1 Power

1.1 Precision of Parameter Estimates

Statistical power refers to the precision of estimates derived from samples, and especially to the relation between precision and sample size. We have seen that when we estimate parameters in a model from a sample, i.e.,

$$\hat{y}_i = b_0$$  \hspace{1cm} (4.1)

The parameter $b_0$ is an estimate of the population parameter $B_0$, with a standard error equal to \( \frac{s}{\sqrt{n}} \) and a $\alpha$% confidence interval equal to,

$$\pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$  \hspace{1cm} (4.2)

Notice that the width of the confidence interval, which is a reflection of the degree of precision in the sample estimate, depends jointly on three considerations: the $\alpha$ level we choose (the more confident we want to be about the location of the population parameter, the wider the interval needs to be); the standard deviation of the sample, which is an estimate of the standard deviation of the population (it is harder to locate population parameters in populations with a lot of variation, so intervals must be wider); and sample size (as the sample size increases, we expect the sample estimate to be closer to the population parameter, so the width of the confidence interval decreases).

In different circumstances, models can estimate the same parameter with very different degrees of precision. Two experiments might both result in the conclusion that $b_0=10$, but in one the confidence interval states that $9.8 \leq B_0 \leq 10.2$, which suggests that we have a pretty good idea about what we are doing; in the other the confidence interval might state that $2.8 \leq B_0 \leq 17.2$, which suggests that we are essentially guessing. Since the goal of collecting data in samples is to estimate population parameters, and since we can never estimate them precisely, we need some way to determine how much precision we can expect from any given research design. These methods are collectively referred to as power analysis.

In conducting a power analysis, the first consideration is to determine how much precision the researcher requires. This is not a statistical question and often throws researchers off for that reason. Instead, the question requires the researcher to make an informed value judgment based on the purpose of the research and the various costs of the resources required to produce a desired degree of precision. One cannot simply say, “I would like to achieve as much precision as possible,” because there is no upper limit to the precision available, and perfection can never be achieved. Presumably researchers to not want to include 10,000 Psych 101 students in their studies. The reason for this is not that they are uninterested in the precision that such a large sample would provide, but rather that the incremental precision that would accrue from so many participants is not justified by the enormous costs in time and money that would be required. Using three
participants would be cheap, but would probably not provide sufficient precision. As is so often the case in statistics, the goal is to find a reasonable middle ground.

All three determinants of precision are under the control of the experimenter. The easiest to manipulate, if not always the easiest to make a decision about, is the degree of confidence in the parameter that is desired in the form of an a level. It has become customary to compute 95% confidence intervals, but this is more a consequence of habit than anything else. If the parameter in question is an indicator of the toxicity of potential drug, we would probably want to be very certain we were estimating it accurately; if it is a measure of a laboratory intervention in the early phases of inquiry, we might be satisfied with much less. For now, we will accept the 95% standard and proceed to the usual subject of decisions about statistical power, sample size. (The third consideration, the variability of the population, is the most difficult to manipulate, although circumstances arise in which it can be. In any case, some more advanced concepts are required so we will not consider it until later chapters.)

Suppose a researcher is investigating how long a string of numbers people can immediately recall after the sequence has been read to them. To conduct a power analysis, the first question the researcher would have to have to address concerns the expectation for the outcome: What, roughly, would he or she expect the answer to be, and how much variation is expected among people? Oh, about seven digits, might be the reply, with a standard deviation of two. And, given a desire to achieve 95% confidence of the location of the population parameter, how much imprecision in this estimate is the researcher willing to settle for? Again, there is no correct answer, but it might seem reasonable to want to be 95% confident of the answer to within ±1 digit. How many participants should the subject collect in order to achieve this degree of precision?

At this point it is not a difficult problem. The confidence interval around whatever value of \( \theta_0 \) is estimated is given by Equation x.y. If we set this equation equal to 1 (the imprecision in either direction the researcher is willing to accept) and solve for \( n \), we obtain,

\[
    n = \left( \frac{t \cdot a \cdot s}{2} \right)^2
\]

The only difficulty is that the value for \( t \), because it depends on \( df \), is a function of \( n \), which is what we are trying to compute. We can solve this problem approximately or more precisely. Working approximately, we can check the critical values of \( t \) for some reasonable sample sizes:

```r
> qt(.975, 2:30)
[8]  2.262157   2.228139   2.200985   2.178813   2.160369   2.144787   2.131450
[15]  2.119905   2.109816   2.100922   2.093024   2.085963   2.079614   2.073873
[22]  2.068658   2.063899   2.059539   2.055529   2.051831   2.048407   2.045230
[29]  2.042272
> mean(qt(.975, 2:30))
[1]  2.288098
```
and conclude that the $t$ value we will be working with will be somewhere around 2.25. Therefore,

$$(2.25*2)^2$$

[1] 20.25

and we would conclude that we would require about 20 participants to achieve our desired degree of precision. Power analysis is often an enterprise where working approximately is appropriate, because it is necessarily based on value judgments about the costs of errors and expectations about one’s as yet uncollected data. S makes it easy to be a little more precise, by taking different $t$ values into account in our calculation. Returning to the original formula x,y, the width of the confidence interval for any $n$ can be calculated as,

nvals_2:30
> nvals
[1] 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
[23] 24 25 26 27 28 29 30
> dfvals_nvals-1
> qt(.975,dfvals)*2/sqrt(nvals)
[1] 0.5683937

or about .57 digits wide. Or if the researcher with 50 participants was satisfied with a confidence interval of ±1 units, how confident could he or she be with the interval? Now we need to solve Equation x,y for the critical $t$ value, which gives us,

$$t_{\alpha/2} = \frac{\sqrt{n}}{s}$$

(4.4)

> sqrt(50)/2
[1] 3.535534
> pt(3.536,49)
[1] 0.9995506

So 1-\(\alpha/2\) equals .9996, from which you can compute that the researcher can be 99.9% confident in the interval.

It is often useful to express information about power and precision in the form of a graph. So, for example, if we wanted to explore the imprecision entailed by a variety of sample sizes in the current example, we might proceed as follows:

```r
nvals_5:100
> int.width_qt(.975,nvals-1)*2/sqrt(nvals)
> plot(nvals,int.width,type="l",xlab="Sample Size")
```

Which gives us the following plot, which we can use to make an informed decision about sample size. Notice that as samples get larger, there are diminishing returns for adding more participants.

**Figure 1.1:** Power as a function of sample size

### 1.2 Precision of Effect Size Estimates

The issues involved in the precision of sample-based effect size estimates are very similar. The only difference is that the magnitude of the effect size depends not only on the estimated value of the parameter and an expectation for the standard deviation, but also on the *a priori* value in the null model. The preliminary discussion in the design stages of the research might go like this.

A researcher has identified a condition under which respondents may be able to recall a longer string of digits than the typical value of 10. The researcher expects that the effect size should be around \(d=+.5\), and is willing to estimate \(d\) to within \(\pm .2\). How many participants are required?

The \(\alpha\)% confidence interval around a signed sample value of \(d\) is given by,

\[
\frac{t}{\sqrt{n}}
\]

If we draw a graph of the relation between \(n\) and precision like we did in the previous section,

```r
nvals_50:150
> int.width_d_qt(.975,nvals-1)/sqrt(nvals)
> plot(nvals,int.width.d,type="l",xlab="Sample Size")
```

Which gives us the following graph, from which it can be seen that the researcher will need almost 100 participants to achieve this degree of accuracy.
1.3 Hypothesis Testing and Power

Classical treatments of statistical power are based on the ability to achieve statistical significance for a given effect size, alpha level and sample size. In the hypothesis testing paradigm, one makes a decision about whether or not to reject the null model. That decision can be either correct or incorrect. Although in practice we do not have any way of determining whether or not any particular decision about the null model is correct (if we could know for sure whether it was correct, there would be no need to conduct the experiment), we can find out about the probability that an error will be made.

In fact, when we make a decision about a null model, we can make two kinds of errors. We can decide to reject the null model when it is actually true, or we can decide not to reject it when it is actually false. We can also, of course, make the right decision. All of the possibilities can be summarized in a two by two table:

<table>
<thead>
<tr>
<th>Researcher's Decision</th>
<th>Don't Reject</th>
<th>Reject</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unknown State</td>
<td>Null Model True</td>
<td>Correct</td>
</tr>
<tr>
<td>Of Nature</td>
<td>Null Model False</td>
<td>Type 2, (\beta) error</td>
</tr>
</tbody>
</table>

Table 1.1: Type I and Type II errors

1.4 Type I Errors

Type I, or alpha, errors, are the easier of the two kinds to think about. We make a Type I error when the null model is true and we reject it anyway. How often does this happen? The rules of hypothesis testing say that we compute the probability of obtaining our data on the assumption that the null model is true, and then decide to reject the null model if that probability is less than \(\alpha\). Therefore, if the null model is indeed true (unknown to us, of course) \(\alpha\)% of the time we will draw a sample that exceeds \(\alpha\) simply because of sampling error. When this happens we will reject the null model, and therefore make an \(\alpha\) error. The \(\alpha\) error we choose to test a hypothesis, therefore, has a very clear meaning: it is the \(\alpha\) error rate we are willing to accept, the percentage of the time we are willing to be incorrect when we reject the null model.

1.5 Type II Error

If the alpha level of a hypothesis test or a confidence interval is an error rate, why not just set it as low as possible? If we set the alpha to .00001 we could be assured that we would practically never make a Type I error. The problem is that there is another kind of error. Type I errors are rejecting the null model when it is true; Type II, or \(\beta\), errors are failures to reject the null hypothesis when they are false. The problem is one of sensitivity, and is inherent in any system designed to detect something in the presence of noise. You know those metal detectors that are used to find coins on a beach? They need to be designed so they beep when there is something significant under the sand, but not when there is
something unimportant, but this goal cannot be achieved unfailingly. The units therefore come with a sensitivity dial. If the sensitivity is turned up, the machine is sure to beep for anything important, but will also beep for some little things the user doesn’t want to find. If the sensitivity is turned down the unit will not beep unnecessarily, but may miss some treasures. The point is that there is no way out of this tradeoff, all the user can do is select a balance that best suits his or her needs.

So it is with Type I and Type II errors. If the “user” dials down the Type I error rate, a price must be paid in terms of an increased Type II rate. Let’s investigate how this works, for both the Type II error rate and it’s opposite, the probability of correctly rejecting the null model when it is false, which is called statistical power and is equal to $1 - \beta$. Computing the Type II error is a little more complex, for the same reasons that complicated the estimates of precision we have already worked through: to compute Type II error, we need to make some assumptions about the characteristics of the effect we are investigating.

To conduct a hypothesis test we assume that the null model is true and compute the probability of obtaining our sample results. What makes computation of Type II error a little tricky is that we have to imagine going through the process of assuming the null model to be true under conditions when we know it to be false. It will be easiest to explain by working an example.

Let’s return to the investigator planning a study of ability to recall strings of numbers under a special condition. The investigator expects that the new conditions will increase the mean number of digits recalled by about a digit, with a standard deviation of 2; this is a way of saying that the investigator expects an effect size of .5. The investigator plans to collect data from 20 participants and use an $\alpha$ level of .05. What will the Type II error rate be?

Once the data are collected, the researcher will compare the obtained sample results to a $t$ value based on the $\alpha$ level and the sample size. With $n=20$ and $\alpha=.05$, the $t$ value required is equal to,

$$> \text{qt (.975, 19)}$$
$$[1] \ 2.093024$$

We will obtain this $t$ value when the sample mean deviates sufficiently from the null model value of 7. We have,

$$2.09 = \frac{\bar{y} - 7}{s/\sqrt{20}} \quad (4.6)$$

Based on our expectation that $s$ will be 2 and solving for $\bar{y}$, we obtain,

$$\bar{y} = 2.09 \frac{2}{\sqrt{20}} + 7 = 7.93 \quad (4.7)$$
So we will reject the null model when \( \bar{y} \) is greater than 7.93. The question is, what percentage of the time will this happen, given our assumption about the magnitude of the effect we are looking for? How often will a sample mean of at least 7.93 occur if the population has a mean of 8 and a standard deviation of 2?

This is a standard sampling distribution problem. The sampling distribution will be \( t \) distributed with a standard deviation equal to \( \frac{2}{\sqrt{20}} = .45 \), so the \( t \) value corresponding to a mean of 7.93 is given by,

\[
t = \frac{7.93 - 8}{\frac{2}{\sqrt{20}}}
\]

so we have,

\[
> \frac{(7.93-8)}{(2/\text{sqrt}(20))}
\]

\[
[1] \ -0.1565248
\]

\[
> \text{pt}(-.157,19)
\]

\[
[1] \ 0.4384506
\]

Under the assumption that the alternative model is true, we will obtain a sample mean less than 7.93 44% of the time. When that happens, we will not be able to reject the null hypothesis, even though we have assumed it is false in conducting the computation. Therefore we expect to make a Type II error 44% of the time. We could also say that our power is equal to 56%.

This result is, unfortunately, typical: although researchers go to some considerable effort to ensure that the \( \alpha \) error rate is maintained at an abstemious .05, small sample sizes frequently ensure that the \( \beta \) error rate is extremely high. Suppose the researcher determined that this Type II error rate was too high; how many participants would be required to reduce it to a more acceptable level, such as 20%?

To solve this problem, we need to take into account that as the sample size increases, the critical value \( t \) required to reject the null model gets closer to 7, and the standard error of the observed mean of \( y \) also gets smaller. This is easiest to accomplish by computing the resulting error rates for a sequence of sample sizes and drawing a graph. Here is a script of S commands to accomplish the task:

**Figure 1.3: Type II error rate and sample size**

From which we can see that a Type II error rate of 20% would be achieved with between 30 and 40 participants.