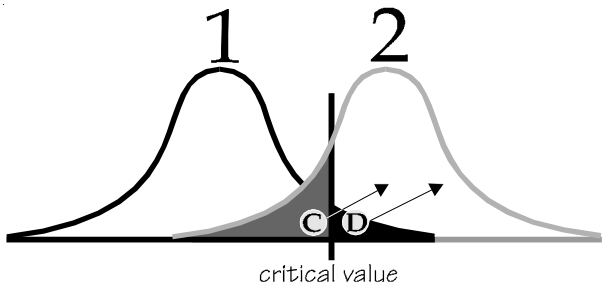
Mean B falls beyond the critical value and is declared "significantly different" from μ. This is a Type 1 error and reflects **box B** (p=α).



In the second diagram at right, we have means C and D which are both **false nulls*** (arrows show they belong with population 2).

Mean C falls to the left of the critical value and is declared "not significantly different" from . This is a Type 2 error, and reflects box C (p=β).

Mean D falls to the right of the critical value and is declared "significantly different" from . This is a correct decision and reflects box D (p=1-β).



Review the decision table and these diagrams until you can see the correspondence between the two.

*Of course, *we can never really know* what the "real world" conditions are, whether the nulls are actually true or false. In the previous examples, I was giving you *hidden information* in order to establish the four possibilities in hypothesis testing. We are left with the tasks of gathering reliable and valid samples of data, applying statistical procedures, and making decisions of outcome based on probability.

But this is still a wonderful mechanism for solving problems. Understanding the

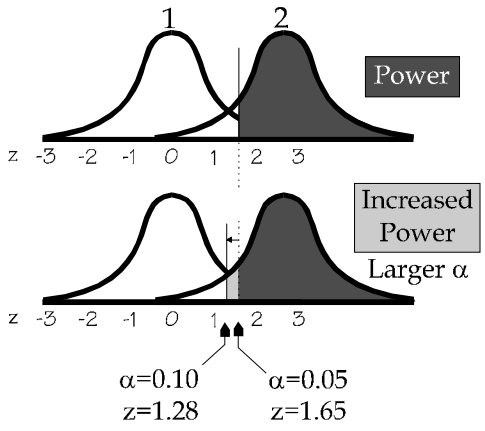
dynamics of hypothesis testing -- error rates, power, z-scores -- is like any other kind of under-the-hood, behind-the-scenes knowledge. It provides insight into how things work, sophistication in what doesn't, and calm assurance that a research design -- whether our own or one found in the literature -- is what it purports to be. Such understanding elevates us from blind user to savvy consumer of research findings. Further, it matures us as a competent research designer.

At the heart of this competence lies the ability to ***improve our chances*** of making a correct decision even before we send out instruments or determine the samples we'll study. A statistician would ask it this way: "How can I increase the power of my study?" Let's take a look at some possibilities.

Increasing Statistical Power

There are four ways to increase the power of your analysis: increase the level of significance (α), increase the difference between population means (μ_1-μ_2), increase the number of subjects studied (N), and decrease the variability of test scores (s). The first two are more theoretical than practical, but they are nonetheless instructive.

Increase α



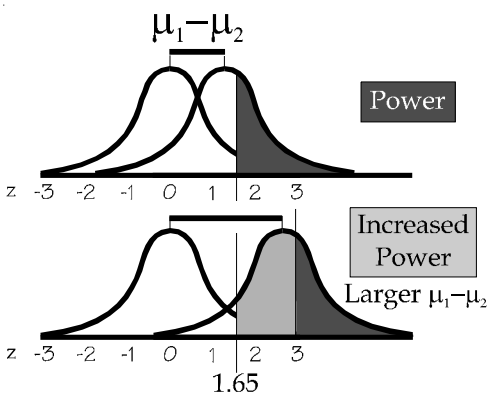
The level of significance (α) directly controls the size of the critical value used to declare a difference "significant." As we increase α, we reduce the critical value. As the critical value is reduced, the likelihood that a mean will be declared "significantly different" from the mean increases. In other words, the probability of declaring a null hypothesis false (power) increases simply because we reduced the critical value. The diagrams at left show how this happens.

Notice the position of the critical value line in the upper diagram at left. It cuts off the curve at z=1.65 (α=0.05).

If we ***increase α to 0.10***, the cut off line moves to the left, to z=1.28. As the critical value line moves to the left, ***the area labelled "power" increases*** by the lighter segment shown in the lower diagram.

However, we have increased power (1-β) by ***increasing α, the Type I error rate.*** This is simply robbing Peter to pay Paul. It does not improve the overall research design to increase the probability of Type 1 errors as we decrease the probability of Type 2 errors . *It is better to remain with the conventional values of 0.05 or 0.01 for α.*

Increase $\mu_1 - \mu_2$



The **power** of a statistical test means is the ***probability of declaring a difference "significant."*** Greater power means nothing more or less than a greater probability of declaring a value "significant." It stands to reason that the probability of declaring a larger difference significant is greater than for declaring a smaller difference significant. The difference between population means in the upper diagram at left is smaller than the difference in the lower diagram. Look at the dramatic difference in power, reflected by the shaded areas in the diagrams.

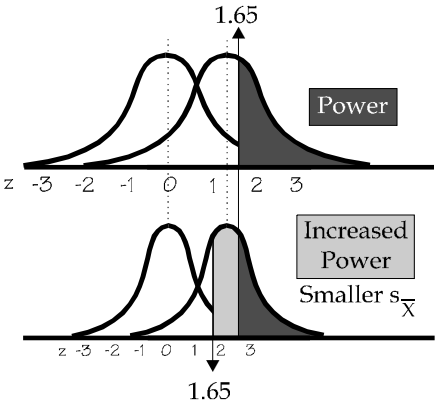
Several years ago a Ford Motor Company commercial featured a road test comparing a Lincoln and a Cadillac. They used ***100 drivers***. The Lincoln won the test (remember, it was a Ford commercial). But interesting to me was that researchers needed 100 persons to show the difference. Why so many? Because the difference between a Cadillac and Lincoln is very small. Had one of the cars been a '67 Chevy taxi cab, the ability to distinguish between cars would have been easier, and the difference could have been firmly established with fewer subjects. It is reasonable to assume that as $\mu_1 - \mu_2$ increases, detecting the difference becomes easier. We can see this very fact in the formula for z itself. The equation for z has the term $\overline{X}-\mu$ ***in the numerator, so that as this difference increases, z increases*** — falling farther out from the mean, and the more likely to be declared significant.

$$z = \frac{\overline{X}-\mu}{s_{\overline{X}}}$$

The problem, of course, is that this discussion is purely theoretical, since we have no control over the size of difference ($\overline{X}-\mu$). So let's turn our attention to elements we do have some control over.

Decrease the Standard Error of the Mean

If we increase z by increasing the numerator, we can by the same formula show that we ***increase z by decreasing the denominator***. In the graphic at right, notice that the difference between the means is the same for both pairs (dotted lines are parallel). The standard error of the mean (variability within the sampling distribution of means) is smaller in the lower diagram, larger in the upper. Now, do you see that "power" (a proportion of the entire curve) is larger in the lower pair than in the upper? *The light-gray area in the lower diagram represents the increase in power* -- the shaded areas in the lower pair cover a larger proportion of the whole (i.e., greater power) than the shaded area in the upper diagram.



$$s_{\overline{X}} = \frac{s}{\sqrt{n}}$$

Reducing the standard error of the mean ($s_{\overline{X}}$) increases z, and therefore the probability of declaring a difference significant. The equation for computing $s_{\overline{X}}$ at left shows the two ways to reduce the standard error. We can either ***decrease s***, the standard deviation of the sample, or ***increase n***, the number of subjects in the group.

Decrease s

The standard error of the mean ($s_{\overline{X}}$) is decreased by **decreasing s**. We do this by *improving the precision and accuracy of the measurements* of our sample(s). By designing better experiments, writing better tests, and using more reliable methods for collecting data, we squeeze some of the noise (extraneous, unsystematic variability) out of our data. It is clear that by gathering data that is more precise, we will be able to detect targeted differences more easily because we are removing unwanted static from the process. Sloppy designs, poor instruments, and awkward data gathering should be replaced by clear designs, accurate and valid instruments, and precise data gathering procedures. ***Decreasing s increases power without increasing Type I error rates.***

Increase n

The second way to decrease $s_{\overline{X}}$ is to increase n by **adding subjects** to our study.

As the number of subjects increases, the more their individual differences ("random noise") cancel each other out and allow true differences to show through.

The size of your sample(s) has a direct influence on the outcome of your study. If you study three approaches to counseling using groups of 10 subjects each, you may not have sufficient statistical power to declare the differences significant, even if they really exist!. The same study, done with three groups of 30, might declare these real differences significant. If you use three groups of 1000, you may find "significant differences" that are, in a practical sense, trivial (see "practical importance" below). Because n is so potent an influence on power, you must use caution in selecting your sample size. You may want to consult an advanced statistics text to determine the size of sample(s) you need for your statistic, but for now, Dr. Curry's "rule of thumb" for sample size (Chapter 7) is a good place to start.

Like Fishing for Minnows

To improve the statistical design of our studies, we must reduce the standard error of the mean by using precise measurements, or increasing sample size, or (preferably) both. Consider with me for a moment the concept of "power" in terms of a minnow net.

If I use a minnow net that has a large mesh, I may *not catch any minnows even if they are present.* This is analogous to conducting a statistical test with low power. The difference may exist in the real world (You really did find the cure for cancer!), but the statistical procedure does not declare it "significant" because the power of the test is too low. (Remember, for a scientist, the only real difference is a statistically significant one).

If I use a minnow net with a fine mesh, I *will catch minnows, if they are there.* This is analogous to conducting a statistical test with high power. If the difference exists, the statistic will declare it to be so. However, if there are no minnows, I will not catch any, no matter how fine the mesh of my net is. Even a high power statistic will not declare "no difference in the real world" as significant.

Maximizing power in a study is a good thing. However, too much power *might declare a trivial difference as significant.* Let's look into this final topic before we close.

Statistical Significance and Practical Importance

If I select a very large sample, I may declare trivial differences "significant." If I select a small sample size, I may declare real differences as "not significant." So why bother with statistical testing?!

The issue is not settled by statistical tests. The *practical importance* of a study is an interpretation of the research study taken as a whole. Practical importance takes into consideration non-statistical factors such as the cost in time and money to implement the procedures in the study. A new reading program may improve reading achievement by 3% (in the context of the experiment, a "significant increase"), but if the cost in materials and personnel is excessive, it may not be "practical" enough to make the change.

Still, hypothesis testing points us in a direction. It provides us with precise tools with which to infer meaning from data. Tools of all kinds can be misused, but this does not invalidate the tool. It encourages us to learn how to use the tools correctly and interpret the results objectively.

Summary

We've come a long way in the last two chapters! We began in Chapter 17 with the standardized z-scale and linked it to the Normal Curve distribution table. We introduced the characteristics of the normal curve. We linked the concepts of z-score and area under the curve. We explained the concept of level of significance (α). We differentiated one- and two-tail statistical tests.

We made the leap from frequency distributions (of scores) to sampling distributions (of means). We related the z-score equation to hypothesis testing with sampling distributions.

In our present chapter we explained and illustrated the concepts of Type I and Type II error rates, as well as power. We tied these concepts to pictorial representations of where these error rates come from, as well as what they mean.

Finally, we described practical ways to improve the statistical design of our studies. ***These two chapters lay the foundation for understanding and using the statistical procedures we'll discuss in the remainder of the book.***

Vocabulary

Alpha	probability of rejecting true null hypothesis (α)
Beta (Type II)	probability of retaining false null hypothesis (ß)
Power	probability of rejecting a false null
Reject null	the decision of statistically significant difference
Retain null	the decision of no statistically significant difference
Type I error rate	probability of rejecting a true null
Type II error rate	probability of retaining a false null

Study Questions

1. Explain the following terms in English:

Type 1 error rate	Practical significance
Type 2 error rate	Statistical significance
Power	Level of significance

2. Draw from memory the 2x2 decision table, labelling the four headings and filling in the boxes with level of significance, error rates and power. Then draw four sets of paired normal curves and identify the areas under the curves which relate to each of the four cells in the decision table.

Sample Test Questions

1. A Type 2 error is the probability of
   A. retaining a true null
   B. rejecting a true null
   C. rejecting a false null
   D. retaining a false null

2. Greater power in a statistic means
   A. more precision
   B. less precision
   C. more differences declared "significant"
   D. fewer differences declared "significant"

3. The best way to increase power in your statistical design is to.
   A. lower the critical value of the test
   B. increase α
   C. increase the standard deviation of scores
   D. use a larger sample

4. You want to be 99% confident of your decision that Sample mean "A" is different from Sample mean "B". Which of the following will allow you to do this?
   A. 1-tail test at α=0.01
   B. 2-tail test at α= 0.99
   C. 1-tail test at α=0.99
   D. 2-tail test at α=0.01

5. T F Results can have statistical significance without having practical significance.

6. T F The best ways to decrease the probability of committing a Type 2 error in your study is to increase n and decrease "noise" in the scores.

7. T F The critical value for a 2-tail test, α=0.01, is $\pm$2.33.