# Appendix E

# Statistical Power and Test Sensitivity

## Contents



*Research reports in the literature are frequently flawed by conclusions that state or imply that the null hypothesis is true. For example, following the finding that the difference between two sample means is not statistically significant, instead of properly concluding from this failure to reject the null hypothesis that the data do not warrant the conclusion that the population means differ, the writer concludes, at least implicitly that there is no difference. The latter conclusion is always strictly invalid, and it is functionally invalid unless power is high.*

—J. Cohen (1988)

The power of a statistical test is the probability that if a true difference or effect exists, the difference or effect will be detected. The power of a test becomes important, especially in sensory evaluation, when a no-difference decision has important implications, such as the sensory equivalence of two formulas or products. Concluding that two products are sensorially similar or equivalent is meaningless unless the test has sufficient power. Factors that affect test power include the sample size, alpha level, variability, and the chosen size of a difference that must be detected. These factors are discussed and worked examples given.

## E.1 Introduction

Sensory evaluation requires experimental designs and statistical procedures that are sensitive enough to find differences. We need to know when treatments of interest are having an effect. In food product development, these treatments usually involve changes in food constituents, the methods of processing, or types of packaging. A purchasing department may change suppliers of an ingredient. Product development may test for the stability of a product during its shelf life. In each of these cases, it is desirable to know when a product has become perceivably different from some comparison or control product, and sensory tests are conducted.

In normal science, most statistical tests are done to insure that a true null hypothesis is not rejected without cause. When enough evidence is gathered to show that our data would be rare occurrences given the null assumption, we conclude that a difference did occur. This process keeps us from making the Type I error

discussed in Appendix A. In practical terms, this keeps a research program focused on real effects and insures that business decisions about changes are made with some confidence.

However, another kind of error in statistical decision making is also important. This is the error associated with accepting the null when a difference did occur. Missing a true difference can be as dangerous as finding a spurious one, especially in product research. In order to provide tests of good sensitivity, then, the sensory evaluation specialist conducts tests using good design principles and sufficient numbers of judges and replicates. The principles of good practice are discussed in Chapter 3. Most of these practices are aimed at reducing unwanted error variance. Panel screening, orientation, and training are some of the tools at the disposal of the sensory specialist that can help minimize unwanted variability. Another example is in the use of reference standards, both for sensory terms and for intensity levels in descriptive judgments.

Considering the general form of the *t*-test, we discover that two of the three variables in the statistical formula are under some control of the sensory scientist. Remember that the *t*-test takes this form:

$$t = \text{difference between means/standard error}$$

and the standard error is the sample standard deviation divided by the square root of the sample size ($N$). The denominator items can be controlled or at least influenced by the sensory specialist. The standard deviation or error variance can be minimized by good experimental controls, panel training, and so on. Another tool for reducing error is partitioning, for example in the removal of panelist effects in the complete block ANOVA ("repeated measures") designs or in the paired *t*-test. As the denominator of a test statistic (like a *F*-ratio or a *t*-value) becomes smaller, the value of the test statistic becomes larger and it is easier to reject the null. The probability of observing the results (under the assumption of a true null) shrinks. The second factor under the control of the sensory professional is the sample size. The sample size usually refers to the number of judges or observations. In some ANOVA models additional degrees of freedom can also be gained by replication.

It is sometimes necessary to base business decisions on acceptance of the null hypothesis. Sometimes we conclude that two products are sensorially similar, or that they are a good enough match that no systematic difference is likely to be observed by regular users of the product. In this scenario, it is critically important that a sensitive and powerful test be conducted so that a true difference is not missed, otherwise the conclusion of "no difference" could be spurious. Such decisions are common in statistical quality control, ingredient substitution, cost reductions, other reformulations, supplier changes, shelf life and packaging studies, and a range of associated research questions. The goal of such tests is to match an existing product or provide a new process or cost reduction that does not change or harm the sensory quality of the item. In some cases, the goal may be to match a competitor's successful product. An equivalence conclusion may also be important in advertising claims, as discussed in Chapter 5.

In these practical scenarios, it is necessary to estimate the power of the test, which is the probability that a true difference would be detected. In statistical terms, this is usually described in an inverse way, first by defining the quantity beta as the long-term probability of missing a true difference or the probability that a Type II error is committed. Then one minus beta is defined as the power of the test. Power depends upon several interacting factors, namely the amount of error variation, the sample size, and the size of the difference one wants to be sure to detect in the test. This last item must be defined and set using the professional judgment of the sensory specialist or by management. In much applied research with existing food products, there is a knowledge base to help decide how much a change is important or meaningful.

This chapter will discuss the factors contributing to test power and give some worked examples and practical scenarios where power is important in sensory testing. Discussions of statistical power and worked examples can also be found in Amerine et al. (1965), Gacula and Singh (1984), and Gacula (1991, 1993). Gacula's writings include considerations of test power in substantiating claims for sensory equivalence of products. Examples specific to discrimination tests can be found in Schlich (1993) and Ennis (1993). General references on statistical power include the classic text by Cohen (1988), his overview article written for behavioral scientists (Cohen, 1992) and the introductory statistics text by Welkowitz et al. (1982). Equivalency testing is also discussed at length by Wellek (2003), Bi (2006), and ASTM (2008). Let the

reader note that many scientific bodies have rejected the idea of using test power as justification for accepting the null, and prefer an approach that proves that any difference lies within a specified or acceptable interval. This idea is most applicable to proving the equivalence of measured variables (like the bioequivalence of drug delivery into the bloodstream). However, this equivalence interval approach has also been taken for simple sensory discrimination testing (see Ennis, 2008; Ennis and Ennis, 2009).

## E.2 Factors Affecting the Power of Statistical Tests

### *E.2.1 Sample Size and Alpha Level*

Mathematically, the power of a statistical test is a function of four interacting variables. Each of these entails choices on the part of the experimenter. They may seem arbitrary, but in the words of Cohen, "all conventions are arbitrary. One can only demand of them that they not be unreasonable" (1988, p. 12). Two choices are made in the routine process of experimental design, namely the sample size and the alpha level. The sample size is usually the number of judges in the sensory test. This is commonly represented by the letter "*N*" in statistical equations. In more complex designs like multi-factor ANOVA, "*N*" can reflect both the number of judges and replications, or the total number of degrees of freedom contributing to the error terms for treatments that are being compared. Often this value is strongly influenced by company traditions or lab "folklore" about panel size. It may also be influenced by cost considerations or the time needed to recruit, screen, and/or train and test a sufficiently large number of participants. However, this variable is the one most often considered in determinations of test power, as it can easily be modified in the experimental planning phase.

Many experimenters will choose the number of panelists using considerations of desired test power. Gacula (1993) gives the following example. For a moderate to large consumer test, we might want to know whether the products differ one half a point on the 9-point scale at most in their mean values. Suppose we had prior knowledge that for this product, the standard deviation is about 1 scale point ($S = 1$), we can find the required number of people for an experiment with 5% alpha and 10% beta (or 90% power). This is given by the following relationship:

$$N = \frac{(Z_\alpha + Z_\beta)^2 S^2}{(M_1 - M_2)^2}$$

$$\frac{(1.96 + 1.65)^2 1^2}{(0.5)^2} \cong 52 \qquad \text{(E.1)}$$

where $M_1$–$M_2$ is the minimal difference we must be sure to detect and $Z_\alpha$ and $Z_\beta$ are the *Z*-scores associated with the desired Type I and Type II error limits. In other words, there are 52 observers required to insure that a one-half point difference in means can be ruled out at 90% power when a non-significant result is obtained. Note that for any fractional *N*, you must round up to the next whole person.

The second variable affecting power is the alpha level, or the choice of an upper limit on the probability of rejecting a true null hypothesis (making a Type I error). Usually we set this value at the traditional level of 0.05, but there are no hard and fast rules about this magical number. In many cases in exploratory testing or industrial practice, the concern over Type II error—missing a true difference—are of sufficient concern that the alpha level for reporting statistical significance will float up to 0.10 or even higher. This strategy shows us intuitively that there is a direct relationship between the size of the alpha level and power, or in other words, an inverse relationship between alpha-risk and beta-risk. Consider the following outcome: we allow alpha to float up to 0.10 or 0.20 (or even higher) and still fail to find a significant *p*-value for our statistical test. Now we have an inflated risk of finding a spurious difference, but an enhanced ability to reject the null. If we still fail to reject the null, even at such relaxed levels, then there probably is no true difference among our products. This assumes no sloppy experiment, good laboratory practices, and sufficient sample size, i.e., meeting all the usual concerns about reasonable methodology. The inverse relationship between alpha and beta will be illustrated in a simple example below.

Because of the fact that power increases as alpha is allowed to rise, some researchers would be tempted to raise alpha as a general way of guarding against Type II error. However, there is a risk involved in this, and that is the chance of finding false positives or spurious

random differences. In any program of repeated testing, the strategy of letting alpha float up as a cheap way to increase test power should not be used. We have seen cases in which suppliers of food ingredients were asked to investigate quality control failures of their ingredient submissions, only to find that the client company had been doing discrimination tests with a lax alpha level. This resulted in spurious rejections of many batches that were probably within acceptable limits.

### E.2.2 Effect Size

The third factor in the determination of power concerns the effect size one is testing against as an alternative hypothesis. This is usually a stumbling block for scientists who do not realize that they have already made two important decisions in setting up the test—the sample size and alpha level. However, this third decision seems much more subjective to most people. One can think of this as the distance between the mean of a control product and the mean of a test product under an alternative hypothesis, in standard deviation units. For example, let us assume that our control product has a mean of 6.0 on some scale and the sample has a standard deviation of 2.0 scale units. We could test whether the comparison product had a value of less than 4.0 or greater than 8.0, or one standard deviation from the mean in a two-tailed test. In plain language, this is the size of a difference that one wants to be sure to detect in the experiment.

If the means of the treatments were two standard deviations apart, most scientists would call this a relatively strong effect, one that a good experiment would not want to miss after the statistical test is conducted. If the means were one standard deviation apart, this is an effect size that is common in many experiments. If the means were less than one half of one standard deviation apart, that would a smaller effect, but one that still might have important business implications. Various authors have reviewed the effect sizes seen in behavioral research and have come up with some guidelines for small, medium, and large effect sizes based on what is seen during the course of experimentation with humans (Cohen, 1988; Welkowitz et al., 1982).

Several problems arise. First, this idea of effect size seems arbitrary and an experimenter may not have any knowledge to aid in this decision. The sensory professional may simply not know how much of a consumer impact a given difference in the data is likely to produce. It is much easier to "let the statistics make the decision" by setting an alpha level according to tradition and concluding that no significant difference means that two products are sensorially equal. As shown above, this is bad logic and poor experimental testing. Experienced sensory scientists may have information at their disposal that makes this decision less arbitrary. They may know the levels of variability or the levels important to consumer rejection or complaints. Trained panels will show standard deviations around 10% of scale range (Lawless, 1988). The value will be slightly higher for difficult sensory attributes like aroma or odor intensity, and lower for "easier" attributes like visual and some textural attributes. Consumers, on the other hand will have intensity attributes with variation in the realm of 25% of scale range and sometimes even higher values for hedonics (acceptability). Another problem with effect size is that clients or managers are often unaware of it and do not understand why some apparently arbitrary decision has to enter into scientific experimentation.

The "sensitivity" of a test to differences involves both power and the overall quality of the test. Sensitivity entails low error, high power, sufficient sample size, good testing conditions, good design, and so on The term "power" refers to the formal statistical concept describing the probability of accepting a true alternative hypothesis (e.g. finding a true difference). In a parallel fashion, Cohen (1988) drew an important distinction between effect size and "operative effect size" and showed how a good design can increase the effective sensitivity of an experiment. He used the example of a paired *t*-test as opposed to an independent groups *t*-test. In the paired design subjects function as their own controls since they evaluate both products. The between-person variation is "partitioned" out of the picture by the computation of difference scores. This effectively takes judge variation out of the picture.

In mathematical terms, this effect size can be stated for the *t*-test as the number of standard deviations separating means, usually signified by the letter "*d*". In the case of choice data, the common estimate is our old friend $d'$ (*d*-prime) from signal detection theory, sometimes signified as a population estimate by the

Greek letter delta (Ennis, 1993). For analyses based on correlation, the simple Pearson's *r* is a common and direct measure of association. Various measures of effect size (such as variance accounted for by a factor) in ANOVAs have been used. Further discussion of effect sizes and how to measure them can be found in Cohen (1988) and Welkowitz et al. (1982).

### *E.2.3 How Alpha, Beta, Effect Size, and* N *Interact*

Diagrams below illustrate how effect size, alpha, and beta interact. As an example, we perform a test with a rating scale, e.g., a just-about-right scale, and we want to test whether the mean rating for the product is higher than the midpoint of the scale. This is the simple *t*-test against a fixed value, and our hypothesis is one tailed. For the simple one-tailed *t*-test, alpha represents the area under the *t*-distribution to the right of the cutoff determined by the limiting *p*-value (usually 5%). It also represents the upper tail of the sampling distribution of the mean as shown in Fig. E.1. The value of beta is shown by the area underneath the alternative hypothesis curve to the left of the cutoff as shaded in Fig. E.1. We have shown the sampling distribution for the mean value under the null as the bell-shaped curve on the left. The dashed line indicates the cutoff value for the upper 5% of the tail of this distribution. This would be the common value set for statistical significance, so that for a give sample size (*N*), the *t*-value at the cutoff would keep us from making a Type I error more than 5% of the time (when the null is true). The right-hand curve represents the sampling distribution for the mean under a chosen alternative hypothesis. We know the mean from our choice of effect size (or how much of a difference we have decided is important) and we can base the variance on our estimate from the sample standard error. When we choose the value for mean score for our test product, the *d*-value becomes determined by the difference of this mean from the control, divided by the standard deviation. Useful examples are drawn in Gacula's (1991, 1993) discussion and in the section on hypothesis testing in Sokal and Rohlf (1981).

In this diagram, we can see how the three interacting variables work to determine the size of the shaded area for beta-risk. As the cutoff is changed by changing the alpha level, the shaded area would become larger or

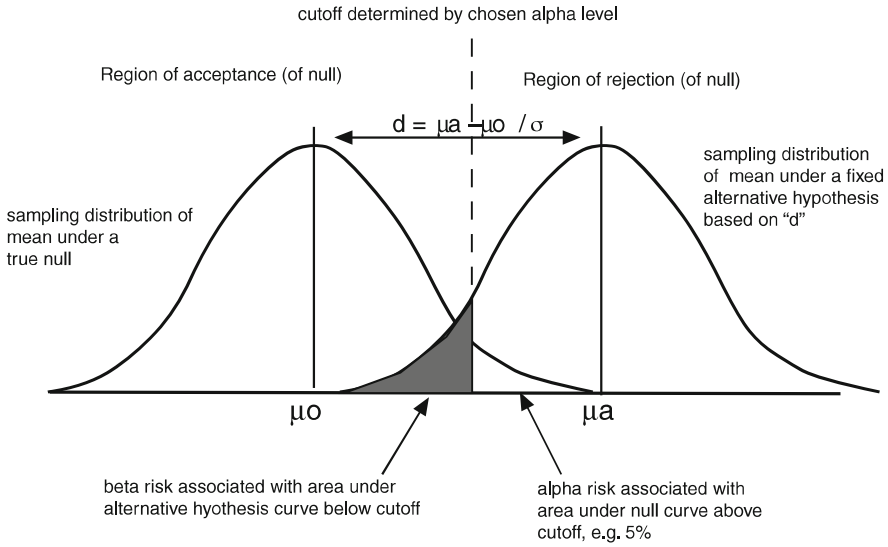


**Fig. E.1** Power shown as the tail of the alternative hypothesis, relative to the cutoff determined by the null hypothesis distribution. The diagram is most easily interpreted as a one-tailed *t*-test. A test against a fixed value of a mean would be done against a population value or a chosen scale point such the midpoint of a just-right scale. The value of the mean for the alternative hypothesis can be based on research, prior knowledge, or the effect size, *d*, the difference between the means under the null and alternative hypotheses, expressed in standard deviation units. Beta is given by the shaded area underneath the sampling distribution for the alternative hypothesis, below the cutoff determined by alpha. Power is one minus beta.

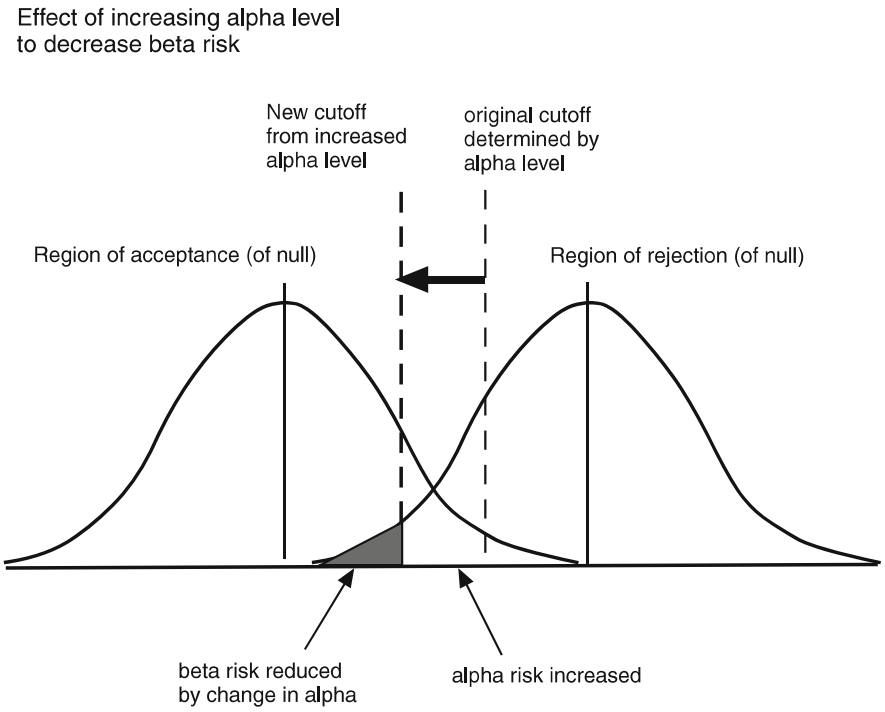


**Fig. E.2** Increasing the alpha level decreases the area associated with beta, improving power (all other variables held equal).

smaller (see Fig. E.2). As the alpha-risk is increased, the beta-risk is decreased, all other factors being held constant. This is shown by shifting the critical value for a significant *t*-statistic to the left, increasing the alpha "area," and decreasing the area associated with beta.

A second influence comes from changing the effect size or alternative hypothesis. If we test against a larger *d*-value, the distributions would be separated, and the area of overlap is decreased. Beta-risk decreases when we choose a bigger effect size for the alternative hypothesis (see Fig. E.3). Conversely, testing for a small difference in the alternative hypothesis would pull the two distributions closer together, and if alpha is maintained at 5%, the beta-risk associated with the shaded area would have to get larger. The chances of missing a true difference are very high if the alternative hypothesis states that the difference is very small. It is easier to detect a bigger difference than a smaller one, all other things in the experiment being equal.

The third effect comes from changing the sample size or the number of observations. The effect of increasing "*N*" is to shrink the effective standard deviation of the sampling distributions, decreasing the standard error of the mean. This makes the distributions taller and thinner so there is less overlap and less area associated with beta. The *t*-value for the cutoff moves to the left in absolute terms.

In summary, we have four interacting variables and knowing any three, we can determine the fourth. These are alpha, beta, "*N*," and effect size. If we wish to specify the power of the test up front, we have to make at least two other decisions and then the remaining parameter will be determined for us. For example, if we want 80% test power (beta = 0.20), and alpha equal to 0.05, and we can test only 50 subjects, then the effect size we are able to detect at this level of power is fixed. If we desire 80% test power, want to detect 0.5 standard deviations of difference, and set alpha at 0.05, then we can calculate the number of panelists that must be tested (i.e., "*N*" has been determined by the specification of the other three variables). In many cases, experiments are conducted only with initial concern for alpha and sample size. In that case there is a monotonic relationship between the other two variables that can be viewed after the experiment to tell us what power can be expected for different effect sizes. These relationships are illustrated below. Various freeware programs are available for estimating power and

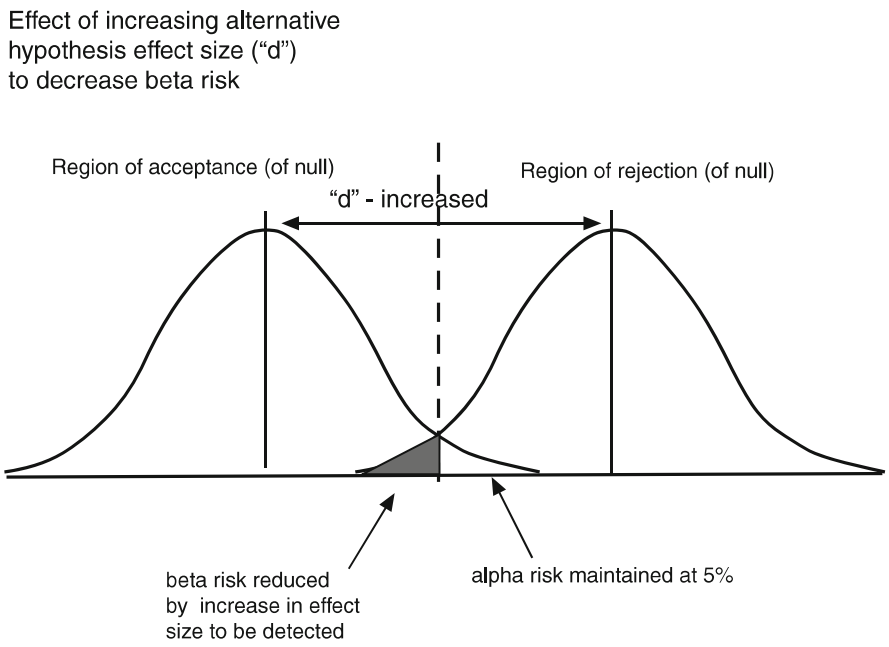


**Fig. E.3** Increasing the effect size that must be detected increases the power, reducing beta. Larger effects (larger *d*, difference between the means of the alternative and null hypotheses) are easier to detect.

sample size (e.g., Erdfelder et al., 1996). Tables for the power of various statistical tests can also be found in Cohen (1988). The R library "pwr" package specifically implements power analyses outlined in Cohen (1988).

## E.3 Worked Examples

### *E.3.1 The* t*-Test*

For a specific illustration, let us examine the independent groups *t*-test to look at the relationship between alpha, beta, effect size, and "*N*." In this situation, we want to compare two means generated from independent groups, and the alternative hypothesis predicts that the means are not equal (i.e., no direction is predicted). Figure E.4 shows the power of the two-tailed independent groups *t*-test as a function of different sample sizes (*N*) and different alternative hypothesis effect sizes (*d*). (Note that *N* here refers to the total sample, not *N* for each group. For very different sample sizes per group, further calculations must be done.) If we set the lower limit of acceptable power at 50%, we can see from these curves that using 200 panelists would allow us to detect a small difference of about 0.3 standard deviations. With 100 subjects this difference must be about 0.4 standard deviations, and for small sensory tests of 50 or 20 panelists (25 or 10 per group, respectively) we can only detect differences of about 0.6 or 0.95 standard deviations, respectively, with 50/50 chance of missing a true difference. This indicates the liabilities in using a small sensory test to justify a "parity" decision about products.

Often, a sensory scientist wants to know the required sample size for a test, so they can recruit the appropriate number of consumers or panelists for a study. Figure E.5 shows the sample size required for different experiments for a between-groups *t*-test and a decision that is two tailed. An example of such a design would be a consumer test for product acceptability, with scaled data and each of the products placed with a different consumer group (a so-called monadic design). Note that the scale is log transformed, since the group size becomes very large if we are looking for small effects. For a very small effect of only 0.2 standard deviations, we need 388 consumers to have a minimal power level of 0.5. If we want to increase power to 90%, the number exceeds

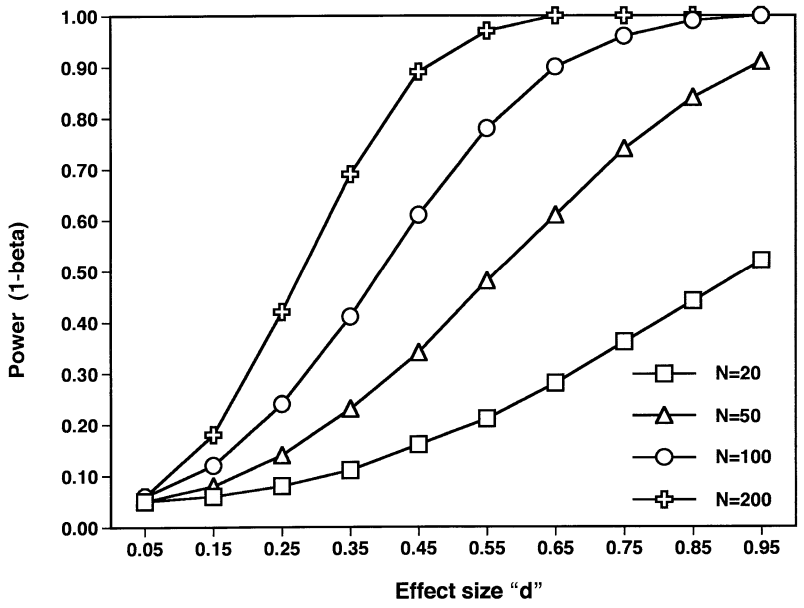


**Fig. E.4** Power of the two-tailed independent groups *t*-test as a function of different sample sizes (*N*) and different alternative hypothesis effect sizes (*d*); the decision is two-tailed at alpha = 0.05. The effect size "*d*" represents the difference between the means in standard deviations. Computed from the GPOWER program of Erdfelder et al. (1996).

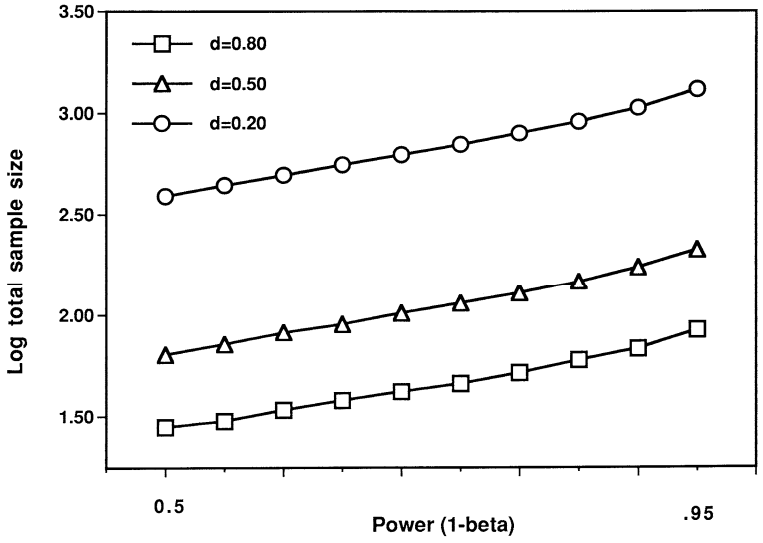


**Fig. E.5** Number of judges required for independent groups *t*-test at different levels of power; the decision is two tailed at alpha = 0.05. Note that the sample size is plotted on a log scale. Computed from the GPOWER program of Erdfelder et al. (1996).

1,000. On the other hand, for a big difference of 0.8 standard deviations (about 1 scale point on the 9-point hedonic scale) we only need 28 consumers for 50% power and 68 consumers for 90% power. This illustrates why some sensory tests done for research and product development purposes are smaller than the corresponding marketing research tests. Market research tests may be aimed at finding small advantages in a mature or optimized product system, and this requires a test of high power to keep both alpha- and beta-risks low.

### *E.3.2 An Equivalence Issue with Scaled Data*

Gacula (1991, 1993) gives examples of calculations of test power using several scenarios devoted to substantiating claims of product equivalence. These are mostly based on larger scale consumer tests, where the sample size justifies the use of the normal distribution (*Z*) rather than the small sample *t*-test. In such an experiment, the calculation of power is straightforward, once

the mean difference associated with the alternative hypothesis is stated. The calculation for power follows this relationship:

$$\text{Power} = 1 - \beta = 1 - \Phi\left[\frac{X_c - \mu_D}{SE}\right] \qquad \text{(E.2)}$$

where $X_c$ represents the cutoff value for a significantly higher mean score, determined by the alpha level. For a one-tailed test, the cutoff is equal to the mean plus 1.645 times the standard error (or 1.96 standard errors for a two-tailed situation). The Greek letter $\Phi$ represents the value of the cumulative normal distribution; in other words we are converting the *Z*-score to a proportion or probability value. Since many tables of the cumulative normal distribution are given in the larger proportion, rather than the tail (as is true in Gacula's tables), it is sometimes necessary to subtract the tabled value from 1 to get the other tail. The parameter $\mu_D$ represents the mean difference as determined by the alternative hypothesis. This equation simply finds the area underneath the alternative hypothesis *Z*-distribution, beyond the cutoff value $X_c$. A diagram of this is shown below.

Here is a scenario similar to one from Gacula (1991). A consumer group of 92 panelists evaluates two products and gives them mean scores of 5.9 and 6.1 on a 9-point hedonic scale. This is not a significant difference, and the sensory professional is tempted to conclude that the products are equivalent. Is this conclusion justified?

The standard deviation for this study was 1.1, giving a standard error of 0.11. The cutoff values for the 95% confidence interval are then 1.96 standard errors, or the mean plus or minus 0.22. We see that the two means lie within the 95% confidence interval so the statistical conclusion of no difference seems to be justified. A two-tailed test is used to see whether the new product is higher than the standard product receiving a 5.9. The two-tailed test requires a cutoff that is 1.96 standard errors above, or 0.22 units above the mean. This sets our upper cutoff value for $X_c$ at 5.9 + 0.22 or 6.12. Once this boundary has been determined, it can be used to split the distributions expected on the basis of the alternative hypotheses into two sections. This is shown in Fig. E.6. The section of the distribution that is higher than this cutoff represents the detection of a difference or power (null rejected) while the section that is lower represents the chance of missing the difference or beta (null accepted).

In this example, Gacula originally used the actual mean difference of 0.20 as the alternative hypothesis. This would place the alternative hypothesis mean at 5.9 + 0.2 or 6.1. To estimate beta, we need to know the area in the tail of the alternative hypothesis distribution to the left of the cutoff. This can be found once we know the distance of the cutoff from out alternative mean of 6.1. In this example, there is a small difference from the cutoff of only 6.1–6.12 or 0.02 units on the original scale, or 0.02 divided by the standard error to give about 0.2 *Z*-score units from the mean of the alternative to the cutoff. Essentially, this mean lies very near to the cutoff and we have split the alternative sampling distribution about in half. The area in the tail associated with beta is large, about 0.57, so power is about 43% (1 minus beta). Thus the conclusion of no difference is not strongly supported by the power under the assumptions that the true mean lies so close to 5.9. However, we have tested against a small difference as the basis for our alternative hypothesis. There is still a good chance that such a small difference does exist.

Suppose we relax the alternative hypothesis. Let us presume that we determined before the experiment that a difference of one-half of one standard deviation on our scale is the lower limit for any practical importance. We could then set the mean for the alternative hypothesis at 5.9 plus one half of the standard deviation (1.1/2 or 0.55). The mean for the alternative now becomes 5.9 + 0.55 or 6.45. Our cutoff is now 6.12–6.45 units away (0.33) or 0.33 divided by the standard error of 0.11 to convert to *Z*-score units, giving a value of 3. This has effectively shifted the expected distribution to the right while our decision cutoff remains the same at 6.12. The area in the tail associated with beta would now be less than 1% and power would be about 99%. The choice of an alternative hypothesis can greatly affect the confidence of our decisions. If the business decision justifies a choice of one half of a scale unit as a practical cutoff (based on one-half of one standard deviation) then we can see that our difference of only 0.2 units between mean scores is fairly "safe" when concluding no difference. The power calculations tell us exactly how safe this would be. There is only a very small chance of seeing this result or one more extreme, if our true mean score was 0.55 units higher. The observed events are

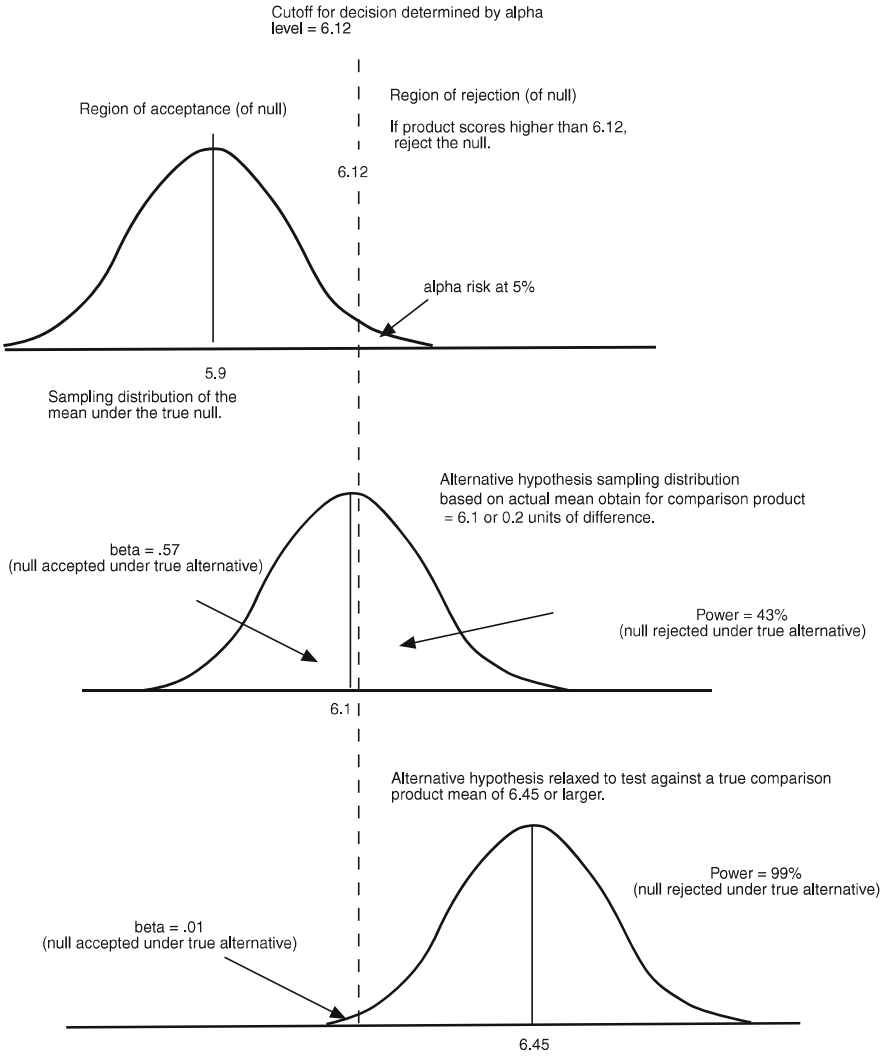


**Fig. E.6** Power first depends upon setting a cutoff value based upon the sample mean, standard error, and the alpha level. In the example shown, this value is 6.12. The cutoff value can then be used to determine power and beta-risk for various expected distributions of means under alternative hypotheses. In Gacula's first example, the actual second product mean of 6.1 was used. The power calculation gives only 43%, which does not provide a great deal of confidence in a conclusion about product equivalence. The lower example shows the power for testing against an alternative hypothesis that states that the true mean is 6.45 or higher. Our sample and experiment would detect this larger difference with greater power.

fairly unlikely given this alternative, so we reject the alternative in favor of the null.

### *E.3.3 Sample Size for a Difference Test*

Amerine et al. (1965) gave a useful general formula for computing the necessary numbers of judges in a discrimination test, based on beta-risk, alpha-risk, and the critical difference that must be detected. This last item is conceived of as the difference between the chance probability, $p_o$ and the probability associated with an alternative hypothesis, $p_a$. Different models for this are discussed in Chapter 5 [see also Schlich (1993) and Ennis (1993)]. For the sake of example, we will take the chance-adjusted probability for 50% correct, which is halfway between the chance probability and 100% detection (i.e., 66.7% for the triangle test).

$$N = \left[\frac{Z_\alpha\sqrt{p_o(1-p_o)} + Z_\beta\sqrt{p_a(1-p_a)}}{|p_o - p_a|}\right]^2 \quad \text{(E.3)}$$

for a one-tailed test (at $a = 0.05$), $Z_\alpha = 1.645$, and if beta is kept to 10% (90% power) then $Z_\beta = 1.28$. The critical difference, $p_o$–$p_a$, has a strong influence on the equation. In the case where it is set to 33.3% for the triangle test (a threshold of sorts) we then require 18 respondents as shown in the following calculation:

$$N = \left[\frac{1.645\sqrt{0.333(.667)} + 1.28\sqrt{0.667(0.333)}}{|0.333 - 0.667|}\right]^2 = 17.03$$

So you would need a panel of about 18 persons to protect against missing a difference this big and limit your risk to 10%. Note that this is a fairly gross test, as the difference we are trying to detect is large. If half of a consumer population notices a difference, the product could be in trouble.

Now, suppose we do not wish to be this lenient, but prefer a test that will be sure to catch a difference about half this big at the 95% power level instead of 90%. Let us change one variable at a time to see which has a bigger effect. First the beta-risk goes from 90 to 95% and if we remain one tailed, $Z\beta$ now equals 1.645. So the numbers become

$$N = \left[\frac{1.645\sqrt{0.333(0.667)} + 1.645\sqrt{0.667(0.333)}}{|0.333 - 0.667|}\right]^2 = 21.55$$

So with the increase only in power, we need four additional people for our panel. However, if we decrease the effect size we want to detect by half the numbers become

$$N = \left[\frac{1.645\sqrt{0.333(0.667)} + 1.28\sqrt{0.667(0.333)}}{|0.167|}\right]^2 = 68.14$$

Now the required panel size has quadrupled. The combined effect of changing both beta and testing for a smaller effect is

$$N = \left[\frac{1.645\sqrt{0.333(0.667)} + 1.645\sqrt{0.667(0.333)}}{|0.167|}\right]^2 = 86.20$$

Note that in this example, the effect of halving the effect size (critical difference) was greater than the effect of halving the beta-risk. Choosing a reasonable alternative hypothesis is a decision that deserves some care. If the goal is to insure that almost no person will see a difference, or only a small proportion of consumers (or only a small proportion of the time) a large test may be necessary to have confidence in a "no-difference" conclusion. A panel size of 87 testers is probably a larger panel size than many people would consider for a triangle test. Yet it is not unreasonable to have this extra size in the panel when the research question and important consequences concern a parity or equivalence decision. Similar "large" sample sizes can be found in the test for similarity as outlined by Meilgaard et al. (1991).

## E.4 Power in Simple Difference and Preference Tests

The scenarios in which we test for the existence of a difference or the existence of a preference often involve important conclusions when no significant effect is found. These are testing situations where acceptance of the null and therefore establishing the power of the test are of great importance. Perhaps for this reason, power and beta-risk in these situations have been addressed by several theorists. The difference testing approaches of Schlich and Ennis are discussed below and a general approach to statistical power is shown in the introductory text by Welkowitz et al. (1982).

Schlich (1993) published risk tables for discrimination tests and offered a SAS routine to calculate beta-risk based on exact binomial probabilities. His article also contains tables of alpha-risk and beta-risk for small discrimination tests and minimum numbers of testers, and correct responses associated with different levels of alpha and beta. Separate tables are computed for the triangle test and for the duo–trio test. The duo–trio table is also used for the directional paired comparison as the tests are both one tailed with chance probability of one-half. The tables showing minimum numbers of testers and correct responses for different levels of beta and alpha in the triangle test are abridged and shown for the triangle test and for the duo–trio test as Tables N1 and N2.

The effect size parameter is stated as the chance-adjusted percent correct. This is based on Abbott's formula, where the chance adjusted proportion, $p_{\mathrm{d}}$,. is based on the difference between the alternative hypothesis percent correct, $p_{\mathrm{a}}$, and the chance percent correct, $p_{\mathrm{o}}$, by the following formula:

$$p_{\mathrm{d}} = \frac{p_{\mathrm{a}} - p_{\mathrm{o}}}{1 - p_{\mathrm{o}}} \tag{E.4}$$

This is the so-called discriminator or guessing model discussed in Chapter 5. Schlich suggests the following guidelines for effect size, that 50% above chance is a large effect (50% discriminators), 37.5% above chance is a medium effect, and 25% above chance is a small effect.

Schlich also gave some examples of useful scenarios in which the interplay of alpha and effect size are driven by competing business interests. For example,

manufacturing might wish to make a cost reduction by changing an ingredient, but if a spurious difference is found they will not recommend the switch and will not save any money. Therefore the manufacturing decision is to keep alpha low. A marketing manager, on the other hand, might want to insure that very few if any consumers can see a difference. Thus they wish to test against a small effect size, or be sure to detect even a small number of discriminators. Keeping the test power higher (beta low) under both of these conditions will drive the required sample size ($N$) to a very high level, perhaps hundreds of subjects, as seen in the examples below.

Schlich's tables provide a crossover point for a situation in which both alpha and beta will be low given a sufficient number of testers and a certain effect size. If fewer than the tabulated number ("$x$") answer correctly, the chance of Type I error will increase should you decide that there is a difference, but the chance of Type II error will decrease should you decide that there is no difference. Conversely, if the number of correct judges exceeds that in the table, the chance of finding a spurious difference will decrease should you reject the null, but the chance of missing a true difference will increase if you accept the null. So it is possible to use these minimal values for a decision rule. Specifically, if the number is less than $x$, accept the null and there will be lower beta-risk than that listed in the column heading. If the number correct is greater than $X$, reject the null and alpha-risk will be lower than that listed. It is also possible to interpolate to find other values using various routines that can be found on the web.

Another set of tabulations for power in discrimination tests has been given by Ennis (1993). Instead of basing the alternative hypothesis on the proportions of discriminators, he has computed a measure of sensory difference or effect size based on Thurstonian modeling. These models take into account the fact that different tests may have the same chance probability level, but some discrimination methods are more difficult than others. The concept of "more difficult" shows up in the signal detection models as higher variability in the perceptual comparisons. The more difficult test requires a bigger sensory difference to obtain the same number of correct judges. The triangle test is more difficult than the three-alternative forced-choice test (3-AFC). In the 3-AFC, the panelist's attention is usually directed to a specific attribute rather than choosing on the odd sample. However, the chance percent correct for both triangle and 3-AFC test is 1/3. The correction for guessing, being based on the chance level, does not take into account the difficulty of the triangle procedure. The "difficulty" arises due to the inherent variability in judging three pairs of differences as opposed to judging simply how strong or weak a given attribute is.

Thurstonian or signal detection models (see Chapter 5) are an improvement over the "proportion of discriminators" model since they do account for the difference in inherent variability. Ennis's tables use the Thurstonian sensory differences symbolized by the lower case Greek letter delta, $\delta$. Delta represents the sensory difference in standard deviations. The standard deviations are theoretical variability estimates of the sensory impressions created by the different products. The delta values have the advantage that they are comparable across methods, unlike the percent correct or the chance-adjusted percent correct. Table E.1 shows the numbers of judges required for different levels of power (80, 90%) and different delta values in the duo–trio, triangle, 2AFC (paired comparison), and 3AFC tests. The lower numbers of judges required for the 2AFC and 3AFC tests arise from their higher sensitivity to differences, i.e., lower inherent variability under the Thurstonian models.

In terms of delta values, we can see that the usual discrimination tests done with 25 or 50 panelists will

**Table E.1** Numbers of judges required for different levels of power and sensory difference for paired comparison (2-AFC), duo–trio, 3-AFC, and triangle tests with alpha = 0.05

| $\delta$ | 2-AFC | Duo–trio | 3-AFC | Triangle |
|---|---|---|---|---|
| 80% power | | | | |
| 0.50 | 78 | 3092 | 64 | 2742 |
| 0.75 | 35 | 652 | 27 | 576 |
| 1.00 | 20 | 225 | 15 | 197 |
| 1.25 | 13 | 102 | 9 | 88 |
| 1.50 | 9 | 55 | 6 | 47 |
| 1.75 | 7 | 34 | 5 | 28 |
| 2.00 | 6 | 23 | 3 | 19 |
| 90% power | | | | |
| 0.50 | 108 | 4283 | 89 | 3810 |
| 0.75 | 48 | 902 | 39 | 802 |
| 1.00 | 27 | 310 | 21 | 276 |
| 1.25 | 18 | 141 | 13 | 124 |
| 1.50 | 12 | 76 | 9 | 66 |
| 1.75 | 9 | 46 | 6 | 40 |
| 2.00 | 7 | 31 | 5 | 26 |

Abstracted from Ennis (1993)

only detect gross differences ($\delta$ >1.5) if the triangle or duo–trio procedures are used. This fact offers some warning that the "non-specific" tests for overall difference (triangle, duo–trio) are really only gross tools that are better suited to giving confidence when a difference is detected. The AFC tests, on the other hand, like a paired comparison test where the attribute of difference is specified (e.g., "pick which one is sweeter") are safer when a no-difference decision is the result.

Useful tables for the power of a triangle test can be found in Chambers and Wolf (1996). A more generally useful table for various simple tests was given by Welkowitz et al. (1982) where the effect size and sample size are considered jointly to produce a power table as a function of alpha. This produces a value we will tabulate as the capital Greek letter delta ($\Delta$, to distinguish it from the lowercase delta in Ennis's tables), while the raw effect size is given by the letter "*d*." $\Delta$ can be thought of as the *d*-value corrected for sample size. The $\Delta$ and *d*-values take the forms shown in Table E.2 for simple statistical tests. Computing these delta values, which take into account the sample size, allows the referencing of power questions to one simple table, (Table E.3). In other words, all of these simple tests have power calculations via the same table.

Here is worked example, using a two-tailed test on proportions (Welkowitz et al., 1982). Suppose a marketing researcher thinks that a product improvement will produce a preference difference of about 8% against the standard product. In other words, he expects that in a preference test, the split among consumers of this product would be something like 46% preferring the standard product and 54% preferring the new product. He conducts a preference test with 400 people, considered by "intuition" to be a hefty sample size and finds no difference. What is the power of the test and what is the certainty that he did miss a true difference of that size?

**Table E.2** Conversion of effect size (*d*) to delta ($\Delta$) value, considering sample size

| Test | *d*-value | $\Delta$ |
|---|---|---|
| One-sample *t*-test | $d = (\mu_1 - \mu_2)/\sigma$ | $\Delta = d\sqrt{N}$ |
| Dependent *t*-test | $d = (\mu_1 - \mu_2)/\sigma$ | $\Delta = d\sqrt{N}$ |
| Independent *t*-test | $d = (\mu_1 - \mu_2)/\sigma$ | $\Delta = d\sqrt{\frac{2N_1N_2}{N_1+N_2}}$ |
| Correlation | $r$ | $\Delta = d\sqrt{N-1}$ |
| Proportions | $\frac{p_o - p_a}{\sqrt{p_o(1-p_o)}}$ | $\Delta = d\sqrt{N}$ |

**Table E.3** Effect size adjusted for sample size to show power as a function of alpha

| Two-tailed alpha | 0.05 | 0.025 | 0.01 | 0.005 |
|---|---|---|---|---|
| One-tailed alpha | 0.10 | 0.05 | 0.02 | 0.01 |
| $\Delta$ | Power | | | |
| 0.2 | 0.11 | 0.05 | 0.02 | 0.01 |
| 0.4 | 0.13 | 0.07 | 0.03 | 0.01 |
| 0.6 | 0.16 | 0.09 | 0.03 | 0.01 |
| 0.8 | 0.21 | 0.13 | 0.06 | 0.04 |
| 1.0 | 0.26 | 0.17 | 0.09 | 0.06 |
| 1.2 | 0.33 | 0.22 | 0.13 | 0.08 |
| 1.4 | 0.40 | 0.29 | 0.18 | 0.12 |
| 1.6 | 0.48 | 0.36 | 0.23 | 0.16 |
| 1.8 | 0.56 | 0.44 | 0.30 | 0.22 |
| 2.0 | 0.64 | 0.52 | 0.37 | 0.28 |
| 2.2 | 0.71 | 0.59 | 0.45 | 0.36 |
| 2.4 | 0.77 | 0.67 | 0.53 | 0.43 |
| 2.6 | 0.83 | 0.74 | 0.61 | 0.51 |
| 2.8 | 0.88 | 0.80 | 0.68 | 0.59 |
| 3.0 | 0.91 | 0.85 | 0.75 | 0.66 |
| 3.2 | 0.94 | 0.89 | 0.78 | 0.70 |
| 3.4 | 0.96 | 0.93 | 0.86 | 0.80 |
| 3.6 | 0.97 | 0.95 | 0.90 | 0.85 |
| 3.8 | 0.98 | 0.97 | 0.94 | 0.91 |
| 4.0 | 0.99 | 0.98 | 0.95 | 0.92 |

Reprinted with permission from Welkowitz et al. (1982), Table H

The *d*-value becomes 0.08 and the delta value is 1.60. Referring to Table E.3, we find that with alpha set at the traditional 5% level, the power is 48% so there is still a 52% chance of Type II error (missing a difference) even with this "hefty" sample size. The problem in this example is that the alternative hypothesis predicts a close race. If the researcher wants to distinguish advantages that are this small, even larger tests must be conducted to be conclusive.

We can then turn the situation around and ask how many consumers should be in the test given the small win that is expected, and the importance of a "no-preference" conclusion? We can use the following relationship for proportions:

$$N = 2\left(\frac{\Delta}{d}\right)^2 \tag{E.5}$$

For a required power of 80% and keeping alpha at the traditional 5% level, we find that a delta value of 2.80 is required. Substituting in our example, we get

$$N = 2\left(\frac{2.80}{0.08}\right)^2 = 2450$$

This might not seem like a common consumer test for sensory scientists, who are more concerned with alpha-risk, but in marketing research or political polling of close races, these larger samples are sometimes justified, as our example shows.

## E.5 Summary and Conclusions

Equations for the required sample sizes for scaled data and for discrimination tests were given by Eqs. (E.1) and (E.3), respectively. The equation for power for scaled data was given in Eq. (E.2). The corresponding equation for choice data from discrimination tests is

$$\text{Power} = 1 - \beta = 1 - \Phi\left[\frac{Z_\alpha\sqrt{p_o(1-p_o)/N} - (p_o - p_a)}{\sqrt{p_a(1-p_a)/N}}\right] \tag{E.6}$$

Table E.4 summarizes these formulae.

A finding of "no difference" is often of importance in sensory evaluation and in support of product research. Many business decisions in foods and consumer products are made on the basis of small product changes for cost reduction, a change of process variables in manufacturing, a change of ingredients or suppliers. Whether or not consumers will notice the change is the inference made from sensory research. In many cases, insurance is provided by performing a sensitive test under controlled conditions. This is the philosophy of the "safety net" approach, paraphrased as follows: "If we do not see a difference under controlled conditions using trained (or selected, screened, oriented, etc.) panelists, then consumers are unlikely to notice this change under the more casual and variable conditions of natural observation." This logic depends upon the assumption that the laboratory test is in fact more sensitive to sensory differences than the consumer's normal experience. Remember that the consumer has extended opportunities to observe the product under a variety of conditions, while the laboratory-based sensory analysis is often limited in time, scope, and the conditions of evaluation.

As stated above, a conclusion of "no difference" based only on a failure to reject the null hypothesis is not logically airtight. If we fail to reject the null, at least three possibilities arise: First, there may have been too much error or random noise in the experiment, so the statistical significance was lost or swamped by large standard deviations. It is a simple matter to do a sloppy experiment. Second, we may not have tested a sufficient number of panelists. If the sample size is too small, we may miss statistical significance because the confidence intervals around our observations are simply too wide to rule out potentially important sensory differences. Third, there may truly be no difference (or no practical difference) between our products. So a failure to reject the null hypothesis is ambiguous and it is simply not proper to conclude that two products are sensorially equivalent simply based on a failure to reject the null. More information is needed.

One approach to this is experimental. If the sensory test is sensitive enough to show a difference in some other condition or comparison, it is difficult to argue that the test was simply not sensitive enough to find any difference in a similar study. Consideration of a track record or demonstrated history of detecting differences with the test method is helpful. In a particular laboratory and with a known panel, it is reasonable to conclude that a tool, which has often shown differences in the past, is operating well and is sufficiently discriminative. Given the history of the sensory procedure under known conditions, it should be possible to use this sort of common sense approach to minimize risk in decision making. In an ongoing sensory testing program for discrimination, it would be reasonable to use a panel of good size (say 50 screened testers), perform a replicated test, and know whether the panel had shown reliable differences in the past.

Another approach is to "bracket" the test comparison with known levels of difference. In other words, include extra products in the test that one would expect to be different. Baseline or positive and negative control comparisons can be tested and if the panel finds significant differences between those benchmark

**Table E.4** Sample size and power formulas (see text for details)

| Form of data | Sample size | Power |
|---|---|---|
| Proportion or frequency | $N = \left[\frac{Z_\alpha\sqrt{p_o(1-p_o)}+Z_\beta\sqrt{p_a(1-p_a)}}{\lvert p_o-p_a\rvert}\right]^2$ | $1-\Phi\left[\frac{Z_\alpha\sqrt{p_o(1-p_o)/N}-(p_o-p_a)}{\sqrt{p_a(1-p_a)/N}}\right]$ |
| Scaled or continuous | $N = \frac{(Z_\alpha+Z_\beta)^2S^2}{(M_1-M_2)^2}$ | $1-\Phi\left[\frac{X_c-\mu_D}{SE}\right] = 1-\Phi\left[\frac{Z_\alpha(SE)-\mu_D}{SE}\right]$ |