Coping with Your Biostatistician

Jay Weedon, Ph.D.
School of Public Health

Statistical consulting

- We are geographically situated in the library
- We provide to faculty, students, residents, fellows:
  - Methodological support for study design
  - Statistical data analysis & write-up
  - Please call me at X7475 for an appointment

How *not* to cope with a statistician

- As they used to say on Mystery Science Theater 3000:

It's Movie sign! 

Movie talking points

- Timing of visits to statistician
  - both pre- & post-data collection
  - but not at the last minute, it's not a small detail!
- Minimize technical detail in your discussion
- How big should the study be?
  - How are sample sizes calculated?
  - The issue of *variability within* study arms
  - Establishing the principle outcome of interest
- Funding your data analysis

When to see the statistician?

- At the time that you're **designing** the study but have not yet started collecting data
  - DON'T WAIT until after data collection to do this!
- At points where data collection problems are apparent or anticipated
- After data collection, so that proper analysis and write-up can be done
  - Data analysis is not always quick and easy, and is never an automated procedure

Why *before* data collection?

- Things that you may not be aware need help:
  - Putting together a study that's sufficiently well designed that it can likely demonstrate something useful *and with minimal ambiguity*
  - Estimating *how much* data needs to be collected
  - Developing a thorough protocol for data collection
  - Constructing a workable **data analysis plan**
  - Designing **surveys** is much harder than you think

Study Design I

- This isn't a research methods presentation but:
  - Observational vs. experimental study?
  - Prospective vs. retrospective study?
  - Case-control vs. cohort study?
  - How many treatment arms to run?
    - Please, not one! Do you need to study dose-response?
  - Equal numbers of subjects in each arm?
    - Or is it better to enroll more subjects in treatments I'm most interested in or where enrollment is easier?
    - Should I enroll more than one control for each case?

Study Design II

- How will subjects be allocated to treatment arms?
  - Simple randomization
  - Block randomization
  - Matching
  - Based on convenience only?
- Should each subject be exposed to just one treatment level?
  - Or would it be more efficient to implement some kind of **crossover** or **repeated measures** design?
- At what optimal **time points** should outcomes be measured?

How many subjects to enroll?

- Statisticians call this question **power analysis**
- Why is it important to know this *in advance*?
- Small studies are problematic in two ways:
  - It's difficult to get statistically significant results from a small study, and without significance, you can't demonstrate that observed effects are "real"
  - If you don't achieve significance, you also can't necessarily say that there *are* no real effects, since a bigger study might have shown the opposite

How does this affect your research?

- If the study is undersized, the IRB may worry that you're imposing on your subjects without any serious chance of *learning* anything useful
- A study too small to show anything definite will likely be deemed by the **funding agency** to be a waste of its money
- Journal reviewers & editors baulk at papers with ambiguous or unwarranted conclusions

So how is sample size calculated?

- Generally, the aim is to achieve high **power**, *i.e.,* probability of seeing statistical significance
- What factors affect reaching this goal?
  - How high you want this probability of success to be
  - What **kind** of test you will use
  - The test's **significance level** (usually set at 0.05)
  - How different the study arms really are
  - How many subjects are in each study arm
  - How much **variability** the **outcome** measure has

Power analysis reversed

- So the point of power analysis is to ask, “Given what I know about all these factors, how powerful can I say my statistical test will be?”
- In practice we usually reverse the question to ask, “If I *set* the test’s power at some particular desired level, then how many subjects would that correspond to?”

How high should the power level be?

- The higher you require it to be, the larger the study needed to guarantee it
  - If you want 100% probability of significance you'll need an infinite-sized study!
- Smaller power levels require smaller samples
  - Customarily acceptable is a test with a power level of 80% or 90%; *i.e.*, that is 80-90% likely to produce significant statistical test results (assuming that there is a real effect to be found!)

What kind of test is used?

- Sometimes given a specific study design we have a choice of statistical test
- One test may be better equipped than another to achieve statistical significance
- But in most cases the study design doesn’t give us too many options
- The statistician is usually the one best able to decide which test is best

The test's significance level

- The basic idea is that you can announce that a test produces a "significant" result only if it yields a "*p*-value" below some threshold value
- Usually this value is preset at 0.05, so unless you can make a strong argument for a different value, you're stuck with that
- The *smaller* this preset criterion, the *larger* the study will need to be

How different the study arms are

- By “arms” I mean “groups to be compared ”
- If the differences among these groups in terms of the outcome of interest are *on average* very large, statistically significant *p*-values are highly likely to occur
- If the average differences are small (*e.g.,* due to a weak treatment or risk factor), you may be very unlikely to get small *p*-values

The number of subjects in each arm

- Almost every researcher understands the inverse relationship between sample size and the *p*-values produced by tests
  - With all other things equal, the bigger the study, the smaller the *p*-values, and therefore the greater likelihood of statistical significance

How much outcome variability?

- What's much less well understood is the role of *variability* in statistical testing
  - If your experiment always leads to exactly the same result, you don't need statistics!
  - Even in the best-controlled lab there will be some variability in outcome scores
  - If subjects are biological organisms (e.g., humans, rats), natural inter-subject variability can be a huge barrier to seeing cause-effect relationships
  - As variability increases, test power decreases

Effect size

- It's customary to combine the measure of average difference among study arms with the measure of variability
- Such a combined measure is called **effect size**
- Common examples
  - [Mean(Treated grp) minus Mean(Control grp)]/SD,
    - measures how many SD units apart the means are
  - Risk difference, risk ratio, odds ratio, hazard ratio
  - Pearson correlation coefficient

Subject attrition

- Many studies require either repeated measurement of subjects, or follow-up to see who reaches some relevant endpoint
- Some subjects will withdraw or be withdrawn or become “lost to follow-up”
- Other subjects will not adhere to study protocols
- It’s almost inevitable that you will lose sample size for these or other reasons

So what to take to the statistician?

- Detailed info on how the study is designed
  - *e.g.*, number of arms, principal outcome measure
- An estimate (maybe coarse!) of effect size, *e.g.,*
  - Projected mean & SD of principal outcome for each study arm
  - Projected prevalence of outcome for each level of risk factor (and also prevalence of risk factor)
  - Projected predictor-outcome correlation
- Estimates of rates of refusals and loss-to-follow-up (where relevant)

Where do you *get* these projections?

- From prior studies
  - Collect as much relevant literature as possible and take it along when you see the statistician
- From some existing archival database
- From pilot data (*i.e.,* a small preliminary study)
- From a guess of what *minimum* effect size other professionals would find interesting
- From rule-of-thumb definitions of “large & small” effect sizes

How do statisticians use all this?

- The various parameters you feed him/her are put together into sample size calculations
- Formulae for these calculations come from many statistical papers published over decades
- Some of them are compiled into book form
  - *e.g.*, Jacob Cohen's (1988) *Statistical Power Analysis for the Behavioral Sciences*
- There exists specialized software, *e.g.,* **nQuery**
- Computer simulation studies are often used

Example 1

- I have 2 dose groups (low, high) plus a control
- I project that mean outcome of groups low & high are 0.4 & 0.7 SD units higher than control
- I want to compare both doses against control
- Based on study design I will use Dunnett tests
- I will use 0.05 as a criterion of significance
- I desire 80% power for 2-tailed tests

  => I'll need 40 observations in each study arm

Example 2

- Does prevalence of obstructive sleep apnea (OSA) differ between white & black children
- I project my cohort to be 80% white, 20% black
- Projected OSA prevalence: White 5%, Black 15%
- Based on study design I'll use a Fisher exact test
- Significance level to be set at 0.05 (2-tailed test)
- I desire 80% power for the test of difference

  => I need to recruit 440 children in total

Proving “no difference”

- Most statistical testing is intended to prove that study arms *differ from one another*
- Occasionally we want to show *no difference*
  - Are outcomes of those receiving a generic drug worse than those receiving a brand-name drug?
- This is called an **equivalence** study or **non-inferiority** study
- Requires a quite different set of calculations
- You must inform the statistician of your intent!

Funding for statistical support

- Some research needs minimal advice on design and half an hour of analysis time
- Some research needs extensive analyst time
- Some grants require explicit statements of statistical support
- Some research needs a methodologist on a **Data & Safety Monitoring Board** (DSMB)
- You should consider including support in your grant proposal for statistical expertise

Biostatistician Do's & Don'ts

- Do invite advice on optimal study design
- Don't wait until the day before your grant proposal/IRB submission is due to talk to us
- Do bring in relevant prior papers
- Don't swamp the meeting with irrelevant detail
- Do ask or suggest how data should be analyzed
- Don't wait until you've collected data then expect the analyst to guess how to treat them