Sample Size Calculations Using SAS, R, and nQuery Software

Toronto Health User Group Meeting
24Oct2019

Jenna Cody
Overview

+ Type I and Type II Error
+ Statistical Power
+ Information needed to calculate sample size
+ Computing sample size calculations by hand: means and proportions
+ Computing sample size calculations in SAS
+ Computing sample size calculations in R
+ Computing sample size calculations in nQuery
+ Comparisons between the software
| Decision based on sample | \( \text{Groups are not different} \)  
\((H_0 \text{ true})\) | \( \text{Groups are different} \) \((H_1 \text{ true})\) |
|-------------------------|---------------------------------|---------------------------------|
| \( \text{Groups are not different} \) \((\text{Accept } H_0)\) | Correct decision \((1-\alpha)\) | Type II error \((\beta)\)  
\(\text{False negative}\) |
| \( \text{Groups are different} \) \((\text{Reject } H_0)\) | Type I error \((\alpha)\)  
\(\text{False positive}\) | Correct decision \((1-\beta)\) |
Graphical depiction of statistical error and power

Examples show normal (left) and chi square (right) distributions

Null hypothesis

Alternative hypothesis

Null hypothesis

Type II Error

Type I Error

Type II Error

Type I Error

Statistical power: Why is it important?

• Can be used to calculate minimum sample size needed to detect a specified effect size
• Similarly, can be used to calculate a minimum effect size likely to be detected given a specified sample size
• Power is used to make comparisons between statistical tests
• Used when designing studies to ensure sample size is large enough to detect a meaningful effect yet small enough that unnecessary resources are not wasted
• Plays a role in determining whether studies are stopped early
• Power analysis improves the chances of conclusive results
Information needed to calculate sample size

Factors that always need to be specified

• Power (1-\(\beta\)): \(\Pr(\text{reject } H_0 \mid H_1 \text{ true})\); correct rejection
• Significance criterion (\(\alpha\)): \(\Pr(\text{reject } H_0 \mid H_0 \text{ true})\); false positive
• Effect size: magnitude of the effect of interest in the population

Other factors that can influence power

• Experimental design: many components of the design can influence power
  - Balanced vs. unbalanced number of observations in each sample group
  - Parametric vs. non-parametric test
  - Crossover vs. parallel group vs. factorial design
• Precision: reduction of measurement error improves statistical power, thus requiring a smaller sample size
• Expected rates of non-completion. In clinical trials, this refers to treatment withdrawals and protocol violations.
Additional background information for computing sample size

• Conventional values: use with discretion—conventions differ based on study design and field of study
  - Statistical power: $1 - \beta = 0.8$ to 0.9 minimum
  - Significance criterion: $\alpha = 0.05$ or less, especially in cases where multiplicity adjustments are required

• Typically calculate based on primary hypothesis of interest
  - Because of this, secondary and exploratory analyses may be underpowered and should not be used to make claims but can influence design of future studies

• If pre-specified, sample size re-estimation can be performed while experiment is ongoing if event rates are lower than anticipated or variability is larger than expected$^{1}$
Computing sample size by hand

Example: 2 sample t-test assuming equal variances. Can approximate with standard normal distribution with large sample sizes (>100)

\[ n = \frac{2\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2} \]

Where:

- \( n \) is the sample size required for each group
- \( z_\gamma \) is the critical value at the point on the standard normal distribution corresponding with the quantile in subscript
- \( \sigma \) is the standard deviation of the population
- \( \Delta \) is the standardized difference between the 2 groups

To find quantile, look up in z table or use functions in SAS or R.

```r
data test;
  q1=quantile("Normal",0.975);
  q2=quantile("Normal",0.8);
run;
```

```r
> qnorm(0.975)
[1] 1.959964
> qnorm(0.8)
[1] 0.8416212
```
Computing sample size by hand

Example: 2 sample test of proportions

\[ n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2[p_1(1 - p_1) + p_2(1 - p_2)]}{(p_1 - p_2)^2} \]

Where:

- \( n \) is the sample size required for each group
- \( z_x \) is the critical value at the point on the standard normal distribution corresponding with the quantile in subscript
- \( p_1 \) is the proportion of events expected to occur in group 1
- \( p_2 \) is the proportion of events expected to occur in group 2
- \((p_1 - p_2)^2\) is the minimum meaningful difference or effect size

To find quantile, look up in z table or use functions in SAS or R.

```r
data test;
q1=quantile("Normal",0.975);
q2=quantile("Normal",0.8);
run;
```

```
q1 q2
1.959964 0.84162
> qnorm(0.975)
[1] 1.959964
> qnorm(0.8)
[1] 0.8416212
```
Computing sample size in SAS

• 2 procedures: PROC POWER and PROC GLMPOWER in the SAS/STAT package
  - Both procedures perform prospective power and sample size analyses

• PROC POWER: used for sample size calculations for tests such as:
  - t tests, equivalence tests, and confidence intervals for means
  - tests, equivalence tests, and confidence intervals for binomial proportions
  - multiple regression
  - tests of correlation and partial correlation
  - one-way analysis of variance
  - rank tests for comparing two survival curves
  - logistic regression with binary response
  - Wilcoxon-Mann-Whitney (rank-sum) test

• PROC GLMPOWER: used for sample size calculations for more complex linear models, and cover Type III tests and contrasts of fixed effects in univariate linear models with or without covariates.

## Inputs for SAS Sample Size Procedures

<table>
<thead>
<tr>
<th>PROC POWER</th>
<th>PROC GLMPOWER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design</td>
<td>Design (including subject profiles and their allocation weights)</td>
</tr>
<tr>
<td>Statistical model and test</td>
<td>Statistical model and contrasts of class effects</td>
</tr>
<tr>
<td>Significance level (alpha)</td>
<td>Significance level (alpha)</td>
</tr>
<tr>
<td>Surmised effects and variability</td>
<td>Surmised response means for subject profiles (i.e. “cell means”) and variability</td>
</tr>
<tr>
<td>Power</td>
<td>Power</td>
</tr>
<tr>
<td>Sample size</td>
<td>Sample size</td>
</tr>
</tbody>
</table>

Not all inputs need to be filled out. Leave result parameter (in this case, sample size) missing by designating it with a missing value in input.

Computing sample size in SAS using the POWER procedure

**Syntax**

```sas
PROC POWER <options> ;
LOGISTIC <options> ;
MULTREG <options> ;
ONECORR <options> ;
ONESAMPLEFREQ <options> ;
ONESAMPLEMEANS <options> ;
ONEWAYANOVA <options> ;
PAIREDFREQ <options> ;
PAIREDMEANS <options> ;
PLOT <plot-options> </ graph-options> ;
TWOSAMPLEFREQ <options> ;
TWOSAMPLEMEANS <options> ;
TWOSAMPLESURVIVAL <options> ;
TWOSAMPLEWILCOXON <options> ;
RUN;
```

- For example, a two-sample t test assuming equal variances can use the following syntax:

  ```sas
  PROC POWER;
  TWOSAMPLEMEANS TEST=DIFF
  GROUPMEANS = mean1 | .
  STDDEV = .
  NTOTAL = .
  POWER = .
  ;
  RUN;
  ```

- Can solve for any of the factors indicated as missing with a “.” but we need to fill in the remaining factors. To calculate sample size, leave NTOTAL as missing.
Computing sample size in SAS using the POWER procedure

Examples: 2 sample t-test for mean difference & Chi-square test for proportion difference

```
proc power;
  twosamplemeans test=diff
groupmeans= 120 | 108
stddev=30
ntotal= .
power=0.8;

  twosamplefreq test=pchi
groupproportions=0.8 | 0.5
power=0.8
ntotal= .;
run;
```
Identify necessary sample size to achieve range of power

Example: 2-sample t test in SAS using PROC POWER

```
plot x=power min=0.8 max=0.95;
```
Computing sample size in SAS using the GLMPOWER procedure

Example: 2-way ANOVA

For example, a 2-way ANOVA can use the following syntax:

```sas
proc glmpower data= dataset;
  class expvar1 expvar2;
  model responsevar = expvar1 | expvar2;
  power
    stddev = .
    ntotal = .
    power = .;
run;
```

- Can solve for any of the factors indicated as missing with a “.” but we need to fill in the remaining factors. To calculate sample size, leave NTOTAL as missing.

PROC GLMPOWER <options> ;
  BY variables ;
  CLASS variables ;
  CONTRAST 'label' effect values <...effect values> </ options> ;
  MODEL dependents = independents ;
  PLOT <plot-options> </ graph-options> ;
  POWER <options> ;
  WEIGHT variable ;
RUN;
Computing sample size in SAS using the GLMPOWER procedure

**Example: 2-way ANOVA**

First, create exemplary data set with expected population means. In this example, these are lab values at each level of treatment and dose.

```sas
data Exemplary;
  do trt = 1 to 2;
    do dose = 1 to 3;
      input lab @@;
      output;
    end;
  end;
data Exemplary;
  do trt = 1 to 2;
    do dose = 1 to 3;
      input lab @@;
      output;
    end;
data Exemplary;
  do trt = 1 to 2;
    do dose = 1 to 3;
      input lab @@;
      output;
    end;
end;
run;
```

```sas
proc glmpower data=Exemplary;
  class trt dose;
  model lab = trt | dose;
  power
    stddev = 5
    ntotal = .
    power = .8;
run;
```

**Fixed Scenario Elements**
- Dependent Variable: lab
- Error Standard Deviation: 5
- Nominal Power: 0.8
- Alpha: 0.05

**Computed N Total**

<table>
<thead>
<tr>
<th>Index</th>
<th>Source</th>
<th>Test DF</th>
<th>Error DF</th>
<th>Actual Power</th>
<th>N Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>trt</td>
<td>1</td>
<td>72</td>
<td>0.828</td>
<td>78</td>
</tr>
<tr>
<td>2</td>
<td>dose</td>
<td>2</td>
<td>36</td>
<td>0.849</td>
<td>42</td>
</tr>
<tr>
<td>3</td>
<td>trt*dose</td>
<td>2</td>
<td>336</td>
<td>0.807</td>
<td>342</td>
</tr>
</tbody>
</table>
Identify necessary sample size to achieve range of power

Example: 2-way ANOVA in SAS using PROC GLMPOWER

```
plot x=power min=.1 max=.9;
```
Computing sample size in R

**Example: 2 sample t-test**

**Syntax**

```r
def = \frac{|\mu_1 - \mu_2|}{\sigma}
\text{where } \mu_1 = \text{mean of group 1}
\mu_2 = \text{mean of group 2}
\sigma = \text{common error variance}
```

Example syntax & output values from our previous example. Similarly to SAS, we can leave the field we want to calculate as blank.

```r
> pwr.t.test(n = , d = .4, sig.level = 0.05, power = .8, type = "two.sample")
```

Two-sample t test power calculation

- n = 99.08032
- d = 0.4
- sig.level = 0.05
- power = 0.8
- alternative = two.sided

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

First, download “pwr” package
Identify necessary sample size to achieve range of power

Example: 2-sample t test in R using plot function

```r
tails = two.sided
effect size d = 0.4
alpha = 0.05

> x <- pwr.t.test(n = , d = .4, sig.level = 0.05, power = .8, type = "two.sample")
> plot(x)
```

Assign power output to an object in R and plot the object.
Computing sample size in R

Different functions needed for each type of test

Syntax for other designs

- **t test with unequal sample sizes**: `pwr.t2n.test(n1 = , n2= , d = , sig.level =, power = )`
- **One-way ANOVA**: `pwr.anova.test(k = , n = , f = , sig.level = , power = )`

\[
f = \sqrt{\frac{\sum_{i=1}^{k} p_i \cdot (\mu_i - \mu)^2}{\sigma^2}}
\]

where 
- \(p_i = n_i / N\),
- \(n_i = \text{number of observations in group } i\),
- \(N = \text{total number of observations}\),
- \(\mu_i = \text{mean of group } i\),
- \(\mu = \text{grand mean}\),
- \(\sigma^2 = \text{error variance within groups}\)

- **Chi-square test**: `pwr.chisq.test(w = , N = , df = , sig.level = , power = )`

\[
w = \sqrt{\frac{\sum_{i=1}^{m} (p_{0i} - p_{1i})^2}{p_{0i}}} 
\]

where 
- \(p_{0i} = \text{cell probability in } i\text{th cell under } H_0\),
- \(p_{1i} = \text{cell probability in } i\text{th cell under } H_1\)

- **Other designs include linear models** (`pwr.f2.test`), **correlations** (`pwr.r.test`), **test of proportions** (`pwr.2p.test`/ `pwr.2p2n.test`/ `pwr.p.test`)
Computing sample size in nQuery

Wizard interface
Computing sample size in nQuery

*Fill in known information. Defines and suggests values.*

<table>
<thead>
<tr>
<th>MTT0-1 / Two Group t-test of Equal Means</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Significance Level, $\alpha$</td>
<td>1</td>
</tr>
<tr>
<td>1 or 2 Sided</td>
<td>2</td>
</tr>
<tr>
<td>Group 1 Mean, $\mu_1$</td>
<td></td>
</tr>
<tr>
<td>Group 2 Mean, $\mu_2$</td>
<td></td>
</tr>
<tr>
<td>Difference in Means, $\mu_1 - \mu_2$</td>
<td></td>
</tr>
<tr>
<td>Common Standard Deviation, $\sigma$</td>
<td></td>
</tr>
<tr>
<td>Effect Size, $\delta =</td>
<td>\mu_1 - \mu_2</td>
</tr>
<tr>
<td>Power (%)</td>
<td></td>
</tr>
<tr>
<td>n per Group</td>
<td></td>
</tr>
</tbody>
</table>

**Test Significance Level, $\alpha$**

Alpha is the probability of rejecting the null hypothesis that the mean equals the specified value when it is true (the probability of a Type I error).

**Suggestion:**

Enter 0.05, a frequent standard.

**Acceptable Entries:**

0.001 to 0.20
Computing sample size in nQuery

Automatically fills in fields once enough information is entered, e.g. Difference in means after Group 1 and Group 2 mean are filled out, Effect size after Difference in means and \( \sigma \) are filled out.

<table>
<thead>
<tr>
<th>Test Significance Level, ( \alpha )</th>
<th>0.050</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 or 2 Sided</td>
<td>2</td>
</tr>
<tr>
<td>Group 1 Mean, ( \mu_1 )</td>
<td>120.000</td>
</tr>
<tr>
<td>Group 2 Mean, ( \mu_2 )</td>
<td>108.000</td>
</tr>
<tr>
<td>Difference in Means, ( \mu_1 - \mu_2 )</td>
<td>12.000</td>
</tr>
<tr>
<td>Common Standard Deviation, ( \sigma )</td>
<td>30.000</td>
</tr>
<tr>
<td>Effect Size, ( \delta =</td>
<td>\mu_1 - \mu_2</td>
</tr>
<tr>
<td>Power (%)</td>
<td>80</td>
</tr>
</tbody>
</table>

Output

A sample size of 100 in each group will have 80% power to detect a difference in means of 12 (the difference between a Group 1 mean, \( \mu_1 \), of 120 and a Group 2 mean, \( \mu_2 \), of 108) assuming that the common standard deviation is 30 using a two group t-test with a 5% two-sided significance level.

Leave the field of interest blank. Once enough information is filled out in the other fields, the result for the blank field will be shown in the section below.
Identify necessary sample size to achieve range of power

Example: 2 sample t-test in nQuery using graph option

Power vs Sample Size

Click here for graph output
Software Comparison

- SAS
  - Ability to calculate sample size for complex linear models and contrasts
  - Blend of user friendly features and advanced options
  - Requires SAS/STAT package
  - Can quickly and easily test a range of values
  - Plots are easily customizable
  - Sample size can be computed in a program so it is easily replicable and “macrotized”
- R
  - Limited in their ability to compute sample size for very complicated models
  - Choose test first and then enter inputs, rather than customizing inputs to influence test
  - Requires more extensive computations by the user for input parameters
  - Plots are most informative
  - Requires pwr package
  - Free and open source
- nQuery
  - Wizard → no programming required
  - Explanations of each input parameter and plain text description of output
  - Great for non-programmers
  - No extensive computations required by the user
  - User-friendly
  - Capabilities for many tests
  - Not free, but documentation is comprehensive
Questions?
Appendix
## PROC POWER Summary of Analyses, Part 1

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Statement</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression: likelihood ratio chi-square test</td>
<td>LOGISTIC</td>
<td></td>
</tr>
<tr>
<td>Multiple linear regression: Type III test</td>
<td>MULTREG</td>
<td></td>
</tr>
<tr>
<td>Correlation: Fisher’s test</td>
<td>ONECORR</td>
<td>DIST=FISHERZ</td>
</tr>
<tr>
<td>Correlation: test</td>
<td>ONECORR</td>
<td>DIST=T</td>
</tr>
<tr>
<td>Binomial proportion: exact test</td>
<td>ONESAMPLEFREQ</td>
<td>TEST=EXACT</td>
</tr>
<tr>
<td>Binomial proportion: test</td>
<td>ONESAMPLEFREQ</td>
<td>TEST=Z</td>
</tr>
<tr>
<td>Binomial proportion: test with continuity adjustment</td>
<td>ONESAMPLEFREQ</td>
<td>TEST=ADJZ</td>
</tr>
<tr>
<td>Binomial proportion: exact equivalence test</td>
<td>ONESAMPLEFREQ</td>
<td>TEST=EQUIV_EXACT</td>
</tr>
<tr>
<td>Binomial proportion: equivalence test</td>
<td>ONESAMPLEFREQ</td>
<td>TEST=EQUIV_Z</td>
</tr>
<tr>
<td>Binomial proportion: test with continuity adjustment</td>
<td>ONESAMPLEFREQ</td>
<td>CI=AGRESTICOULL</td>
</tr>
<tr>
<td>Binomial proportion: confidence interval</td>
<td>ONESAMPLEFREQ</td>
<td>CI=JEFFREYS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CI=EXACT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CI=WALD</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CI=WALD_CORRECT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CI=WILSON</td>
</tr>
<tr>
<td>One-sample test</td>
<td>ONESAMPLEMEANS</td>
<td>TEST=T</td>
</tr>
<tr>
<td>One-sample test with lognormal data</td>
<td>ONESAMPLEMEANS</td>
<td>TEST=T DIST=LOGNORMAL</td>
</tr>
<tr>
<td>One-sample equivalence test for mean of normal data</td>
<td>ONESAMPLEMEANS</td>
<td>TEST=EQUIV</td>
</tr>
<tr>
<td>One-sample equivalence test for mean of lognormal data</td>
<td>ONESAMPLEMEANS</td>
<td>TEST=EQUIV DIST=LOGNORMAL</td>
</tr>
<tr>
<td>Confidence interval for a mean</td>
<td>ONESAMPLEMEANS</td>
<td>CI=T</td>
</tr>
</tbody>
</table>
### PROC POWER Summary of Analyses, Part 2

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Statement</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-way ANOVA: one-degree-of-freedom contrast</td>
<td>ONEWAYANOVA</td>
<td>TEST=CONTRAST</td>
</tr>
<tr>
<td>One-way ANOVA: overall test</td>
<td>ONEWAYANOVA</td>
<td>TEST=OVERALL</td>
</tr>
<tr>
<td>McNemar exact conditional test</td>
<td>PAIREDFREQ</td>
<td></td>
</tr>
<tr>
<td>McNemar normal approximation test</td>
<td>PAIREDFREQ</td>
<td>DIST=NORMAL</td>
</tr>
<tr>
<td>Paired test</td>
<td>PAIREDMEANS</td>
<td>TEST=DIFF</td>
</tr>
<tr>
<td>Paired test of mean ratio with lognormal data</td>
<td>PAIREDMEANS</td>
<td>TEST=RATIO</td>
</tr>
<tr>
<td>Paired additive equivalence of mean difference with normal data</td>
<td>PAIREDMEANS</td>
<td>TEST=EQUIV_DIFF</td>
</tr>
<tr>
<td>Paired multiplicative equivalence of mean ratio with lognormal data</td>
<td>PAIREDMEANS</td>
<td>TEST=EQUIV_RATIO</td>
</tr>
<tr>
<td>Confidence interval for mean of paired differences</td>
<td>PAIREDMEANS</td>
<td>Cl=DIFF</td>
</tr>
<tr>
<td>Pearson chi-square test for two independent proportions</td>
<td>TWOSAMPLEFREQ</td>
<td>TEST=PCHI</td>
</tr>
<tr>
<td>Fisher’s exact test for two independent proportions</td>
<td>TWOSAMPLEFREQ</td>
<td>TEST=FISHER</td>
</tr>
<tr>
<td>Likelihood ratio chi-square test for two independent proportions</td>
<td>TWOSAMPLEFREQ</td>
<td>TEST=LRCHI</td>
</tr>
<tr>
<td>Two-sample test assuming equal variances</td>
<td>TWOSAMPLEMEANS</td>
<td>TEST=DIFF</td>
</tr>
<tr>
<td>Two-sample Satterthwaite test assuming unequal variances</td>
<td>TWOSAMPLEMEANS</td>
<td>TEST=DIFF_SATT</td>
</tr>
<tr>
<td>Two-sample pooled test of mean ratio with lognormal data</td>
<td>TWOSAMPLEMEANS</td>
<td>TEST=RATIO</td>
</tr>
<tr>
<td>Two-sample additive equivalence of mean difference with normal data</td>
<td>TWOSAMPLEMEANS</td>
<td>TEST=EQUIV_DIFF</td>
</tr>
<tr>
<td>Two-sample multiplicative equivalence of mean ratio with lognormal data</td>
<td>TWOSAMPLEMEANS</td>
<td>TEST=EQUIV_RATIO</td>
</tr>
<tr>
<td>Two-sample confidence interval for mean difference</td>
<td>TWOSAMPLEMEANS</td>
<td>Cl=DIFF</td>
</tr>
<tr>
<td>Log-rank test for comparing two survival curves</td>
<td>TWOSAMPLESURVIVAL</td>
<td>TEST=LOGRANK</td>
</tr>
<tr>
<td>Gehan rank test for comparing two survival curves</td>
<td>TWOSAMPLESURVIVAL</td>
<td>TEST=GEHAN</td>
</tr>
<tr>
<td>Tarone-Ware rank test for comparing two survival curves</td>
<td>TWOSAMPLESURVIVAL</td>
<td>TEST=TARONEWARE</td>
</tr>
</tbody>
</table>