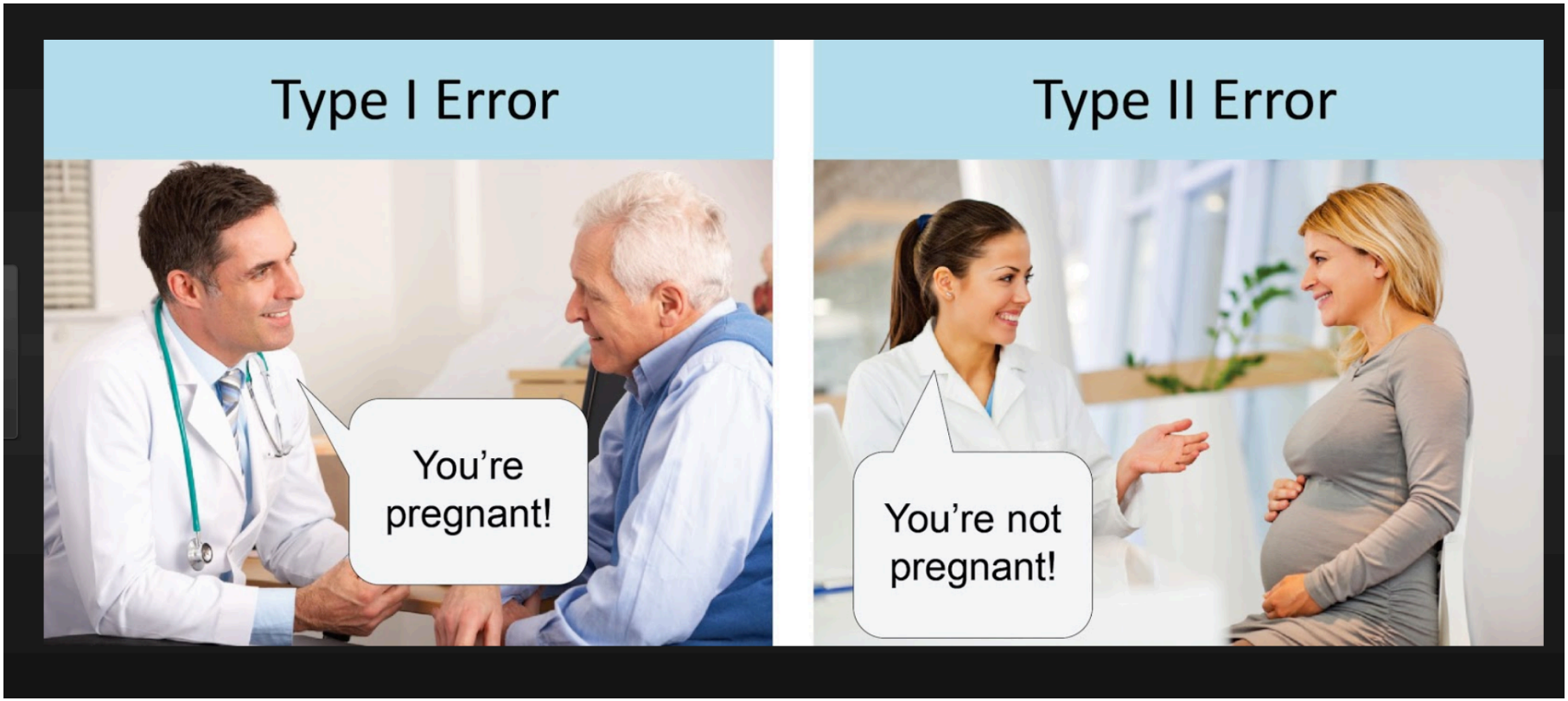
humor...


Type I Error
You're pregnant!
Type II Error
You're not pregnant!

Power

As discussed previously, power is the probability of rejecting the null hypothesis when the null is false.

Power depends on the effect size (how far from the truth the null is), the sample size, and α, the probability of rejecting when the null hypothesis is true. Power calculations usually assume that the model (such as the distribution) is correct, but only the parameters of the model described by the null hypothesis are incorrect.

Usually, if the null hypothesis is false, the probability of rejecting the null hypothesis is larger than α (assuming you have a size α test. Note that a conservative test might have size $< \alpha$.

Power

Why is power used?

One of the most common applications of power analysis is determining an appropriate sample size to make a study likely to find a statistically significant effect. Knowing what sample size might be needed is useful for planning, especially the cost of a study, and can be important when applying for grant funding.

Although statisticians often think of "more is better" in terms of sample size, in biomedical studies with side effects, there is an ethical component to sample size calculations. If it is unkonwn whether a treatment will have adverse side effects, or there are risks for a certain treatment, then having more observations than necessary for the purpose of the study means that more people are subjected to unnecessary risk.

For grant purposes and also for cases using expensive treatments or experiments, sample size calculations are also useful to make sure that a study isn't wasteful in terms of spending more money than necessary.

Power

In some cases, power can be calculated theoretically. In other cases, power calculations are complicated and must be simulated. Power calculations can be slow because to estimate a probability you might need approximately 10,000 simulated hypothesis tests (which require simulated data sets), to estimate power within 1%. This only gives you the power for a particular choice of α, n, and the effect size.

To estimate the power as a function of n or the effect size, a simulation requires simulating hundreds of thousands of data sets to get a smooth power curve.

Power

A case where the power curve can be calculated analytically, consider the following:

Let $X_1, \ldots, X_{50}$ be i.i.d. normal with mean μ and known variance $\sigma^2 = 1$. Let the null and alternative hypotheses be

$$H_0 : \mu \leq 1$$

$$H_1 : \mu > 1$$

If we are testing using $\alpha = 0.05$, we can think of the power as a function of μ since $n = 50$ is fixed.

Power

To calculate the power, we need the probability of rejecting the null hypothesis. The test statistic for this problem is

$$Z_{obs} = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\overline{X} - 1}{1/\sqrt{50}}$$

Note that when $\mu > 1$, Z_{obs} does not have a standard normal distribution, but does still follow a normal distribution. If $\mu > 1$, the distribution of Z_{obs} is normal with mean and variance

Power

$$
\begin{aligned}
E[Z_{obs}] &= E\left[\frac{\overline{X}-1}{1/\sqrt{50}}\right] \\
&= (\mu-1)\sqrt{50} \\
Var(Z_{obs}) &= Var(\sqrt{50}(\overline{X}-1)) \\
&= 50 \cdot Var(\overline{X}) \\
&= 50 \cdot Var\left(\frac{1}{50}\sum_{i=1}^{50} X_i\right) \\
&= \frac{50}{50^2}\sum_{i=1}^{50} Var(X_i) \\
&= \frac{50}{50} = 1
\end{aligned}
$$

Power

Since $Z_{obs} \sim N(\sqrt{50}(\mu - 1), 1)$, we can calculate probabilities that Z_{obs} takes different values. Note that $Z_{obs} - E[Z_{obs}]$ has a $N(0, 1)$ distribution.
For a one-sided test, reject H_0 at the $\alpha = .05$ level if $Z_{obs} \geq 1.645$.

$$\begin{aligned} P(Z_{obs} > 1.645) &= P\left(Z_{obs} - (\mu - 1)\sqrt{50} > 1.645 - (\mu - 1)\sqrt{50}\right) \\ &= P(Z > 1.645 - (\mu - 1)\sqrt{50}) \\ &= 1 - \Phi(1.645 - (\mu - 1)\sqrt{50}) \end{aligned}$$

Thus, the power function is

$$Power(\mu) = 1 - \Phi(1.645 - (\mu - 1)\sqrt{50})$$

If we were doing a two-sided hypothesis test, how should we calculate the power function?

Power

The power can now be plotted as a function of μ.

```
> mu <- seq(1,3,.01)
> power50 <- 1-pnorm(1.645-(mu-1)*sqrt(50))
> power40 <- 1-pnorm(1.645-(mu-1)*sqrt(40))
> power30 <- 1-pnorm(1.645-(mu-1)*sqrt(30))
> power20 <- 1-pnorm(1.645-(mu-1)*sqrt(20))
> power10 <- 1-pnorm(1.645-(mu-1)*sqrt(10))
> plot(mu,power50,type="l",lwd=2)
> plot(mu,power50,type="l",lwd=2,ylim=c(0,1))
> points(mu,power40,type="l",lwd=2)
> points(mu,power30,type="l",lwd=2)
> points(mu,power20,type="l",lwd=2)
> points(mu,power10,type="l",lwd=2)
```

Power

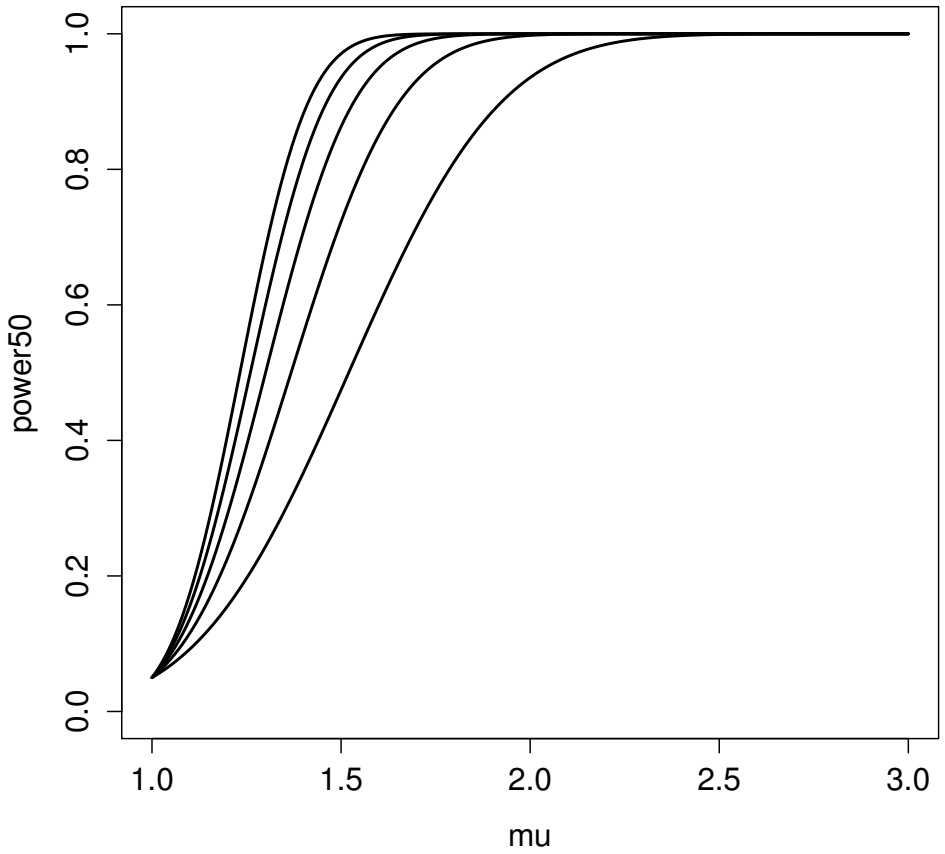


Figure : Power for 1-sided test when $n = 10, 20, \ldots, 50$.

Power

Alternatively, you can plot the power as a function of the sample sizes for selected values of μ.

```
> n <- seq(5,100,1)
> power1 <- 1-pnorm(1.645-(1-1)*sqrt(n))
> power1.2 <- 1-pnorm(1.645-(1.2-1)*sqrt(n))
> power1.1 <- 1-pnorm(1.645-(1.1-1)*sqrt(n))
> power1.3 <- 1-pnorm(1.645-(1.3-1)*sqrt(n))
> power1.4 <- 1-pnorm(1.645-(1.4-1)*sqrt(n))
> power1.5 <- 1-pnorm(1.645-(1.5-1)*sqrt(n))
```

Power

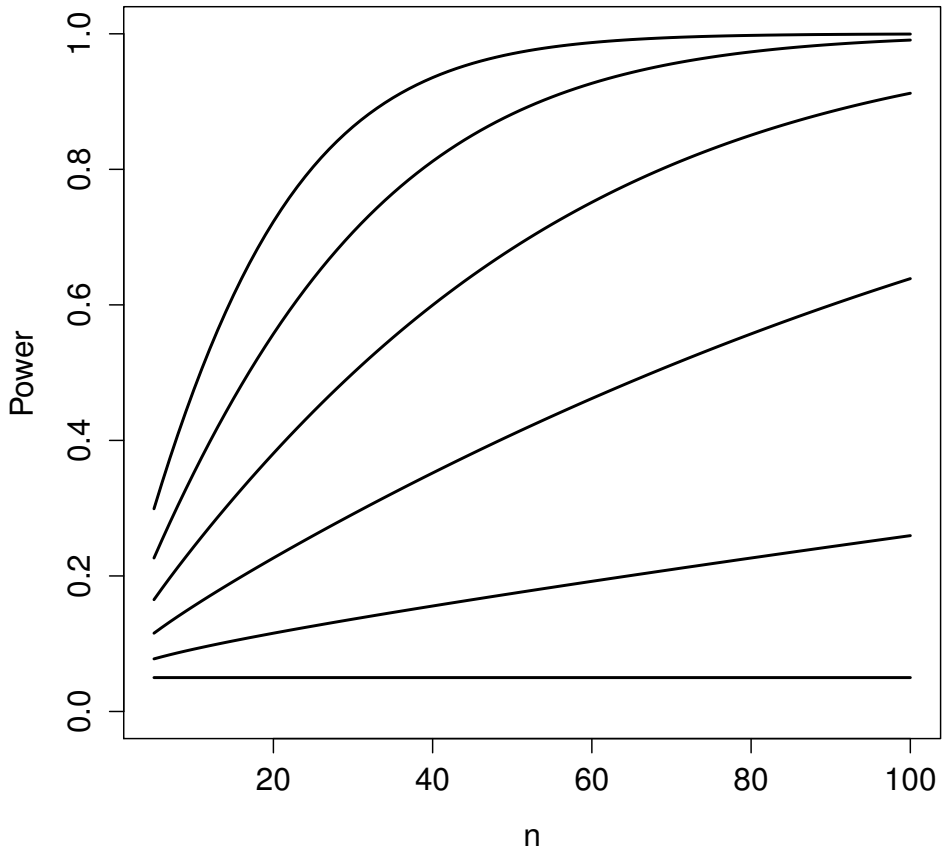


Figure : Power for 1-sided test when $\mu = 1, 1.1, \ldots, 1.5$.

Power

A nice thing about power curves is that it summarizes trends. For this problem, if the effect size is small (e.g. 1.1), the power increases very slowly with the sample size, with the power being only 0.259 with $n = 100$.

Of course, you can also use functions in R to get the power. The function `power.t.test()` is especially useful. This will give slightly different results because it uses the t distribution, but results will be quite similar for larger sample sizes. Power based on the t-test is much more likely to be useful in practice. For this problem you can use

```
> power.t.test(n=100,delta=.1,sd=1,type="one.sample",
alternative="one.sided")

     One-sample t test power calculation

              n = 100
          delta = 0.1
             sd = 1
      sig.level = 0.05
          power = 0.2573029
    alternative = one.sided
```

Power

In consulting situations, you can often use an online power calculator, at least for examples like t-tests and tests of proportions. I recommend using this for working with many biomedical people who might not be familiar with R or interested in learning R.

An example is here: https://www.stat.ubc.ca/~rollin/stats/ssize/

Power

Basing your sample size calculation on previous studies has some risks. In particular, a small study might have gotten lucky in observing a larger effect than is really there, and this could make you overestimate your power.

Also, publication bias could mean that the effect size is overestimated.

Again, there is uncertainty in the estimate of the standard deviation, so if this is underestimated (which is likely if the significance was accidentally inflated), then again you might be overestimating your power or underestimating the sample needed for a given study.

Nevertheless, the standard is to design a sample size based on 80% power. Estimating the uncertainty in this estimated power is difficult however, and the result is that many studies that are designed to have 80% power actually have less power. To me, 80% seems awfully low (a high risk of failure to observe the effect), but requiring higher power would be costly in

Power

Another issue is clinical significance. If there is an effect, such as a reduction in pain for different pain treatments, then an effect might be considered too small to be of clinical significance even if it is measurable given very large samples. It might be more difficult to argue that some level of reducing the probability of death is "clinically insignficant".

In a case where there is minimum effect size for clinical significance, then in the absence of knowing the effect size, you might calculate the sample size needed to show that there is a clinically significant difference when the difference is the minimum necessary to meet clinical signficance. If pain is measured on a scale of 1–10, maybe a difference of pain of less than 1.0 points is considered clinically insignficant. In this case, if the true mean difference in pain is 0.5 points, then this might not be considered important. If the study has acceptable power for detecting a 1.0 difference in pain but has low power for detecting a 0.5 point difference in pain, then this might be acceptable.

Power

Other examples where it might be important to pay attention to practical versus statistical significance include

- improvements in educational testing (if SAT score improves by < 10, is this a meaningful improvement?
- patient satisfaction scores (if on a scale of 1–10), how important is a difference of 0.1?

Effect size

Power analysis often involves the term "effect size". Usually, under the null hypothesis, the effect size (sometimes abbreviated ES) is 0 and is the value of the difference between the true value of a parameter and the hypothesized value.

The actual effect size can have a couple of definitions. One is that the actual effect size under the alternative hypothesis is then the difference between the true parameter and the hypothesized value of the parameter. Another possibility is that the ES is the difference divided by the standard deviation of the test statistic. The latter definition was more useful when power had to be looked up in textbooks that published tables of power analyses using standardized effect sizes based on a single standard deviation.
For the two-sample t-test, the effect size is often called Cohen's d:

$$d = \frac{|\overline{\mathbf{x}} - \overline{\mathbf{y}}|}{s}$$

where s is the pooled sample standard deviation. (Remove absolute values for a one-sided test.)

Effect size

Cohen considers $d = 0.2$ to be small, $d = 0.5$ to be medium, and $d = 0.8$ to be large. Note that d is quite similar to the coefficient of variation for a single random variable $E[X]/SD(X)$.

Cohen (Jacob Cohen) wrote a book in 1969 called "Statistical Power Analysis for the Behavioral Sciences" where he defines many of these ideas. It is quite influential still (second edition was 1988). The book printed lots of typed tables of power values to be used as references. I don't think there is a single plot in the book, but there are over 100 pages of tables for a 500-page text. This is also the Cohen of Cohen's kappa.

Effect size

An effect size can depend on the type of problem. Instead of a difference in means, the ES might be a relative risk, an odds ratio, or a difference in proportions.

It is good practice to report effect sizes in addition to p-values. A small p-value does not necessarily mean that there is a large effect size. Instead p-values depend on a combination of the effect size (relative to the null hypothesis), the variability in the sample, and the sample size. Small p-values can be obtained for small effects with large sample sizes, and large p-values can occur when the effect size is clinically meaningful but the sample size and sample variability was too large to conclude "statistical significance". This could mean that the observed effect size would be practically signficant, but might have been observed by chance in the study.

Power: sample size determination

Sample sizes for studies aren't usually completely under the researcher's control, but they are analyzed as though they are fixed parameters rather than random variables. In planning a study, you also have to estimate the number of people that drop out of the study, or the proportion that will enroll after you attempt to recruit.

If you recruit people to be in a study for example using flyers around campus, the hospital, etc., then you might have historical data to predict what a typical sample size would be based on how long and widely you advertise. Study designers can therefore often indirectly control the sample size.

Power: sample size determination

Random sample sizes might be worth considering, however, For the t-test example, you might have better power to reject the null hypothesis if your sample sizes are equal for the two groups than if they are unequal. For example, suppose you are recruiting for testing whether a drug reduces headaches, and you recruit both men and women. Suppose you suspect that the drug is more effective for men than women.

If you recruit people for the study, you might not be in direct control of how many men versus women volunteer to be in the study. Suppose 55 women volunteer to be in the study and 45 men volunteer. You could make the sample sizes equal by randomly dropping data from 10 of the women, but this would be throwing away information. It is better to use information from all 90 study participants, although you might have less power with 45 men versus 55 women than with 50 d for each sex.

Power: sample size determination

On the other hand, if for your study, you are collecting expensive information, such as doing MRIs for each participant, you might decide to accept the first n women volunteers and the first n men volunteers. A power analysis could help you decide whether it was important to have a balanced design or not.

Power: effect of unbalanced designs

How could we simulate the effect of unbalanced versus balanced designs? Assuming we knew that there were a fixed number of participants (say $n = 100$), we could compare the effect of a particular unbalanced design (for example 45 versus 55) versus the balanced design (50 per group). We could also let the number of men versus women in each iteration of a simulation be a binomial random variable, so that the degree of imbalance is random.

Power: determining effect size

In addition to graphing power as a function of sample size, it is common to plot power as a function of the effect size for a fixed sample size. Ultimately, power depends on three variables: α, n, and the effect size such as $\mu_1 - \mu_2$ for the two-sample t-test example. We usually fix two of these variables and plot power as a function of the other variable.

The t-test example is easy to modify to plot power as a function of the effect size for a given sample size (say, $n = 20$).

Power: determining effect size

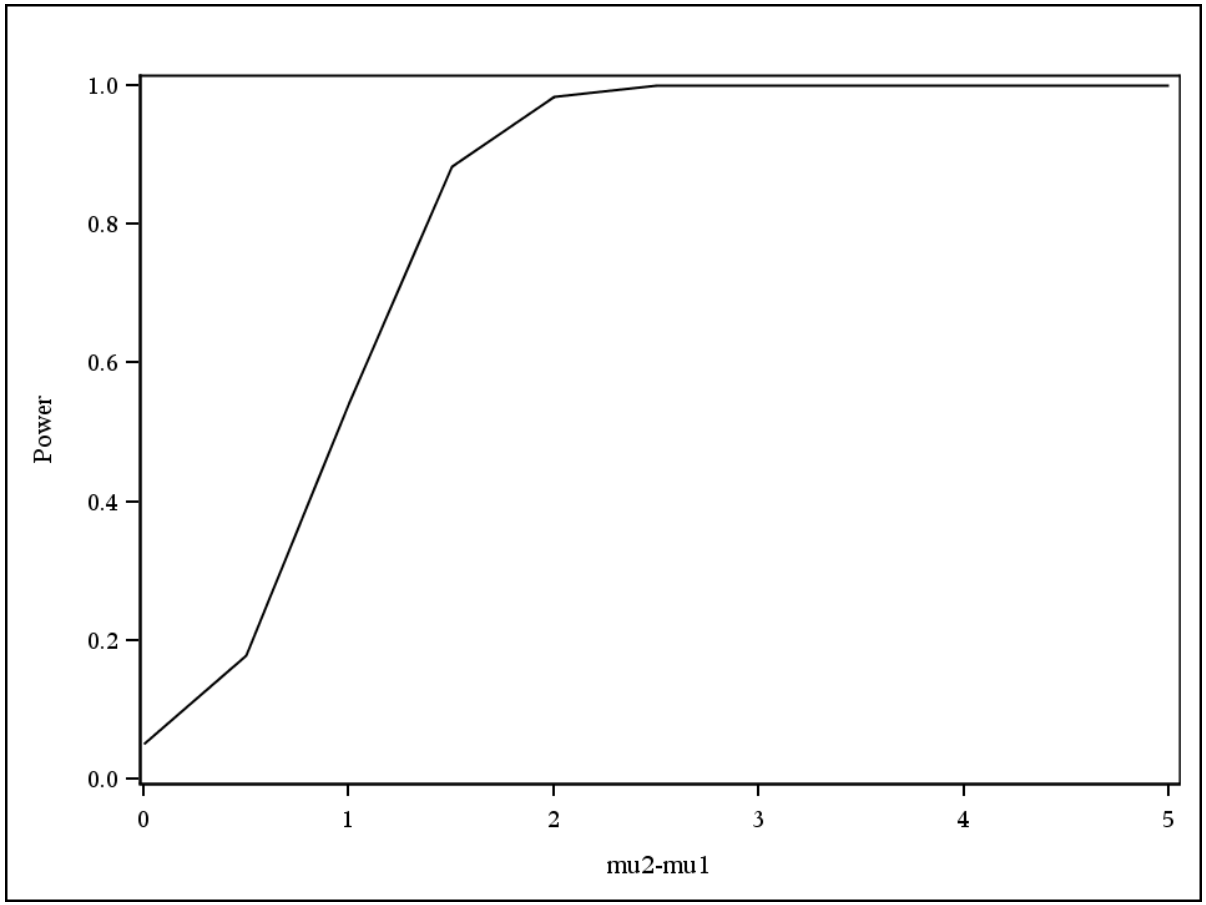
```
%let n=10;
%let iter=1000;

data sim;
  /* generate two exponentials for each
     combination of i and j */
  do effect = 0 to 5 by .5;
    do i=1 to &iter;
      do j=1 to &n;
      x = rannorm(3014*&n + &iter);
      group = "A";
      output;
      x = rannorm(2013*&n + &iter)+effect;
      group = "B";
      output;
      end;
    end;
  end;
  /* i is the iteration */
  keep group x i effect;
run;

ods output TTests=pvalues;
ods select TTests;
proc ttest data=sim;
  by effect i;
  class group;
  var x;
run;
```

Power: determining effect size

```
data reject;
  set pvalues;
  if Probt <= .05 then reject=1;
  else reject=0;
run;

proc means data=reject;
  by effect;
  where variances="Equal";
  var reject;
  output out=power;
run;

title " ";
ods pdf file = "power2.pdf";
proc sgplot data=power;
  where _STAT_ = "MEAN";
  series x=n y=reject;
  yaxis label="Power";
  xaxis label="Sample size";
run;
ods pdf close;
```

Power: determining effect size

1



Power: plotting both sample size and effect size

```
%let n=10;
%let iter=1000;

data sim;
  /* generate two exponentials for each
     combination of i and j */
  do n=10,20,30;
    do effect = 0 to 3 by .5;
      do i=1 to &iter;
        do j=1 to n;
        x = rannorm(3014*n + &iter);
        group = "A";
        output;
        x = rannorm(2013*n + &iter)+effect;
        group = "B";
        output;
        end;
      end;
    end;
  end;
  /* i is the iteration */
  keep group x n effect i;
run;

ods output TTests=pvalues;
ods select TTests;
proc ttest data=sim;
  by n effect i;
  class group;
  var x;
run;
```

Power: plotting both sample size and effect size

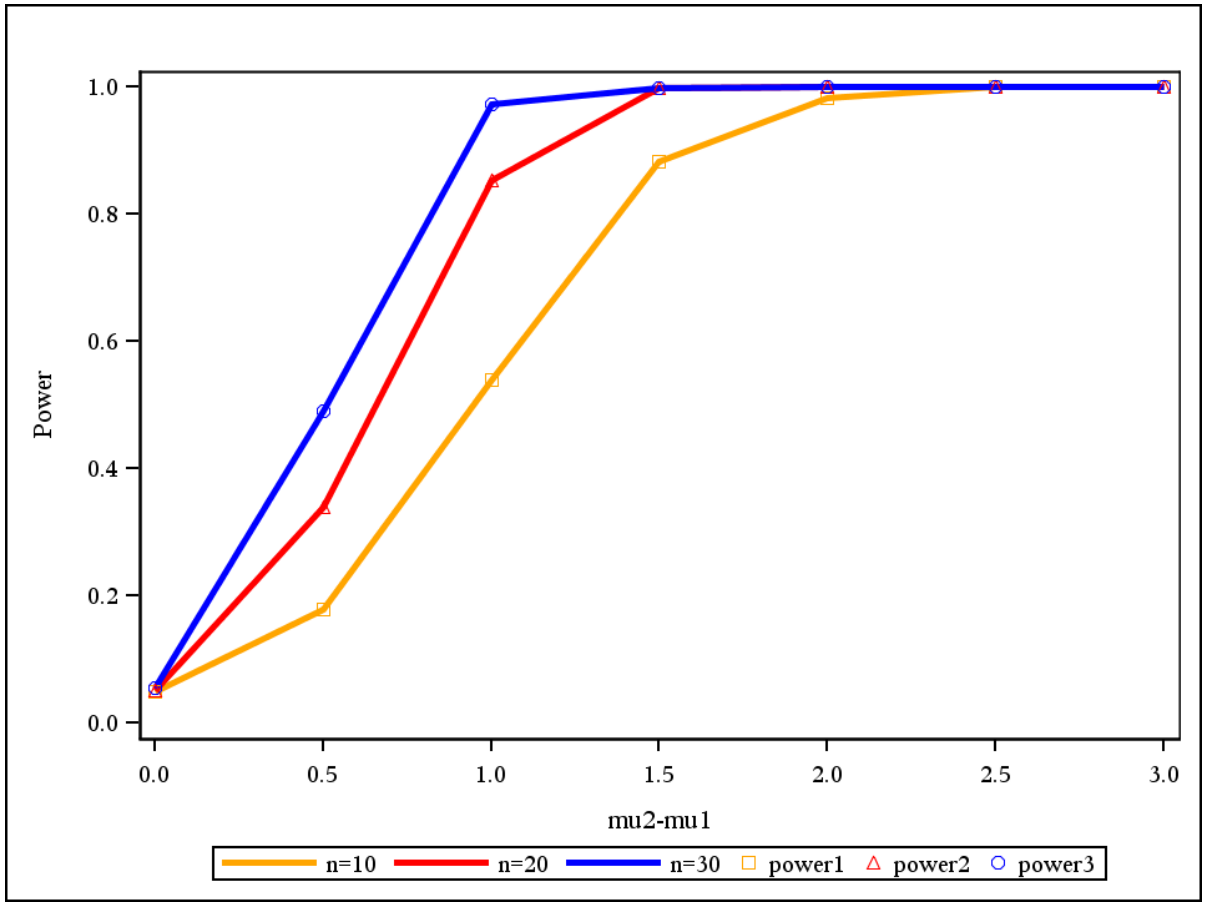
```
data reject;
  set pvalues;
  if Probt <= .05 then reject=1;
  else reject=0;
run;

proc means data=reject noprint;
  by n effect;
  where variances="Equal";
  var reject;
  output out=power;
run;

data power;
  set power;
  if _STAT_ = "MEAN";
  if n=10 then power1 = reject; else power1=.;
  if n=20 then power2 = reject; else power2=.;
  if n=30 then power3 = reject; else power3=.;
run;
```

Power: plotting both sample size and effect size

```
title " ";
ods pdf file = "power3.pdf";
proc sgplot data=power;
  where _STAT_ = "MEAN";
  series x=effect y=power1 / lineattrs=(color=orange thickness=3)
                             legendlabel="n=10";
  series x=effect y=power2 / lineattrs=(color=red thickness=3)
                             legendlabel="n=20";
  series x=effect y=power3 / lineattrs=(color=blue thickness=3)
                             legendlabel="n=30";
  scatter x=effect y=power1 / markerattrs=(color=orange symbol=square size=8)
                              legendlabel="";
  scatter x=effect y=power2 / markerattrs=(color=red symbol=triangle size=8)
                              legendlabel="";
  scatter x=effect y=power3 / markerattrs=(color=blue symbol=circle size=8)
                              legendlabel="";
  yaxis label="Power";
  xaxis label="mu2-mu1";
run;
ods pdf close;
```

Power: determining effect size

1



Power: determining effect size

Note that the data set `sim` that has all of my simulated data has 840,000 observations. SAS is still reasonably fast, and the log file gives information about how long it took.

```
NOTE: SAS Institute Inc., SAS Campus Drive, Cary, NC USA 27513
NOTE: The SAS System used:
      real time               22.28 seconds
      cpu time                9.43 seconds
```

We could make the plots smoother by incrementing the effect size by a smaller value (say .01), although this will generate 50 times as many observations. When simulations get this big, you start having to plan them – how long will they take (instead of 30s, will it take 25min?, 25 days?), how much memory will they use, and so on, even though this is a very simple simulation.

Length of simulations

The log file also breaks down how long each procedure took. Much of the time was actually due to generating the PDF file with ODS. From the log file:

```
NOTE: The data set WORK.SIM has 840000 observations and 5 variables
NOTE: DATA statement used (Total process time):
      real time           0.19 seconds
      cpu time            0.18 seconds
NOTE: The data set WORK.PVALUES has 42000 observations and 9 variabl
NOTE: The PROCEDURE TTEST printed pages 1-21000.
NOTE: PROCEDURE TTEST used (Total process time):
      real time           9.12 seconds
      cpu time            8.97 seconds
...
NOTE: PROCEDURE SGPLOT used (Total process time):
      real time           12.44 seconds
      cpu time            0.19 seconds
```

Length of simulations

When designing simulations, there are usually tradeoffs. For example, suppose I don't want my simulation to take any longer than it already has. If I want smoother curves, I could double the number of effect sizes I used, but then to keep the simulation the length of time, I might have to use fewer iterations (say 500 instead of 1000). This would increase the number of data points at the expense of possibly making the curve more jittery, or even not monotonically increasing. There will usually be a tradeoff between the number of iterations and the number of parameters you can try in your simulation.

Length of simulations for R

If you want to time R doing simulations, the easiest way is to run R in batch mode. In Linux or Mac OS X, you can go to a terminal, and at the shell prompt, type

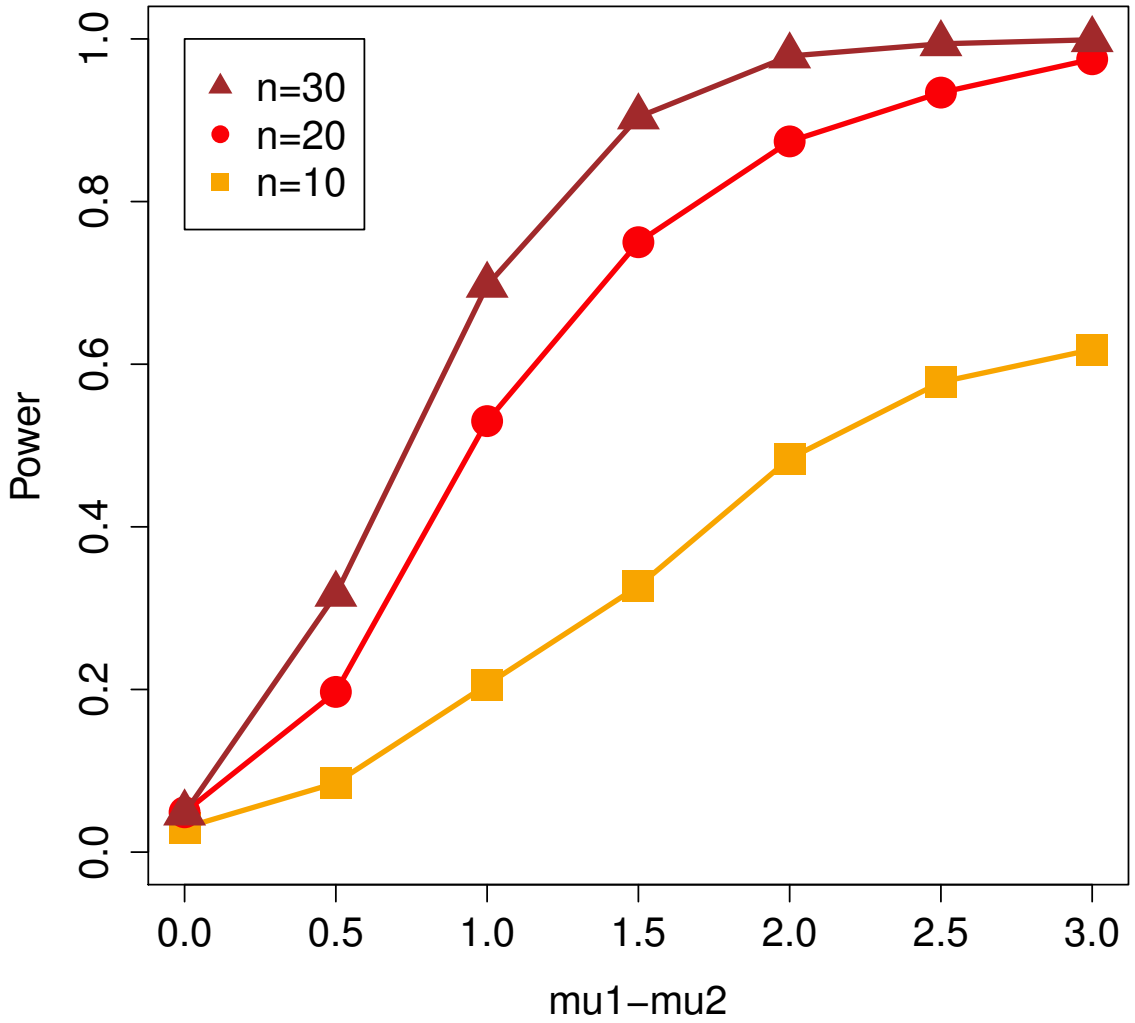
```
time R CMD BATCH myprogram.r
```

and it will give a similar print out of real time versus cpu time for your R run.

Power in R

Here's an example of testing the power of the t-test to distinguish two exponential popoulations with different means. Here the rate in the first exponential sample is 1, and the rate in the second exponential varies from 1 to 3. The sample size varies from 10 to 30.

Power in R



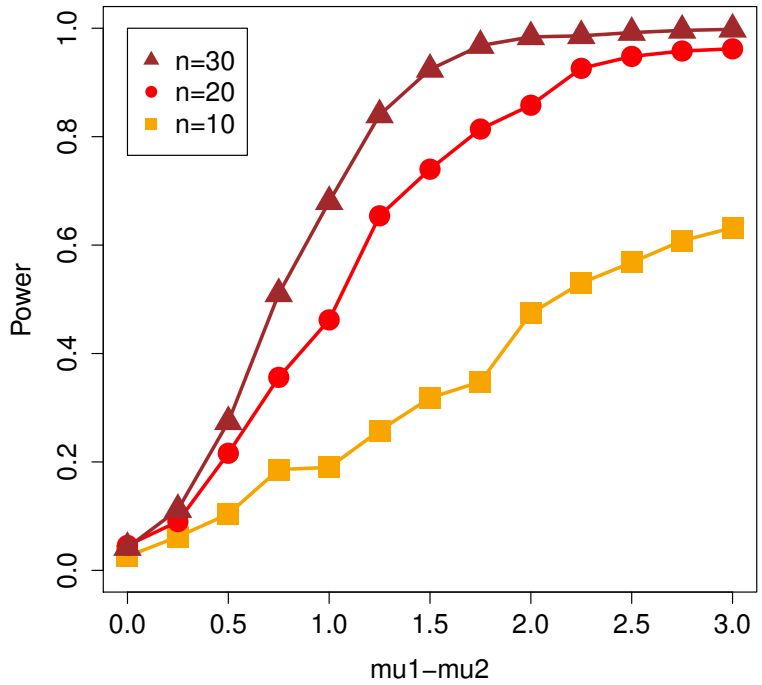
Power: determining effect size

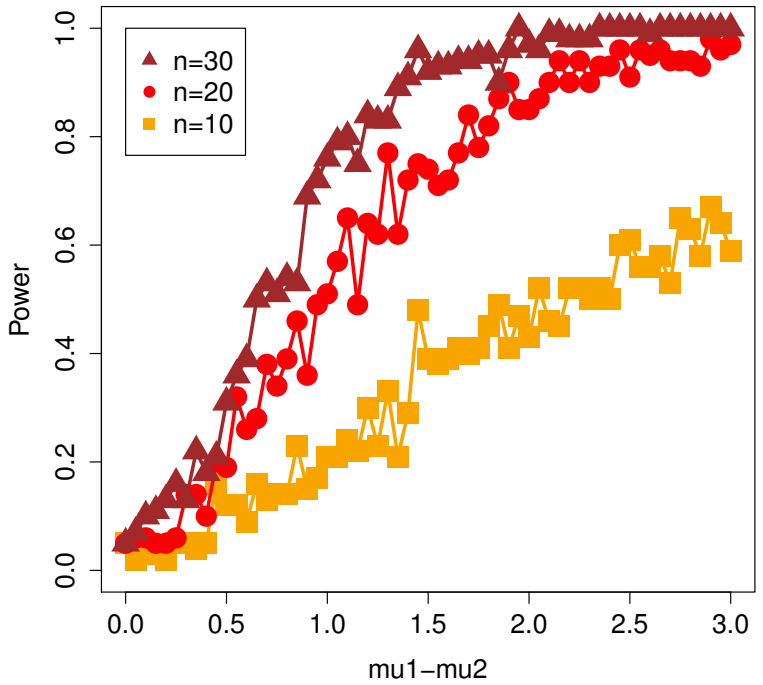
```
I <- 1000
effect <- seq(0,3,0.5)
n <- c(10,20,30)
nrows <- I*length(effect)*length(n)
results <- matrix(data=NA,nrow=nrows,ncol=4)
row <- 0
for(k in 1:length(n)) {
  for(j in 1:length(effect)) {
    for(i in 1:I) {
       row <- row+1
       x <- rexp(n[k],1)
       y <- rexp(n[k],1+effect[j])
       pvalue <- t.test(x,y)$p.value
       results[row,1] <- i
       results[row,2] <- effect[j]
       results[row,3] <- n[k]
       results[row,4] = (pvalue <= 0.05)
    }
  }
}

# extract columns from simulated data
c2 <- results[,2]
c3 <- results[,3]
c4 <- results[,4]

mycol = c("orange","red","brown")
```

Power: determining effect size

```
mycol = c("orange","red","brown")

postscript(file="powerR.eps",height=7,width=7)
plot(c(0,max(effect)),c(0,1),type="n",xlab="mu1-mu2",ylab="Power",cex.lab=1.4,cex.axis=1.4)
for(k in 1:3) {
  #meanp has average power
  meanp <- NULL
  for(j in 1:length(effect)) {
    p1 <- c4[c2==effect[j] & c3==n[k]]
    meanp <- c(meanp,mean(p1))
  }
  points(effect,meanp,type="l",col=mycol[k],lwd=3)
  points(effect,meanp,pch=(14+k),cex=2.5,col=mycol[k])
}
legend(0,1,legend=c("n=30","n=20","n=10"),pch=c(15,16,17),col=mycol,cex=1.4)
dev.off()
```

Power:tradeoff between number of parameters and number of iterations (500 vs 100 iterations)





Using Power to select methods

As mentioned before, power analyses are useful for determining which method is preferable when there are multiple methods available to analyze data.

As an example, to consider the two sample t-test again when we have exponential data. Suppose we wish test $H_0 : \mu = 2$ when $\lambda = 1$, so that the null hypothesis is false. Since the assumptions of the test are false, researchers might prefer using a nonparametric test.

Using Power to select methods

As an alternative, you can use a permutation test or other nonparametric test. Here we might wish to see which method is most powerful. If you can live with the inflated type 1 error for the t-test (or adjust for it by using a smaller α-level, then you might prefer it if is more powerful.

A number of nonparametric procedures are implemented in PROC NPAR1WAY, as well as PROC MULTTEST. In addition, there are macros floating around the web that can do permutation tests without using these procedures.

Using power to select methods

Here we'll try PROC NPAR1WAY and just one nanparametric method, the Wilcoxon rank-sum test (also called the Mann-Whitney test). The idea is to pool all of the data, then rank them. Then calculate the sum of the ranks for group A versus group B. The two sums should be approximately equal, with greater differences in the sums of the ranks being evidence that the mean for one group is larger than the mean for the other group.

Using power to select methods

Note that there are many other methods we could have selected such as a median test or a permutation test. This is just to illustrate, and we are not necessarily finding the most powerful method.

Power: comparing methods

```
data sim;
  /* generate two exponentials for each
     combination of i and j */
  do n = 5,10,15,20,25,30;
    do i=1 to &iter;
      do j=1 to n;
      x = ranexp(3014*n + &iter)*2;
      group = "A";
      output;
      x = ranexp(2013*n + &iter);
      group = "B";
      output;
      end;
    end;
  end;
  /* i is the iteration */
  keep group x n i;
run;

ods output TTests=pvalues;
ods select TTests;
proc ttest data=sim;
  by n i;
  class group;
  var x;
run;
```

Power: comparing methods

```
ods output WilcoxonTest = wilcoxonp;
ods select WilcoxonTest;
proc npar1way data=sim wilcoxon ;
  by n i;
  class group;
  var x;
run;

data reject1;
  set pvalues;
  if Probt <= .05 then reject=1;
  else reject=0;
run;

data reject2;
  set WilcoxonTest;
  where name = PT2_WIL;
  if nvalue1 <= 0.05 then reject=1;
  else reject=0;
  keep n i reject;
run;
```

Power: comparing methods

```
proc means data=reject1 noprint;
  by n;
  where variances="Equal";
  var rejectT;
  output out=power1;
run;

proc means data=reject2 noprint;
  by n;
  var rejectW;
  output out=power2;
run;

data power;
  merge power1 power2;
  by n;
run;

title "Final Results";
proc print data=power;
  where _STAT_ = "MEAN";
run;
```

Power: comparing methods

Obs	n	i	Variable	Name1	Label1	cValue1	nValue1
59977	30	998	x				.
59978	30	998	x		t Approximation		.
59979	30	998	x	PTR_WIL	One-Sided Pr > Z	0.0072	0.007217
59980	30	998	x	PT2_WIL	Two-Sided Pr > |Z|	0.0144	0.014434
59981	30	999	x	_WIL_	Statistic	989.0000	989.000000

Final Results 13073

Obs	n	_TYPE_	_FREQ_	_STAT_	reject T	reject W
4	5	0	1000	MEAN	0.106	0.050
9	10	0	1000	MEAN	0.219	0.157
14	15	0	1000	MEAN	0.396	0.303
19	20	0	1000	MEAN	0.511	0.391
24	25	0	1000	MEAN	0.629	0.515
29	30	0	1000	MEAN	0.723	0.612

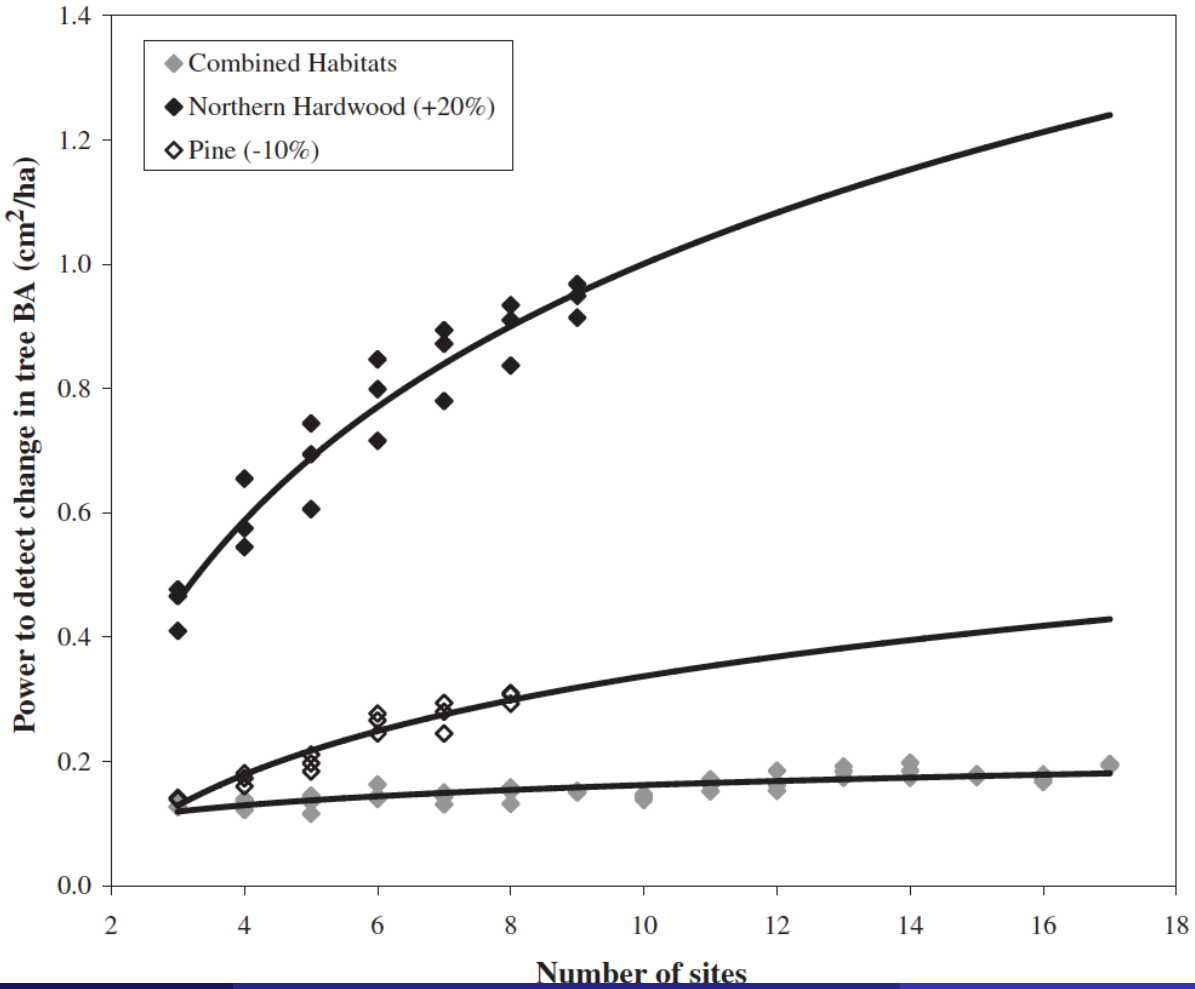
Power: comparing methods

For these parameter values (exponentials with means of 1 and 2), the t-test was more powerful than the Wilcoxon test at all sample sizes. The Wikipedia article on the Mann-Whitney test says: "It [The Wilcoxon or Mann-Whitney test] has greater efficiency than the t-test on non-normal distributions, such as a mixture of normal distributions, and it is nearly as efficient as the t-test on normal distributions."

Given our limited simulation, we have some reason to be a little bit skeptical of this claim. Still, we only tried one combination of parameters. It is possible that for other parameters or other distributions, the t-test is less powerful. Also, the t-test has inflated type 1 error, so the comparison might be a little unfair. We could re-run the experiment using $\alpha = .01$ for the t-test and $\alpha = .05$ for the Wilcoxon to make sure that both had controlled type 1 error rates.

Power: comparing methods

Here's an example from an empirical paper,



Power: comparing methods

Table 2. Power (percentage of P values < 0.1; power values ≥ 80% are in boldface type) of the US Forest Service Forest Inventory and Analysis (FIA) and the Plant Ecology laboratory of the University of Wisconsin–Madison (PEL) sampling methods for detecting changes in vegetation at Apostle Islands National Lakeshore, Wisconsin.

		FIA		PEL			Power (%) to detect 20% change			Min. no. of sites sampled to attain 80% power‡	
Level of analysis	Vegetation parameter	Area sampled per site (m^2)	No. of sites*	Area sampled per site (m^2)	No. of sites*	σ†	FIA	PEL	Additional power of PEL at 20% change	FIA	PEL
Ground layer quadrats	Forb frequency	12.0	19	40.0	20	0.15	45.6	**99.9**	+54.3	ns	7
	Forb richness	12.0	19	40.0	20	0.20	47.9	**94.5**	+46.6	ns	13
	Fern frequency	12.0	16	40.0	17	0.20	10.4	**96.0**	+85.6	ns	11
	Fern ally frequency	12.0	17	40.0	19	0.20	7.4	57.4	+50.0	ns	ns
	Clintonia borealis frequency	12.0	11	40.0	11	0.25	16.9	18.9	+2.00	ns	ns
	Woody understory frequency	12.0	20	40.0	20	0.20	74.3	**99.4**	+25.1	ns	10
	Taxus canadensis frequency	12.0	15	40.0	16	0.35	71.3	**99.8**	+28.5	ns	6
Shrub quadrats – microplots	Shrub richness	160.0	20	53.9	20	0.25	**93.9**	**99.9**	+6.00	14	6
	No. of shrub stems	160.0	20	53.9	20	0.25	**88.8**	**100**	+11.2	16	5
	No. of *Taxus canadensis* stems	53.9	15	160.0	15	0.25	21.8	55.7	+33.9	ns	ns
	No. of *Acer spicatum* stems	53.9	20	160.0	20	0.25	15.3	59.0	+43.7	ns	ns
Sapling quadrats – microplots	Sapling density (no./m^2)	53.9	17	160.0	20	0.3	10.5	54.9	+44.4	ns	ns
	Sapling BA (no./m^2)	53.9	17	160.0	20	0.3	1.50	22.9	+21.4	ns	ns
Tree subplots	Tree density (no./m^2)	673.0	20	~3282	20	0.10	**99.6**	**100**	+0.40	8	4–5
	Tree BA	673.0	20	~3282	20	0.10	**98.5**	**100**	+1.50	10	<5
	Betula papyrifera density (no./m^2)	673.0	12	NA§	—	0.10	33.8	NA	NA	ns	NA

Comparing power among three sampling methods for monitoring forest vegetation

Sarah E. Johnson, E.L. Mudrak, E.A. Beever, S. Sanders, and D.M. Waller

Can. J. For. Res. **38**: 143–156 (2008)

Speed: comparing methods

For large analyses, speed and/or memory might be an issue for choosing between methods and/or algorithms. This paper compared using different methods within SAS based on speed for doing permutation tests.

Table 2. Power (percentage of P values < 0.1; power values ≥ 80% are in boldface type) of the US Forest Service Forest Inventory and Analysis (FIA) and the Plant Ecology laboratory of the University of Wisconsin–Madison (PEL) sampling methods for detecting changes in vegetation at Apostle Islands National Lakeshore, Wisconsin.

		FIA		PEL			Power (%) to detect 20% change			Min. no. of sites sampled to attain 80% power‡	
Level of analysis	Vegetation parameter	Area sampled per site (m^2)	No. of sites*	Area sampled per site (m^2)	No. of sites*	σ†	FIA	PEL	Additional power of PEL at 20% change	FIA	PEL
Ground layer quadrats	Forb frequency	12.0	19	40.0	20	0.15	45.6	**99.9**	+54.3	ns	7
	Forb richness	12.0	19	40.0	20	0.20	47.9	**94.5**	+46.6	ns	13
	Fern frequency	12.0	16	40.0	17	0.20	10.4	**96.0**	+85.6	ns	11
	Fern ally frequency	12.0	17	40.0	19	0.20	7.4	57.4	+50.0	ns	ns
	Clintonia borealis frequency	12.0	11	40.0	11	0.25	16.9	18.9	+2.00	ns	ns
	Woody understory frequency	12.0	20	40.0	20	0.20	74.3	**99.4**	+25.1	ns	10
	Taxus canadensis frequency	12.0	15	40.0	16	0.35	71.3	**99.8**	+28.5	ns	6
Shrub quadrats – microplots	Shrub richness	160.0	20	53.9	20	0.25	**93.9**	**99.9**	+6.00	14	6
	No. of shrub stems	160.0	20	53.9	20	0.25	**88.8**	**100**	+11.2	16	5
	No. of *Taxus canadensis* stems	53.9	15	160.0	15	0.25	21.8	55.7	+33.9	ns	ns
	No. of *Acer spicatum* stems	53.9	20	160.0	20	0.25	15.3	59.0	+43.7	ns	ns
Sapling quadrats – microplots	Sapling density (no./m^2)	53.9	17	160.0	20	0.3	10.5	54.9	+44.4	ns	ns
	Sapling BA (no./m^2)	53.9	17	160.0	20	0.3	1.50	22.9	+21.4	ns	ns
Tree subplots	Tree density (no./m^2)	673.0	20	~3282	20	0.10	**99.6**	**100**	+0.40	8	4–5
	Tree BA	673.0	20	~3282	20	0.10	**98.5**	**100**	+1.50	10	<5
	Betula papyrifera density (no./m^2)	673.0	12	NA§	—	0.10	33.8	NA	NA	ns	NA

Comparing power among three sampling methods for monitoring forest vegetation

Sarah E. Johnson, E.L. Mudrak, E.A. Beever, S. Sanders, and D.M. Waller

Use of macros for simulations

The author of the previous paper provides an appendix with lengthy macros to use as more efficient substitutes to use as replacements for SAS procedures such as PROC NPAR1WAY and PROC MULTTEST, which from his data could crash or not terminate in a reasonable time.

In addition to developing your own macros, a common use of macros is to use macros written by someone else that have not been incorporated into the SAS language. You might just copy and paste the macro into your code, possibly with some modification, and you can use the macro even if you cannot understand it. Popular macros might eventually get replaced by new PROCs or new functionality within SAS. This is sort of the SAS alternative to user-defined packages in R.

From Macro to PROC

An example of an evolution from macros to PROCS is for bootstrapping. For several years, to perform bootstrapping, SAS users relied on macros often written by others to do the bootstrapping. In bootstrapping, you sample you data (or the rows of your data set) with replacement and get a new dataset with the same sample size but some of the values repeated and others omitted. For example if your data is

`-3 -2 0 1 2 5 6 9` bootstrap replicated datas set might be

```
-2 -2 1 5 6 9 9 9
-3 0 1 1 2 5 5 6
```

etc.

From Macro to Proc

Basically to generate the bootstrap data set, you generate random n random numbers from 1 to n, with replacement, and extract those values from your data. This was done using macros, but now can be done with PROC SURVEYSELECT. If you search on the web for bootstrapping, you still might run into one of those old macros.

Newer methods might still be implemented using macros. A webpage from 2012 has a macro for Bootstrap bagging, a method of averaging results from multiple classification algorithms.

`http://statcompute.wordpress.com/2012/07/14/a-sas-macro-for-bootstrap-aggregating-bagging/`

There are also macros for searching the web to download movie reviews or extract data from social media. Try searching on "SAS macro 2013" for interesting examples.