

Chapter 9

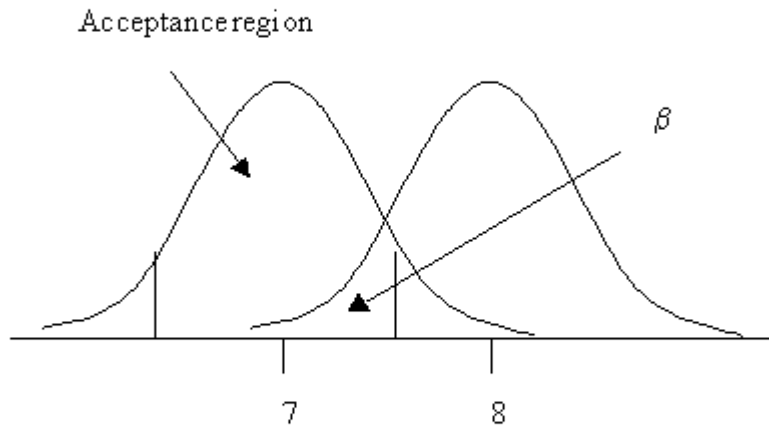
Hypothesis Testing III: Power

Power and the Calculation of β Error

Now that you understand the fundamentals of hypothesis testing, it is possible to fully examine the relationship between Type I and Type II errors. As previously stated, β or Type II errors occur when you fail to reject the null hypothesis when it is false, and another hypothesis is actually true. A clear relationship exists between Type I and Type II errors. As you reduce the probability of committing a Type I error by expanding your confidence limits, you increase the probability of accepting a null hypothesis when it is false – making a Type II error. Likewise, if you decrease the probability of making a Type II error by shrinking your confidence limits, you increase the probability you will reject a null hypothesis when it is true – a Type I error.

This relationship will be easier to understand with an example. Let us begin with two hypothetical distributions, both of which have $\sigma=0.5$ but one which has $\mu=7$ and the other $\mu=8$. Given $\mu=7$, let's say we wish to evaluate the null hypothesis $H_0 : Y_i = 7$ where $H_1 : Y_i = 8$ is an alternate hypothesis. As illustrated in Figure 9.1, there is an overlap between the two distributions.

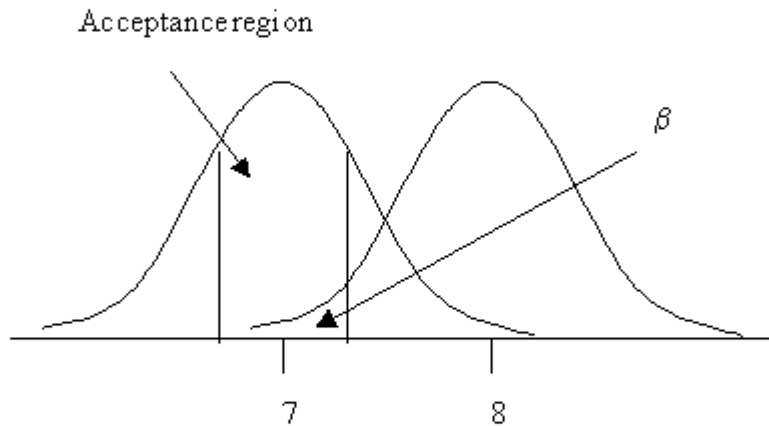
Figure 9.1.



A portion of the distribution with $\mu = 8$ falls within the acceptance region of $H_0 : Y_i = 7$. This area of overlap is β ; the probability of failing to reject the null hypothesis when it is false and $H_1 : Y_i = 8$ is true. If we reject $H_0 : Y_i = 7$, then we conclude that $Y_i \neq 7$ and the overlap is not important to our interpretation. However, if we fail to reject $H_0 : Y_i = 7$, then we must ask how likely is it that Y_i is actually equal to 8.0 instead of 7.0.

If we reduce α , thereby tightening our area of acceptance, we will decrease the amount of overlap between the distribution of $\mu = 8$ and the acceptance region for $H_0 : Y_i = 7$ (Figure 9.2).

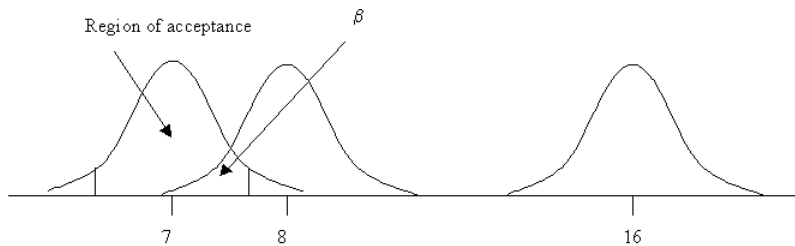
Figure 9.2.



We will also increase the chances of committing a Type I error. If, however, we decrease α to limit that chance of our committing a Type I error, we greatly increase our probability of committing a Type II error.

It should be clear that β is unique for each alternate hypothesis. While only one α level is present for each hypothesis test, β must be calculated for each and every alternate hypothesis. For example, we may have a large amount of overlap between $H_0 : Y_i = 7$ and $H_1 : Y_i = 8$. We may have very little overlap, though, between $H_0 : Y_i = 7$ and $H_2 : Y_i = 16$. β then may be very large between $H_0 : Y_i = 7$ and $H_1 : Y_i = 8$ but almost non-existence between $H_0 : Y_i = 7$ and $H_2 : Y_i = 16$ (Figure 9.3). In fact, the only reason we consider it only “almost” non-existent is because the two distributions are considered hypothetically asymptotic to infinity – an assumption of the theoretical normal distribution.

Figure 9.3.



There is no good rule of thumb dictating which error is more troublesome, or what levels of error are acceptable. Ideally, we would like the probability of making either a Type I or Type II error to be quite small. However, this is not always possible. We must therefore weigh the results of making both types of errors within the context of our specific analysis.

Calculating β

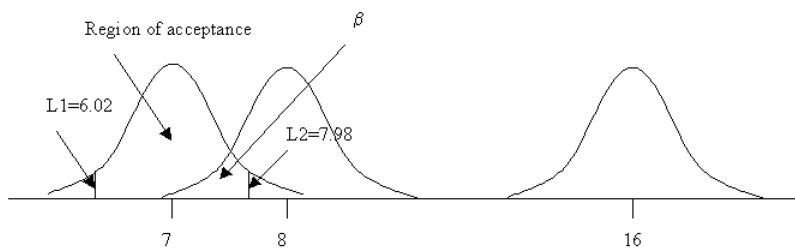
Calculating β is simple. First, we use the z- or the t-distributions to determine the area of the distribution specified in the alternate hypothesis that is within the acceptance limits of our null hypothesis. For example, all three of our hypothetical distributions above ($\mu=7$, $\mu=8$, and $\mu=16$) have standard deviations of 0.5. We can calculate the upper and lower confidence limits for $H_0 : Y_i = 7$ using an $\alpha=0.05$ as $7 \pm 1.96\sigma$.

L_1 is therefore equal to $7.0 - 1.96(0.5) = 6.02$.

L_2 is equal to $7.0 + 1.96(0.5) = 7.98$.

Figure 9.4 illustrates these confidence limits, and their relationships to the means of the other two distributions being considered.

Figure 9.4.



We now have both α (specified by us at .05) and the area of acceptance for $H_0 : Y_i = 7$.

To determine our two β values for $H_1 : Y_i = 8$ and $H_2 : Y_i = 16$ we must identify the areas of $\mu=8$ and $\mu=16$ that fall within the acceptance region for $H_0 : Y_i = 7$. This can be calculated by computing a z-score for the upper limit of the acceptance region of $H_0 : Y_i = 7$ and then solving for the area of the alternate distributions lower than L_2 .

Our z-scores would be $z = \frac{L_2 - \mu}{\sigma}$ or $z = \frac{7.98 - 8.0}{.5} = -.04$ and

$$z_2 = \frac{7.98 - 16.0}{.5} = -16.04.$$

(Remember that positive and negative z-values reflect which only whether the variate we are comparing with a μ , in this case L_2 , if greater or less than μ). Our z-value for our first example states that there are only 0.04 standard deviation units between 8.0 and the upper confidence level for $\mu=7$. In other words, the overlap is quite large. Using Table

A, areas under the normal curve, we find that the tabled value for .04 corresponds to an area under the curve of only .0160. Subtracting .0160 from .50 gives us the area under the curve where our two distributions overlap, or a β of .4840 for $H_1 : Y_i = 8$. In other words, 48.40 % of the distribution represented by $\mu = 8.0$ is within the area of acceptance.

For our second example where $H_2 : Y_i = 16$, our z is 16.04. 16.04 standard deviation units from a mean represents a large distance, and the tabled value is off the chart and not easily calculated. We do know, however, that our chance of making a β error is extremely small, as much less than 0.01% of the distribution represented by $\mu = 16.0$ is within the area of acceptance.

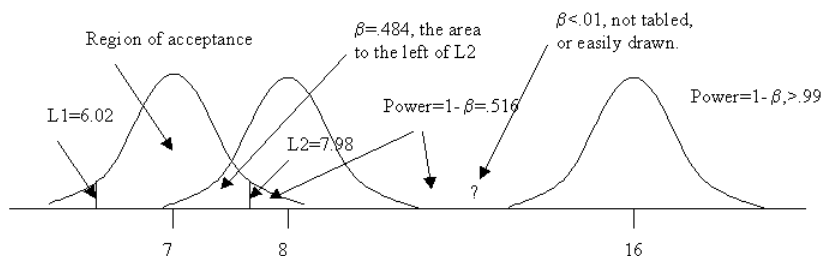
The results of our calculation of β have a couple of interesting consequences. First, if we do not reject the null hypothesis, we cannot conclude that the Y_i likely is not a part of the distribution $\mu = 8$ because the probability of a β error is high. Similarly, can we conclude that Y_i is likely not a part of the distribution $\mu = 16$ since the probability of a β error is very low? The answer is a qualified yes, which leads us to our next topic: power.

An Introduction to the Concept of Power

The power of a statistical test is derived using the formula $\text{Power} = 1 - \beta$. It represents the ability of a particular statistical test to correctly differentiate between two distributions, samples, or populations. Formally, the power of the test is defined as *the probability of rejecting the null hypothesis when the null hypothesis is false and the alternate hypothesis is true*. The greater the power, the more certainty you have that a statistical test can accurately differentiate between the members of different distributions.

The preceding example considering $H_0 : Y_i = 7$ and alternative hypotheses illustrate the relationship between β and power (Figure 9.5).

Figure 9.5.



Using the formula $\text{Power} = 1 - \beta$ we find that the power between the hypotheses $H_0 : Y_i = 7$ and $H_1 : Y_i = 8$ is .516 and the power between the hypotheses $H_0 : Y_i = 7$ and $H_2 : Y_i = 16$ is $>.99$. Thus, we can accurately differentiate only 51.6 percent of the variates belonging to the distribution $\mu=8$ when evaluating the null hypothesis $H_0 : Y_i = 7$. The rest of the time we will be unable to determine if the variates belong to $\mu=7$ or $\mu=8$. In contrast, we can easily differentiate between the variates associated with $\mu=7$ and $\mu=16$ when evaluating $H_0 : Y_i = 7$.

As you can see from the example, if the degree of the overlap of the two distributions is large, the power of a statistical test differentiating between the alternate hypotheses is small (e.g., $H_0 : Y_i = 7$ and $H_1 : Y_i = 8$). If the degree of the overlap of the two distributions is small, the power of a statistical test differentiating between the alternate hypotheses is large (e.g., $H_0 : Y_i = 7$ and $H_2 : Y_i = 16$).

While the example above used the z score, power can just as easily be determined using the t-distribution when samples of means are being compared, or when the population parameter σ is not known.

Increasing the Power of a Test

Ideally, we would like our tests to be as powerful as possible. The ways you can increase the power of your test are the same as the means of tightening your confidence limits previously discussed, as the power of your test is directly controlled by the size of your confidence limits. The best way to improve the power of a statistical test without changing α is to increase the sample size. Increased sample sizes will provide us with a tighter distribution, especially when dealing with a distribution of means. The acceptance areas for the same mean will be tighter, and the alternate distributions will include less overlap. The probability of making a Type II error will decrease and the power of the test will increase.

Additionally, we can occasionally increase a test's power by redefining the null hypothesis. For example, when comparing skeletal populations, we can often distinguish males and females based upon non-metric traits. A distribution of skeletons including both males and females will have a higher standard deviation than either of the distributions of only males or only females. When comparing skeletal populations, we could thus increase the power of our statistical tests by differentiating between males and females, when appropriate.

Calculating Power: An Archaeological Example

An archaeological example will help to demonstrate Power's utility and the impact that a consideration of β can have on our research. The following table (Table 9.1) is a list of the coefficient of variation for the maximum ceramic vessel diameter in samples from groups with specialist and non-specialist production (Crown 1995: 150-151).

Table 9.1. Coefficients of variation for ceramics produced by specialists and non-specialists (from Crown 1995:150-151).

Corrected CV of Maximum Diameter	Group	Form
Specialists		
.04	Paradijon	Small-medium cooking vessels
.03	Paradijon	Medium cooking vessels
.04	Paradijon	Medium-large cooking vessels
.11	Paradijon	Small flower pots
.10	Paradijon	Medium flower pots
.08	Paradijon	Large flower pots
.07	Paradijon	Extra large flower pots
.15	Amphlett Island	Small household cooking vessels
.08	Amphlett Island	Ceremonial cooking vessels
.06	Amphlett Island	Ceremonial cooking vessels
.18	Amphlett Island	Large household cooking vessels
.15	Sacoj Grande	Medium cooking vessels
.06	Sacoj Grande	Medium cooking vessels
.07	Sacojito	Medium water containers
.05	Sacojito	Large water containers

.02	Durazno	Small water containers
.03	Durazno	Medium-large water containers
.05	Durazno	Medium-large water containers
.14	Ticul	Plant pots
.06	Ticul	Decorative vessels
.18	Ticul	Small food bowl
Non-specialists		
.12	Kalinga	Medium vegetable pots
.10	Kalinga	Medium rice bowls
.13	Goodenough Island	Small cooking vessels
.12	Goodenough Island	Small cooking vessels
.12	Goodenough Island	Small cooking vessels
.16	Shipibo-Conibo	Small cooking vessels
.22	Shipibo-Conibo	Medium cooking vessels
.12	Shipibo-Conibo	Large cooking vessels
.16	Shipibo-Conibo	Water containers
.18	Shipibo-Conibo	Water containers
.18	Shipibo-Conibo	Water containers

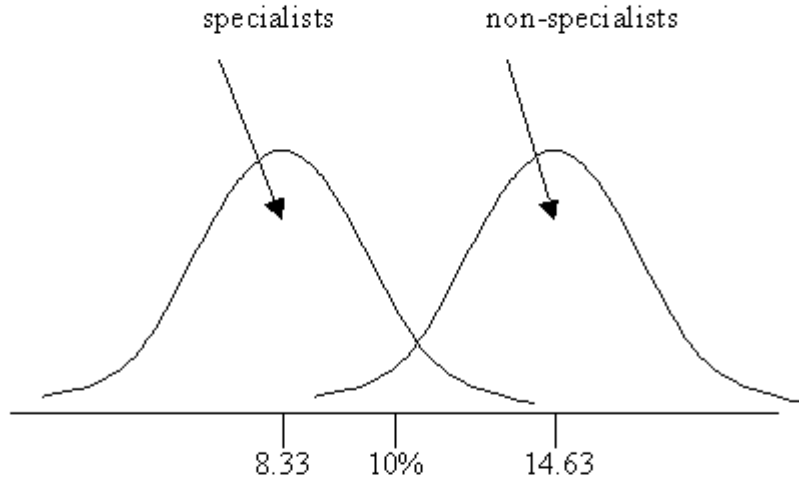
These data have been used to argue that ceramics produced by specialists and non-specialists can be differentiated by the amount of variation within the morphological attributes of the vessels. Specifically, many authors (e.g., Benco 1988; Crown 1994:116, 1995:148-149; Longacre et al. 1988) have suggested that ceramics produced by specialists will be characterized by coefficients of variation less than 10% while those produced by non-specialists will have coefficients of variation greater than 10%. This breaking point does seem intuitively meaningful (most samples made by specialists do have coefficients of variation less than 10% while those made by non-specialists tend to have coefficients of variation greater than 10%), but how certain are we about our conclusions derived using this rule of thumb? We don't know. Perhaps the 10% cut-off provides a very accurate cut-off, but perhaps it doesn't. It *looks like* it ought to be somewhat accurate, but we know there are exceptions and we don't know how often we will be wrong when using 10% as a demarcation. But shouldn't we?

The answer is, of course, yes; we should know what percentage of the time we are likely to come to a wrong conclusion. We need to take another step to ascertain the probability that we could incorrectly classify a sample made by specialists or non-specialists—the probability of making a β error. With β we can calculate the power of the test.

Using the z-score we can easily determine the probability of committing a β error. The β error is the portion of the distribution of specialists with a CV greater than 10% and the proportion of non-specialists with a CV less than 10%, i.e., the area of each distribution in the acceptance region of the other distribution. The average coefficient of variation for the ceramics produced by specialists is 8.33% while the standard deviation is 4.97%. Likewise, the mean and standard deviation of the coefficients of variation produced by

non-specialists are 14.63% and 3.64% respectively. The relationships between the two means and the 10% cutoff value is depicted in Figure 9.6.

Figure 9.6

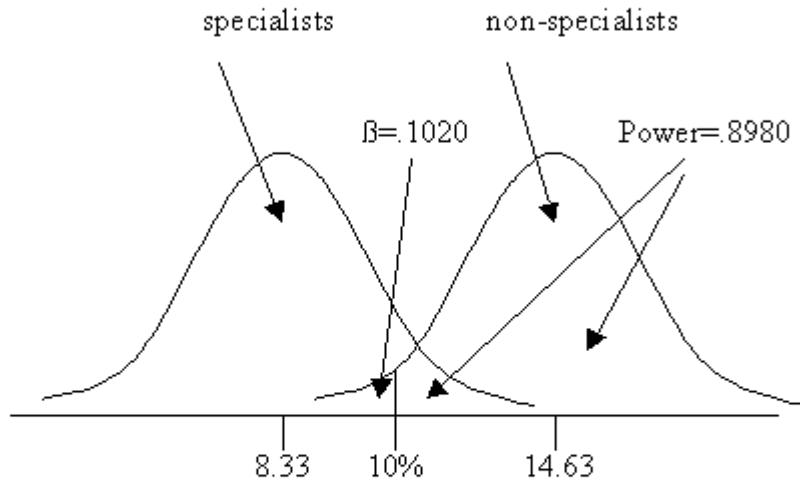


Using the z we can calculate the power of our tests. First, let's determine the probability of concluding that a ceramic was made by a non-specialist, when in actuality it had been made by a specialist. To do so, we use the z -score to calculate the proportion of the non-specialist distribution that is below 10%.

$$z = \frac{Y_i - \mu}{\sigma} = \frac{10 - 14.63}{3.64} = -1.27$$

Appendix A shows us that a z -score of -1.27 corresponds to an area of 0.3980. β is therefore $0.5 - 0.3980 = 0.1020$. The power of the test is $1 - \beta = 1 - 0.1020 = 0.8980$.

These relationships are depicted in Figure 9.7.



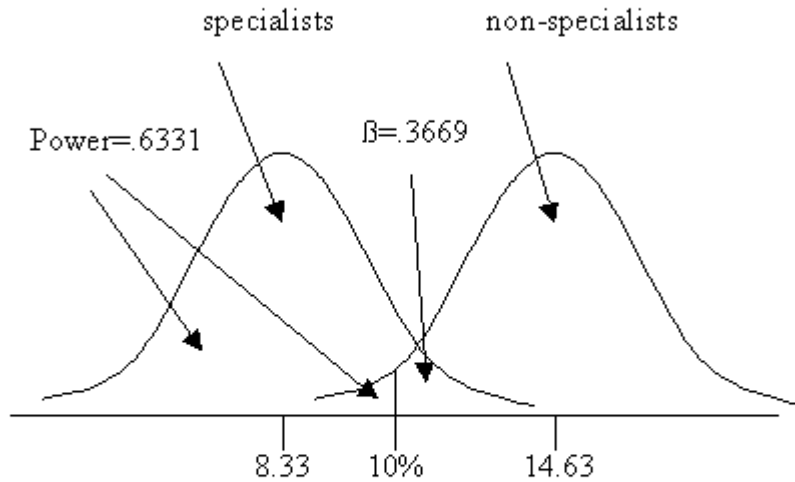
Clearly, the cut-off point of a coefficient of variation of 10% is fairly powerful when distinguishing specialist from non-specialist production. Only 10% of the time will pottery produced by non-specialist production be incorrectly identified as the products of specialists. Roughly 90% of the time this test will be able to successfully differentiate ceramics made by specialists from ceramics made by non-specialists. Is this sufficient power? That is up to the investigator conducting the analysis.

Continuing, we must look at this relationship from the other direction, and use the z-score to determine the probability of concluding that a ceramic was made by a specialist, when in actuality it had been made by a non-specialist. Now, we use the z-score to calculate the proportion of the specialist distribution that is *above* 10%. In this case, the β error is the portion of the distribution of ceramics manufactured by specialists with a coefficient of variation greater than 10% (i.e., the area of specialist distribution in the acceptance region of the non-specialists). Using the z-score, the probability of a variate between \bar{Y} and 10% for the specialists is:

$$z = \frac{10 - 8.333}{4.973} = \frac{1.667}{4.998} = 0.34 .$$

The z-score 0.34 corresponds to an area of 0.1331 from the mean. To determine the area in the distribution of specialists that is greater than 10% then, we must subtract 0.1331 from 0.5. The result β error, 0.3669, is the probability of incorrectly concluding that a non-specialist made a sample of pots when they were actually manufactured by specialists. In other words, 36.69% of the time we would expect to incorrectly conclude ceramic vessels made by specialists were made by non-specialists. The power of the test is $1 - \beta = 1 - 0.3669 = 0.6331$.

These relationships are depicted in Figure 9.8.



In this instance, the 10% cut-off does not produce a very powerful test. A large portion of the time, we will incorrectly conclude that vessels made by specialists were made by non-specialists.

The implications for our use of the coefficients of variation of ceramic morphological attributes to identify ceramics made by specialists and non-specialists are clear. First, we know that we are 90% certain given this data that a coefficient of variation below the cut-off point of 10% represents ceramics made by specialists. We also know, though, that we are far less certain that ceramics with coefficients of variation greater than 10% were

made by non-specialists. Almost 40% of the coefficients of variation of samples of the maximum vessel width of ceramics manufactured by specialists are larger than 10%. This large β error is probably greater than most researchers would feel comfortable with, and would probably prompt most researchers to construct a more powerful test. This could be done in a number of ways.

We could increase the cut-off value from 10% to some higher number such as 12%. However, while this would increase the power of our test for non-specialist production, it would decrease the power of our test when evaluating specialists production. In other words, increasing the cut-off point would decrease the probability we would incorrectly conclude ceramics made by specialists were made by non-specialists, but would increase the probability that we would incorrectly conclude ceramics made by non-specialists were made by specialists.

We could also define a middle area of coefficients of variation which may represent either specialist or non-specialist production. For example, we could continue to conclude that ceramics with a coefficient of variation less than 10% were made by specialists, but then conclude that ceramics with a coefficient of variation between 10% and 12% could be made by either specialists and non-specialists and ceramics with a coefficient of variation greater than 12% were made by non-specialists. Finally, we could keep the 10% cut-off point but rely on corroborating evidence and logical arguments to strengthen our claims that specific samples of ceramics with a coefficient of variation greater than 10% were manufactured by non-specialists. While we do not argue for the appropriateness of any one of these alternatives, such an understanding of the possible errors associated with the use of coefficients of variation to identify the nature of ceramic production is only possible when the power of a statistical test is determined.

Power Curves

While the example provided above presents a real world archaeological example of the use of power, archaeologists often do not have specific alternate hypotheses to compare to our null hypotheses. Even in these cases power can provide insight into the sensitivity and usefulness of our statistical tests.

Using *hypothetical* alternate distributions, archaeologists can calculate a *power curve* that describes the relationship between the null hypothesis being tested and a *range* of alternate hypotheses. This is extremely useful, especially when dealing with certain types of data that have realistic limits to the values that are possible.

Using our example of $H_0 : Y_i = 7$, we will demonstrate the construction of a power curve by comparing $H_0 : Y_i = 7$ to a variety of other possible values for μ (Table 9.2). The power for each of the hypothetical alternate values is computed using the z-score as outlined above. Because the alternate means are hypothetical, we do not have a standard deviation for each to use in the equation. Since we really intend for the hypothetical means to represent possible real world alternate hypotheses, we can use the standard

deviation of our original hypothesis as an estimate of the standard deviations for all of the related distributions.

Table 9.2.

μ_0	deviation	β	power (1- β)
5.0	$\frac{5.5 - 6.02}{.50} = -1.04$.0207	.9793
5.5	$\frac{5.5 - 6.02}{.50} = -1.04$.1492	.8508
6.0	$\frac{6.0 - 6.02}{.50} = -.04$.4840	.5160
6.5	$\frac{6.5 - 6.02}{.50} = .96$.8315	.1685
7.0	$\frac{7.0 - 6.02}{.50} = 1.96$.9500	.0500
7.5	$\frac{7.5 - 7.98}{.50} = .96$.8315	.1685
8.0	$\frac{8.0 - 7.98}{.50} = .04$.4840	.5160
8.5	$\frac{8.5 - 7.98}{.50} = 1.04$.1492	.8508
9.0	$\frac{9.0 - 7.98}{.50} = 2.04$.0207	.9793

Figure 9.9 illustrates the power curve for the hypothetical alternative hypotheses.



Notice that in Figure 9.9 the power of the test is greatest in the tails of the power curve and decreases rapidly as the values of the alternate means approach μ . This is a result of the relationship noted earlier that the power of a test decreases as the amount of overlap between distributions increases. Also notice the effects of sample size on the power of the tests.

While the implications of the power curve may not be intuitively obvious, it may provide valuable insights into the strength of our statistical tests. If our test used to evaluate the null hypothesis is weak over the entire range of possible alternate hypotheses, we must conclude that our test itself is weak and prone to Type II errors. A failure to reject a null hypothesis is not significant, because the variate could be a member of many different distributions other than the one specified in the null hypothesis. If weak, once again we can increase the power of our test in three ways. We can increase our sample size, re-define our null hypothesis, or increase our alpha level (with the associated consequences).

Putting it All Together: A Final Overview of Hypothesis Testing

Now that we understand the use of power and the importance of β , we have all of the tools necessary to efficiently evaluate hypotheses. The process of hypothesis testing can be broken into six steps:

Steps to Hypothesis Testing

1. State the statistical hypothesis in English, and as a formal statistical hypothesis. Be clear in the presentation of what is being measured, how it is being measured, and the validity of the measurement. Be sure that you clearly specify the relationship between the question being asked in the research and the statistical hypothesis.

2. Considering the consequences of making Type I and Type II errors and the relationship of these errors to power. With these in mind, set alpha. State the consequences of making Type I and Type II errors.
3. Select the appropriate statistical procedure.
4. Define the region of rejection, or critical value.
5. Perform the computations and make the appropriate statistical decision.
6. State the decision regarding the null hypothesis in statistical terms and in English. If you fail to reject the null hypothesis, assess the power of the test. If the test is not powerful, come to no conclusions regarding the original hypothesis being tested. If possible, increase n to increase power. Power may also be increased by redefining the original hypothesis to increase the distance between the null hypothesis and other hypotheses. Go to step 1.

Using these steps, you are now able to complete powerful statistical analyses. Now let us review the types of hypotheses tests we may use.

Types of Hypothesis Tests:

Determining if a single observation comes from a population.

To test if a single observation Y_i comes from a population with a mean of μ . We use the familiar formula:

$$z = \frac{Y_i - \mu}{\sigma}$$

If either population parameter is unknown, we may substitute \bar{Y} and/or s, and use the t-distribution as follows.

$$t = \frac{Y_i - \bar{Y}}{s}$$

Determining if a sample mean comes from a population.

Where our population parameters are known, use the following procedure to determine if a sample \bar{Y} could have been drawn from a population:

$$z = \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}}$$

If our population parameter $\sigma_{\bar{Y}}$ is unknown, we substitute $s_{\bar{Y}}$ and use the t-distribution:

$$t = \frac{\bar{Y} - \mu}{s_{\bar{Y}}}$$

Comparing two sample means.

When μ is unknown, there is really no way to determine if a sample mean \bar{Y} can be drawn from a population with a mean of μ . If we substitute a sample \bar{Y} for μ , we are actually comparing two sample means. The comparisons of two sample means is probably the most common statistical comparison done in archaeology. It is so common, and so associated with the t-statistic, that this comparison is sometimes called a t-test, although as we have seen we can use the t distribution in many ways.

We use the following to test for the difference between two means:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{s_{\bar{Y}_1 - \bar{Y}_2}}$$

where:

$$s_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

and

$$s_p = \sqrt{\frac{\sum y_1^2 + \sum y_2^2}{n_1 + n_2 - 2}}$$

Where $s_{\bar{Y}_1 - \bar{Y}_2}$ is the *pooled standard error*, and s_p the *pooled standard deviation*. The pooled standard deviation and pooled standard error allow us to see if the differences between the two means is greater than we might expect when expressed in standard deviation units. s_p allows us to take into account the standard deviation of each sample when calculating the standard error.

Paired T-tests

One frequently misused application of the t-test is the calculation of multiple t-tests to determine if differences exist between multiple means. In the archaeological literature it is common to see multiple comparisons of three or more means. For example, researchers frequently test the hypothesis $H_o : \mu_1 = \mu_2 = \mu_3$ by examining all of the following null

hypotheses: $H_o : \mu_1 = \mu_2$, $H_o : \mu_1 = \mu_3$, and $H_o : \mu_2 = \mu_3$. To use a two-sample test in a multi-sample comparison is *not* valid. The t-test is designed to determine if there is a significant difference between two means. The more means that are compared, the greater likelihood we are likely to erroneously conclude that a significant difference exists between any pair when they could have, in fact, come from the same distribution. In other words, by doing more than one comparison, we are increasing the probability that we will be making a Type I error.

Pearson (1942) calculated the probability of committing a Type I error by using multiple t tests to determine differences in all possible pairs of means. Table 9.3 presents these probabilities.

Number of Means	Alpha					
	0.20	0.10	0.05	0.02	0.01	0.001
2	0.20	0.10	0.05	0.02	0.01	0.001
3	0.41	0.23	0.13	0.05	0.03	0.003
4	0.58	0.36	0.21	0.09	0.05	0.006
5	0.71	0.47	0.29	0.13	0.07	0.009
10	0.96	0.83	0.63	0.37	0.23	0.034
20	1.00	0.98	0.92	0.71	0.52	0.109
∞	1.00	1.00	1.00	1.00	1.00	1.000

To illustrate this table, if we compare 2 means at alpha =.05, the true probability of making a Type I error is 5 times out of 100. However, if we test three means, two at a time, the probability of wrongly concluding that two of the means are estimates of different population parameters increases to .13. By the time 20 means are compared, we are virtually assured (.92 probability) of committing a Type I error. For every level of alpha, our probability greatly increases for each additional comparison. The t-test is clearly inappropriate here. Procedures introduced in the next chapter associated with the analysis of variance (ANOVA) are the appropriate procedure.