9

Power and the computation of sample size

A statistical test will not be able to detect a true difference if the sample size is too small compared with the magnitude of the difference. When designing experiments, the experimenter should try to ensure that a sufficient amount of data are collected to be reasonably sure that a difference of a specified size will be detected. R has methods for doing these calculations in the simple cases of comparing means using one- or two-sample *t* tests and comparing two proportions.

9.1 The principles of power calculations

This section outlines the theory of power calculations and sample-size choice. If you are practically inclined and just need to find the necessary sample size in a particular situation, you can safely skim this section and move quickly to subsequent sections that contain the actual R calls.

The basic idea of a hypothesis test should be clear by now. A test statistic is defined, and its value is used to decide whether or not you can accept the (null) hypothesis. Acceptance and rejection regions are set up so that the probability of getting a test statistic that falls into the rejection region is a specified significance level (α) if the null hypothesis is true. In the present context, it is useful to stick to this formulation (as opposed to the use of p-values), as rigid as it might be.

P. Dalgaard, *Introductory Statistics with R*,
DOI: 10.1007/978-0-387-79054-1_9, © Springer Science+Business Media, LLC 2008

Since data are sampled at random, there is always a risk of reaching a wrong conclusion, and things can go wrong in two ways:

- The hypothesis is correct, but the test rejects it (type I error).
- The hypothesis is wrong, but the test accepts it (type II error).

The risk of a type I error is the significance level. The risk of a type II error will depend on the size and nature of the deviation you are trying to detect. If there is very little difference, then you do not have much of a chance of detecting it. For this reason, some statisticians disapprove of terms like "acceptance region" because you can never prove that there is no difference — you can only fail to prove that there is one.

The probability of rejecting a false hypothesis is called the *power* of the test, and methods exist for calculating or approximating the power in the most important practical situations. It is inconvenient to talk further about these matters in the abstract, so let us move on to some concrete examples.

9.1.1 *Power of one-sample and paired t tests*

Consider the case of the comparison of a sample mean to a given value. For example, in a matched trial we wish to test whether the difference between treatment A and treatment B is zero using a paired t test (described in Chapter 5).

We call the true difference δ. Even if the null hypothesis is not true, we can still work out the distribution of the test statistic, provided the other model assumptions hold. It is called the *noncentral t distribution* and depends on a noncentrality parameter as well as the usual degrees of freedom. For the paired t test, the noncentrality parameter ν is a function of δ, the standard deviation of differences σ, and the sample size n and equals

$$\nu = \frac{\delta}{\sigma/\sqrt{n}}$$

That is, it is simply the true difference divided by the standard error of the mean.

The cumulative noncentral t distribution is available in R simply by adding an `ncp` argument to the `pt` function. Figure 9.1 shows a plot of `pt` with `ncp=3` and `df=25`. A vertical line indicates the upper end of the acceptance region for a two-sided test at the 0.05 significance level. The plot was created as follows:

```
> curve(pt(x,25,ncp=3), from=0, to=6)
> abline(v=qt(.975,25))
```

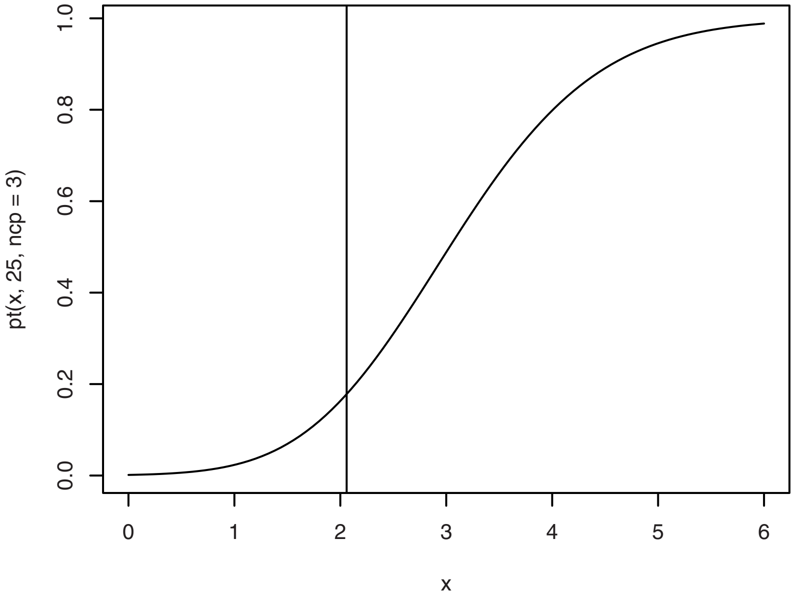

Figure 9.1. The cumulative noncentral t distribution with $\nu = 3$ and 25 degrees of freedom. The vertical line marks the upper significance limit for a two-sided test at the 0.05 level.

The plot shows the main part of the distribution falling in the rejection region. The probability of getting a value in the acceptance region can be seen from the graph as the intersection between the curve and the vertical line. (Almost! See Exercise 9.4.) This value is easily calculated as

```
> pt(qt(.975,25),25,ncp=3)
[1] 0.1779891
```

or roughly 0.18. The power of the test is the opposite, the probability of getting a significant result. In this case it is 0.82, and it is of course desirable to have the power as close to 1 as possible.

Notice that the power (traditionally denoted β) depends on four quantities: δ, σ, n, and α. If we fix any three of these, we can adjust the fourth to achieve a given power. This can be used to determine the necessary sample size for an experiment: You need to specify a desired power ($\beta = 0.80$ and $\beta = 0.90$ are common choices), the significance level (usually given by convention as $\alpha = 0.05$), a guess of the standard deviation, and δ, which is known as the "minimal relevant difference" (MIREDIF) or "smallest meaningful difference" (SMD). This gives an equation that you can solve for n. The result will generally be a fractional number, which should of course be rounded up.

You can also work on the opposite problem and answer the following question: Given a feasible sample size, how large a difference should you reasonably be able to detect?

Sometimes a shortcut is made by expressing δ relative to the standard deviation, in which case you would simply set σ to 1.

9.1.2 *Power of two-sample t test*

Procedures for two-sample t tests are essentially the same as for the one-sample case, except for the calculation of the noncentrality parameter, which is calculated as

$$\nu = \frac{\delta}{\sigma\sqrt{1/n_1 + 1/n_2}}$$

It is generally assumed that the variance is the same in the two groups; that is, using the Welch procedure is not considered. In sample-size calculations, one usually assumes that the group sizes are the same since that gives the optimal power for a given total number of observations.

9.1.3 *Approximate methods*

For hand calculations, the power calculations can be considerably simplified by assuming that the standard deviation is known, so that the t test is replaced by a test in the standard normal distribution. The practical advantage is that the approximate formula for the power is easily inverted to give an explicit formula for n. For the one- and two-sample cases, this works out as

$$n = \left(\frac{\Phi_{\alpha/2} + \Phi_\beta}{\delta/\sigma}\right)^2 \qquad \text{one-sample}$$

$$n = 2 \times \left(\frac{\Phi_{\alpha/2} + \Phi_\beta}{\delta/\sigma}\right)^2 \qquad \text{two-sample, each group}$$

with the Φ_x denoting quantiles on the normal distribution. This is for two-sided tests. For one-sided tests, use α instead of $\alpha/2$.

These formulas are often found in textbooks, and some computer programs implement them rather than the more accurate method described earlier. They do have the advantage of more clearly displaying theoretical properties such as the proportionality of δ and $1/\sqrt{n}$ for a given power.

However, they become numerically unsatisfactory when the degrees of freedom falls below 20 or so.

9.1.4 Power of comparisons of proportions

Suppose you wish to compare the morbidity between two populations and have to decide the number of persons to sample from each population. That is, you plan to perform a comparison of two binomial distributions as in Section 8.2 using `prop.test` or `chisq.test`.

For binomial comparisons, exact power calculations become unwieldy, so we rely on normal approximations to the binomial distribution. The power will depend on the probabilities in both groups, not just their difference. As for the *t* test, the group sizes are assumed to be equal. The theoretical derivation of the power proceeds along the same lines as before by calculating the distribution of $\hat{p}_1 - \hat{p}_2$ when $p_1 \neq p_2$ and the probability that it falls outside the range of values compatible with the hypothesis $p_1 = p_2$. Assuming equal numbers in the two groups, this leads to the sample-size formula

$$n = \left(\frac{\Phi_{\alpha/2}\sqrt{2p(1-p)} + \Phi_{\beta}\sqrt{p_1(1-p_1) + p_2(1-p_2)}}{|p_2 - p_1|} \right)^2$$

in which $p = (p_1 + p_2)/2$.

Since the method is only approximate, the results are not reliable unless the expected number in each of the four cells in the 2 × 2 table is greater than 5.

9.2 Two-sample problems

The following example is from Altman (1991, p. 457) and concerns the influence of milk on growth. Two groups are to be given different diets, and their growth will be measured. We wish to compute the sample size that with a power of 90%, using a two-sided test at the 1% level, can find a difference of 0.5 cm in a distribution with a standard deviation of 2 cm. This is done as follows:

```
> power.t.test(delta=0.5, sd=2, sig.level = 0.01, power=0.9)

     Two-sample t test power calculation

              n = 477.8021
          delta = 0.5
```

```
             sd = 2
      sig.level = 0.01
          power = 0.9
    alternative = two.sided

 NOTE: n is number in *each* group
```

`delta` stands for the "true difference", and `sd` is the standard deviation. As is seen, the calculation may return a fractional number of experimental units. This would, of course, in practice be rounded up to 478. In the original reference, a method employing nomograms (a graphical technique) is used and the value obtained is 450. The difference is probably due to difficulty in reading the value off the nomogram scale. To know which power you would actually obtain with 450 in each group, you would enter

```
> power.t.test(n=450, delta=0.5, sd=2, sig.level = 0.01)

     Two-sample t test power calculation

              n = 450
          delta = 0.5
             sd = 2
      sig.level = 0.01
          power = 0.8784433
    alternative = two.sided

 NOTE: n is number in *each* group
```

The system is that exactly four out of five arguments (`power`, `sig.level`, `delta`, `sd`, and `n`) are given, and the function computes the missing one (defaults exist to set `sd=1` and `sig.level=0.05` — if you wish to have those calculated, explicitly pass them as `NULL`). In addition, there are two optional arguments: `alternative`, which can be used to specify one-sided tests; and `type`, which can be used to specify that you want to handle a one-sample problem. An example of the former is

```
> power.t.test(delta=0.5, sd=2, sig.level = 0.01, power=0.9,
+ alt="one.sided")

     Two-sample t test power calculation

              n = 417.898
          delta = 0.5
             sd = 2
      sig.level = 0.01
          power = 0.9
    alternative = one.sided

 NOTE: n is number in *each* group
```

9.3 One-sample problems and paired tests

One-sample problems are handled by adding `type="one.sample"` in the call to `power.t.test`. Similarly, paired tests are specified with `type="paired"`; although these reduce to one-sample tests by forming differences, the printout will be slightly different.

One pitfall when planning a study with paired data is that the literature sometimes gives the intra-individual variation as "standard deviation of repeated measurements on the same person" or similar. These may be calculated by measuring a number of persons several times and computing a common standard deviation within persons. This needs to be multiplied by $\sqrt{2}$ to get the standard deviation of differences, which `power.t.test` requires for paired data. If, for instance, it is known that the standard deviation within persons is about 10, and you want to use a paired test at the 5% level to detect a difference of 10 with a power of 85%, then you should enter

```
> power.t.test(delta=10, sd=10*sqrt(2), power=0.85, type="paired")

     Paired t test power calculation

              n = 19.96892
          delta = 10
             sd = 14.14214
      sig.level = 0.05
          power = 0.85
    alternative = two.sided

 NOTE: n is number of *pairs*, sd is std.dev. of
       *differences* within pairs
```

Notice that `sig.level=0.05` was taken as the default.

9.4 Comparison of proportions

To calculate sample sizes and related quantities for comparisons of proportions, you should use `power.prop.test`. This is based on approximations with the normal distribution, so do not trust the results if any of the expected cell counts drop below 5.

The use of `power.prop.test` is analogous to `power.t.test`, although `delta` and `sd` are replaced by the hypothesized probabilities in the two groups, `p1` and `p2`. Currently, it is not possible to specify that one wants to consider a one-sample problem.

An example is given in Altman (1991, p. 459) in which two groups are administered or not administered nicotine chewing gum and the binary outcome is smoking cessation. The stipulated values are $p_1 = 0.15$ and $p_2 = 0.30$. We want a power of 85%, and the significance level is the traditional 5%. Inserting these values yields

```
> power.prop.test(power=.85,p1=.15,p2=.30)

     Two-sample comparison of proportions power calculation

              n = 137.6040
             p1 = 0.15
             p2 = 0.3
      sig.level = 0.05
          power = 0.85
    alternative = two.sided

 NOTE: n is number in *each* group
```

9.5 Exercises

9.1 The `ashina` trial was designed to have 80% power if the true treatment difference was 15% and the standard deviation of differences within a person was 20%. Comment on the sample size chosen. (The power calculation was originally done using the approximative formula. The imbalance between the group sizes is due to the use of an open randomization procedure.)

9.2 In a trial comparing a binary outcome between two groups, find the required number of patients to find an increase in the success rate from 60% to 75% with a power of 90%. What happens if we reduce the power requirement to 80%?

9.3 Plot the density of the noncentral t distribution for `ncp=3` and `df=25` and compare it with the distribution of $t + 3$, where t has a central t distribution with `df=25`.

9.4 In two-sided tests, there is also a risk of falling into the rejection region on the opposite side of the true value. The power calculations in R only take this into account if you set `strict=TRUE`. Discuss the consequences.

9.5 It is occasionally suggested to choose n to "make the true difference significant". What power would result from choosing n by such a procedure?