APPENDIX

A  The Logic of Power Calculations

A.1 What is “power”?  

The concept of “power” originates from the theory of statistical tests. The purpose of hypothesis testing is to verify the plausibility of the null hypothesis in the light of experimental data. Depending on the outcome, the null hypothesis either will or will not be rejected. With power we mean the probability of correctly rejecting a false null hypothesis. The primary purpose of power calculations is to determine the sample size that allows us to achieve the desired level for power.

To make the above statements clearer we have to go back to the basics of the theory of statistical tests.

A.2 Hypothesis testing  

A.2.1 Significance tests  

The common approach to significance testing is to start with a hypothesis about a population parameter, called the null hypothesis. Data are then collected and the plausibility of the null hypothesis is determined in light of the data via a suitable test statistic. The measure used to assess how different the data are from what would be expected under the assumption that the null hypothesis is true is the p-value. It is defined as the probability of observing a value of the test statistics that is at least as unlikely under the null hypothesis as the observed one. The criterion used for rejecting the null hypothesis is the significance level indicated by the Greek letter \( \alpha \). Traditionally, either the 0.05 level (called the 5% level) is used or the 0.01 level (1% level), although the choice of levels is largely subjective. The p-value is compared with the significance level; if it is less than or equal to \( \alpha \), then the null hypothesis is rejected; otherwise the null hypothesis it is not rejected.\(^1\)

Example 1.0: Suppose that a pharmaceutical industry wants to assess the effect of a new hypertensive drug they produce. The null hypothesis they want to test is that \( \mu_2 - \mu_1 = 0 \) where \( \mu_1 \) is the mean level of blood pressure of patients treated with a well established drug and \( \mu_2 \) is the mean level for patients treated with the new drug. Thus, the null hypothesis concerns the parameter \( \delta = \mu_2 - \mu_1 \), and the null hypothesis is that the parameter equals zero. The significance level is fixed at 0.05.

A possible experimental design\(^2\) for verifying the above null hypothesis is to administer the new drug to a group of \( N \) patients who achieved good results with the traditional treatment after a suitable wash-out period has passed. Denote by \( y_{i1} \) and \( y_{i2} \) the blood pressure measured respectively before and after the new treatment for the \( i \)th patient. Let \( d_i = y_{i2} - y_{i1} \) be the corresponding change in response. Suppose it is sensible to assume a Gaussian model for the distribution of blood pressure in both periods. Then under the null hypothesis \( d_i \) has a normal distribution with mean 0 and some unknown standard deviation \( \sigma \). Let \( \bar{d} \) and \( s_d \) be the sample mean and standard deviation of the differences \( d_i \). Under the null hypothesis, and provided that the \( N \) subjects are independent, the statistic \( T = \frac{\bar{d}}{s_d \sqrt{N}} \) has a Student t distribution with \( N-1 \) degrees of freedom.\(^3\)

Suppose that for a group of \( N = 40 \) patients we obtained the following values: \( \bar{d} = 3.67 \) and \( s_d = 12.42 \). The observed value of the test statistic is \( t_{obs} = 1.87 \). The probability of observing a result at least as extreme as the observed one is \( Pr( T \geq t_{obs} ) = 0.035 \), where \( T \) is a student t distribution with \( N-1=39 \) degrees of freedom. The null hypothesis is hence rejected at the specified 5% level.

The lower the significance level, the more the data must diverge from the null hypothesis to be significant. Therefore, the 1% level is more conservative than the 5% level.

Example 1.1 (cont.): Note that if the significance level in the above example were 0.01, we would not reject the null hypothesis. That is, at the 5% level the evidence provided by the data is strong enough to say that the value \( \bar{d} = 3.67 \) is not simply due to randomness, but originates from a distribution with mean value larger than zero. At the 1% level, we cannot exclude that \( \bar{d} = 3.67 \) is a random observation from a central Student t distribution with 39 degrees of freedom.

When the null hypothesis is rejected; the outcome is said to be “statistically significant”; otherwise it is said to be “statistically not significant”.

A.2.2 Null hypothesis and alternative hypothesis  

In order to compute the p-value we need to identify the values of the test statistic that are “at least as unlikely under the null hypothesis as the observed one”. Hence, in hypothesis testing we are not only asked to formulate the null hypothesis, but also must have an idea of which hypothesis, called the alternative hypothesis, the null hypothesis should be contrasted with. Traditionally the symbols \( H_0 \) and \( H_1 \) are used to indicate respectively the null hypothesis and the alternative hypothesis.

Example 1.2 (cont.): For calculating the p-value in Example 1.1 we considered all values larger than \( t_{obs} \). This is, because the new drug is supposed to be as effective in lowering the blood pressure as the established one. Implicitly, we identified failure of the new treatment with its inability of lowering the

---

\(^1\) Note that not rejecting a null hypothesis does not automatically imply that we accept it. It only means that the data do not provide enough evidence to reject it. An experiment might yield evidence sufficient to reject the null hypothesis. But no experiment can demonstrate that the null hypothesis is true, as the true value of the population parameter remains unknown.

\(^2\) This is for illustrative purposes only. Better designs are available.

\(^3\) This test statistic is known as paired t test.
blood pressure. If this is the case, the mean \( \delta = \mu_2 - \mu_1 \) would not be zero, but larger. The notation is: \( H_0: \delta = 0 \) against \( H_1: \delta > 0 \). Under \( H_1 \) the t statistic would still follow a Student t distribution with \( N-1 \) degrees of freedom but centered at a value larger than zero. It follows that the values that are unlikely under the null hypothesis, but likely if this fails, are in the right-hand tail of the distribution as shown in Figure a). The p-value we computed is a one-tailed p-value and corresponds to the shaded area in the figure.

The null hypothesis is the one we are interested in. It is quite often a hypothesis of no difference as in our guiding example. That is why the word “null” is used, although there are many occasions when the parameter is not hypothesized to be 0. The alternative hypothesis includes all possible distributions except the null. For the example just given, these are the distributions that are shifted to the right with respect to the curve given in Figure a). Some examples are given in Figure b) where the bold curve corresponds to the null distribution.

Although the interest usually focuses on the null hypothesis, the formulation of the alternative hypothesis is important to determine how to calculate the p-value.

Example 1.3 (cont.): Suppose that the new drug may not only fail in lowering the blood pressure of the patients \( (\delta > 0) \) but also in that it lowers it too much \( (\delta < 0) \). In this case \( H_0: \delta = 0 \), while \( H_1: \delta \neq 0 \). Under the second scenario \( (\delta < 0) \) we may observe values that fall into the left-hand tail of the reference distribution. As shown in Figure c), the p-value is now given by the sum of the two tail probabilities identified by the shaded area, i.e., \( 2 \cdot Pr (|T| > t_{0.05}) = 0.070 \), where \( T \) is still a student t distribution with \( 39 \) degrees of freedom. This time the null hypothesis cannot be rejected at the 5% level. Instead of using a one-tailed p-value we used a two-tailed p-value.

A.2.3 Type I and Type II errors

Hypothesis testing is a method of inferential statistics and hence subject to randomness. There are two kinds of errors that may occur: (1) a true null hypothesis can be incorrectly rejected and (2) a false null hypothesis can fail to be rejected.

Example 1.4 (cont.): Suppose that a second experiment on further \( N = 40 \) patients yielded the following values: \( \bar{y} = 4.37 \) and \( s_y = 9.1 \). The observed value of the test statistic is equal to \( t_{obs} = 3.04 \). The two-sided p-value is now 0.0021. The data are highly significant, and the null hypothesis can be rejected at the 1% level. Which experiment should we trust?

If we reject the null hypothesis when it is true we make a Type I error. The probability of making a Type I error is denoted by \( \alpha \). Note that it is the same than the significance level of the test. If we do not reject the null hypothesis when the alternative hypothesis is true, we make a Type II error. The probability of making a Type II error is denoted by \( \beta \). The possible scenarios are summarized in the table below.

A Type II error is only an error in the sense that an opportunity to reject the null hypothesis correctly was lost. It is not an error in the sense that an incorrect conclusion was drawn since no conclusion is drawn when the null hypothesis is not rejected. A Type I error, on the other hand, is an error in every sense of the word. A conclusion is drawn that the null hypothesis is false when, in fact, it is true. This explains why Type I errors are generally considered more serious than Type II errors, and why the probability of a Type I error, the significance level, is set by the experimenter.

A.2.4 Power

It may happen that the null hypothesis is the reverse of what the experimenter actually believes: researchers very frequently put forward a null hypothesis in the hope that they can discredit it. To make this clearer, let us slightly change our guiding example.

Example 2.0: Suppose that a researcher is interested in whether the time to respond to a tone is affected by the consumption of alcohol. The null hypothesis concerns again the parameter \( \delta = \mu_2 - \mu_1 \), where \( \mu_2 \) is the mean time to respond after consuming alcohol and \( \mu_1 \) is the mean time to respond otherwise, and the null hypothesis is that the parameter equals zero. However, the experimenter probably expects alcohol to have a harmful effect. If the experimental data show a sufficiently large effect of alcohol, then the null hypothesis that alcohol has no effect can be rejected. In this view, making a Type II error is more serious than making a Type I error.

As mentioned in the beginning, the power of a test is the probability of correctly accepting the alternative hypothesis when it is true. This probability equals \( 1 - \beta \). That is, it equals the complement of the Type II error rate. If the power of an experiment is low, then there is a good chance that the experiment will be inconclusive. That is why it is so important to consider power in the design of experiments.

Several methods are available for estimating the power of an experiment before the
Example 1.7 (cont.): Let us suppose again that $\delta = 2$, $\sigma_d = 10$, $N = 40$, and that we want to test $H_0: \delta = 0$ against $H_1: \delta > 0$. The significance level $\alpha$ this time is 1%. By construction, the null hypothesis is not rejected for values of the t statistic smaller than 2.43 corresponding to the 99% quantile of a central Student t distribution with 39 degrees of freedom. The Type II error probability becomes $\beta = \Pr (T_{39} \leq 2.43) = 0.067$ where $T_{39}$ is a Student t distribution with $N-1 = 39$ degrees of freedom centered at $\Delta = 1.26$. The power is equal to $1 - \beta = 0.933$.

Requiring very strong evidence to reject the null hypothesis makes it very unlikely that a true null hypothesis will be rejected. However, it increases the chance that a false null hypothesis will not be rejected. A graphical way of summarizing the Type I and Type II error rates of a particular test statistic is by using a ROC (receiver operating characteristics) curve as shown in Figure e). The x-axis reports the Type I error rate (or significance level) $\alpha$, while on the y-axis we find $1-\beta$, that is the power of the test statistic.

The farther away the curve is from the straight line (corresponding to a random guess), the better is the test statistic. $^5$

If the power is too low, then the experiment can be re-designed by changing one of the factors that determine power.

A.3.2 Factors that affect power

The factors affecting power will be illustrated in terms of the hypertensive drug example. Notice that the same considerations apply to test statistics in general.

a) Difference between population means

The size of the difference between the population means, $\delta$ in our example, is an important factor in determining the power of a test statistic. Naturally, the more the means differ from each other, the easier it is to detect the difference. As shown in Figure f), the farther the value of $\delta$ is from zero, keeping the standard deviation $\sigma_d$ fixed, (hence the larger $\Delta$ is), the smaller is the Type II error rate ($\beta$), and, therefore, the larger is the power for a fixed value $\alpha$. The experimenter usually has little control over the size of the effect, although sometimes he/she can choose the levels of the independent variable to increase the size of the effect.

$^5$ This corresponds to finding the optimal test statistic, that is, the test statistic that minimizes the Type II error rate for a given significance level. However, as this problem is very hard to solve, optimal test only exist for very special problems. One example is the paired t test statistic used as guiding example in the text.
b) Standard deviation

The larger the standard deviation $\sigma$, the lower the power. Suppose we keep $\delta$ fixed at 2 and let $\sigma$ assume the values 10, 15, 20 and 30. The corresponding ROC curves are shown in Figure g). There are ways that an experimenter can reduce variance to increase power. A second way to reduce variance is by using a within-subjects design. In these designs, the overall level of performance of each person is subtracted out.

c) Significance level

A further important factor affecting power is the significance level $\alpha$. The more conservative (lower) $\alpha$ is chosen, the lower the power. As we saw in Example 1.7, using the 1% level will result in lower power than using the 5% level. The cost of stronger protection against Type I errors is more Type II errors.

d) Sample size

Increasing N, the sample size, increases power. In fact, the cut-off point and the variance of the distribution of the test statistic change as N increases, while the mean of the distribution does not. The effect of sample size on power for the hypertensive drug example is shown in more detail in Figure h). Choosing a sample size is a difficult decision. There is a trade-off between the gain in power and the time and cost of testing a large number of subjects.

e) Other factors

Several other factors affect power. One factor is whether the population distribution is normal. Deviations from the assumption of normality usually lower power. A second factor affecting power is the type of statistical procedure used. Some of the distribution-free tests are less powerful than other tests when the distribution is normal, but more powerful when the distribution is highly skewed. One-tailed tests are more powerful than two-tailed tests as long as the effect is in the expected direction. Otherwise, their power is zero. The choice of an experimental design can have a profound effect on power. Within-subject designs are usually much more powerful than between-subject designs.

A.3.3 Sample size calculations

Although there are several factors that may affect the power of a test statistic, most of them cannot be easily influenced by the experimenter. The difference of sample means $\delta$ is usually the quantity he/she is interested in detecting, rather than a means to regulate the power of the test statistic. The significance level $\alpha$ is usually kept fixed to the two common values .05 or .01. Also, it may not always be possible to reduce the standard deviation $\sigma$, by changing the experimental design. The only quantity that can easily be accessed is the sample size N. Specified the power one wishes to achieve, the sample size needed for that level of power can be estimated. This is exactly what power calculations do.

Example 1.8 (cont.): Given the significance level $\alpha$, let us determine the sample size N that gives a power of 1- $\beta$ against $H^1: \delta > 0$ in the hypertensive drug example.

As mentioned several times, the null hypothesis is not rejected at the level $\alpha$ if the observed value of the t statistic is smaller than $z_{1-\alpha}$, the 1- $\alpha$ quantile of the central Student t distribution with N-1 degrees of freedom. Power is defined as 1- $\beta$ = $\Pr (T > z_{1-\alpha})$, where $T_d$ is a Student t distribution with N-1 degrees of freedom centered at $\delta = N \delta / \sigma_d$. This yields an equation in N whose solution is approximated by

$$N = (z_{1-\alpha} + z_{\beta})^2 \sigma_d^2 / \delta^2 = (z_{1-\alpha} + z_{\beta})^2 / CV^2,$$

(*)

where $CV=\delta / \sigma_d$ is the coefficient of variation of the d's.6

If instead of performing a one-sided test we performed a two-sided test ($H: \delta \neq 0$), formula (*) would be changed into

$$N = (z_{1-\alpha/2} + z_{\beta})^2 \sigma_d^2 / \delta^2 = (z_{1-\alpha/2} + z_{\beta})^2 / CV^2.$$

(**)

As mentioned at the end of Section A.2.4, two-sided tests are less powerful than one-sided, and a larger sample size is consequently needed to achieve the same level of power.

As illustrated by the above example, the determination of the sample size depends on the specified significance level $\alpha$ and the desired power $\beta$ through the two quantities $z_{1-\alpha}$ and $z_{\beta}$, but also on the effect size, measured in terms of the mean $\delta$ and the standard deviation $\sigma_d$ of the response. Hence, the first and most difficult step in choosing the sample size is to estimate the size of the effect. Estimating the effect size includes estimating the population variance as well as the population means. Information may be retrieved from different sources.

1. If there are published experiments similar to the one to be conducted, then the effects obtained in these published studies can be used as a guide. There is, however, a need for caution, since there is a tendency for published studies to contain overestimates of effect sizes. Often previous studies are not sufficiently similar to a new study to provide a valid basis for estimating the effect size. In this case, it is possible to specify the minimum effect size that is considered important.

---

6 The details are omitted.
2. Frequently, it is easiest to specify the effect size in terms of the number of standard deviations separating the population means. Thus, one might find it easier to estimate that the population mean for the experimental group is 0.5 standard deviations above the population mean for the control group, that is to estimate the coefficient of variation CV, than to estimate the two population means and the population variance.

As shown by formulae (*) and (**), the power remains the same for any experiment in which the coefficient of variation CV is constant, that is, where the difference between population means is the same number of population standard deviations apart. **Power curves** are used to report the power of the test statistic as a function of the different values of the parameter under the alternative hypothesis. Figure i) report the power curves for testing $H_0: \delta = 0$ against $H_1: \delta > 0$ in the hypertensive drug example. On the x-axis we put the effect size defined through the coefficient of variation CV, while the y-axis gives the sample size N. The five curves correspond to the powers 0.5, 0.8, 0.85, 0.9, 0.95 and 0.99; the significance level $\alpha$ is 5%.

A.4 Final remarks

It is important to keep in mind that power is not about whether or not the null hypothesis is true; in fact it is assumed to be false. The focus is on the probability that the data gathered in an experiment will be sufficient to reject the null hypothesis. Power calculation try to answer the question: If the null hypothesis is false with specified population means and standard deviation, what is the probability that the data from the experiment will be sufficient to reject the null hypothesis? If the experimenter discovers that the probability of rejecting the null hypothesis is low (low power) even if the null hypothesis is false to the degree expected, then it is likely that the experiment should be redesigned. Otherwise, considerable time and expense will go into a project that has a small chance of being conclusive even if the theoretical ideas behind it are correct.

A question that arises spontaneously is: How much power is enough power? Obviously, the more power the better. However, in some experiments it is very time consuming and expensive to run each subject. In these experiments, the experimenter usually must accept less power than is typically found in experiments in which subjects can be run cheaply and easily. In any case, power below 0.25 would almost always be considered too low and power above 0.80 would be considered satisfactory. Keep in mind that a Type II error is not necessarily so bad since a failure to reject the null hypothesis does not mean that the research hypothesis should be abandoned. If the results are suggestive, further experiments should be conducted to see if the existence of the effect can be confirmed. The power of the two experiments taken together will be greater than the power of either one.

Last but not least, one should keep in mind that power calculations are often based upon approximations. These may be approximations to the solution of the equation defining the sample size as was the case in Example 1.8. Or we may need to approximate, and more generally simply, the reference model. This is for instance the case if power calculations are to be performed for a regression model. As an example take a between-subjects design where the interest focuses on a particular covariate, while the remaining ones are included to take account of subject-specific response patterns. In these situations it is quite common to resort to a simpler model that allows one to do the calculations necessary to identify the sample size.