Advanced Power: After Power
What Then?

By
Dan Ayers
Senior Associate in Biostatistics
What do you actually get?

- One measure for this is the probability that a random subject selected from a treated group will have a larger (smaller) value than a random subject from the untreated (or control) population.

- Simple case
  - Control Tumor Volume (mm³) ~ N(1200, 587)
  - Treatment 1 Tumor Volume (mm³) ~ N(826, 587)
  - Treatment 2 Tumor Volume (mm³) ~ N(533, 587)
What do you actually get?

Control group TV ~ N(1200, 300)
No Effect Treatment TV ~ N(1200, 300)

Pr(Y<X) = 0.5
Pr(X<Y) = 0.5
Odds (Y<X) = 1

Tumor Volume (mm³)
What do you actually get?

Control group TV $\sim$ N(1200, 300)
Treatment 1 TV (E=1/2 S.D) $\sim$ N(1050, 300)

$\Pr(Y<X) = 0.64$

$\Pr(X<Y) = 0.36$

Odds $(Y<X) = 1.78$

Sample size $= 63$ per group
What do you actually get?

Control group TV $\sim N(1200, 300)$

Treatment 1 TV (E=1 S.D) $\sim N(900, 300)$

$\Pr(Y < X) = 0.76$

$\Pr(X < Y) = 0.24$

Odds $(Y < X) = 3.3$

Sample size = 17 per group
### Decision Error in Experiments

#### Truth

<table>
<thead>
<tr>
<th>Study Says</th>
<th>No Effect</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Effect</td>
<td>correct</td>
<td>False negative</td>
</tr>
<tr>
<td>Effect</td>
<td>False positive</td>
<td>correct</td>
</tr>
</tbody>
</table>
Of course, we don’t know “the truth”, that’s why we are doing the experiment.

Can still control the 2 types of error using statistical methods (i.e. power calculations)
- False positive = Type I error ($\alpha$)
- False negative = Type II error ($\beta$)

Overall, we won’t be wrong very often. Right?
What do you actually get?

- Is the type I and type II error rate approach a good decision approach theoretically?

Simple case

- 5% and 30% of 1000 investigators make true hypotheses and each test at type I and type II errors of 5% and 20%
- How many correct conclusions (true positives and true negatives) will we make?
- How many true differences are detected?
Thus out of 1000 hypotheses, 88 will have statistically significant evidence of a true effect in their favor. Of these, 40 (45%) are in fact true.
Thus out of 1000 hypotheses, 275 will have statistically significant evidence of a true effect in their favor. Of these, 240 (87%) are in fact true.
Type I error = 5%, Power=80%

Percent of Hypothesized Differences That Are True

Percent of 1000 Tests

True Differences Discovered
Correct Conclusions Made
Problem

- Investigators consistently face the requirement to replicate research, including studies that were fully powered to have a high probability of detecting meaningful effects
  - Basic science – Publication requirements of reviewers
  - Clinical science – FDA IND approval
Why Repeat Testing

- Basic science experiments highly sensitive to known and unknown sources of variability
  - Most scientists have experienced (perhaps many times) the situation where results have been completely reversed upon repeated testing
  - Relatively low chance a result will be confirmed by others
  - Repeatability illustrates PI expertise and robustness of experimental results.

- Referee’s (journals) require replication of even fully powered studies
  - Replicate once or twice to get confirmation of the statistical outcome or
  - Replicate only statistically significant results.
To Repeat or Not to Repeat, That is the Question!

- Should the IACUC allow investigators to repeat fully powered experiments 2 or 3 times?

- Is it appropriate for journals to require replicated experiments?
"We do not have any written policy on the question raised. The requirement is fundamentally that all experiments are hopefully reproducible, and have been replicated sufficiently to be statistically significant and to justify presentation of the data and conclusions. The specific requirements will vary with each manuscript and with the opinion of the reviewers in any specific case."
Decision Trees for Testing Behavior

- Repeat as planned regardless of outcome
  - Single repeat
  - 2 repeats

- Repeat only if first results statistically significant
  - Single repeat
  - 2 repeats
3 Experiments
No Treatment Effect
(only error is a false positive)

Probability of 2 Correct Conclusions in 2 Experiments = 0.90
Probability of 2 Correct Conclusions in 3 Experiments = 0.95
Probability of 3 Correct Conclusions in 3 Experiments = 0.86
Three Experiments
True Treatment Effect
(only error is a false negative)

Probability of 2 Correct Conclusions in 2 Experiments = 0.64
Probability of 2 Correct Conclusions in 3 Experiments = 0.77
Probability of 3 Correct Conclusions in 3 Experiments = 0.51
3 Experiments if First is Statistically Significant
No Treatment Effect
(only error is a false positive)

Probability of 2 Correct Conclusions in 2 Experiments = 0
Probability of 2 Correct Conclusions in 3 Experiments = 0.045
Probability of 3 Correct Conclusions in 3 Experiments = 0
3 Experiments if First is Statistically Significant
True Treatment Effect
(only error is a false negative)

Probability of 2 Correct Conclusions in 2 Experiments = 0.64
Probability of 2 Correct Conclusions in 3 Experiments = 0.77
Probability of 3 Correct Conclusions in 3 Experiments = 0.51
Thus out of 1000 hypotheses, 940 will have statistically significant evidence of a true effect in their favor. Of these, 32 (3.4%) are in fact true.
Solution: Better Study Design

- Blind replication of studies is an ill conceived ad hoc practice that significantly hinders scientific discovery.

- Hypothesis testing and reproducibility are goals that can easily be met by experimental design.

- In this environment, the two-sample t-test should be retired.
Reproducibility By Design

- Focus is not to repeat entire experiments, but to account for and estimate nuisance sources of variability by design

- Simple Example:
  - Combining data from 3 sources (e.g., cell lines revived at 3 time periods over a year)
# Normalized qRTPCR Data

<table>
<thead>
<tr>
<th></th>
<th>Gene 1</th>
<th></th>
<th>Gene 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WT</td>
<td>Mutant</td>
<td>WT</td>
</tr>
<tr>
<td></td>
<td>0.0001069</td>
<td>0.00003296</td>
<td>0.0002474</td>
</tr>
<tr>
<td></td>
<td>0.0007165</td>
<td>0.00002966</td>
<td>0.0002090</td>
</tr>
<tr>
<td></td>
<td>0.0007848</td>
<td>0.00003769</td>
<td>0.0003767</td>
</tr>
<tr>
<td></td>
<td>0.0001624</td>
<td>0.0003840</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0001342</td>
<td></td>
<td>0.0003971</td>
</tr>
<tr>
<td></td>
<td>0.00013842</td>
<td>0.00003794</td>
<td>0.0009806</td>
</tr>
<tr>
<td></td>
<td>0.00027723</td>
<td>0.00003863</td>
<td>0.0009798</td>
</tr>
<tr>
<td></td>
<td>0.00025064</td>
<td>0.00003857</td>
<td>0.0008997</td>
</tr>
<tr>
<td></td>
<td>0.00013400</td>
<td></td>
<td>0.0007913</td>
</tr>
<tr>
<td></td>
<td>0.00018913</td>
<td></td>
<td>0.0007527</td>
</tr>
<tr>
<td></td>
<td>0.00017758</td>
<td>0.00004626</td>
<td>0.0016572</td>
</tr>
<tr>
<td></td>
<td>0.00012128</td>
<td>0.00002702</td>
<td>0.0014788</td>
</tr>
<tr>
<td></td>
<td>0.00012925</td>
<td></td>
<td>0.0011352</td>
</tr>
<tr>
<td></td>
<td>0.00012800</td>
<td></td>
<td>0.0012355</td>
</tr>
</tbody>
</table>
## Analysis

<table>
<thead>
<tr>
<th>Gene</th>
<th>Percent of Random Error Explained by Time</th>
<th>Ignoring Time</th>
<th>Accounting for Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31%</td>
<td>p&lt;0.0001</td>
<td>p=0.04</td>
</tr>
<tr>
<td>2</td>
<td>63%</td>
<td>p=0.09</td>
<td>p&lt;0.0001</td>
</tr>
</tbody>
</table>
Summary

- Statisticians and cancer biologists should intimately understand what a power calculation is providing.
- Overall hypothesis testing framework seems to work well.
- Neither type I or type II error have the impact that high rates of correct hypotheses have on having a high success yield from the hypothesis testing framework.
- The practice of confirming only a statistically significant initial result is disastrous.
- Statisticians and biologists should realize the simple two-sample t-test represents a design that is probably inappropriate. Most experiments are more complex and require a statistical design and analysis that suits the complexity.