Power and Sample Size

Overview

- Introduce Concept of Power via Correlation Coefficient ($\rho$) Example
- Discuss Factors Contributing to Power
- Practical:
  - Simulating data as a means of computing power
  - Using Mx for Power Calculations
Simple example

Investigate the linear relationship between two random variables X and Y: \( \rho = 0 \) vs. \( \rho \neq 0 \) using the Pearson correlation coefficient.

Sample subjects at random from population
Measure X and Y
Calculate the measure of association \( \rho \)
Test whether \( \rho \neq 0 \).

How to Test \( \rho \neq 0 \)

Assume data are normally distributed
Define a null–hypothesis (\( \rho = 0 \))
Choose an \( \alpha \) level (usually .05)
Use the (null) distribution of the test statistic associated with \( \rho = 0 \)
\[
t = \rho \sqrt{\frac{(N-2)}{(1-\rho^2)}}
\]
How to Test $\rho \neq 0$

Sample N=40
$r=.303$, $t=1.867$, $df=38$, $p=.06 \alpha = .05$

Because observed $p > \alpha$, we fail to reject $\rho = 0$

Have we drawn the correct conclusion that $p$ is genuinely zero?

---

DOGMA

$\alpha =$ type I error rate
probability of deciding $\rho \neq 0$
(while in truth $\rho=0$)

$\alpha$ is often chosen to equal .05...why?
N=40, r=0, nrep=1000, central
t(38),
$\alpha=0.05$ (critical value 2.04)

Observed non-null
distribution ($\rho=.2$) and
null distribution
In 23% of tests that $\rho=0$, $|t|>2.024$ ($\alpha=0.05$), and thus correctly conclude that $\rho = 0$.

The probability of correctly rejecting the null-hypothesis ($\rho=0$) is $1-\beta$, known as the power.

Hypothesis Testing

- Correlation Coefficient hypotheses:
  - $h_0$ (null hypothesis) is $\rho=0$
  - $h_a$ (alternative hypothesis) is $\rho \neq 0$
    - Two-sided test, where $\rho > 0$ or $\rho < 0$ are one-sided
- Null hypothesis usually assumes no effect
- Alternative hypothesis is the idea being tested
Summary of Possible Results

<table>
<thead>
<tr>
<th>H-0 true</th>
<th>H-0 false</th>
</tr>
</thead>
<tbody>
<tr>
<td>accept H-0</td>
<td>reject H-0</td>
</tr>
<tr>
<td>1-(\alpha)</td>
<td>(\alpha)</td>
</tr>
<tr>
<td>(\beta)</td>
<td>1-(\beta)</td>
</tr>
</tbody>
</table>

\(\alpha\)=type 1 error rate
\(\beta\)=type 2 error rate
1-\(\beta\)=statistical power

STATISTICS

<table>
<thead>
<tr>
<th>Reality</th>
<th>Rejection of H_0</th>
<th>Non-rejection of H_0</th>
</tr>
</thead>
<tbody>
<tr>
<td>H_0 true</td>
<td>Type I error at rate (\alpha)</td>
<td>Nonsignificant result (1-(\alpha))</td>
</tr>
<tr>
<td>H_A true</td>
<td>Significant result (1-(\beta))</td>
<td>Type II error at rate (\beta)</td>
</tr>
</tbody>
</table>
Power

- The probability of rejecting the null-hypothesis depends on:
  - the significance criterion ($\alpha$)
  - the sample size ($N$)
  - the effect size (NCP)

“The probability of detecting a given effect size in a population from a sample of size $N$, using significance criterion $\alpha$”

Standard Case

$P(T)$

Sampling distribution if $H_0$ were true

$\text{Sampling distribution if } H_A \text{ were true}$

$\alpha 0.05$

$POWER = 1 - \beta$

Effect Size (NCP)
Impact of less conservative $\alpha$

Sampling distribution if $H_0$ were true

$P(T)$

$\alpha = 0.1$

Sampling distribution if $H_A$ were true

POWER = $1 - \beta$

Impact of more conservative $\alpha$

Sampling distribution if $H_0$ were true

$P(T)$

$\alpha = 0.01$

Sampling distribution if $H_A$ were true

POWER = $1 - \beta$

$\beta \uparrow$ $\beta \downarrow$ $t_{\alpha}$
Impact of increased sample size

- Reduced variance of sampling distribution if $H_A$ is true.
- Sampling distribution if $H_0$ is true.

$$P(T)$$

$$\alpha = 0.05$$

Impact of increase in Effect Size

- Reduced variance of sampling distribution if $H_A$ were true.
- Sampling distribution if $H_0$ were true.

$$P(T)$$

$$\alpha = 0.05$$

$$\text{Effect Size (NCP)} \uparrow$$
Summary: Factors affecting power

- Effect Size
- Sample Size
- Alpha Level
- <Beware the False Positive!!!>
- Type of Data:
  - Binary, Ordinal, Continuous
- Research Design

Uses of power calculations

- Planning a study
- Possibly to reflect on ns trend result
- No need if significance is achieved
- To determine chances of study success
Power Calculations via Simulation
- Simulate Data under theorized model
- Calculate Statistics and Perform Test
- Given $\alpha$, how many tests $p < \alpha$
- Power = (#hits)/(#tests)

Practical: Empirical Power 1
- Simulate Data under a model online
- Fit an ACE model, and test for C
- Collate fit statistics on board
Practical: Empirical Power 2

- First get
  http://www.vipbg.vcu.edu/neale/gen619/power/power-raw.mx and put it into your directory
- Second, open this script in Mx, and note both places where we must paste in the data
- Third, simulate data (see next slide)
- Fourth, fit the ACE model and then fit the AE submodel

Practical: Empirical Power 3

- Simulation Conditions
  - 30% A²  20% C²  50% E²
  - Input:
    - A 0.5477 C of 0.4472 E of 0.7071
    - 350 MZ 350 DZ
    - Simulate and use “Space Delimited” option at
      - http://statgen.iop.kcl.ac.uk/workshop/unisim.html or click here in slide show mode
    - Click submit after filling in the fields and you will get a page of data
Practical: Empirical Power 4

- With the data page, use ctrl-a to select the data, control-c to copy, switch to Mx (e.g. with alt-tab) and in Mx control-v to paste in both the MZ and DZ groups.
- Run the ace.mx script with the data pasted in and modify it to run the AE model.
- Report the -2log-likelihoods on the whiteboard
- Optionally, keep a record of A, C, and E estimates of the first model, and the A and E estimates of the second model

Simulation of other types of data

- Use SAS/R/Matlab/Mathematica
- Any decent random number generator will do
- See http://www.vipbg.vcu.edu/~neale/gem619/power/sim1.sas
R

- R is in your future
- Can do it manually with rnorm
- Easier to use mvrnorm

```
library (MASS)
mvrnorm(n=100,c(1,1),matrix(c(1,.5,.5 ,1),2,2),empirical=FALSE)
```

- runmx at Matt Keller’s site:
  - [http://www.matthewckeller.com/html/m](http://www.matthewckeller.com/html/m)

Mathematica Example

```
In[32]:= (mu={1,2,3,4};
sigma={{1,1/2,1/3,1/4},{1/2,1/3,1/4,1/5},{1/3,1/4,1/5,1/6},{1/4,1/5,1/6,
1/7}};
Timing[Table[Random[MultinormalDistribution[mu,sigma]],{1000}]][[1]])

Out[32]= 1.1 Second

In[33]:= Timing[RandomArray[MultinormalDistribution[mu,sigma],1000]][[1]]

Out[33]= 0.04 Second

In[37]:= TableForm[RandomArray[MultinormalDistribution[mu,sigma],10]]

Obtain mathematica from VCU
[http://www.ts.vcu.edu/faq/stats/mathematica.html](http://www.ts.vcu.edu/faq/stats/mathematica.html)
Theoretical Power Calculations

- Based on Stats, rather than Simulations
- Can be calculated by hand sometimes, but Mx does it for us
- Note that sample size and alpha-level are the only things we can change, but can assume different effect sizes
- Mx gives us the relative power levels at the alpha specified for different sample sizes

We will use the power.mx script to look at the sample size necessary for different power levels

In Mx, power calculations can be computed in 2 ways:
  - Using Covariance Matrices (We Do This One)
  - Requiring an initial dataset to generate a likelihood so that we can use a chi-square test
Power.mx 1

! Simulate the data
! 30% additive genetic
! 20% common environment
! 50% nonshared environment

#NGroups 3

G1: model parameters
Calculation
Begin Matrices;
  X lower 1 1 fixed
  Y lower 1 1 fixed
  Z lower 1 1 fixed
End Matrices;

Matrix X 0.5477
Matrix Y 0.4472
Matrix Z 0.7071

Begin Algebra;
  A = X*X ;
  C = Y*Y ;
  E = Z*Z ;
End Algebra;

End

Power.mx 2

G2: MZ twin pairs
Calculation
Matrices = Group 1
Covariances A+C+E | A+C _
  A+C | A+C+E /
Options MX%E=mzsim.cov
End

G3: DZ twin pairs
Calculation
Matrices = Group 1
H Full 1 1
Covariances A+C+E | H@A+C _
  H@A+C | A+C+E /
Matrix H 0.5
Options MX%E=dzsim.cov
End
! Second part of script
! Fit the wrong model to the simulated data
! to calculate power

#NGroups 3
G1 : model parameters
Calculation
Begin Matrices:
   X lower 1 1 free
   Y lower 1 1 fixed
   Z lower 1 1 free
End Matrices;

Begin Algebra:
   A = X’X ;
   C = Y’Y ;
   E = Z’Z ;
End Algebra;
End

G2 : MZ twins
Data NInput_vars=2 NObservations=350
CMatrix Full File=mzsim.cov
Matrices= Group 1
Covariances A+C+E  |   A+C _        
                   |   A+C    |   A+C+E /
Option Risduals
End

G3 : DZ twins
Data NInput_vars=2 NObservations=350
CMatrix Full File=dzsim.cov
Matrices= Group 1
H Full 1 1
Covariances A+C+E  |  H@A+C _  
                   |  H@A+C    |  A+C+E /
Matrix H 0.5
Option Risduals

! Power for alpha = 0.05 and 1 df
Option Power= 0.05,1
Model Identification

- Necessary Conditions
- Sufficient Conditions
- Algebraic Tests
- Empirical Tests

Necessary Conditions

- Number of Parameters < or = Number of Statistics
- Structural Equation Model usually count variances & covariances to identify variance components
- What is the number of statistics/parameters in a univariate ACE model? Bivariate?
Sufficient Conditions

- No general sufficient conditions for SEM

- Special case: ACE model
  - Distinct Statistics (i.e. have different predicted values)
    - $VP = a^2 + c^2 + e^2$
    - $CMZ = a^2 + c^2$
    - $CDZ = .5 a^2 + c^2$

Sufficient Conditions 2

- Arrange in matrix form

- \[
\begin{array}{ccc}
1 & 1 & 1 \\
1 & 1 & 0 \\
.5 & 1 & 0 \\
\end{array}
\begin{array}{c}
a^2 \\
c^2 \\
e^2 \\
\end{array}
\begin{array}{c}
VP \\
CMZ \\
CDZ \\
\end{array}
\]

- $A \times x = b$

- If $A$ can be inverted then can find $A^{-1}b$
Sufficient Conditions 3

Solve ACE model

\[
\text{Begin Matrices;}
A \text{ full } 3 \times 3
b \text{ full } 3 \times 1
\text{End Matrices;}
\]

Matrix A

\[
\begin{pmatrix}
1 & 1 & 1 \\
1 & 1 & 0.5 \\
1 & 0 & 0
\end{pmatrix}
\]

Labels

Col A  A  C  E
Row A  VP  CMZ  CDZ

\[
\text{Matrix } b
\]

\[
\begin{pmatrix}
! \text{ Data, essentially}
\end{pmatrix}
\]

Labels

Col B  Statistic
Row B  VP  CMZ  CDZ

\[
\text{Begin Algebra;}
C = A^{-1};
x = A^{-1} \times b;
\text{End Algebra;}
\]

Labels

Row x  A  C  E

Sufficient Conditions 4

- What if not soluble by inversion?
- Empirical:
  - 1 Pick set of parameter values \( T_1 \)
  - 2 Simulate data
  - 3 Fit model to data starting at \( T_2 \) (not \( T_1 \))
  - 4 Repeat and look for solutions to step 3 that are perfect but have estimates not equal to \( T_1 \)
- If equally good solution but different values, reject identified model hypothesis
Conclusion

- Power calculations relatively simple to do
- Curse of dimensionality
- Different for raw vs summary statistics
- Simulation can be done many ways
- No substitute for research design