

A Power Analysis

Suppose we are interested in determining whether alewife (a small lake fish) prefer to eat eggs or plankton. We want to gather together some alewife, put them in aquariums for 24 hours with both eggs and plankton to eat, and examine the contents of their stomachs to see which they have chosen. The null hypothesis is that they have no preference; that is,

$$\begin{aligned} H_0 : p &= 0.5 \\ H_a : p &\neq 0.5 \end{aligned}$$

where p is the probability that the fish prefer eggs to plankton.

How many fish do we need in this experiment? The answer depends on how large a difference we want to detect and how willing we are to make an error in our decision.

We can simulate this situation in the classroom by considering an analogous question:

Is the probability of obtaining a head on a spin of a penny equal to 0.5?

Spinning a Penny

Suppose we spin a penny 25 times. Is this the same as spinning 25 pennies once each? We will consider this question shortly. To begin, let's spin a single penny 25 times.

The results are: in 25 spins, I observe 16 Tails and 9 Heads. In this case, $\hat{p} = \frac{16}{25} = 0.64$.

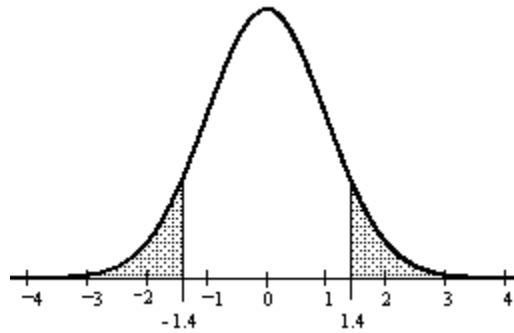
Because both $np = 25(0.5) = 12.5$ and $n(1-p) = 25(0.5) = 12.5$ are greater than 5, the z -test will be used to test the hypothesis:

$$\begin{aligned} H_0 : p &= 0.5 \\ H_a : p &\neq 0.5 \end{aligned}$$

Based on our observation $\hat{p} = \frac{16}{25} = 0.64$, can we reject the null hypothesis at the $\alpha = 0.05$ significance level? The test statistic is

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.64 - 0.5}{\sqrt{\frac{0.5(0.5)}{25}}} = 1.4$$

Since this was a two-sided test, the p -value associated with this test statistic is 0.1615.



Therefore, we fail to reject the null hypothesis that the probability of getting a head on a spin of a penny is 0.5.

Practical Significance

But 0.64 is almost two-thirds. Is the true value of p equal to 0.64? This difference may be practically significant even though it was not statistically significant. We want to consider the question, "What sample size is needed to statistically detect a practical difference". Determining what is a practical difference is not a question for the statistician, but for the scientist. The statistics should always support the science.

In a test of a hypothesis, the Type I Error (\mathbf{a}), Type II Error (\mathbf{b}), and the Sample Size (n) are interrelated. Once two of these are specified, the third is also determined and beyond your control. Traditionally, experimenters have set \mathbf{a} and n and tolerated whatever \mathbf{b} resulted. Now more consideration is being given to the control of Type II error and power at the design phase of the study.

To see how this works, we will first develop the power function associated with our test of hypothesis. Recall, we set $\mathbf{a} = 0.05$ and $n = 25$.

Power

The power of a test is simply the probability of rejecting the null hypothesis. This probability, and thus the power, varies with our selection of the Sample Size (n), the accepted probability of a Type I Error (\mathbf{a}), and the difference between the true value of p and the tested value of p from the null hypothesis. As a rule of thumb, we would like the power to be greater than 0.8.

The power function, sometimes denoted

$$\mathbf{y}(p), \text{ is } \mathbf{y}(p) = P(\text{Reject } H_0 \mid p).$$

In the coin spinning setting, this is

$$y(p) = P(\text{Reject } H_0 | p) = P\left(\left|\frac{\hat{p}-0.5}{\sqrt{\frac{0.5(0.5)}{25}}}\right| > 1.96 | p\right)$$

In words, this is the probability of obtaining a value of \hat{p} that will generate a z-score greater than 1.96 or less than -1.96 from the distribution with true parameter p . If $p = 0.5$, then the test statistic has an approximate standard normal distribution, and the power is 0.05. (The power function is always equal to the significance level at the value of the hypothesized parameter.)

However, if $p \neq 0.5$ then the null hypothesis is not true, and the test statistic $\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$

is no longer approximately standard normal. Can we rewrite the inequality so that the quantity on the left-hand side is approximately standard normal?

$$P(\text{Reject } H_0 | p) = P\left(\left|\frac{\hat{p}-0.5}{\sqrt{\frac{0.5(0.5)}{25}}}\right| > 1.96 | p\right) = 1 - P\left(\left|\frac{\hat{p}-0.5}{\sqrt{\frac{0.5(0.5)}{25}}}\right| \leq 1.96 | p\right).$$

Rather than compute the probability that our test statistics is outside the interval $(-1.96, 1.96)$ we can compute the probability that the test statistic is inside the interval $[-1.96, 1.96]$. The sum of these two must be one.

Now, we can simplify and re-write the absolute value expression so that

$$\begin{aligned} 1 - P\left(\left|\frac{\hat{p}-0.5}{\sqrt{\frac{0.5(0.5)}{25}}}\right| \leq 1.96 | p\right) &= 1 - P\left(0.5 - 1.96\sqrt{\frac{0.5(0.5)}{25}} \leq \hat{p} \leq 0.5 + 1.96\sqrt{\frac{0.5(0.5)}{25}} | p\right) \\ &= 1 - P(0.304 \leq \hat{p} \leq 0.696 | p). \end{aligned}$$

This value is the probability that we have an observed value \hat{p} that is sufficiently different from the hypothesized value of $p = 0.5$ that we reject the null hypothesis. The probability of generating observed value in this range from the true distribution with parameter p is given by

$$= 1 - P \left(\frac{0.304 - p}{\sqrt{\frac{p(1-p)}{25}}} \leq z \leq \frac{0.696 - p}{\sqrt{\frac{p(1-p)}{25}}} \right).$$

An Example

For example, suppose $p = 0.75$. Then the power is

$$P(\text{Reject } H_0 \mid p = 0.75) = 1 - P \left(\frac{0.304 - 0.75}{\sqrt{\frac{0.75(0.25)}{25}}} \leq z \leq \frac{0.696 - 0.75}{\sqrt{\frac{0.75(0.25)}{25}}} \right).$$

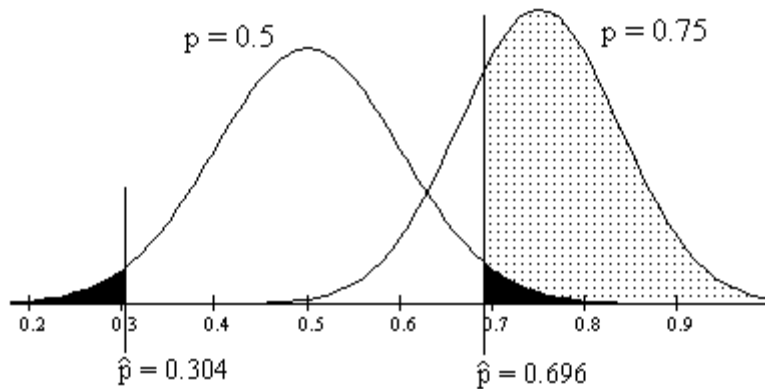
Simplifying these expressions gives

$$\begin{aligned} P(\text{Reject } H_0 \mid p = 0.75) &= 1 - P(-5.15 \leq z \leq -0.624) \\ &= 1 - 0.2663 \\ &= 0.7337 \end{aligned}$$

The probability of rejecting the null hypothesis when $p = 0.75$ under the conditions of this experiment ($\alpha = 0.05$, $n = 25$) is 0.7337. The power of this test when $p = 0.75$ is, therefore,

$$y(0.75) = 0.7337.$$

This probability can be visualized by comparing the distribution of the null hypothesis with $p = 0.5$ and the true distribution with $p = 0.75$. In the figure below, the area shaded in black under the curve of the null hypothesis is $\alpha = 0.05$. Values of \hat{p} less than 0.696 will be considered consistent with the null hypothesis while values of \hat{p} greater than 0.696 will be considered inconsistent with the null hypothesis and cause us to reject the null hypothesis. The probability that we achieve a value of \hat{p} greater than 0.696 from the true distribution is shaded in gray in the figure below. The area under the curve on the right but below the vertical line at $\hat{p} = 0.696$ is \mathbf{b} or the probability of making a Type II error. Notice that it is almost impossible to generate a value of \hat{p} less than 0.304 from the distribution on the right. The consequence of this is that, although the computations involve both extremes, the overwhelming majority of the probability comes from just one end of the distribution.



If, instead of spinning the coin 25 times, we had spun it 50 times, with proportional results, (32 out of 50) what would be the power of the test against the alternative $p = 0.75$?

Repeating the computations above with $n = 50$, we have

$$\begin{aligned}
 P(\text{Reject } H_0 \mid p) &= P\left(\left|\frac{\hat{p} - 0.5}{\sqrt{\frac{0.5(0.5)}{50}}}\right| > 1.96 \mid p\right) \\
 &= 1 - P\left(0.5 - 1.96\sqrt{\frac{0.5(0.5)}{50}} \leq \hat{p} \leq 0.5 + 1.96\sqrt{\frac{0.5(0.5)}{50}} \mid p\right) \\
 &= 1 - P(0.3614 \leq \hat{p} \leq 0.6386 \mid p)
 \end{aligned}$$

If we have an observed \hat{p} larger than 0.6386 or smaller than 0.3614, we will reject the null hypothesis that $p = 0.5$. How likely is it to generate such \hat{p} from 50 draws from a population where $p = 0.75$?

$$1 - P(0.3614 \leq \hat{p} \leq 0.6386 \mid p) = 1 - P\left(\frac{0.3614 - p}{\sqrt{\frac{p(1-p)}{50}}} \leq z \leq \frac{0.6386 - p}{\sqrt{\frac{p(1-p)}{50}}}\right)$$

$$\begin{aligned}
 P(\text{Reject } H_0 \mid p = 0.75) &= 1 - P\left(\frac{0.3614 - 0.75}{\sqrt{\frac{0.75(0.25)}{50}}} \leq z \leq \frac{0.6386 - 0.75}{\sqrt{\frac{0.75(0.25)}{50}}}\right) \\
 P(\text{Reject } H_0 \mid p = 0.75) &= 1 - P(-6.346 \leq z \leq -1.820)
 \end{aligned}$$

$$= 1 - 0.0344$$

$$= 0.9656$$

As before, it is unlikely to generate $\hat{p} < 0.3614$ when $p = 0.75$, so we focus on the right endpoint of the interval, where $\hat{p} > 0.6386$. In this example, with $n = 50$, we need a value of \hat{p} greater than 0.6386 to reject the null hypothesis at the $\alpha = 0.05$ significance level. This value is indicated in the figure below by the vertical line. The area under the distribution of the null hypothesis shaded in black represents $\alpha = 0.05$, while the area shaded in gray under the true distribution represents the power of the test when $p = 0.75$, $n = 50$, and $\alpha = 0.05$.

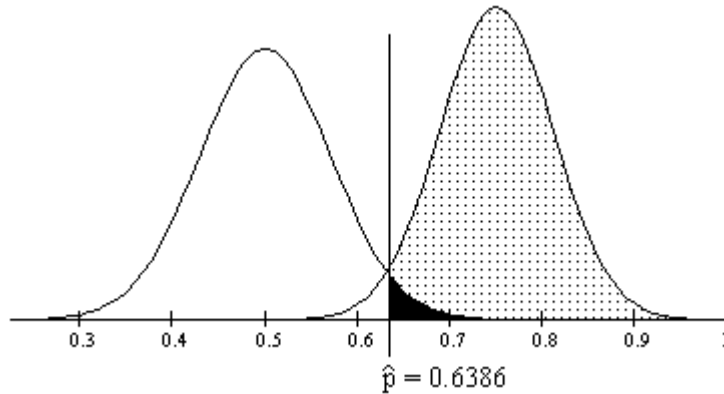


Figure : Graphical representation of $y(\alpha = 0.05, p = 0.75, n = 50) = 0.9656$

Generating the Power Function

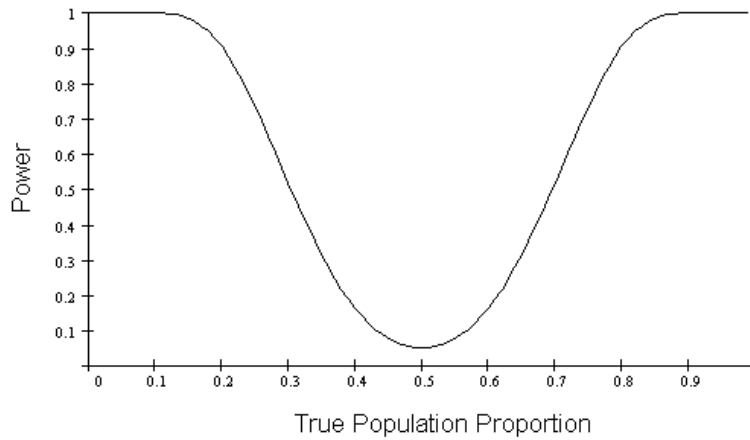
As can be seen in this example, the power of the test for a given value of p is a function of the number of experimental units n .

Viewed another way, for a given value of n , the power is a function of the value of population proportion p . In this example, we assumed the true value was 0.75 and computed the probability of rejecting the null hypothesis of $p = 0.5$. A short calculator activity will compute the power for various values of the population parameter with fixed n .

To create the power curve for $H_0 : p = 0.5$ with $n = 25$, type in the following commands:

<code>seq(X,X,.01,.99,.01) → L1</code>	<i>this defines the domain from $p = 0.01$ to $p = 0.99$</i>
<code>(.304 - L1)/√(L1*(1-L1)/25) → L2</code>	<i>places in List 2 the lower z-scores</i>
<code>(.696 - L1)/√(L1*(1-L1)/25) → L3</code>	<i>places in List 3 the upper z-scores</i>
<code>seq(normalcdf(L2(X),L3(X)),X,1,98) → L4</code>	<i>computes area between z-scores; values in List 4</i>
<code>1 - L4 → L5</code>	<i>computes power of the test; values in List 5</i>

Now plot the values in List 5 against the values in List 1 for the power curve.



For $n = 50$

$$\text{seq}(X,X,.01,.99,.01) \rightarrow L1$$

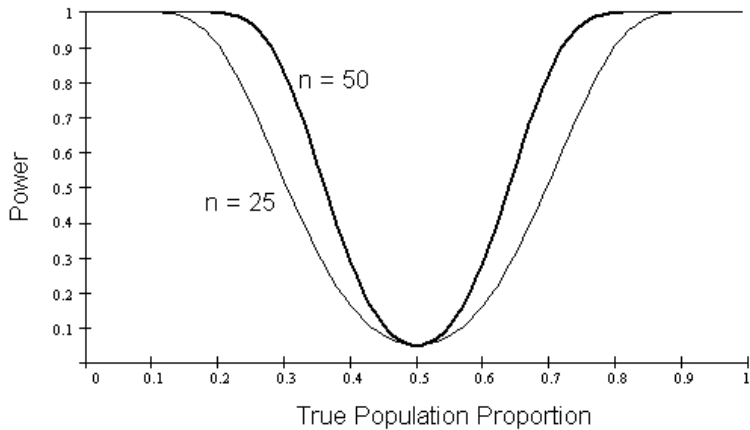
$$(.3614 - L1) / \sqrt{(L1 * (1 - L1) / 50)} \rightarrow L2$$

$$(.6386 - L1) / \sqrt{(L1 * (1 - L1) / 50)} \rightarrow L3$$

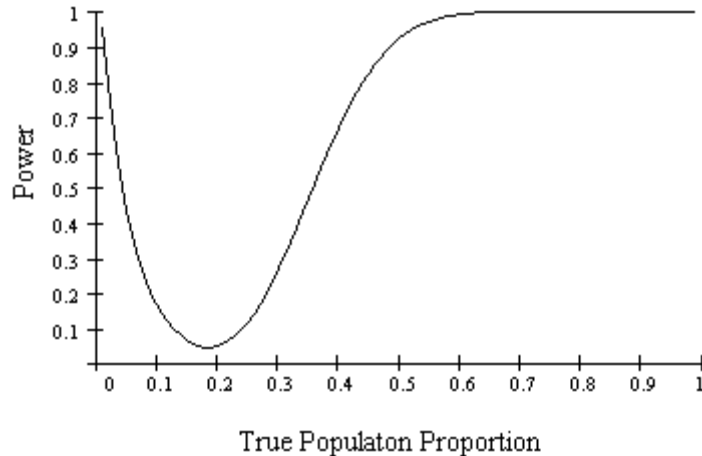
$$\text{seq}(\text{normalcdf}(L2(X),L3(X)),X,1,.98) \rightarrow L4$$

$$1 - L4 \rightarrow L5$$

The power curve for $n = 50$ is plotted on the same axis as the power curve for $n = 25$ for comparison. Notice that for both functions, $y(0.5) = 0.05 = \alpha$.



The symmetry of the curves is a result of the fact that the null hypothesis is $p = 0.5$. If, instead, we had $H_0 : p = 0.2$, we would see the following asymmetric power curve. As before, the power at $p = 0.2$ is $y(0.2) = 0.05 = \alpha$.

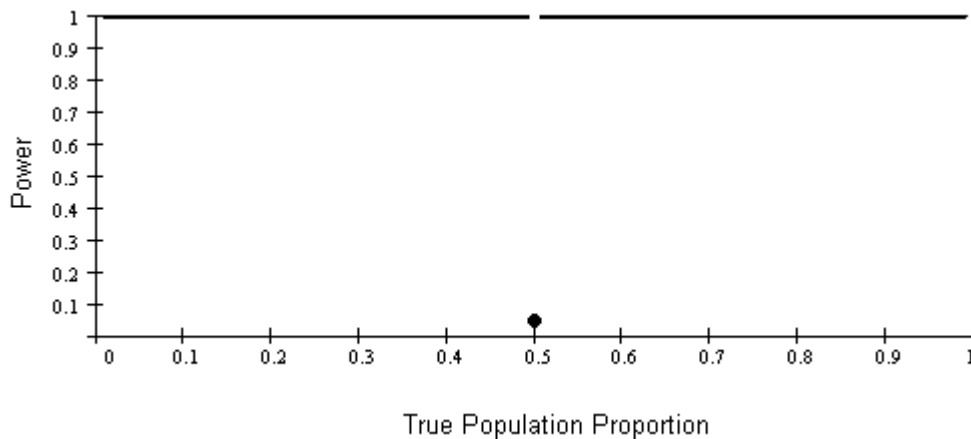


Back to the fish

How far from 0.5 do you need to be to think that the proportion is biologically significant? This is not a statistical question. It is a biological question that must be answered by the scientist. The statistician can use the biologist's judgment and help design an experiment that has a good chance of detecting a biologically significant result if it exists. In our particular example, the researcher decided to collect the fish from the lake instead of using aquaria because the sample size needed for good power exceeded the number of aquaria available. However, the analysis then became more complex because eggs and plankton were not equally plentiful in the lake.

Return to the Pennies

Suppose it is important to determine departures of 0.1 or more from the hypothesize value of p ; that is, if $p \geq 0.6$ or $p \leq 0.4$, we want to be confident that we will reject the null hypothesis. Ideally, we want a power function that looks like this:



We want the power function to have a value of 1 if the true population proportion is not 0.5, and α if the true population proportion is 0.5. Unfortunately, we cannot achieve this power function.

Again, assume that the significance level is $\alpha = 0.05$. We need to decide how "confident" we want to be and how large a difference from the hypothesized proportion we believe is practically significant. As an example, suppose we believe that any departure of 0.1 or more would be worth knowing about. We want to determine the sample size needed to find this practically significant result statistically significant. More explicitly, suppose we want the probability of detecting departures of 0.1 or more from the hypothesized value of p to be 0.8 or more. What sample size do we need to take?

We want

$$P\left(\left|\frac{\hat{p}-0.5}{\sqrt{\frac{0.5(0.5)}{n}}}\right| > 1.96 \mid |p-0.5| \geq 0.1\right) \geq 0.8.$$

As before, we rewrite the inequality. Instead of computing the probability of finding the extremes (less than -1.96 or greater than 1.96), we work with the bounded interval instead (between -1.96 and 1.96).

$$1 - P\left(\left|\frac{\hat{p}-0.5}{\sqrt{\frac{0.5(0.5)}{n}}}\right| \leq 1.96 \mid |p-0.5| \geq 0.1\right) \geq 0.8$$

whenever

$$1 - P\left(0.5 - 1.96\sqrt{\frac{0.5(0.5)}{n}} \leq \hat{p} \leq 0.5 + 1.96\sqrt{\frac{0.5(0.5)}{n}} \mid |\hat{p}-0.5| \geq 0.1\right) \geq 0.8$$

This inequality is satisfied if we find the sample size n needed for a power of 0.8 when $p = 0.6$ or $p = 0.4$. We will consider the case when $p = 0.6$. Thus, we want to find n such that

$$1 - P\left(\frac{0.5 - 1.96\sqrt{\frac{0.5(0.5)}{n}} - 0.6}{\sqrt{\frac{0.4(0.6)}{n}}} \leq z \leq \frac{0.5 + 1.96\sqrt{\frac{0.5(0.5)}{n}} - 0.6}{\sqrt{\frac{0.4(0.6)}{n}}}\right) \geq 0.8$$

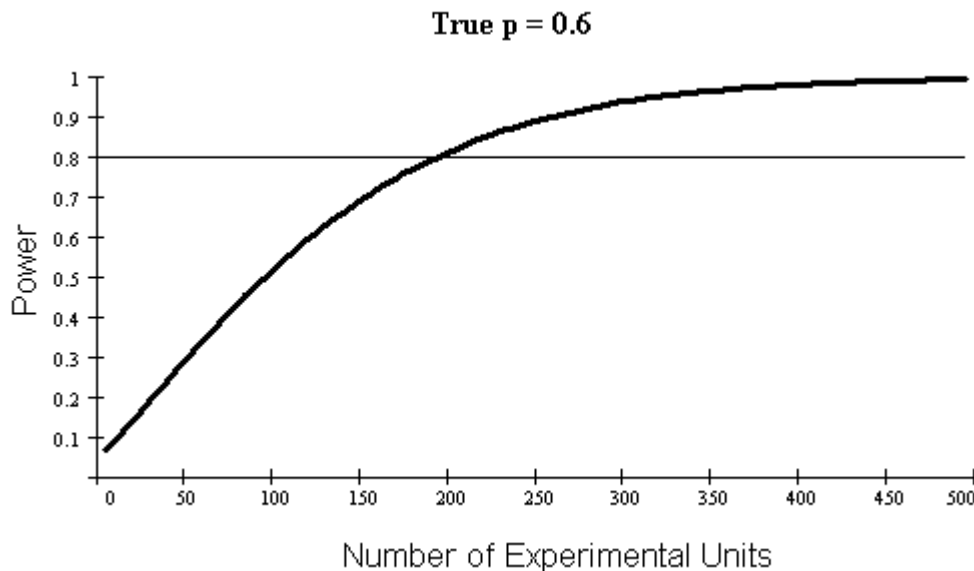
$$1 - P\left(\left(-1.96\sqrt{\frac{0.25}{n}} - 0.1\right)\sqrt{\frac{n}{0.24}} \leq z \leq \left(1.96\sqrt{\frac{0.25}{n}} - 0.1\right)\sqrt{\frac{n}{0.24}}\right) \geq 0.8$$

$$1 - P\left(-2.0004 - 0.2041\sqrt{n} \leq z \leq 2.0004 - 0.2041\sqrt{n}\right) \geq 0.8$$

By trial and error, we can find the smallest n such that the above probability is true. A calculator exercise will evaluate the power as a function of n . The smallest n is 194. We can use techniques similar to those creating the power curves to consider the power of this test for different values of n . The following sequence of steps will create values for the power of this test for values of n from 5 to 500 in steps of 5.

- seq(X,X,5,500,5) → L1 *this defines the domain from 5 to 500, steps of 5*
- 2.0004 - 0.204√(X) → L2 *places in List 2 the lower z-scores*
- 2.0004 - 0.204√(X) → L3 *places in List 3 the upper z-scores*
- seq(normalcdf(L2(X),L3(X)),X,1,100) → L4 *computes area between z-scores; values in List 4*
- 1 - L4 → L5 *power of the test stored in List 5*

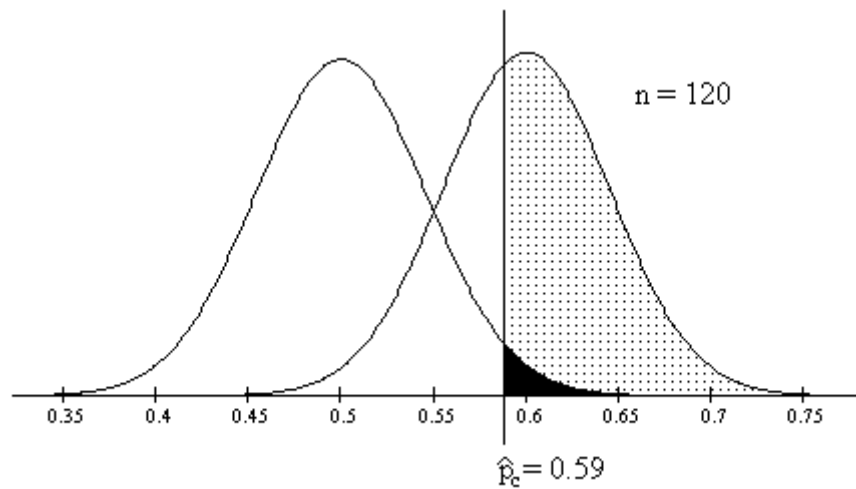
Now if we plot the values of $L5$ against the values of $L1$, we will have the power plotted as a function of n , in units of 5. When $n > 194$, the power of the test is greater than 0.8; that is, when $n > 194$, the probability of detecting departures of 0.1 or more from the hypothesized value of 0.5 is greater than 0.8.



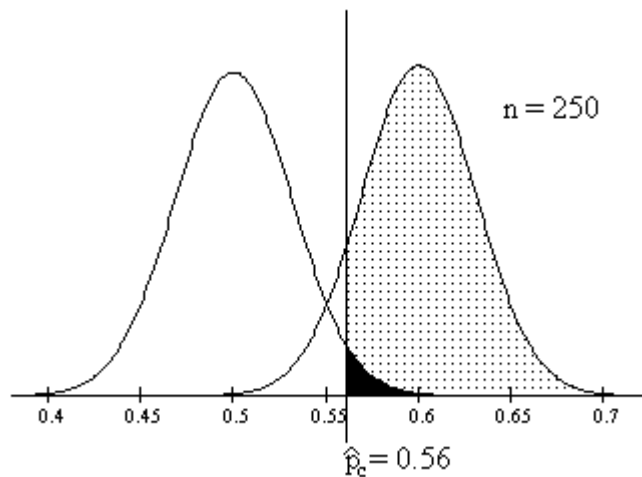
Although there are two inequalities to be considered when simplifying

$$1 - P\left(0.5 - 1.96\sqrt{\frac{0.5(0.5)}{n}} \leq \hat{p} \leq 0.5 + 1.96\sqrt{\frac{0.5(0.5)}{n}} \mid |\hat{p} - 0.5| \geq 0.1\right) \geq 0.8,$$

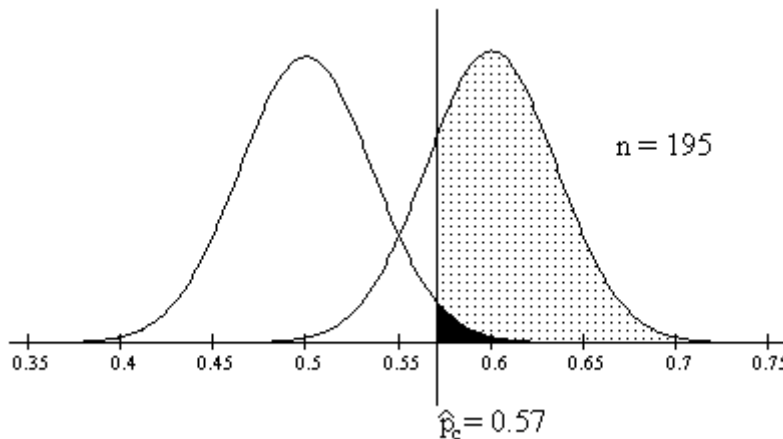
it is really only the right inequality that has a significant probability when $p > 0.5$. When $p = 0.6$, we need to compute the probability of getting a value of \hat{p} that would cause us to reject the null hypothesis. We will denote the cut-off value of \hat{p} with \hat{p}_c . The appropriate value of \hat{p}_c is given by $0.5 + 1.96\sqrt{\frac{0.5(0.5)}{n}}$ for different values of n . For example, if $n = 120$, then $\hat{p}_c = 0.59$. The alpha level of 0.05 is shaded black, while the power is dotted. This area is not large enough to be 0.8, so $n = 120$ is not a large enough sample.



For example, if $n = 250$, then $\hat{p}_c = 0.56$. The alpha level of 0.05 is shaded black, while the power is dotted. This area is more than 0.8, so $n = 250$ is a larger sample than is necessary.



For example, if $n = 195$, then $\hat{p} = 0.57$. The alpha level of 0.05 is shaded black, while the power is dotted. This area is approximately 0.8, so $n = 195$ satisfies the power requirements of the problem without going overboard.



Scope of Inference

Recall that the purpose of the study was to determine whether the probability of obtaining a head from a spin of a coin differed from 0.5. Every action one takes in conducting an experiment affects the analysis and/or conclusions from the study. In this case, I always used the same 1999 penny. What impact does that have on my conclusions? Can I draw inference about any other penny?

The scope of inference refers to the population to which inference can reasonably be drawn based on the study. This population is the population from which the random sample used in the study was drawn. If only one penny was used, we consider the results a random sample of results possible with this penny. We can comment only on this penny. If someone suggests that our penny is special and other pennies wouldn't spin similarly, we have no counter-argument. We only have information about that one penny.

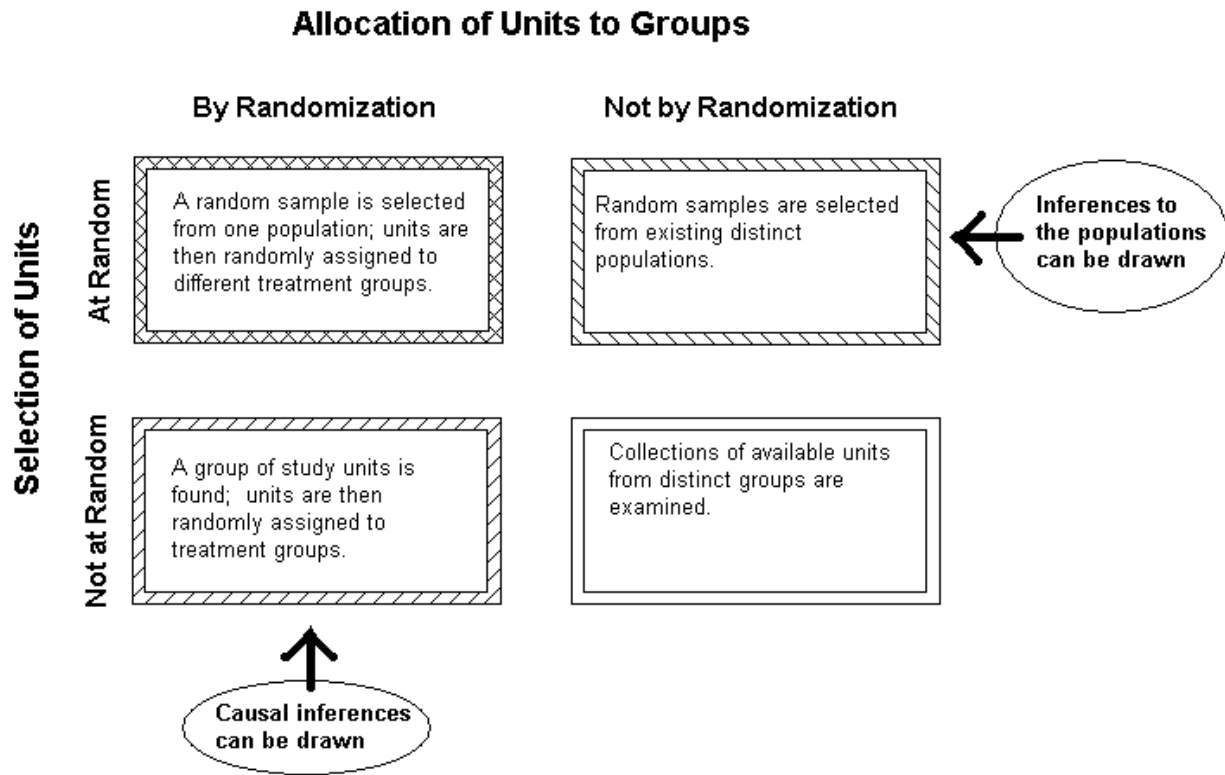
What are the advantages to increasing the scope of inference? What are the advantages to decreasing the scope of inference?

If we had taken a random sample of pennies from a collection of 1999 pennies, we introduce the variation inherent in those pennies. Some are worn more than others, may be nicked or otherwise altered. This variation makes it more difficult for us to find a difference if one exists. However, if we do find a significant difference from $p = 0.5$, we can say something about 1999 pennies, not just one special penny.

If we had taken a random sample from pennies of all ages, even more variation is put into the system. The more variation, the more difficult to achieve a significant result, but if we do, we can make statements about the spinning probabilities of pennies of all years, not just 1999 pennies or a single penny.

Randomization and Inference

The diagram below from Ramsey and Schafer (1997) illustrates the essential role randomization plays in the scope of statistical inference. There are four fundamental compartments in the diagram below formed by partitioning the Selection of Units and the Allocation of Units to Group into two distinct groups, namely, *By Randomization* and *Not by Randomization*.



Ramsey, Fred L. and Daniel W. Schafer. 1997. *The Statistical Sleuth: A Course in Methods of Data Analysis*. Duxbury Press: Belmont CA.

Suppose a dentist wants to know if a daily dose of 500 mg of vitamin C will result in fewer canker sores in the mouth than taking no vitamin C.

We will consider 4 scenarios that correspond to the 4 divisions shown above.

Case 1) No Randomization in Selection of Units and No Randomization in Allocation of Units to Treatments

The dentist, working through the local dental society, convinces all of the dental patients in town with appointments the first two weeks in December to be subjects in an experiment. He divides them into two groups, those who take at least 500 mg of vitamin C each day and those who don't. He then asks them how often they have canker sores in

their mouth and checks their patients records to see who has complained about canker sores. He compares the proportion of those who take vitamin C daily and complain of canker sores with the proportion of those who don't take vitamin C and complain of canker sores. There is a significant difference in the two proportions, with a significantly smaller proportion of those taking vitamin C having canker sores. What can we conclude?

Case 2) No Randomization in Selection of Units but Randomization in Allocation of Units to Treatments

A dentist, working through the local dental society, convinces all of the dental patients in town with appointments the first two weeks in December to be subjects in an experiment. He randomly assigns half of them to take 500 mg of vitamin C each day and the other half to abstain from taking vitamin C for three months. At the end of this time he determines the proportion of each group that has suffered from canker sores during those three months. There is a significant difference in the two proportions, with a significantly smaller proportion of those taking vitamin C having canker sores. What can we conclude?

Case 3) Randomization in Selection of Units but No Randomization in Allocation of Units to Treatments

The dentist, working through the local dental society, selects a random sample of dental patients in town and convinces them to be subjects in an experiment. He divides them into two groups, those who take at least 500 mg of vitamin C each day and those who don't. He then asks them how often they have canker sores in their mouth and checks their patients records to see who has complained about canker sores. He compares the proportion of those who take vitamin C daily and complain of canker sores with the proportion of those who don't take vitamin C and complain of canker sores. There is a significant difference in the two proportions, with a significantly smaller proportion of those taking vitamin C having canker sores. What can we conclude?

Case 4) Randomization in Selection of Units and Randomization in Allocation of Units to Treatments

The dentist, working through the local dental society, selects a random sample of dental patients in town and convinces them to be subjects in an experiment. He randomly assigns half of them to take 500 mg of vitamin C each day and the other half to abstain from taking vitamin C for three months. At the end of this time he determines the proportion of each group that has suffered from canker sores during those three months. There is a significant difference in the two proportions, with a significantly smaller proportion of those taking vitamin C having canker sores. What can we conclude?

Conclusions

Case 1) Since the patients do not represent a random sample from any population, it is not possible to make any inference about this result holding for a larger population. Since the study was observational, with subjects not randomly assigned to treatments, no causal inference can be made. We just know that for these patients, those who take vitamin C have fewer canker sores than those who don't. We don't know why, and we don't know if this result would be consistent with another group.

Case 2) Since the patients do not represent a random sample from any population, it is not possible to make any inference that this result would hold for a larger population. However, the treatments were randomly assigned to the subjects, so (assuming other factors were controlled or randomized) the difference in proportions having canker sores can be attributed to the vitamin C. We don't know if this result would be consistent with another group, but we believe we know why, for this group, the proportions differ.

Case 3) Since the patients selected were a random sample of dental patients in town, we can infer that the results observed in this experiment would be consistent with results from the whole population of dental patients in this town. However, since the study was observational, with subjects not being randomly assigned to treatments, no causal inference can be made. We believe that for the population of dental patients in this town, that those taking vitamin C have fewer canker sores than those who didn't. We don't know if it is the vitamin C that causes this reduction or some other confounding variable. We cannot conclude that for the general population, those taking vitamin C have fewer canker sores, since the sample was only of dental patients. To the extent that the dental patients in this town are representative of dental patients in general, we can infer that dental patients who take vitamin C tend to have fewer canker sores than those who don't.

Case 4) Since the patients selected were a random sample of dental patients in town, we can infer that the results observed in this experiment would be consistent with results from the whole population of dental patients in this town. Moreover, the treatments were randomly assigned to the subjects, so (assuming other factors were controlled or randomized) the difference in proportions having canker sores can be attributed to the vitamin C. We believe that for the population of dental patients in this town, that those taking vitamin C have fewer canker sores than those who don't. Also, we believe that the reduction in canker sores is a consequence of taking the vitamin C. We cannot conclude that for the general population, those taking vitamin C have fewer canker sores, since the sample was only of dental patients. To the extent that the dental patients in this town are representative of dental patients in general, we can infer that dental patients who take vitamin C tend to have fewer canker sores than those who don't, as a result of taking the vitamin C.

