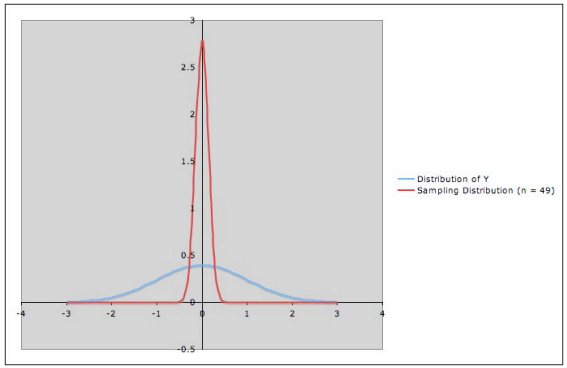


- The theorem that depends on the model assumptions will tell us enough so that it is possible to do the following:
  - If we specify a probability (we'll use .95 to illustrate), we can find a number *a* so that

(*) The probability that $\bar{Y}_n$ lies between μ - a and μ + a is approximately 0.95.

Caution: It is important to get the reference category straight here. This amounts to keeping in mind what is a random variable and what is a constant. Here, $\bar{Y}_n$ *is the random variable* (that is, the *sample* is varying), whereas μ *is constant*.

Note: The z-procedure for confidence intervals is only an approximate procedure; that is why the "approximately" is in (*) and below. Many procedures are "exact"; we don't need the "approximately" for them.

  - A little algebraic manipulation allows us to restate (*) as

(**) The probability that μ lies between $\bar{Y}_n$ - a and $\bar{Y}_n$ + a is approximately 0.95

Caution: It is again important to get the reference category correct here. It hasn't changed: it is still the *sample* that is varying, *not* μ. So the probability refers to $\bar{Y}_n$, *not* to μ.

Thinking that the probability refers to μ *is a* **common mistake** *in interpreting confidence intervals*.

It may help to restate (**) as

(***) The probability that the interval from $\bar{Y}_n$ - a to $\bar{Y}_n$ + a contains μ is approximately 0.95.

- We are now faced with two possibilities (assuming the model assumptions are indeed all true):

1) The sample we have taken is one of the approximately 95% for which the interval from $\bar{Y}_n$ - a to $\bar{Y}_n$ + a *does* contain μ. ☺

2) Our sample is one of the approximately 5% for which the interval from $\bar{Y}_n$ - a to $\bar{Y}_n$ + a does *not* contain μ. ☹

Unfortunately, *we can't know which of these two possibilities is true*. ☹

- Nonetheless, we calculate the values of $\bar{Y}_n$ - a and $\bar{Y}_n$ + a *for the sample we have*, and call the resulting interval an approximate *95% confidence interval for* μ.
  - We can say that *we have obtained the confidence interval by using a procedure that, for approximately 95% of all simple random samples from Y, of the given size, produces an interval containing the parameter we are estimating*.
  - Unfortunately, we can't know whether or not the sample we have used is one of the approximately 95% of "good" samples that yield a confidence interval containing the true mean μ, or whether the sample we have is one of the approximately 5% of "bad" samples that yield a confidence interval that does not contain the true mean μ.
  - We can just say that we have used a procedure that "works" about 95% of the time.
  - Various web demos can demonstrate.

In general: We can follow a similar procedure for many other situations to obtain confidence intervals for parameters.

- Each type of confidence interval procedure has its own model assumptions.
  - If the model assumptions are not true, we are not sure that the procedure does what is claimed.
  - However, some procedures are robust to some degree to some departures from models assumptions -- i.e., the procedure works pretty closely to what is intended if the model assumption is not too far from true.
  - Robustness depends on the particular procedure; there are no "one size fits all" rules.

- We can decide on the "level of confidence" we want;
  - E.g., we can choose 90%, 99%, etc. rather than 95%.
  - Just which level of confidence is appropriate depends on the circumstances. (More later)
- The *confidence level* is *the percentage of samples for which the procedure results in an interval containing the true parameter*. (Or approximate percentage, if the procedure is not exact.)
- However, *a higher level of confidence will produce a wider confidence interval*. (See demo)
  - i.e., less certainty in our estimate.
  - So *there is a trade-off between degree of confidence and degree of certainty*.
- Sometimes the best we can do is a procedure that only gives *approximate* confidence intervals.
  - i.e., the sampling distribution can be described only approximately.
  - i.e., there is one more source of uncertainty.
  - This is the case for the large-sample z-procedure.
- If the sampling distribution is not symmetric, we can't expect the confidence interval to be symmetric around the estimate.
  - There may be slightly different procedures for calculating the endpoints of the confidence interval.
- There are variations such as "upper confidence limits" or "lower confidence limits" where we are only interested in estimating how large or how small the estimate might be.

FREQUENTIST HYPOTHESIS TESTS AND P-VALUES

We'll now continue the discussion of hypothesis tests.

Recall: Most commonly used frequentist hypothesis tests involve the following elements:

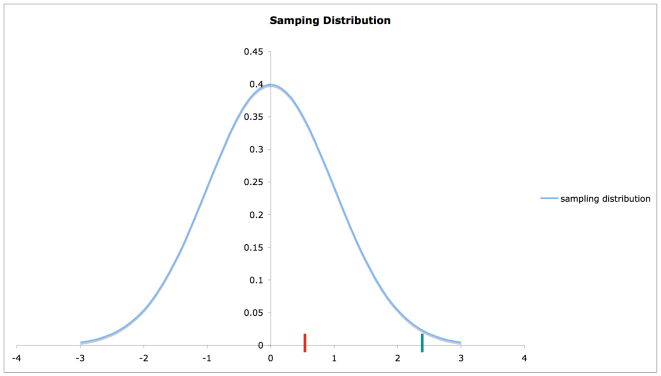
1. Model assumptions
2. Null and alternative hypothesis
3. A test statistic (something calculated by a rule from a sample)
   - This needs to have the property that extreme values of the test statistic cast doubt on the null hypothesis.
4. A mathematical theorem saying, "If the model assumptions and the null hypothesis are both true, then the sampling distribution of the test statistic has this particular form."

The exact details of these four elements will depend on the particular hypothesis test.

We will use the example of a *one-sided t-test for a single mean* to illustrate the general concepts of of p-value and hypothesis testing as well as sampling distribution for a hypothesis test.

- We have a random variable Y that is *normally distributed*. (This is one of the *model assumptions*.)
- Our *null hypothesis* is: The population mean μ of the random variable Y is μ_0.
- For simplicity, we will discuss a *one-sided alternative hypothesis*: The population mean μ of the random variable Y is greater than μ_0. (i.e., $\mu > \mu_0$)
- Another *model assumption* says that samples are *simple random samples*. We have data in the form of a simple random sample of size n.

- To understand the idea behind the hypothesis test, we need to put our sample of data on hold for a while and consider *all* possible simple random samples of the same size n from the random variable Y.
  - For any such sample, we could calculate its sample mean $\bar{y}$ and its sample standard deviation s.
  - We could then use $\bar{y}$ and s to calculate the *t-statistic* t $= \dfrac{\bar{y} - \mu_0}{s/\sqrt{n}}$
  - Doing this for all possible simple random samples of size n from Y gives us a new random variable, T_n. Its distribution is called a *sampling distribution*.
  - The mathematical theorem associated with this inference procedure (one-sided t-test for population mean) tells us that *if the null hypothesis is true*, then the sampling distribution has what is called the *t-distribution with n degrees of freedom*. (For large values of n, the t-distribution looks very much like the standard normal distribution; but as n gets smaller, the peak gets slightly smaller and the tails go further out.)
- *Now consider where the t-statistic for the data at hand lies on the sampling distribution.* Two possible values are shown in red and green, respectively, in the diagram below.
  - *Remember that this picture depends on the validity of the model assumptions and on the assumption that the null hypothesis is true.*



If the t-statistic lies at the *red* bar (around 0.5) in the picture, *nothing is unusual*; our data are consistent with the null hypothesis.

But if the t-statistic lies at the *green* bar (around 2.5), then the data would be fairly *unusual* -- assuming the null hypothesis is true.

So ***a t-statistic at the green bar would cast some reasonable doubt on the null hypothesis***.

A t-statistic even further to the right would cast even more doubt on the null hypothesis.

Note: A little algebra will show that if $t = \dfrac{\bar{y} - \mu_0}{s/\sqrt{n}}$ is unusually large, then so is $\bar{y}$, and vice-versa

p-values

We can quantify the idea of how unusual a test statistic is by the *p-value*. The general definition is:

> p-value = the probability of obtaining a test statistic *at least as extreme* as the one from the data at hand, assuming the model assumptions and the null hypothesis are all true.

Recall that we are only considering samples, from the same random variable, that fit the model assumptions *and of the same size as the one we have.*

So the definition of p-value, if we spell everything out, reads

> p-value = the probability of obtaining a test statistic *at least as extreme* as the one from the data at hand, *assuming*
> - *the model assumptions are all true, and*
> - *the null hypothesis is true, and*
> - *the random variable is the same (including the same population), and*
> - *the sample size is the same.*

Comment: The preceding discussion can be summarized as follows:
If we obtain an unusually small p-value, then (at least) one of the following must be true:

- At least one of the model assumptions is not true (in which case the test may be inappropriate).
- The null hypothesis is false.
- The sample we have obtained happens to be one of the small percentage that result in a small p-value.

The interpretation of "at least as extreme as" depends on the alternative hypothesis:

- For the *one-sided alternative hypothesis* $\mu > \mu_0$, "at least as extreme as" means "at least as great as".
    - Recalling that the probability of a random variable lying in a certain region is the area under the probability distribution curve over that region, we conclude that for this alternative hypothesis, the p-value is the area under the distribution curve to the *right* of the test statistic calculated from the data.
    - Note that, in the picture, the p-value for the t-statistic at the green bar is much less than that for the t-statistic at the red bar.
- Similarly, for the *other one-sided alternative,* $\mu < \mu_0$, the p-value is the area under the distribution curve to the *left* of the calculated test statistic.
    - Note that for this alternative hypothesis, the p-value for the t-statistic at the green bar would be much greater than the t-statistic at the red bar, but both would be large as p-values go.
- For the two-sided alternative $\mu \neq \mu_0$, the p-value would be the area under the curve to the right of the absolute value of the calculated t-statistic, *plus* the area under the curve to the left of the negative of the absolute value of the calculated t-statistic.
    - Since the sampling distribution in the illustration is symmetric about zero, the two-sided p-value of, say the green value, would be twice the area under the curve to the right of the green bar.

Note that, *for samples of the same size*, *the smaller the p-value, the stronger the evidence against the null hypothesis*, since a smaller p-value indicates a more extreme test statistic.

Thus, if the p-value is small enough (*and* assuming all the model assumptions are met), *rejecting the null hypothesis in favor of the alternate hypothesis* can be considered a rationale decision.

Comments:

1. How small is "small enough" is a judgment call.

2. "Rejecting the null hypothesis" does *not* mean the null hypothesis is false or that the alternate hypothesis is true.

3. Comparing p-values for samples of different size is a **common mistake**.

- In fact, larger sample sizes are more likely to detect a difference, so are likely to result in smaller p-values than smaller sample sizes, even though the context being examined is exactly the same.

We will discuss these comments further later.

MISINTERPRETATIONS AND MISUSES OF P-VALUES

Recall:

p-value = the probability of obtaining a test statistic at least as extreme as the one from the data at hand, assuming:

- the model assumptions for the inference procedure used are all true*, and*
- the null hypothesis is true*, and*
- the random variable is the same (including the same population)*, and*
- the sample size is the same.

Notice that this is a *conditional probability*: The probability that something happens, given that various other conditions hold. *One* **common mistake** *is to neglect some or all of the conditions*.

Example A: Researcher 1 conducts a clinical trial to test a drug for a certain medical condition on 30 patients all having that condition.

- The patients are randomly assigned to either the drug or a look-alike placebo (15 each).
- Neither patients nor medical personnel know which patient takes which drug.
- Treatment is exactly the same for both groups, except for whether the drug or placebo is used.
- The hypothesis test has null hypothesis "proportion improving on the drug is the same as proportion improving on the placebo" and alternate hypothesis "proportion improving on the drug is greater than proportion improving on the placebo."
- The resulting p-value is p = 0.15.

Researcher 2 does another clinical trial on the *same drug*, with the *same placebo*, and *everything else the same* except that 200 patients are randomized to the treatments, with 100 in each group. The same hypothesis test is conducted with the new data, and the resulting p-value is p = 0.03.

Are these results contradictory? No -- since the sample sizes are different, the p-values are not comparable, even though everything else is the same. (In fact, *a larger sample size typically results in a smaller p-value*; more later).

Example B: Researcher 2 from Example A does everything as described above, but for convenience, his patients are all from the student health center of the prestigious university where he works.

- He *cannot* claim that his result applies to patients other than those of the age and socio-economic background, etc. of the ones he used in the study, because his sample was taken from a smaller population.

Example C: Researcher 2 proceeds as in Example A, with a sample carefully selected from the population to which he wishes to apply his results, but he is testing for equality of the means of an outcome variable for the two groups.

- The hypothesis test he uses requires that the variance of the outcome variable for each group compared is the same.
- He doesn't check this, and in fact the variance for the treatment group is twenty times as large as the variance for the placebo group.
- He is *not* justified in rejecting the null hypothesis of equal means, no matter how small his p-value.

Another **common misunderstanding** *of p-values is the belief that the p-value is "the probability that the null hypothesis is true".*

- This is essentially a case confusing a conditional probability with the reverse conditional probability: In the definition of p-value, "the null hypothesis is true" is the condition, not the event.
- The basic assumption of frequentist hypothesis testing is that the null hypothesis is either true (in which case the probability that it is true is 1) or false (in which case the probability that it is true is 0).

Note: In the *Bayesian* perspective, it makes sense to consider "the probability that the null hypothesis is true" as having values other than 0 or 1.

- In that perspective, we consider "states of nature;" in different states of nature, the null hypothesis may have different probabilities of being true.
- The goal is then to determine the probability that the null hypothesis is true, given the data.
- This is the *reverse conditional probability* from the one considered in frequentist inference (the probability of the data given that the null hypothesis is true).

Type I and II Errors and Significance Levels

Type I Error:

Rejecting the *null* hypothesis when it is in fact true is called a ***Type I error***.

Significance level:

Many people decide, before doing a hypothesis test, on a maximum p-value for which they will reject the null hypothesis. This value is often denoted α (alpha) and is also called the ***significance level***.

When a hypothesis test results in a p-value that is less than the significance level, the result of the hypothesis test is called *statistically significant*.

Confusing statistical significance and practical significance is a **common mistake**.

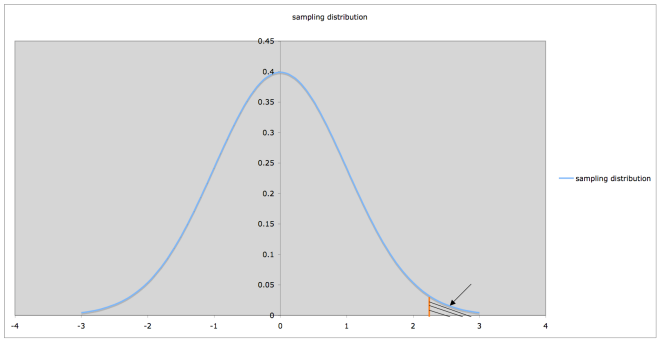
Example: A large clinical trial is carried out to compare a new medical treatment with a standard one. The statistical analysis shows a statistically significant difference in lifespan when using the new treatment compared to the old one.

- However, the increase in lifespan is at most three days, with average increase less than 24 hours, and with poor quality of life during the period of extended life.
- Most people would not consider the improvement practically significant.

Caution: The larger the sample size, the more likely a hypothesis test will detect a small difference. Thus *it is especially important to consider practical significance when sample size is large.*

Connection between Type I error and significance level:

A significance level α corresponds to a certain value of the test statistic, say t_α, represented by the orange line in the picture of a sampling distribution below (the picture illustrates a hypothesis test with alternate hypothesis "$\mu > 0$").



- Since the shaded area indicated by the arrow is the p-value corresponding to t_α, that p-value (shaded area) is α.
- To have p-value less than α, a t-value for this test must be to the right of t_α.
- So the probability of rejecting the null hypothesis when it is true is the probability that $t > t_\alpha$, which we have seen is α.
- In other words, *the probability of Type I error is α.*
- Rephrasing using the definition of Type I error:
  The significance level α is the probability of making the wrong decision when the <u>null</u> hypothesis is true.
- *Note*:
  - α is also called the *bound on Type I error*.
  - Choosing a value α is sometimes called *setting a bound on Type I error*.

Pros and Cons of Setting a Significance Level:

- Setting a significance level (*before* doing inference) has the *advantage* that the analyst is not tempted to chose a cut-off on the basis of what he or she hopes is true.
- It has the *disadvantage* that it neglects that some p-values might best be considered borderline.
  - *This is one reason why it is important to report p-values when reporting results of hypothesis tests. It is also good practice to include confidence intervals corresponding to the hypothesis test.*
  - For example, if a hypothesis test for the difference of two means is performed, *also* give a confidence interval for the difference of those means.
  - If the significance level for the hypothesis test is .05, then use confidence level 95% for the confidence interval.

Type II Error

Not rejecting the *null* hypothesis when in fact the *alternate* hypothesis is true is called a ***Type II error***.

- Example 2 below provides a situation where the concept of Type II error is important.
- *Note*: "The alternate hypothesis" in the definition of Type II error may refer to the alternate hypothesis in a hypothesis test, or it may refer to a "specific" alternate hypothesis.

Example: In a t-test for a sample mean μ, with null hypothesis "μ = 0" and alternate hypothesis "μ > 0":

- We might talk about the Type II error relative to the *general alternate hypothesis "μ > 0"*.
- Or we might talk about the Type II error relative to the *specific alternate hypothesis "μ = 1"*.
- Note that *the specific alternate hypothesis is a special case of the general alternate hypothesis*.

In practice, people often work with Type II error relative to a *specific* alternate hypothesis.

- In this situation, the probability of Type II error relative to the specific alternate hypothesis is often called β.
- In other words, β is the probability of making the *wrong* decision when the *specific alternate* hypothesis is *true*.
- The specific alternative is considered since it is more feasible to calculate β than the probability of Type II error relative to the general alternative.
- See the discussion of power below for related detail.

Considering both types of error together:

The following table summarizes Type I and Type II errors:

		Truth (for population studied)	
		Null Hypothesis True	Null Hypothesis False
Decision (based on sample)	Reject Null Hypothesis	*Type I Error*	*Correct Decision*
	Don't reject Null Hypothesis	*Correct Decision*	*Type II Error*

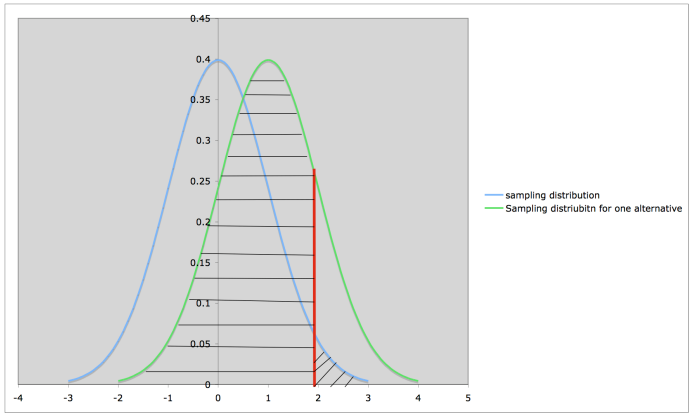
An analogy that can be helpful in understanding the two types of error is to consider a defendant in a trial.

- The null hypothesis is "defendant is not guilty."
- The alternate is "defendant is guilty."
- A Type I error would correspond to convicting an innocent person.
- Type II error would correspond to setting a guilty person free.
- This could be more than just an analogy if the verdict hinges on statistical evidence (e.g., a DNA test), and where rejecting the null hypothesis would result in a verdict of guilty, and not rejecting the null hypothesis would result in a verdict of not guilty.
- The analogous table would be:

		Truth	
		Not Guilty	Guilty
Verdict	Guilty	*Type I Error* -- Innocent person goes to jail (and maybe guilty person goes free)	*Correct Decision*
	Not Guilty	*Correct Decision*	*Type II Error* -- Guilty person goes free

The following diagram illustrates the Type I error and the Type II error

- against the specific alternate hypothesis "μ =1"
- in a hypothesis test for a population mean μ,
- with null hypothesis "μ = 0,"
- alternate hypothesis "μ > 0",
- and significance level α= 0.05.



In the diagram,

- The blue (leftmost) curve is the *sampling* distribution of the test statistic assuming the null hypothesis ""μ = 0."
- The green (rightmost) curve is the *sampling* distribution of the test statistic assuming the specific alternate hypothesis "μ =1".
- The vertical red line shows the cut-off for rejection of the null hypothesis:
    - The null hypothesis is rejected for values of the test statistic to the *right* of the red line (and *not* rejected for values to the *left* of the red line).
- The area of the diagonally hatched region to the *right* of the red line and under the *blue* curve is the probability of type I error (α).
- The area of the horizontally hatched region to the *left* of the red line and under the *green* curve is the probability of Type II error against the specific alternative (β).

Deciding what significance level to use:

This should be done *before* analyzing the data -- preferably before gathering the data. There are (at least) two reasons why this is important:

1) The significance level desired is one criterion in deciding on an appropriate sample size.
    - See discussion of Power below.

2) If more than one hypothesis test is planned, additional considerations need to be taken into account.
    - See discussion of Multiple Inference below.

The choice of significance level should be based on the consequences of Type I and Type II errors:

1. If the *consequences of a Type I error are serious or expensive*, a very *small* significance level is appropriate.

Example 1: Two drugs are being compared for effectiveness in treating the same condition.

- Drug 1 is very affordable, but Drug 2 is extremely expensive.
- The null hypothesis is “both drugs are equally effective.”
- The alternate is “Drug 2 is more effective than Drug 1.”
- In this situation, a Type I error would be deciding that Drug 2 is more effective, when in fact it is no better than Drug 1, but would cost the patient much more money.
- That would be undesirable from the patient’s perspective, so a *small* significance level is warranted.

2. If the consequences of a Type I error are not very serious (and especially *if a Type II error has serious consequences*), then a *larger* significance level is appropriate.

Example 2: Two drugs are known to be equally effective for a certain condition.

- They are also each equally affordable.
- However, there is some suspicion that Drug 2 causes a serious side effect in some patients, whereas Drug 1 has been used for decades with no reports of the side effect.
- The null hypothesis is "the incidence of the side effect in both drugs is the same".
- The alternate is "the incidence of the side effect in Drug 2 is greater than that in Drug 1."
- Falsely rejecting the null hypothesis when it is in fact true (Type I error) would have no great consequences for the consumer.
- But a Type II error (i.e., failing to reject the null hypothesis when in fact the alternate is true, which would result in deciding that Drug 2 is no more harmful than Drug 1 when it is in fact more harmful) could have serious consequences from a public health standpoint.
- So setting a large significance level is appropriate.

Comments:

- Neglecting to think adequately about possible consequences of Type I and Type II errors (and deciding acceptable levels of Type I and II errors based on these consequences) before conducting a study and analyzing data is a **common mistake** in using statistics.
- Sometimes there may be serious consequences of each alternative, so some compromises or weighing priorities may be necessary.
  - The trial analogy illustrates this well: Which is better or worse, imprisoning an innocent person or letting a guilty person go free?
  - *This is a value judgment; value judgments are often involved in deciding on significance levels.*
  - *Trying to avoid the issue by always choosing the same significance level is itself a value judgment.*
- Different people may decide on different standards of evidence.
  - This is another reason why *it is important to report p-values even if you set a significance level.*
  - It is *not* enough just to say, “significant at the .05 level,” “significant at the .01 level,” etc.
- Sometimes different stakeholders have different interests that compete (e.g., in the second example above, the developers of Drug 2 might prefer to have a smaller significance level.)
- See Wuensch (1994) for more discussion of considerations involved in deciding on reasonable levels for Type I and Type II errors.
- See also the discussion of Power below.
- Similar considerations hold for setting confidence levels for confidence intervals; see http://www.ma.utexas.edu/users/mks/statmistakes/conflevel.html.

POWER OF A STATISTICAL PROCEDURE

Overview

The ***power*** of a statistical procedure can be thought of as *the probability that the procedure will detect a true difference of a specified type.*

- As in talking about p-values and confidence levels, the reference for "probability" is the sample.
- Thus, power is the probability that a randomly chosen sample
    - satisfying the model assumptions
    - will give evidence of a difference of the specified type when the procedure is applied,
    - if the specified difference does indeed occur in the population being studied.
- Note also that power is a conditional probability: the probability of detecting a difference, *if* indeed the difference does exist.

In many real-life situations, there are reasonable conditions that we would be interested in being able to detect or that would not make a practical difference.

Examples:

- If you can only measure the response to within 0.1 units, it doesn't really make sense to worry about falsely rejecting a null hypothesis for a mean when the actual value of the mean is within less than 0.1 units of the value specified in the null hypothesis.
- Some differences are of no practical importance -- for example, a medical treatment that extends life by 10 minutes is probably not worth it.

In cases such as these, neglecting power could result in one or more of the following:

- Doing much more work or going to more expense than necessary
- Obtaining results which are meaningless
- Obtaining results that don't answer the question of interest.

Elaboration

For many *confidence interval procedures*, power can be defined as:

> The probability (again, the reference category is "samples") that the procedure will produce an interval with a half-width of at least a specified amount.

For a *hypothesis test*, power can be defined as:

> The probability (again, the reference category is "samples") of rejecting the null hypothesis under a specified condition.

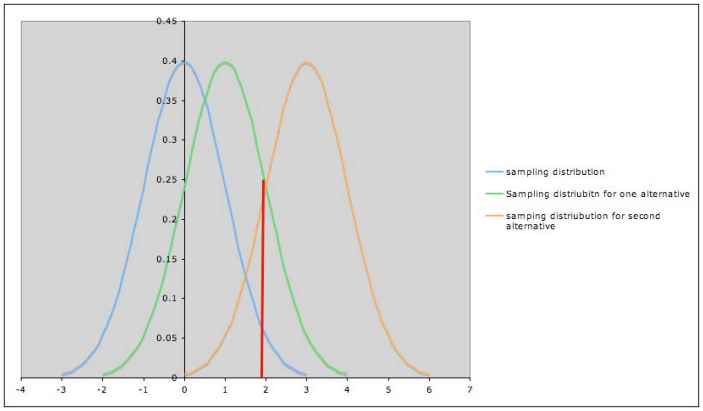
Example: For a one-sample t-test for the mean of a population, with null hypothesis Ho: $\mu = 100$, you might be interested in the probability of rejecting Ho when $\mu \geq 105$, or when $|\mu - 100| > 5$, etc.

As with Type I Error, we may think of power for a hypothesis test in terms of *power against a specific alternative* rather than against a general alternative.

Example: If we are performing a hypothesis test for the mean of a population, with null hypothesis H_0: $\mu = 0$ and alternate hypothesis $\mu > 0$, we might calculate the power of the test *against the specific alternative* H_1: $\mu = 1$, or against the specific alternate H_3: $\mu = 3$, etc. The picture below shows three sampling distributions:

- The sampling distribution assuming H_0 (blue; leftmost curve)
- The sampling distribution assuming H_1 (green; middle curve)
- The sampling distribution assuming H_3 (yellow; rightmost curve)

The red line marks the cut-off corresponding to a significance level $\alpha = 0.05$.



- Thus the area under the *blue* curve to the right of the red line is 0.05.
- The area under the *green* curve the to right of the red line is the probability of rejecting the null hypothesis ($\mu = 0$) if the specific alternative H_1: $\mu = 1$ is true.
  - In other words, this area is *the power of the test against the specific alternative H_1: $\mu = 1$.*
  - We can see in the picture that in this case, this power is greater than 0.05, but noticeably less than 0.50.
- Similarly, the area under the *yellow* curve the to right of the red line is *the power of the test against the specific alternative H_3: $\mu = 3$.*
  - Notice that the power in this case is much larger than 0.5.

This illustrates the general phenomenon that *the farther an alternative is from the null hypothesis, the higher the power of the test to detect it.* (See Claremont WISE Demo)

Note: For most tests, it *is* possible to calculate the power against a specific alternative, at least to a reasonable approximation. It is *not* usually possible to calculate the power against a general alternative, since the general alternative is made up of infinitely many possible specific alternatives.

Power and Type II Error

Recall: The Type II Error rate β of a test against a specific alternate hypothesis test is represented in the diagram above as the area under the sampling distribution curve for that alternate hypothesis and to the *left* of the cut-off line for the test. Thus

β + (Power of a test against a specific alternate hypothesis)

= total area under sampling distribution curve

= 1,

so

Power = 1 - β

Power and Sample Size

Power will depend on sample size as well as on the specific alternative.

- The picture above shows the sampling distributions for one particular sample size.
- If the sample size is larger, the sampling distributions will be narrower.
- This will mean that sampling distributions will have less overlap, and the power will be higher.
- Similarly, a smaller sample size will result in more overlap of the sampling distributions, hence in lower power.
- This dependence of power on sample size allows us, *in principle*, to figure out beforehand what sample size is needed to detect a specified difference, with a specified power, at a given significance level, if that difference is really there.
- See Claremont University's Wise Project's Statistical Power Applet for an interactive demonstration of the interplay between sample size and power for a one-sample z-test.

In practice, details on figuring out sample size will vary from procedure to procedure. Some considerations involved:

- The difference used in calculating sample size (i.e., the specific alternative used in calculating sample size) should be decided on the base of practical significance and/or "worst case scenario," depending on the consequences of decisions.
- Determining sample size to give desired power and significance level will usually require some estimate of parameters such as variance, so will only be as good as these estimates.
    - These estimates usually need to be based on previous research or a pilot study.
    - It is wise to use a conservative estimate of variance (e.g., the upper bound of a confidence interval from a pilot study), or to do a sensitivity analysis to see how the sample size estimate depends on the parameter estimate.
- Even when there is a good formula for power in terms of sample size, "inverting" the formula to get sample size from power is often not straightforward.
    - This may require some clever approximation procedures.
    - Such procedures have been encoded into computer routines for many (not all) common tests.
    - See John C. Pezzullo's Interactive Statistics Pages for links to a number of online power and sample size calculators.

- Good and Hardin (2006, p. 34), *Common Errors in Statistics*, Wiley, p. 34) report that using the default settings for power and sample size calculations is a **common mistake** made by researchers.
- For discrete distributions, the "power function" (giving power as a function of sample size) is often saw-toothed in shape.
    - A consequence is that software may not necessarily give the optimal sample size for the conditions specified.
    - Good software for such power calculations will also output a graph of the power function, allowing the researcher to consider other sample sizes that might give be better than the default given by the software.

Common Mistakes involving Power:

1. ***Accepting a null hypothesis when a result is not statistically significant, without taking power into account.***

- Since power typically increases with increasing sample size, practical significance is important to consider.
- Looking at this from the other direction: Power decreases with decreasing sample size.
- Thus *a small sample size may not be able to detect an important difference*.
- If there is strong evidence that the power of a procedure will indeed detect a difference of practical importance, then accepting the null hypothesis is appropriate.
- Otherwise "accepting the null hypothesis" is *not appropriate* -- all we can legitimately say then is that we fail to reject the null hypothesis.

2. ***Neglecting to do a power analysis/sample size calculation before collecting data***

- Without a power analysis, you may end up with a result that does not really answer the question of interest.
- You might obtain a result that is not statistically significant, but is not able to detect a difference of practical significance.
- You might also waste resources by using a sample size that is larger than is needed to detect a relevant difference.

3. ***Confusing retrospective power and prospective power***.

- Power as defined above for a hypothesis test is also called *prospectiv*e or *a priori* power.
- It is a conditional probability, P(reject H_0 | H_a), calculated *without using the data to be analyzed*.
- *Retrospective* power is calculated after the data have been collected and analyzed, using the data.
- Retrospective power can be used legitimately to estimate the power and sample size for a *future* study, but *cannot* legitimately be used as describing the power of the study from which it is calculated.
- See Hoenig and Heisley (2001) and Wuensch et al (2003) for more discussion and further references.

REFERENCES

Good and Hardin (2006), *Common Errors in Statistics*, Wiley

Hoenig, J. M. and D. M. Heisley (2001) "The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis," *The American Statistician* 55(1), 19-24

Wuensch, K. L. (1994). Evaluating the Relative Seriousness of Type I versus Type II Errors in Classical Hypothesis Testing, http://core.ecu.edu/psyc/wuenschk/StatHelp/Type-I-II-Errors.htm

Wuensch, K.L. et al (2003), "Retrospective (Observed) Power Analysis, Stat Help website, http://core.ecu.edu/psyc/wuenschk/stathelp/Power-Retrospective.htm