The Art of Data Analysis:

Good Advice, Counters to Bad Advice, and Some Examples

James A. Wiley
San Francisco State University
June 2, 2009

Data Analysis Versus Statistics

- Data analyst: Finds the most effective ways to express measurements, state hypotheses, formulate models, and design comparisons for a given scientific problem (Data Doctor)
- Statistician: Can effectively separate "noise" from systematic variation in many ways and contexts (Priest of Uncertainty)

Principle 1: Everything is a Finding

- Findings that violate your expectations are still findings
- It is as important to explain why your hypothesis failed as it is to explain why it succeeded
- Small correlations can be beautiful

Principle 2: Make educated guesses about what you will find

- To make an educated guess, you have to come up with reasons ("theory")
- If your guess is wrong, you will learn something by trying to understand an unexpected result

Principle 3: There is no "best" way to analyze your data

- Think of multiple methods/tests/models as different "windows" through which to look at your data
- Multiple options about measurement, grouping and stratification of cases, and types of comparisons are available in most studies
- You will have more confidence in a result which shows up in several kinds of data analysis and statistical testing

Principle 4: If you have a strong *a priori* hypothesis, give it a simple test first

- If you don't see your hypothesis confirmed in a simple comparison, you will usually not find it to be confirmed in a more complex comparison (e.g., one with lots of adjustments)

Principle 5: Work first with data that are as close to observations as possible

- Individual responses rather than scales, data for short time units rather than long time units, specific quantitative variables rather than indexes that combine several variables
- you will need to learn how to effectively combine pieces (or not) to pursue your problem
- Some findings are revealed more in fine-grained data than in aggregated data and vice versa---try lumping and splitting

Principle 6: Keep a log/diary of your data analyses

- A diary will reveal how you approach the data analysis task and what improvements you might consider
- It will help you identify the reasons you made certain decisions --- useful when you write up the results

Principle 7: Learn how to get help from data analysis/statistics specialists

- Pick a specialist who is interested in finding an answer to your research question
- Learn how to tell a specialist what you want

Some Characteristic Data Analysis/Statistics Problems

- How do I know whether or not my proposed measurement is a good one?
- How many subjects/observations do I need for my study?
- What is the evidence for a causal relation in my study?
- What is the best way to study changes over time in my longitudinal study?

How do I know whether or not my proposed measurement is a good one?

Reliability and Validity of Measurements I

- if there's no change in the thing being measured, you should get the same result every time you take a measurement (Reliability)
- if you have multiple but somewhat different measures of the same thing, they should be associated in a consistent way (Internal Consistency)
- if two or more researchers rate observations on the same occasion, their measurements should agree (Agreement)

Reliability and Validity of Measurements II

- if there is a "gold standard" for the thing being measured, your measures should more or less agree with the gold standard (Validity)
- if you can argue *a priori* that some persons should be "high" on the measurement and others should be "low", your measure should predict who is in the high group and who is in the low group (Predictive Validity)

Reliability and Validity of Measurements III

- if you expect your measure to correlate highly with another measurement (of some distinct but related concept), you should observe a sizable correlation (Construct Validity)
- If you expect your measure to be weakly correlated with a measurement of some other concept, you should observe a small correlation (Discriminant Validity)
- Slight differences in the way you design the measurement protocol shouldn't produce very large differences in the results (Stability)

Internal Consistency in Measurement of Attitudes Toward Technology in the 1992 Eurobarometer Survey

Would you please tell me how much you agree or disagree: strongly agree, agree to some extent, disagree to some extent, or strongly disagree

- "Even if it brings no immediate benefits, scientific research which advances the frontiers of knowledge is necessary and should be supported by the government"
- "New inventions will always be found to counteract any harmful consequences of scientific and technological development"
- "Only by applying the most modern technology can our economy become more competitive"
- "Scientific and technological progress will help cure illnesses such as AIDS and cancer"
- "The benefits of science are greater than any harmful effects it may have"

Cronbach's Alpha for Internal Consistency

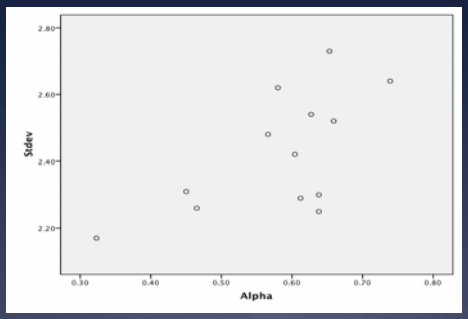
k= the number of items

X_i = the ith component (item score)

Y = the total score = sum of Xs

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum_i^k Var\ (X_i)}{Var\ [Y]}\right] = \frac{common\ \ variation}{total\ \ variation}$$

Country	Cronbach's alpha	Mean Range 5-20	Std. Deviation
All N=4,459	0.586	15.6	2.51
France N=367	0.627	15.4	2.54
Belgium N=356	0.580	14.9	2.62
Netherlands N=383	0.465	15.0	2.26
West Germany N=380	0.739	15.7	2.64
Italy N=332	0.612	15.8	2.29
Luxembourg N=166	0.450	15.4	2.31
Denmark N=392	0.323	14.6	2.17
Ireland N=288	0.659	15.5	2.52
United Kingdom N=488	0.604	14.8	2.42
Greece N=313	0.638	16.6	2.30
Spain N=312	0.566	16.0	2.48
Portugal N=312	0.653	15.9	2.73
East Germany N=370	0.638	16.7	2.25



How many subjects/observations do I need for my study?

Implications of Sample Size

- Generally, more is better, but there are exceptions
- You need to know how large a sample is needed to make sure that an expected finding isn't missed (planning the sample size needed in advance)
- You a proposing a secondary analysis of existing data (is the existing sample large enough?)

Power in words

- A (random) sample is selected from among a larger population of observations
- Sampling gives an approximation of what is happening in that population
- The approximation gets better as the sample size increases
- Most routine statistical tests are set up as tests of no difference (no relationship) in the population: these are called "null hypotheses"

Power in words (continued)

- The typical significance level is the probability that the test will lead to rejection of the null hypothesis when it's true: this circumstance is called a Type I error
- The power of the significance test is the probability that the test will lead to rejection of the null hypothesis when in fact there is a non-zero difference or relationship is of a specific size. This size is known as the "effect size"

Ingredients of a power calculation

- The nature of the null hypothesis and the level of significance (α)
- The sample size and how the sample is distributed over the number of groups
- The "effect size", whose definition depends on the nature of the significance test

A Three Group Comparison

- Research Question: The general question to be addressed is the extent to which specific immunological and psychosocial variations among the three sampled groups – natural viral suppressors, non-progressors [viral load < X and CD4 count >Y with standard anti-viral medications], and progressors [viral load > X and CD4 count <Y whether or not on anti-viral medications] are consistent with expectations based on our theory of immune disregulation and our prior research.

Test of significance level α	.05	.05	.05	.05	.01	.01	.01	.01
Number of Groups	3	3	3	3	3	3	3	3
Variance of means (between groups)	10	10	10	10	10	10	10	10
Standard deviation within groups	10.0	11.5	14.1	20.0	10.0	11.5	14.1	20.0
Effect Size= ratio of variance between groups to variance within groups	.100	.075	.050	.025	.100	.075	.050	.025
Power = probability of accepting the hypothesis of effect size=0 when effect size is as above	.99	.99	.94	.68	.99	.96	.83	.44
Sample size in each group	100	100	100	100	100	100	100	100

G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences FRANZ FAUL Christian-Albrechts-Universität Kiel, Kiel, Germany EDGAR ERDFELDER Universität Mannheim, Mannheim, Germany AND ALBERT-GEORG LANG AND AXEL BUCHNER Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany

G*Power (Erdfelder, Faul, & Buchner, 1996) was designed as a general stand-alone power analysis program for statistical tests commonly used in social and behavioral research. G*Power 3 is a major extension of, and improvement over, the previous versions. It runs on widely used computer platforms (i.e., Windows XP, Windows Vista, and Mac OS X 10.4) and covers many different statistical tests of the t, F, and 2 test families. In addition, it includes power analyses for z tests and some exact tests. G*Power 3 provides improved effect size calculators and graphic options, supports both distribution-based and design-based input modes, and offers all types of power analyses in which users might be interested. Like its predecessors, G*Power 3 is free.

Behavior Research Methods 2007, 39 (2), 175-191

http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/download-and-register

What is the evidence for a causal relation in my study?

Criteria for Causal Inference

- Evidence of a relationship between putative cause and effect (association)
- Change in cause is followed by change in effect (time order)
- A mechanism that explains the linkage between cause and effect ("theory")
- Other factors are not responsible for the linkage (ruling out confounding)

The Randomized Experimental Paradigm

Randomized Group	Before	After
Intervention Group	Pre-intervention measurements	Post-Intervention measurement
Control Group	Pre-measurement	Post-measurement

Inference about Intervention Effects In An Experiment

- What direct control of administration of an intervention means
- What a control group accomplishes
- What randomization accomplishes
- Why do a pre-measurement?
- The counterfactual

Typical Non-Experimental Situation

- There is an association between potential cause and effect
- "Theory" says there should be an association and supplies some reasons (weak versus strong theory)
- Time order is unclear, especially in cross-sectional studies but also in many longitudinal designs
- It is very difficult to rule out alternative explanations for the observed relationship

Ways of Dealing with Confounding

- Holding some factors constant by dividing the observations into groups (stratification)
- Control by matching on potential confounding factors and/or "propensity scores"
- Control by adjustment in multivariate models

Does Victim Participation in Domestic Violence Prosecution Increase the Risk of Subsequent Violence?

- Study of 1,000 incidents of domestic violence leading to arrest of a male perpetrator in a Midwestern county
- Main question: does victim participation in prosecution of cases compromise her safety?
- Data collected from police incident reports, prosecutor files, court records, and emergency room records

Variables in the Analysis

- Key independent variable= match between victim and prosecutor re prosecution of domestic violence
- Key dependent variable= police records of violence after the disposition of the case
- Potential confounders: prior domestic violence arrest of perpetrator, level of violence in the index event

Victim-Prosecutor Match and Subsequent Violence

match4 * polevent Crosstabulation

			polevent		
			0	1	Total
match4	P_pro V_pro	Count	174	252	426
		% within match4	40.8%	59.2%	100.0%
	P_nonpro V_pro	Count	61	125	186
		% within match4	32.8%	67.2%	100.0%
	P_pro V_nonpro	Count	41	80	121
		% within match4	33.9%	66.1%	100.0%
	P_nonpro V_nonpro	Count	73	85	158
		% within match4	46.2%	53.8%	100.0%
	Total	Count	349	542	891
		% within match4	39.2%	60.8%	100.0%

Match with Controls in a Logistic Regression Analysis

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	match4			7.383	3	.061	
	match4(1)	.192	.190	1.021	1	.312	1.212
	match4(2)	.523	.225	5.401	1	.020	1.686
	match4(3)	.495	.251	3.891	1	.049	1.641
	priorarrestrc	.256	.156	2.688	1	.101	1.291
	danger	.093	.142	.427	1	.514	1.097
	Constant	.050	.175	.080	1	.777	1.051

Is There Such a Thing as Over-control?

- No simple answer
- A strategy: add controls judiciously; examine hypothesized causal relation in a context (e.g., a group within the sample) which allows for a relatively clean assessment; look for collateral evidence of mechanism at work
- The special case of chains of causation you haven't considered in your analysis

What is the best way to study changes over time in my longitudinal study?

Modern Longitudinal Analysis

- Focus on factors that influence the shape trajectories of outcomes over time. This derives from "growth models" rather than "panel models"
- Factors that affect trajectories include stable characteristics (gender, ethnicity) as well as time-varying characteristics (marital status, snapshots of behavior)
- Software to study trajectories is readily available (Mplus, HLM, GEE under SPSS)

Approximating sample size in longitudinal studies

- N= the number of subjects (number of clusters)
- k= the number of occasions each subject is observed (cluster size)
- r = intraclass correlation (ratio of between person variance to total variance; =1 when no within person variance and =0 when no between person variance)
- Effective sample size, *Eff*(N) is approximately
- *Eff*(N)= Nk/[1+r(k-1)]
- *Eff*(N)= Nk when r=0 *Eff*(N)= N when r=1

Research Questions

- Are there different patterns of homelessness in welfare populations across time and what baseline variables predict membership in these patterns?
- For patterns of homelessness than change across time, are there time varying predictors that are associated with this change?

Welfare Client Longitudinal Study

Year	Data	# TANF women
2001	Representative sample of TANF & GA applicants (n=1510, RR=86%)	455
2002	1st Year Follow up of Recipients: Oversample of substance abusers (n=718, RR=86%)	400
2003	2nd Year Follow-up (RR=85%)	388
2004	3rd Year Follow-Up (RR=82%)	371
2005	4th Year Follow-up (RR=81%)	Data not yet available

Total N of TANF WOMEN with Available For Longitudinal Analyses = 419

Indicators of Homelessness

Respondents indicated whether they had lived in any number of places in the past 12 months. For a given respondent-year, severity of homeless was classified into:

a) Had own place the entire past year
b) Was doubled-up for some period in the past year (lived with family, friends, etc). For each doubled-up dwelling location, data were also available on whether or not the primary dweller of the household was happy (i) that the respondent was living there or (ii) less than happy
c) Were literally homeless (slept in a bus station, car, shelter, etc.)

From these three categories a dichotomous homelessness measure was constructed for each respondent-year as follows:

The respondent-year was classified as Not Homeless if: a) or b(i)
The respondent-year was classified as Homeless if: b(ii) or c)

	Baseline	1st FU	2nd FU	3rd FU
% Homeless	32.7%	28.4%	22.2%	19.7%

Latent Class Growth Curve Modeling

For K separate groups (k = 1,…, K > 1), latent growth curves estimated the probability of homelessness across time as:

$$\text{logit}\left(\frac{p_i^k}{1-p_i^k}\right) = \beta_0^k + \beta_1^k \cdot t + \beta_2^k \cdot t^2$$

Where each respondent can only belong to a single group (similar to clustering of objects) and the grouping of respondents is done empirically at the same time as the estimation of the growth parameters β. The value of K must be specified a-priori.

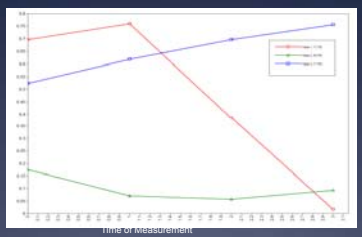
Latent Class Growth Curve Parameters

Table 1: Parameter estimates for each of the 3 Homelessness profiles

	Homelessness Profile		
	Low Homelessness n=296	Descending Homelessness N=71	High Homelessness n=52
Parameter Estimates			
Intercept	1.21**	3.71**	3.48***
Linear	.23	-1.41***	.68
Quadratic	.03	.39	-.41

*, **, *** Indicates a significant coefficient at the .05, .01, .001 level
Model BIC = 1638.45 for 3 profile solution, Entropy = .60, Pearson Model χ^2 p-value=.48

A Three Class LCGC Analysis Solution



Group 1 – Descending HL Group, 17.3%
Group 2 – Low HL Group, 64.9%
Group 3 – High HL Group, 17.8%