Linking Aims, Hypotheses and Analysis: Why Coherence in Grant Proposals is Crucial

Presented as a CCTS Professional Development Workshop

Richard T. Campbell, PhD
Professor Emeritus of Biostatistics and Sociology
Member, IHRP Methodology Research Core
IHRP Fellow

Overview

- In the past 25 years there has been a continuing revolution in the application of statistical methods for health research.
- The revolution stems, largely, from developments in computing.
- As a result, NIH grant reviewers, some of whom are statisticians, require increasingly sophisticated analysis plans.
- A key issue is that the analysis plan links tightly to aims and hypotheses

A personal point of view

- My own training (1968-73) was heavily statistical but with a strong substantive focus.
- It was assumed that investigators did all of their own statistical analysis.
- That's rarely possible these days, but you owe it to your self to work hard to understand the statistics.
- In particular, learn how to write general linearized model equations; they are often the core on an analysis section.

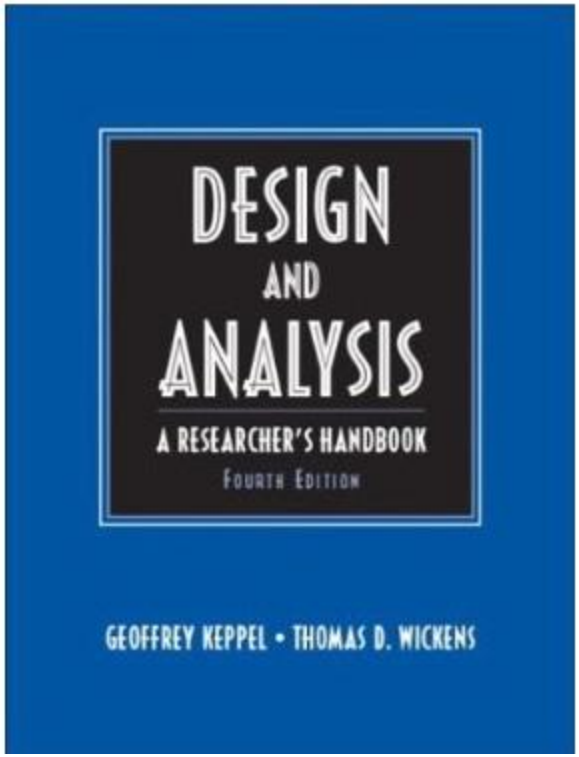
Statistical education is behind the curve

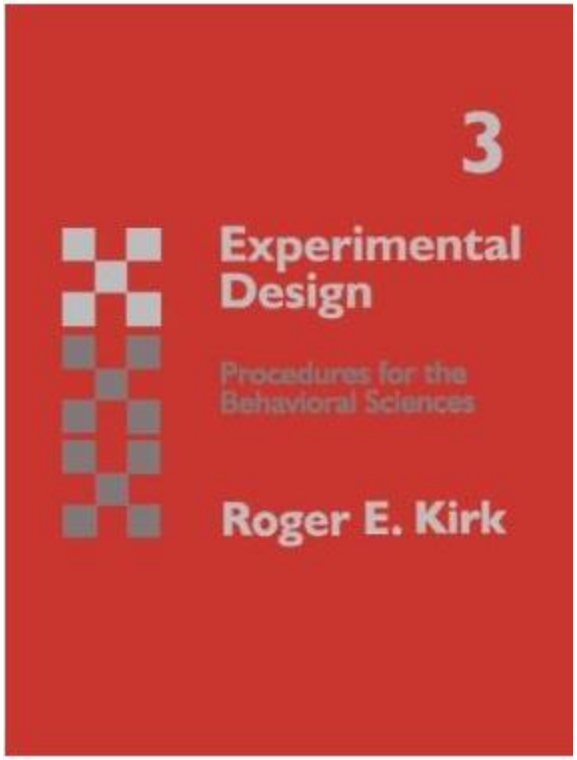
- Newer methods are just now finding their way into text books.
- Many instructors are behind the curve themselves.
- Courses are often generic, not recognizing issues appropriate for different substantive fields, e.g. measurement.

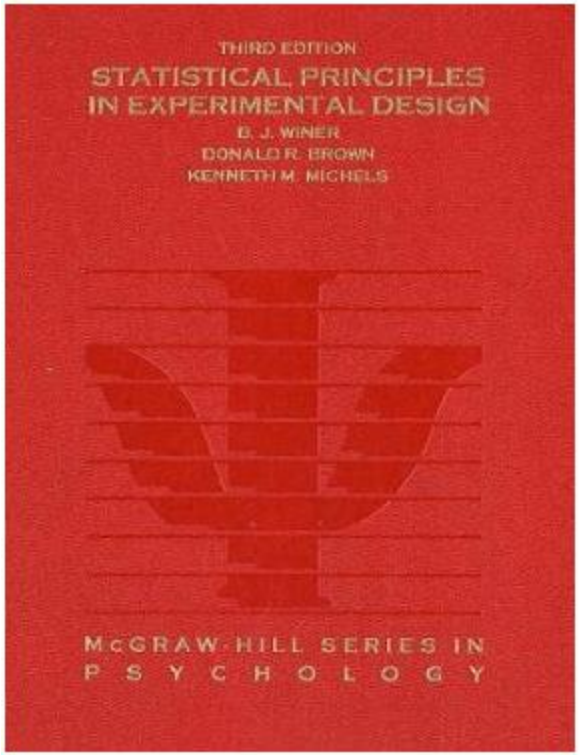
The standard stat curriculum

- Stat 1: Classical hypothesis testing
- Stat 2: Anova or regression
- Maybe Stat 3: longitudinal, multilevel
- What my advisor does (whether it's the best thing to do or not)

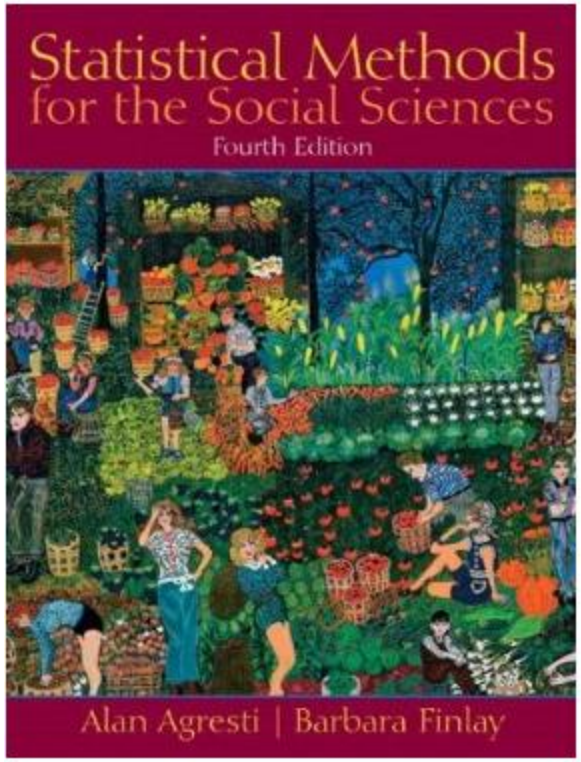
Standard texts for psych-influenced courses

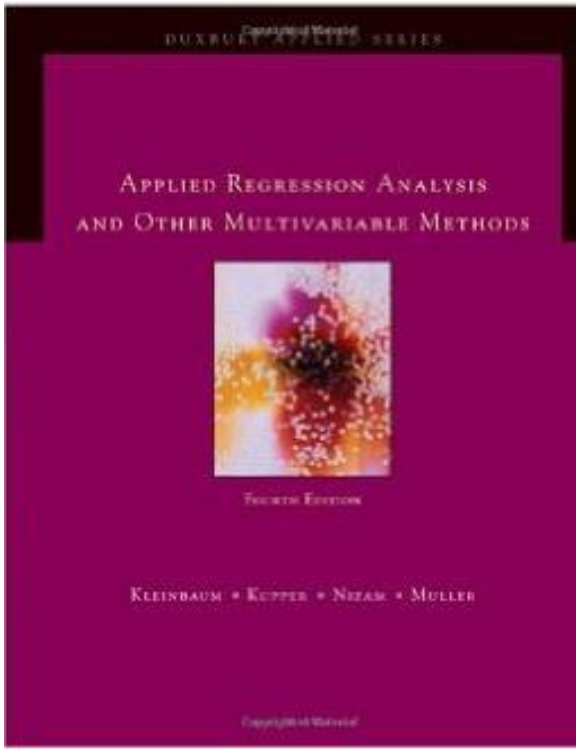




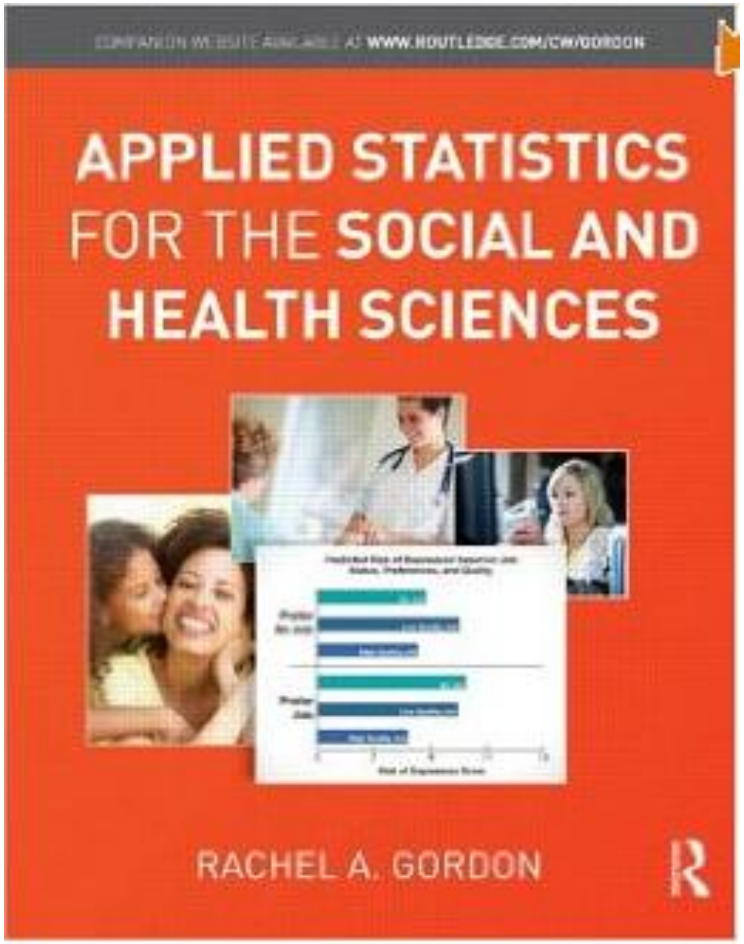


Or you may have used one of these if course was in sociology or public health





Plug for great book by local author



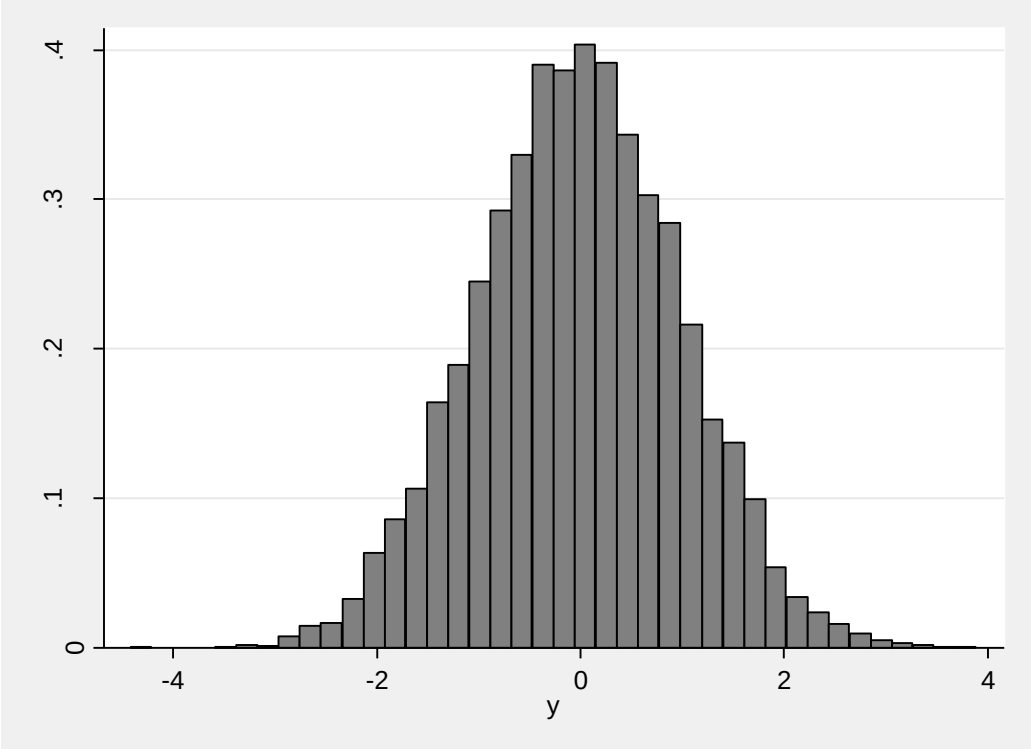
Contact Information

Sociology Office: 4140A BSB
Sociology Phone: 312-413-3857
Email: ragordon@uic.edu
Office: Rice Bldg, Suite 525
Phone: 312-413-0295

A world that no longer exists

In many substantive areas there were shared understandings, paradigms and conventions about how to do things. The next few slides show some examples. The computing/statistics revolution changed all that.

We assumed all variables looked like this



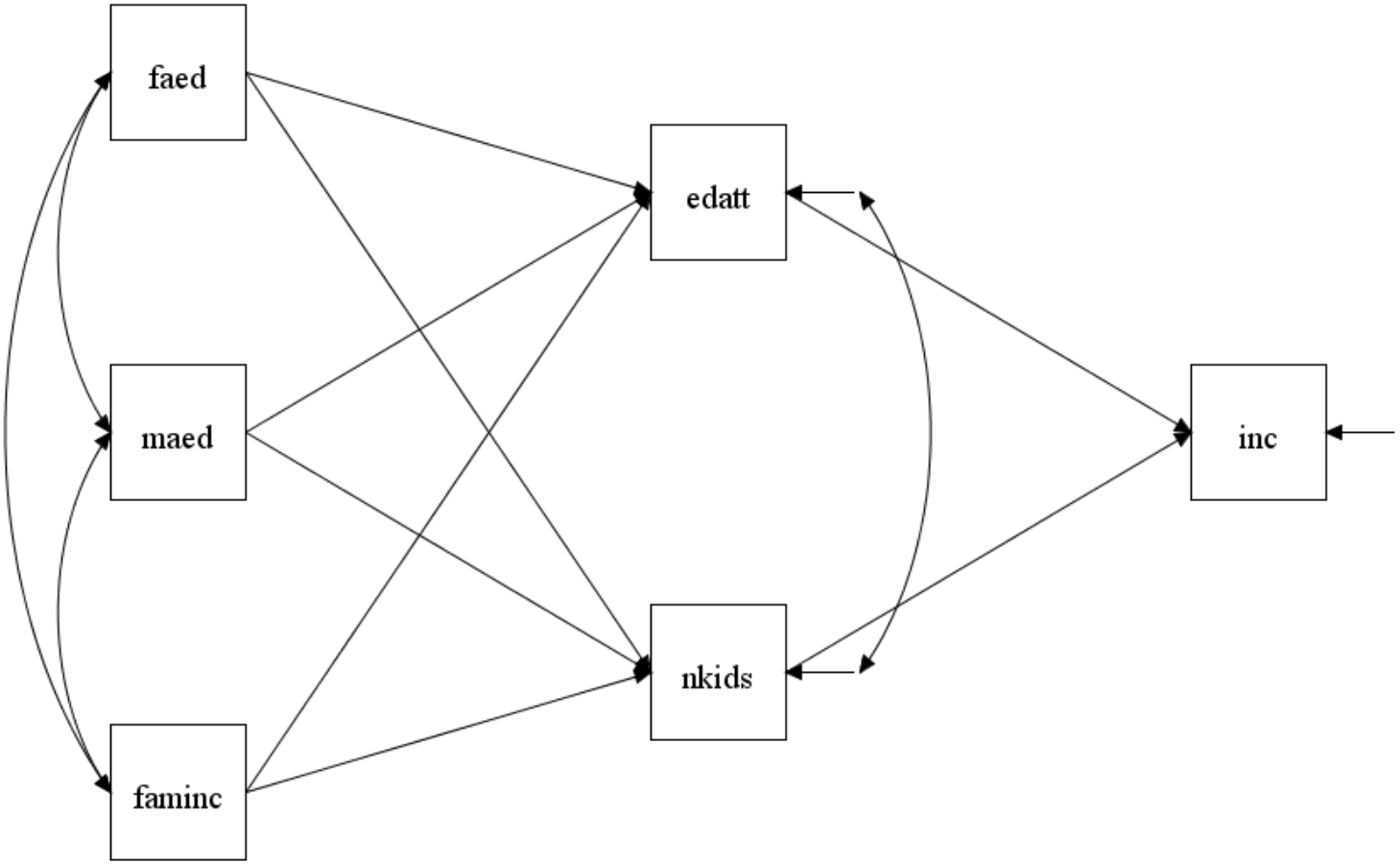
And we could use simple ANOVA

```
anova systolic drug disease drug#disease
```

Number of obs =	58		R-squared	=	0.4560
Root MSE	= 10.5096		Adj R-squared	=	0.3259

Source	Partial SS	df	MS	F	Prob > F
Model	4259.33851	11	387.212591	3.51	0.0013
drug	2997.47186	3	999.157287	9.05	0.0001
disease	415.873046	2	207.936523	1.88	0.1637
drug#disease	707.266259	6	117.87771	1.07	0.3958
Residual	5080.81667	46	110.452536		
Total	9340.15517	57	163.862371		

Or maybe a simple path diagram



Longitudinal data could be analyzed in a simple ANOVA framework

```
. anova score person drug, repeated(drug)

                           Number of obs =      20     R-squared     =  0.9244
                           Root MSE      = 3.06594     Adj R-squared =  0.8803

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |         1379     7         197       20.96     0.0000
                         |
                  person |        680.8     4       170.2       18.11     0.0001
                    drug |        698.2     3  232.733333       24.76     0.0000
                         |
                Residual |        112.8    12         9.4
              -----------+----------------------------------------------------
                   Total |      1491.8    19  78.5157895

Repeated variable: drug

                                    Huynh-Feldt epsilon        =  1.0789
                                    *Huynh-Feldt epsilon reset to 1.0000
                                    Greenhouse-Geisser epsilon =  0.6049
                                    Box's conservative epsilon =  0.3333

                                      ------------ Prob > F ------------
                  Source |     df       F    Regular    H-F      G-G        Box
              -----------+----------------------------------------------------
                    drug |      3    24.76     0.0000   0.0000    0.0006     0.0076
                Residual |     12
```

Threats to validity could be dealt with by quasi-experimental design

TABLE 1

SOURCES OF INVALIDITY FOR DESIGNS 1 THROUGH 6

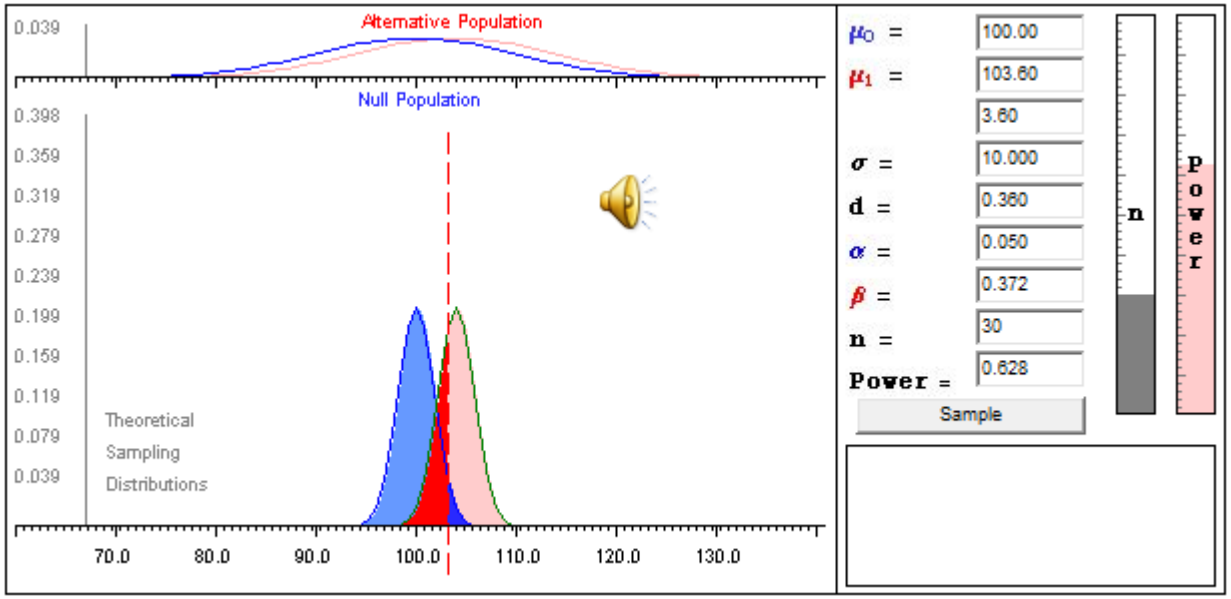
	Sources of Invalidity											
	Internal								External			
	History	Maturation	Testing	Instrumentation	Regression	Selection	Mortality	Interaction of Selection and Maturation, etc.	Interaction of Testing and X	Interaction of Selection and X	Reactive Arrangements	Multiple-X Interference
Pre-Experimental Designs:												
1. One-Shot Case Study X O	−	−				−	−			−		
2. One-Group Pretest-Posttest Design O X O	−	−	−	−	?	+	+	−	−	−	?	
3. Static-Group Comparison X O --- O	+	?	+	+	+	−	−	−		−		
True Experimental Designs:												
4. Pretest-Posttest Control Group Design R O X O R O O	+	+	+	+	+	+	+	+	−	?	?	
5. Solomon Four-Group Design R O X O R O O R X O R O	+	+	+	+	+	+	+	+	+	?	?	
6. Posttest-Only Control Group Design R X O R O	+	+	+	+	+	+	+	+	+	?	?	

Measurement was simple.

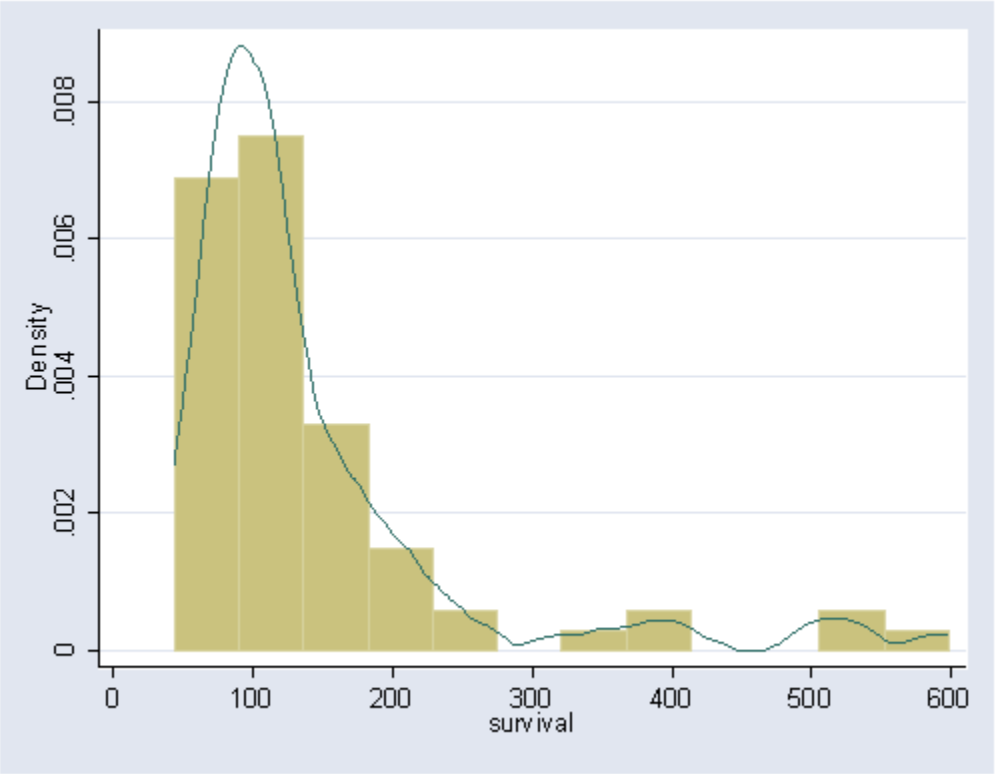
- Cronbach's alpha > .6 was all you had to show.
- If someone else used the measure and got it published, it was good for all time.

But then things changed and analyses which were once acceptable became unacceptable or at least subject to serious challenge.

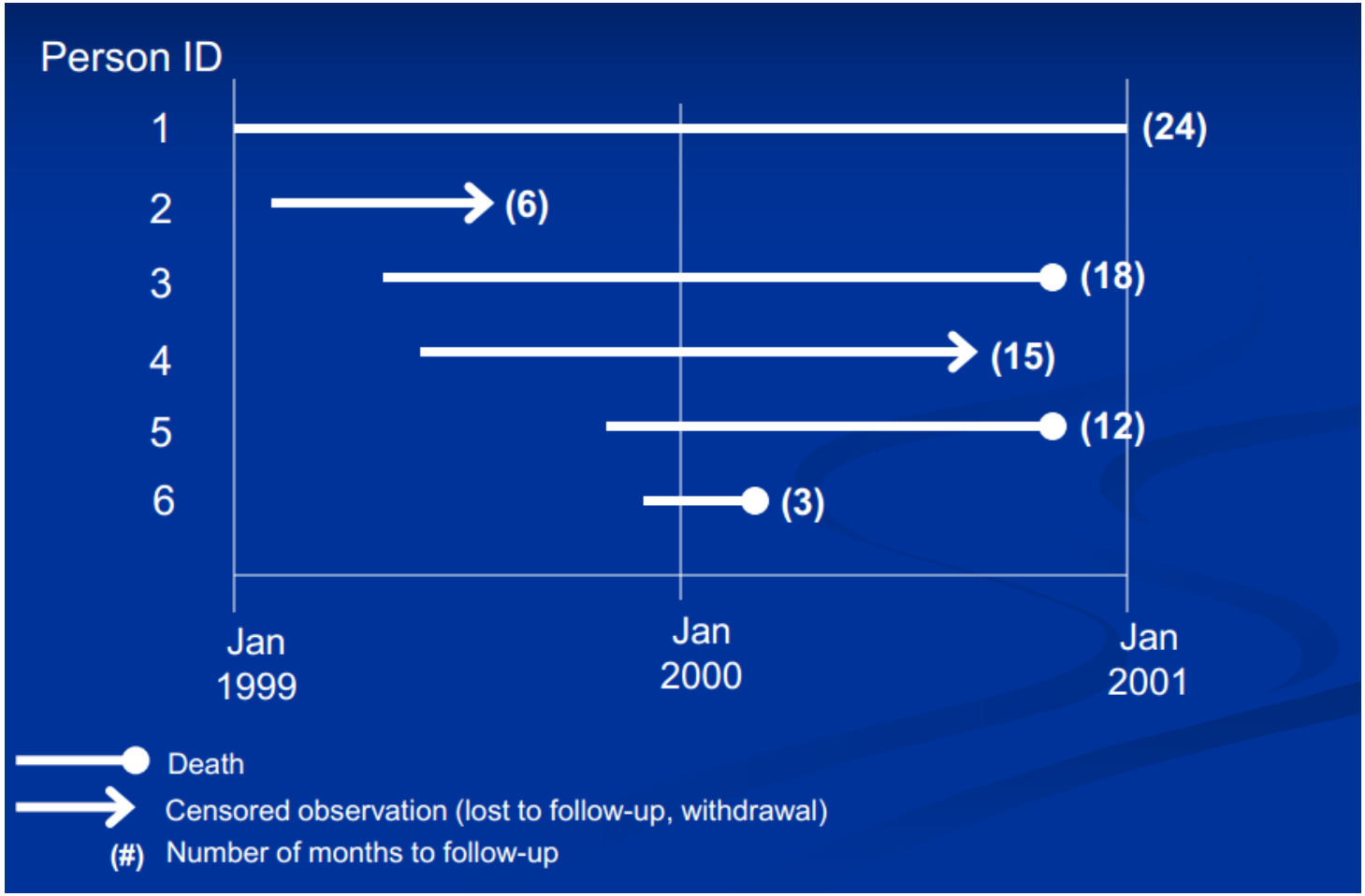
Power analyses are now *de rigueur*



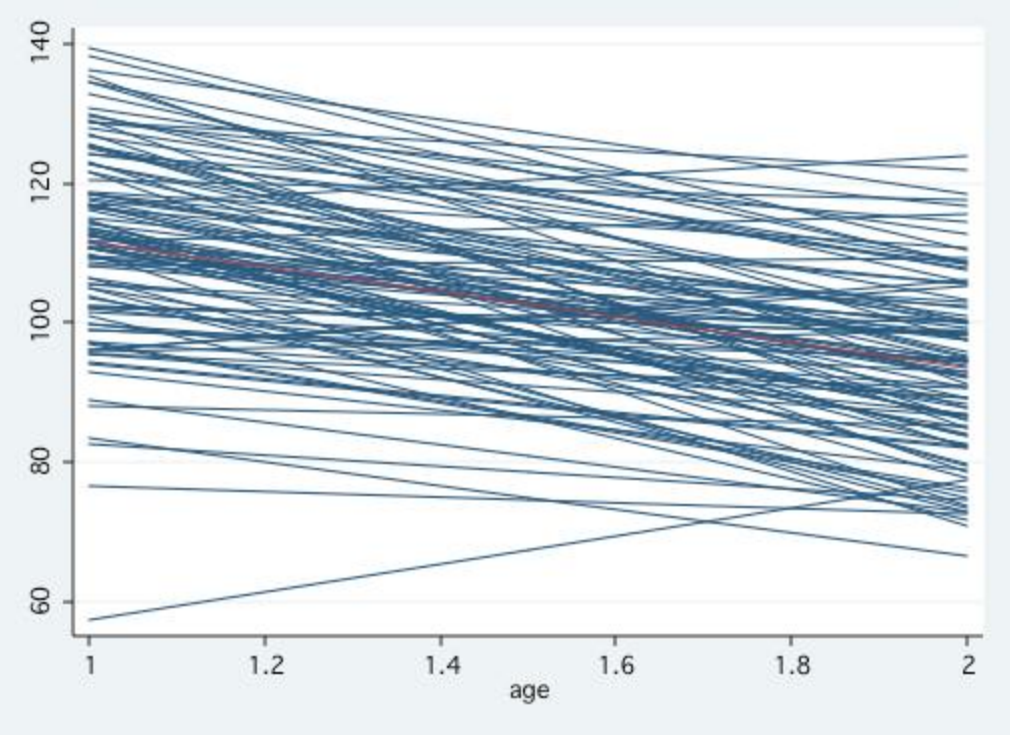
You can’t ignore non-normal distributions.



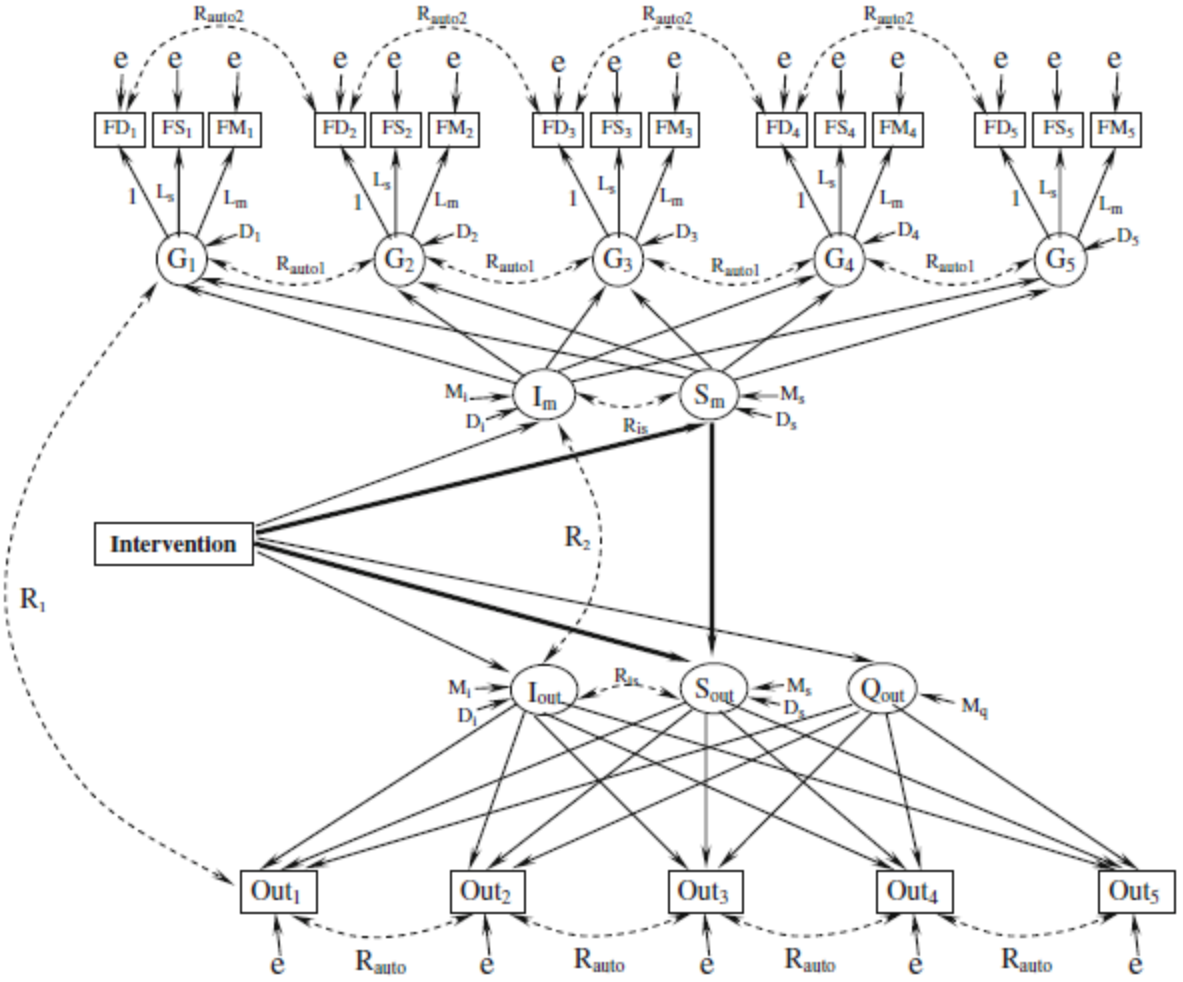
If your outcome involves time, you may need survival analysis.



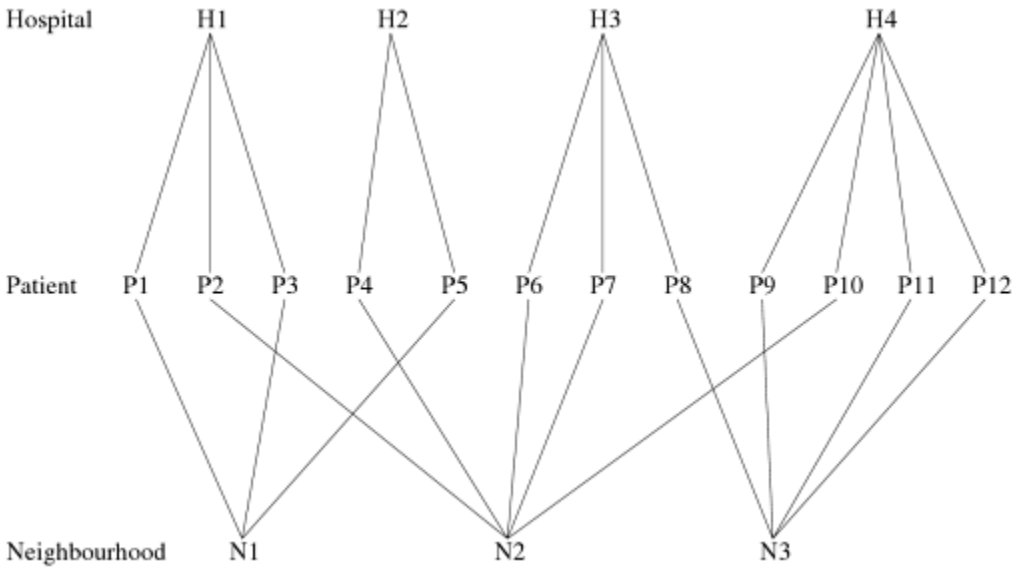
Longitudinal data present real challenges.



Path diagrams have become a bit more complex.



Multilevel data structures are common.



Data are analyzed using a generalized linear model.

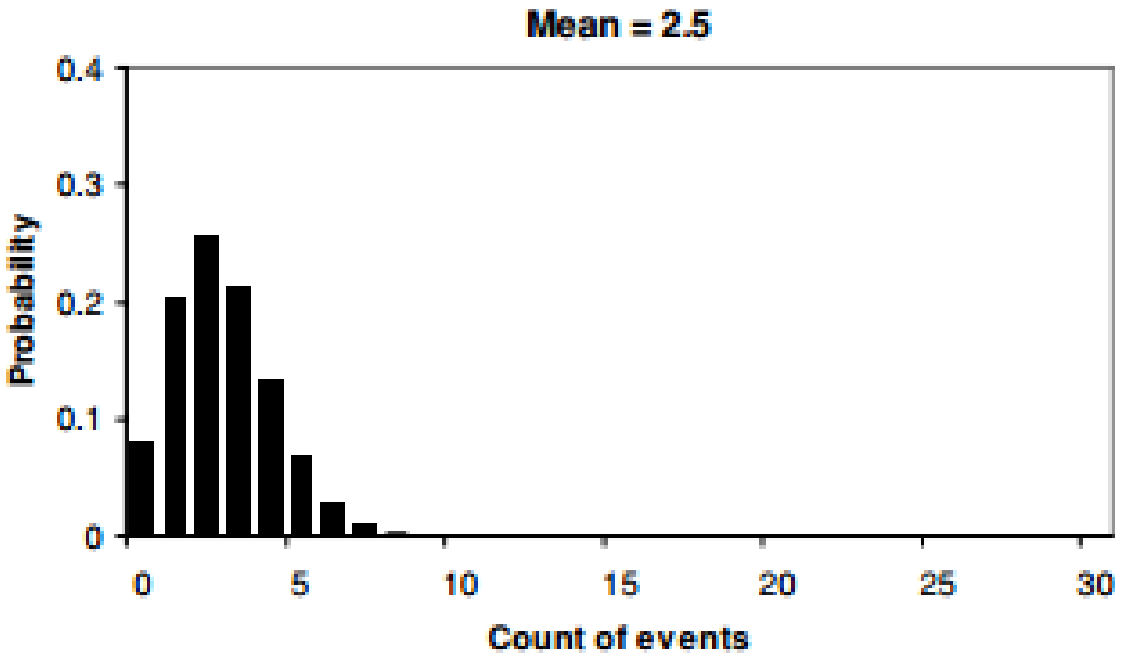
$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 Drug1 + \beta_1 Drug2 + \beta_3 Drug3 +$$

$$\beta_4 Disease1 + \beta_5 Disease2 + \beta_6 Drug1 * Disease1 +$$

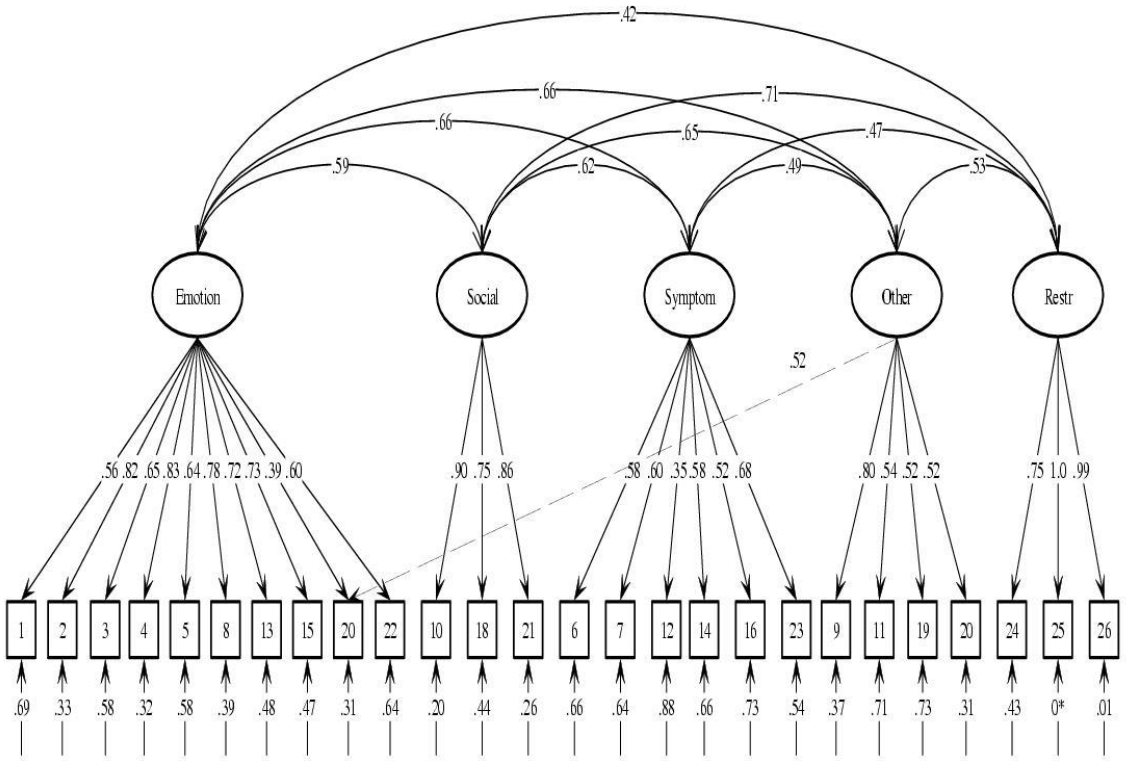
$$\ldots + \ldots. + \beta_{11} Drug3 * Disease2$$

systolic	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
drug						
2	-1.333333	6.363903	-0.21	0.835	-14.14321	11.47654
3	-13	7.431438	-1.75	0.087	-27.95871	1.958708
4	-15.73333	6.363903	-2.47	0.017	-28.54321	-2.923461
disease						
2	-1.083333	6.783944	-0.16	0.874	-14.7387	12.57204
3	-8.933333	6.363903	-1.40	0.167	-21.74321	3.876539
drug#disease						
2 2	6.583333	9.783943	0.67	0.504	-13.11072	26.27739
2 3	-.9	8.999918	-0.10	0.921	-19.0159	17.2159
3 2	-10.85	10.24353	-1.06	0.295	-31.46916	9.769157
3 3	1.1	10.24353	0.11	0.915	-19.51916	21.71916
4 2	.3166667	9.301675	0.03	0.973	-18.40663	19.03997
4 3	9.533333	9.202189	1.04	0.306	-8.989712	28.05638
_cons	29.33333	4.290543	6.84	0.000	20.69692	37.96975

Event counts require Poisson regression or one of its variants such as negative binomial.



Scale construction requires confirmatory factor analysis and/or item response theory (IRT).



You can't assert causation by just controlling on confounders.

Course announcement in current edition of *Amstat News*

Section on Statistics in Epidemiology Sponsors Short Course

1 MAY 2013 4 VIEWS NO COMMENT

The Section on Statistics in Epidemiology will sponsor the short course "Causal Inference and Its Application in Health Sciences" at the Joint Statistical Meetings. Instructors are Miguel Hernán of Harvard and Dylan Small of the Wharton School.

The first half of this short course, led by Hernán, presents a framework for causal inference from observational studies and recent methodological developments, with an emphasis on complex longitudinal data. The second half, led by Small, focuses on instrumental variable methods for causal inference in clinical trials and observational studies to control for unmeasured confounding. Software for structural models and instrumental variable methods will be discussed and real examples will be used for illustration.

So, you know you need a statistician.

But sometimes the interaction doesn't go so well.

And so, you feel a little put upon.

What we have got here is a failure to communicate.

(With a nod to *Cool Hand Luke)*

(With a nod to *Cool Hand Luke)*

In a nutshell…..

- Principal investigators frequently don't understand the statistical methods they propose to use.
- Statistical collaborators often don't fully understand what the PI is trying to do.
- It is glaringly obvious to reviewers when the Aims and the analysis sections of the proposal don't articulate.

Aims and hypotheses don't match analysis plan.

- Problem:
  - Nuances of investigator's ideas are not expressed clearly.
  - Even if they are, analysis section doesn't pick up on them.
- Solution:
  - Make sure analysis section speaks explicitly to each aim/hypothesis and I mean *explicitly.*
  - Ideally, there should be equations corresponding to each proposed analysis.
  - But, in addition, there should be a clear verbal explanation of what is going on. You are speaking to two kinds of reviewers.

You can't understand the statistics.

- Problem: the analysis section is so "mathy" that you can't understand it.
- Solution: This is indeed a problem, your proposal will be read by non-stats types
  - Make sure that there a clear verbal explanation of what you are doing in addition to relevant eqs.
  - Graphs are a *big* help.
  - Avoid statistical jargon and abbreviations where possible

Analysis not appropriate for nature of outcome variable.

- Problem:
  - Your outcome variable is not normally distributed or continuous but you say you are going to do regression.
- Solution:
  - You can now model virtually any kind of outcome in a GLM framework:
    - continuous but non-normal: transformations
    - dichotomous, ordered, multinomial: various forms of logistic
    - count: Poisson, negative binomial, zero inflated binomial etc
    - time to event: survival
  - Be explicit about the correct model for the type of outcome and that you are aware of alternatives.

Theory section says (or hints at) interaction and/or non-linearity but hypotheses and analysis plan doesn't recognize it

- Problem:
  - Your aims say you want to "compare men and women."
  - Compare what: mean outcomes, difference in effects of an intervention (interaction), measurement properties, what?
- Solution:
  - Convey these ideas to your stats person clearly; graphs help a great deal.
  - Make sure that you specify equations which test the actual hypothesis.

Potential missing data problems ignored or treated glibly.

- Problem:
  - Analysis section says nothing about missing data but it's obvious there will be some.
  - Vague statement: "We will use multiple imputation if necessary."
- Solution:
  - Anticipate amount if missing data from prior studies or simple pilots.
  - Missing data is different from attrition/loss to follow up.
  - Simple methods, e.g. listwise deletion often work best, but justify what you do.
  - Multiple imputation is tricky, you should not refer to it unless you know *exactly* how it will be used.

Longitudinal analysis incompletely specified / not linked to aims.

- Problem:
  - You measure outcomes at several time points.
  - You are interested in the trajectory over time, but particularly the last time point.
  - You are not sure which of several analytic methods you should use.
- Solution (sort of):
  - Know the difference between random effect and GEE models and which you are using.
  - Last time point analysis can use traditional methods, but what about attrition?
  - Are trajectories and differences therein what you are really interested in?

Analysis section focuses on details rather than larger picture.

- Problem:
  - Stats write up seems focused on detailed exposition of methods. Lots of formulas and discussion of assumptions, computational methods etc.
- Solution:
  - Use references and appended papers to statistician's own work.
  - Clearly separate technical material from flow of verbal discussion.
  - Distinguish between standard methods and anything that is innovative in the proposal. Reference the former and fully explain the latter. (Secondary Aim?)

Power analysis

- Problem:
  - Power analysis is very brief and cryptic or extremely detailed and overwhelming.
- Solution:
  - Be brief, but,
    - Specify plausible effect sizes.
    - Use plausible range of variance estimates.
    - Allow for missing data and attrition.
    - Specify software used.
    - Should be just a few paragraphs long in most cases.

Things that you can do to maximize the statistician's contribution

Educate your statistician prior to your first meeting.

- Provide one or two published papers in your field of study that use methods similar to yours. The more s/he knows about your substantive area the better.
- Send a list of specific questions prior to a face to face meeting.
- Be honest about what you know and don't know. Try to establish a common ground of communication.

Ask an experienced person to read your proposal in draft with an eye toward coherence.

- If your reader doesn't understand the stats you have a problem.
- Don't be afraid to ask for analysis section revisions.
- Obviously, you need to give a collaborator time to do this.
- But it's better to ask, even very close to deadline, than to let it go.

The analysis section should speak to each aim/hypothesis explicitly.

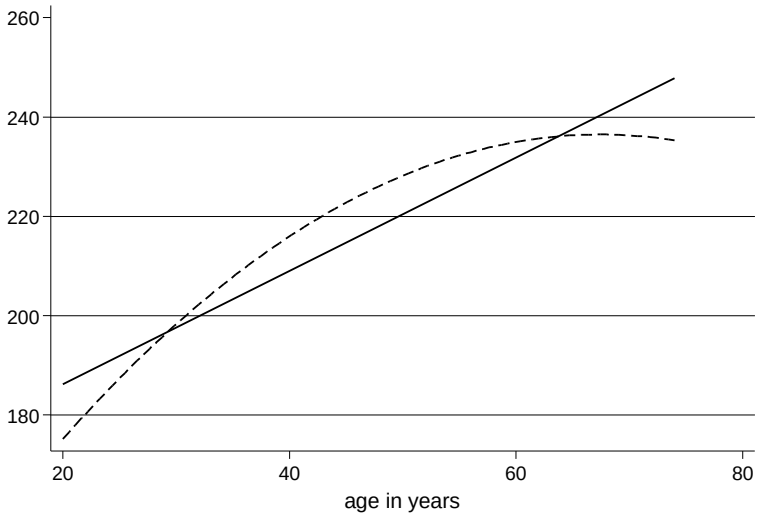
- Try to talk through this at the beginning. Don't just send the aims etc to the stats person and expect them to deal with them.
- Help the statistician understand exactly what you are trying to accomplish. His or her questions may help you achieve greater clarity.
- Remember that you are, until the grant is actually funded, getting free consultation.
- Still, it's to the statistician's advantage to get things right at the beginning.

Do not ignore measurement issues.

- Non-statistical reviewers tend to focus on reliability and validity.
- Many statistics people are not familiar with measurement theory. You may have to do this on your own.
- You may need to write this yourself, but someone has to do it.
- Brief reference to Cronbach's alpha is not enough unless the measure is very well established.
- Spend some time learning modern methods of scale construction.

Use graphs and other visual devices.

- Conceptual diagrams
- Flow charts showing project sequence
- Time lines
- Graphs of expected results.



A little statistical innovation goes a long way.

- Ideally, your analysis should push the envelope a little, but not too far. These might be secondary aims.
- Examples:
  - Trajectory analysis that uses structural equation methods to deal with measurement error.
  - Sensitivity analysis that tests importance of assumptions.
  - Propensity scores to deal with non-random assignment.

Two very useful articles

- These articles appeared in *AmStat News* recently and contain many of the points I made today.
- I wrote this presentation before reading them, honest!
- Overview of NIH review process
- Tips for Writing a Statistical Analysis Plan