Power Analysis Using R

S. P. Blomberg

January 7, 2014

Introduction

In a series of papers in the early 20th century, J. Neyman and E. S. Pearson developed a decision-theoretic approach to hypothesis-testing (Neyman and Pearson, 1928a,b, 1933a,b). The theory was later extended and generalised by Wald (e.g. Wald, 1939, 1950). For a full account of the theory, see Lehmann and Romano (2005).

A centrepiece of the Neyman-Pearson approach is the idea of there being two possible types of error when considering rejecting a hypothesis based on data (See Table 1):

Table 1

	H_0 False	H_0 True
Reject H_0	Correct	Type I Error
Accept H_0	Type II Error	Correct

Thus, a rational approach to hypothesis testing will seek to reject a hypothesis if it is false and accept a hypothesis when it is true. The two types of error are rejecting the null hypothesis when it is true (Type I) and accepting the null hypothesis when it is false (Type II). In the Neyman-Pearson theory, it is usual to fix the Type I error probability (α) at some constant (often at 0.05, but not necessarily), and then choose a test which minimises the Type II error probability (β), conditional on α. The (null) hypothesis is then either rejected when the associated p-value for the test is less than α, or otherwise accepted.

The Neyman-Pearson theory has come under a lot of criticism since its formulation. (For a recent critique from a biologist and a psychologist, see Hurlbert and Lombardi (2009)) Although much maligned, it is still used to justify and compare statistical testing procedures, whether or not scientists accept the paradigm for their everyday data analyses. For our purposes, the theory introduces the concept of "power" of a test, where power is defined as P(Reject H_0|H_0 is false). That is, 1 minus the Type II error probability, or $1 - \beta$. (The first person to promote power as a concept was "Student" a.k.a W. S. Gosset, head brewer for the Guinness brewery (Ziliak and McCloskey, 2008). Interestingly, R. A. Fisher (the other main architect of classical statistics) was against the whole *idea* of power (Kruskal, 1980).)

Factors affecting the Power of a Test

There are several factors affecting the power of a test:

- There is a trade-off between α and β such that greater power can be gained by accepting a less stringent condition for the Type I error. In other words, power is increased when α is increased. The shortcoming of setting a higher α is that Type I errors will be more likely. This may not be desireable.
- Power can be increased by increasing the effect size. The reasoning is that any test will have trouble rejecting the null hypothesis if the null hypothesis is only "slightly" wrong. If the effect size of an experiment (such as the difference between a treatment and a control) is large, then it is easier to detect, and the null hypothesis will be soundly rejected.
- The use of covariates or blocking variables can also increase the power of a test, as "controlling" for other, nuisance, variables can make the test more precise.
- Increasing the sample size in an experiment or observational study also increases the power to detect a response.

It is important to carefully consider sample sizes before doing any experimental work. This is because data are expensive to collect, and may involve an appreciable amount of time and money expended by the researcher. Also, especially in medical research, there are ethical issues to consider. Experiments should be designed so that the minimum number of subjects (e.g. rats, dogs, or monkeys) are used in order to arrive at valid scientific conclusions. For these reasons, it is important to take a quantitative approach to maximising power in an experiment and to design experimental protocols that are efficient in their use of resources (subjects).

Power Analysis in R

Here we describe approaches to power analysis and sample size selection using the freely available R environment for statistical computing and graphics. Because R is free, researchers can use it anywhere in the world, and are not limited by restrictive license agreements. Thus, encouraging researchers and students to use R for statistical analyses effectively democratises the process of study design and data analysis.

Obtaining R

The easiest way to get R is to download it from the R web site. Upon visiting the official web site, you will see a sidebar with a link to CRAN, the Comprehensive R Archive Network. Click on it and choose a local mirror site (e.g. http://cran.ms.edu.au). R is available for Linux, Windows, and MacOS X. Download and install the appropriate version for your system.

Built-in R Functions for Power Analysis

R comes with a wide variety of built-in functions. For power analysis, we need to look to the `stats` package that comes with R. Try searching for functions for power analysis in the `stats` package. Your output should look approximately like this:

```
help.search("power", package="stats")
```

```
stats::power            Create a Power Link Object
stats::power.anova.test
                        Power Calculations for Balanced One-Way
                        Analysis of Variance Tests
stats::power.prop.test
                        Power Calculations for Two-Sample Test for
                        Proportions
stats::power.t.test     Power calculations for one and two sample t
                        tests
stats::print.power.htest
                        Print method for power calculation object
```

We can see that there are 5 functions with "power" in their name. However, it is clear that `power.anova.test`, `power.prop.test`, and `power.t.test` are the most appropriate for our study of power analysis. We first examine `power.t.test`.

Power of the t-test

Recall that one use of the t-test is used to test for differences between two sample means, drawn from two Normal populations with unknown variance. The null hypothesis is that the samples were drawn from a single population. i.e. H_0: the sample means are not different. Now, examine the help page for `power.t.test`, part of which is presented below:

```
?power.t.test
```

```
power.t.test              package:stats              R Documentation

Power calculations for one and two sample t tests

Description:

     Compute power of test, or determine parameters to obtain target
     power.

Usage:

     power.t.test(n = NULL, delta = NULL, sd = 1, sig.level = 0.05,
                  power = NULL,
                  type = c("two.sample", "one.sample", "paired"),
                  alternative = c("two.sided", "one.sided"),
                  strict = FALSE)
```

Notice that `power.t.test` accepts 8 *arguments*. Arguments `type`, `alternative` and `strict` describe the behaviour of the t-test. That is, whether the test is a two-sample, one-sample, or paired t-test, and whether it is a two-sided or one-sided test. Further, we are given the option of including both tails in the power calculation for the two-sided test, by setting `strict=TRUE`. Note that the default is `strict=FALSE`.

The first 5 arguments determine the type of analysis. To use the function, you specify 4 of the first 5 arguments, and the unspecified argument is the one that is output from the computation. Here's an example:

```
power.t.test(n=6, .2, sd=.1, power=NULL)

     Two-sample t test power calculation

              n = 6
          delta = 0.2
             sd = 0.1
      sig.level = 0.05
          power = 0.8764176
    alternative = two.sided

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

Note that we have left the `sig.level` argument at its default (0.05). We have specified the sample size, and the two arguments that contribute to the effect size (`delta`, the difference between the means and `sd`, the common standard deviation. More on effect sizes below). `power` was set to `NULL`, as this is the value we are trying to compute. The output is printed, including the estimated power (0.88 in this case).

1 Exercises

1. Calculate the power of a two-sample, two-sided t-test when n=12, delta=1.5, sd=2.3. Use a significance level of 0.05.

2. Using the the same t-test, calculate the sample size needed to attain a power of 0.85, with delta=6, sd=4.5. Use a significance level of 0.05.

3. `power.anova.test`: Calculate the sample size needed for a one-way ANOVA with between-group variance = 3, within-group variance=5, and 4 groups. Use a significance level of 0.05.

4. Calculate the power of a one-way ANOVA with 4 groups, within-group variance = 12, between-group variance=4, 20 subjects in each group. Use a significance level of 0.05.

5. `power.prop.test` Calculate the power of the test comparing the proportions in two groups (0.5, 0.4), with 20 in each group. Use a significance level of 0.05.

6. Calculate the sample size necessary to detect a difference in proportions where group 1 has a proportion of .6 and group 2 has a proportion of 0.8. Use power=0.85, and a significance level of 0.01.

Power Analysis in R Add-on Packages

One of the most important aspects of R is its modularity. Currently, there are over 3000 add-on packages for R on CRAN. The two that we will be dealing with are packages `pwr`, which contains more general functions for power analysis and sample size calculations, and `powerSurvEpi`, which contains functions for power and sample size determination for some survival models.

The first step is to install `pwr` and `powerSurvEpi` from CRAN. Execute the following command:

```
install.packages(c("pwr", "powerSurvEpi"))
```

You only need to execute this command once, as the install is permanent (although it may only stay on the UQ system for the duration of your computing session, and will be lost when you log off).

Package `pwr`

We will consider functions in the `pwr` package first. Load the package into R using the following command:

```
library(pwr)
```

and then look at the diverse array of functions provided in the package:

```
help(package=pwr)
```

The important part of the output is below:

```
ES.h                    Effect size calculation for proportions
ES.w1                   Effect size calculation in the chi-squared test
                        for goodness of fit
ES.w2                   Effect size calculation in the chi-squared test
                        for association
cohen.ES                Conventional effects size
pwr-package             Basic power calculations pwr
pwr.2p.test             Power calculation for two proportions (same
                        sample sizes)
pwr.2p2n.test           Power calculation for two proportions
                        (different sample sizes)
pwr.anova.test          Power calculations for balanced one-way
                        analysis of variance tests
pwr.chisq.test          power calculations for chi-squared tests
pwr.f2.test             Power calculations for the general linear model
pwr.norm.test           Power calculations for the mean of a normal
                        distribution (known variance)
```

```
pwr.p.test              Power calculations for proportion tests (one
                        sample)
pwr.r.test              Power calculations for correlation test
pwr.t.test              Power calculations for t-tests of means (one
                        sample, two samples and paired samples)
pwr.t2n.test            Power calculations for two samples (different
                        sizes) t-tests of means
```

Notice that there are many more functions for power analysis in `pwr` than in built-in package `stats`. There are additional functions for χ^2 tests (`pwr.chisq.test`), Pearson's product-moment correlation (`pwr.r.test`), and unbalanced two-sample t-tests (`pwr.t2n.test`), among others. However, they work in the same way as the previous examples from `stats`. You specify the values for all the arguments except one (which is left as NULL), and that unspecified variable is computed by the function.

2 Exercises

7. Calculate the power of the Pearson's product moment correlation test, where r = 0.8, n = 20, and significance level is 0.05.

8. Calculate the sample size necessary to detect a correlation of r = 0.6, with a power of 0.85 and significance level = 0.05.

Effect Sizes

It was mentioned previously that increasing the effect size (the standardised "difference" between treatment groups) results in an increased power. However, calculation of effect sizes varies from test to test, depending on the underlying distribution of the test statistic. Frequently, we do not know the likely effect size that may occur in an experiment. The best approach is then to do a pilot experiment on a small scale to estimate the likely effect size. In the absence of pilot data, Cohen (1988) provides standard measures of effect size, classified as "small", "medium", and "large" for a variety of tests. These effect sizes are built into the `pwr` package, using the function `cohen.ES`. Although these "standard" effect sizes are somewhat arbitrary, they can provide a first guide for sample size estimation. Note, however, that a pilot experiment is the recommended way to estimate effect sizes for an experimental study.

3 Exercises

9. Use `cohen.ES` to extract "small", "medium", and "large" effect sizes for χ^2, Pearson's r and proportion tests.

10. Use the above effect sizes to calculate sample sizes with power = 0.8, and sig.level = 0.05, using the following functions from the `pwr` package: `pwr.chisq.test`, `pwr.r.test`, `pwr.p.test`.

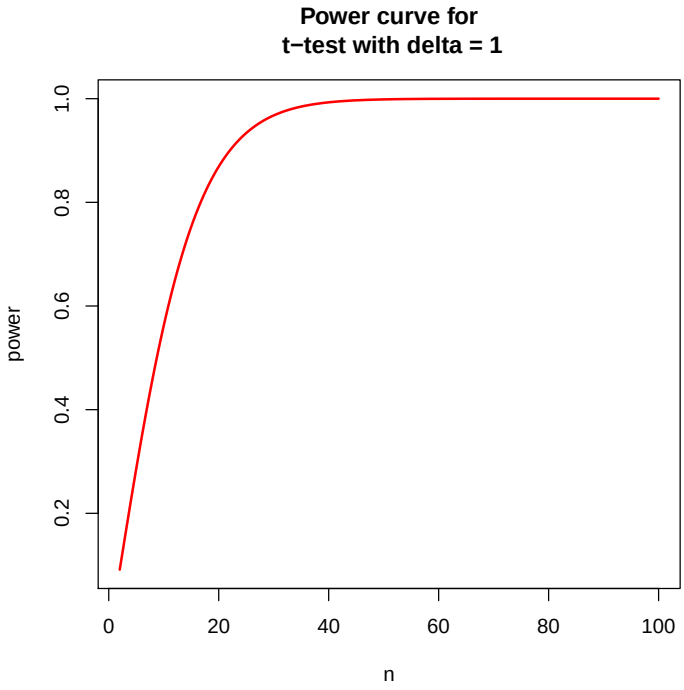
11. Calculate the power of the above tests with sample sizes 10, 50, 100.

12. Calculate the detectable effect size for the above tests when power = 0.8 and n = 10, 50, 100.

Power Curves

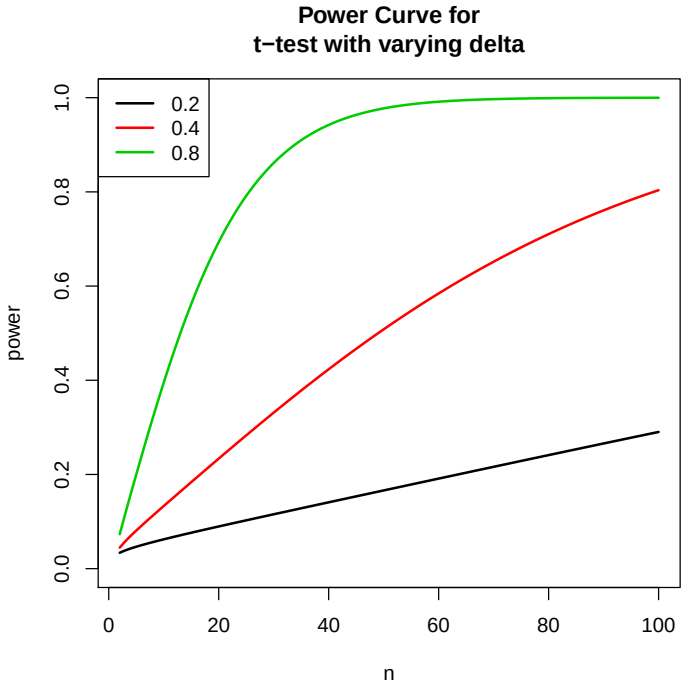
Calculating specific values of power, sample size or effect size can be illuminating with regard to the statistical restrictions on experimental design and analysis. But frequently a graph tells the story more completely and succinctly. Here we show how to draw power, sample size, and effect size curves using the above functions in R:

```
nvals <- seq(2, 100, length.out=200)
powvals <- sapply(nvals, function (x) power.t.test(n=x, delta=1)$power)
plot(nvals, powvals, xlab="n", ylab="power",
main="Power curve for\n t-test with delta = 1",
lwd=2, col="red", type="l")
```



If we are unsure of our effect size, we can also alter delta to see the effect of both effect size and sample size on power:

```
deltas <- c(0.2, 0.4, 0.8)
plot(nvals, seq(0,1, length.out=length(nvals)), xlab="n", ylab="power",
main="Power Curve for\nt-test with varying delta", type="n")
for (i in 1:3) {
powvals <- sapply(nvals, function (x) power.t.test(n=x, delta=deltas[i])$power)
lines(nvals, powvals, lwd=2, col=i)
}
legend("topleft", lwd=2, col=1:3, legend=c("0.2", "0.4", "0.8"))
```



4 Exercises

13. Make a graph of the relationship between effect size and power, for a sample size of 5, 10, and 20, using `power.anova.test`. Use 4 groups, with the within-group variance equal to 1, and the between-group variance varying between 0.1 and 1.2.

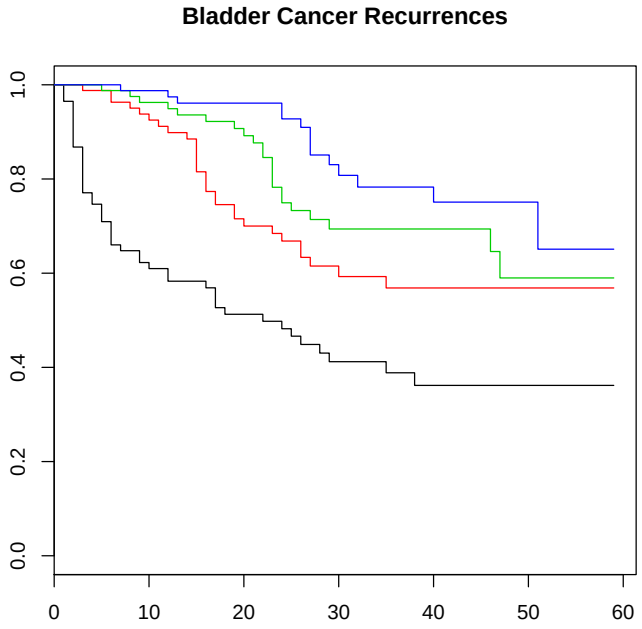
5 Power of Cox Regression

In an important paper, statistician Sir David Cox developed an approach to survival analysis which did not depend on the precise form of the baseline hazard function (Cox, 1972). Instead, Cox made a far weaker assumption: that the hazard functions were proportional for different treatment groups over the duration of the study. That is, he assumed proportional hazards. Hence, Cox's method is often referred to as Cox Proportional Hazards regression, or Cox regression.

Consider the following survival curves based on recurrences of bladder cancer in 85 patients (Wei et al., 1989). The curves look quite similar, although the black curve looks like it could be steeper than the other three.

```
library(survival)
tst <-coxph(Surv(stop, event) ~ strata(enum), bladder)
```

```
plot(survfit(tst), col=1:4, main="Bladder Cancer Recurrences",
mark.time=FALSE)
```



How many subjects do we need for such a survival study? We can answer this question using the function `powerCT` in package `powerSurvEpi`. The function only works for two groups, so we use only the groups corresponding to the black and red lines as a pilot data set, for illustration. We are interested in the power of the study with 100 Experimental subjects and 100 Control subjects, and we assume a relative risk (hazard ratio) of 1.6. That is, our effect size is 1.6, corresponding to the treatment group having a 1.6 times the expected risk of dying, compared to the control group:

```
library(powerSurvEpi)
bl2 <- bladder[bladder$enum==1|bladder$enum==2,]
bl2$enum2 <- ifelse(bl2$enum==1, "C", "E")
bl3 <- bl2[,c("stop", "event", "enum2")]
PCT100 <- powerCT(Surv(stop, event)~enum2, dat=bl3, nE=100,
nC=100, RR=1.6)
print(PCT100$power)
```

```
[1] 0.7507715
```

It seems our power is quite low. What if we increase the sample size in each treatment by 50?

```
PCT150 <- powerCT(Surv(stop, event)~enum2, dat=bl3, nE=150,
nC=150, RR=1.6)
print(PCT150$power)
```

```
[1] 0.8978767
```

This is a good improvement! We can also use the `ssizeCT` function to compute sample sizes, again based on pilot data. We assume that there are to be equal numbers in each group. ie k=1:

```
ss <- ssizeCT(Surv(stop, event)~enum2, dat=bl3,
power=0.85, RR=1.6, k=1)
print(ss$ssize)
```

```
 nE  nC
130 130
```

So if we are to have a relative risk of 1.6 and a power of 0.85, then we need at least 130 subjects in each treatment to detect a difference at $\alpha = 0.05$.

6 Exercises

14. Construct a graph of the relationship between power and sample size using the pilot data (`bl3`), for a relative risk of 0.8, 1.4, 1.8. and sample sizes ranging from 10 to 200.

15. Balanced designs (equal numbers in treatment groups) have more power than unbalanced designs. Test the effect of having different numbers in each group by examining the relationship between power and sample size for different values of k.

Power Analysis by Simulation

Frequently, the complexity of our experimental designs means that we must go far beyond what can be accomplished with standard software, such as the built-in power functions and the `pwr` package. Fortunately, R can be easily programmed to produce power analyses for any experimental design. The general approach is:

1. Simulate data under the null hypothesis (for checking Type I error probabilities) and also for different effect sizes, to estimate power.

2. Fit the model to the simulated data.

3. Record whether the analysis of the simulated data set was significant, using the usual tests.

4. Store the significance level in a vector.

5. Repeat from step 1. a large number of times.

6. Tabulate how many simulations produced a significant result, and hence calculate power.

Here is an example: Suppose we wish to conduct a study with two fixed factor, leading us to a 2-way analysis of variance (ANOVA), with two levels for each factor. We could simulate data under the null hypothesis (no difference between means) using the following code:

```
## First set up the design
reps <- 1000
design <- expand.grid(A=c("a", "b"), B=c("c","d"), reps=1:10)
pvals <- vector("numeric", length=reps)
## simulate data under the null hypothesis.
for (i in 1:reps) {
  design$response <- rnorm(40) # random data with zero mean
  ## fit the model
  fit <- aov(response~A*B, data=design)
  ## record the p-value for the interaction term.
  ## You could record other values.
  ## Save the p value
  pvals[i] <- summary(fit)[[1]][[5]][3]
}
Type1 <- length(which(pvals < 0.05))/reps
print(Type1)
```

```
[1] 0.034
```

It appears that the Type I error rate is acceptable for the 2-factor ANOVA interaction term. Now to calculate some power statistics. We will calculate power for a difference between means of 2 units, for sample sizes 5, 10, 20.

```
ssize <- c(5, 10, 20)
pvals <- matrix(NA, nrow=reps, ncol=3)
## First set up the design
for (j in 1:3) {
  reps <- 1000
  design <- expand.grid(reps=1:ssize[j], A=c("a", "b"), B=c("c","d"))

  ## simulate data under the null hypothesis.

  for (i in 1:reps) {
    design$response <- c(rnorm(3*ssize[j]), rnorm(ssize[j], mean=2))
    ## fit the model
    fit <- aov(response~A*B, data=design)
    ## record the p-value for the interaction term.
    ## You could record other values.
    ## Save the p value
    pvals[i,j] <- summary(fit)[[1]][[5]][3]
  }
}
Power <- apply(pvals, 2, function (x) length(which(x < 0.05))/reps)
names(Power) <- as.character(ssize)
print(Power)
```

```
    5    10    20
0.538 0.852 0.994
```

We see that the power is too low for a sample size of 5, but it increases to an acceptable level for 10 replicates per treatment.

7 Exercises

16. Construct a graph of the relationship between power and sample size for a multiple regression model with 3 predictor variables, over a range of 1 to 10 for each predictor. For the effect size, let the residual error have $\sigma = 5$, and $\beta_1 = 0.1$, $\beta_2 = 1$ and $\beta_3 = 5$. Try varying the effect size to examine its effects on power in this case.

Further Reading

Cohen (1988) is the classic text for non-statisticians. Hoenig and Heisey (2001) provides cautionary advice against doing *post hoc* power analyses. Further readings are below, but are mostly of historical interest.

References

Cohen J., 1988. Statistical Power Analysis for the Behavioural Sciences. Routledge, 2nd ed.

Cox D.R., 1972. Regression models and life-tables. Journal of the Royal Statistical Society. Series B (Methodological), 34:187–220.

Hoenig J.M., Heisey D.M., 2001. The abuse of power: The pervasive fallacy of power calculations for data analysis. The American Statistician 55:19–24.

Hurlbert S.H., Lombardi C.M., 2009. Final collapse of the neyman-pearson decision theoretic framework and rise of the neofisherian. Ann. Zool. Fennici 46:311–349.

Kruskal W., 1980. The significance of fisher: A review of r. a. fisher: The life of a scientist. Journal of the American Statistical Association 75:1019–1030.

Lehmann E.L., Romano J.P., 2005. Testing Statistical Hypotheses. Springer Texts in Statistics, Springer, 3rd ed.

Neyman J., Pearson E.S., 1928a. On the use and interpretation of certain test criteria for purposes of statistical inference: Part i. Biometrika 20A:175–240.

Neyman J., Pearson E.S., 1928b. On the use and interpretation of certain test criteria for purposes of statistical inference: Part ii. Biometrika 20A:263–294.

Neyman J., Pearson E.S., 1933a. On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London. Series A 231:289–337.

Neyman J., Pearson E.S., 1933b. The testing of statistical hypotheses in relation to probabilities a priori. Mathematical Proceedings of the Cambridge Philosophical Society 29:492–510.

Wald A., 1939. Contributions to the theory of statistical estimation and testing hypotheses. Annals of Mathematical Statistics 10:299–326.

Wald A., 1950. Statistical Decision Functions. New York: John Wiley and Sons.

Wei L.J., Lin D.Y., Weissfeld L., 1989. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. Journal of the American Statistical Association 84:1065–1073.

Ziliak S.T., McCloskey D.N., 2008. The cult of statistical significance : how the standard error costs us jobs, justice, and lives. Ann Arbor: University of Michigan Press.