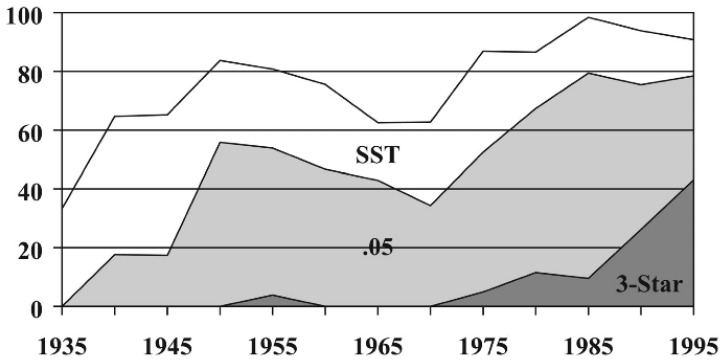
210A Week 10 Notes

Statistical versus Substantive Significance

We have a tendency to treat statistical significance as an indicator of how important an effect is. This is not appropriate, though statistical significance can be a crude proxy for importance.[1] Part of the problem is that if variables are on different scales it can be hard to compare effect sizes so it's tempting to rely on statistical significance as a common scale. Worse yet, with with some advanced regression techniques like logit or negative-binomial it can be hard to interpret what the effects actually mean which makes relying on statistical significance tempting. This should be avoided and if you need help interpreting effect strength you should use standardized coefficients or calculate vignettes insead of relying on statistical significance.

Nonetheless there is a tendency to fetishize statistical significance such that in practice $p<.05$ is taken to mean something other than "there is a 5% or less chance that these results could occur even if the null were true." Unfortunately, $p<.05$ is often taken as a magic number such that .051 is treated as "not significant" and .049 as "significant" when in fact both of them mean the same thing, that the risk of Type I error is about 1 in 20. This tendency is reinforced by the increasing popularity of asterixes to represent statistical significance, forcing significance level from a quantitative gradation (Type I error is more or less plausible) to a categorical distinction (the finding is or is not statistically significant).

Figure 1. Percent of articles (calculated using five-year intervals) over time that use statistical significance testing (statistical significance test), statistical significance testing and alpha = .05 (.05), and statistical significance testing, alpha = .05, and the 3-star system (3-Star)



Source: Leahey, Erin. 2005. "Alphas and Asterisks: The Development of Statistical Significance Testing Standards in Sociology." *Social Forces* 84:1-24.

Statistical significance only tells us how likely the patterns in the sample are to generalize to the population, it doesn't tell us whether we should care. If

[1] Ziliak, Stephen T. and Dierde N. McCloseky. 2004. "Size Matters: The Standard Error of Regressions in the American Economic Review." *Journal of Socio-Economics* 33:527-546.

we found a statistically significant result that one group had a 50% chance of some event and another group had a 50.01% chance of the same event, these groups are substantively similar. Such statistically significant but substantively trivial results are possible because statistical significance is a function of both the parameter and its standard error. Since standard error is an inverse function of the root of n, statistical significance increases with the root of n (at least absent publication bias). Therefore you should especially take statistical significance with skepticism for large N studies, it may only mean that a trivial effect generalizes.

Reifiying Data

> They are ourselves, I replied; and they see only the shadows of the images which the fire throws on the wall of the den; to these they give names, and if we add an echo which returns from the wall, the voices of the passengers will seem to proceed from the shadows. –Plato
>
> For we know in part, and we prophesy in part. But when that which is perfect is come, then that which is in part shall be done away. When I was a child, I spake as a child, I understood as a child, I thought as a child: but when I became a man, I put away childish things. For now we see through a glass, darkly; but then face to face: now I know in part; but then shall I know even as also I am known. – St. Paul

It is tempting to treat our data as the thing itself, but we must always be conscious that our computer files do not have facts about social processes but measures of facts and these measures are themselves the result of social processes. These measures may not accurately reflect the underlying reality, or worse yet, there may be no underlying reality for them to reflect. So we can't really say that college education is correlated with liking opera, only that in surveys self-reported college education is correlated with claiming to like opera. These sound similar but there are lots of reasons they might be decoupled.

Opinion polls are intensely artificial constructs. It is somewhat absurd to say that there is an underlying reality as to how many Americans think the Second Amendment guarantees an individual right to bear arms versus a right of the states to maintain militias for the simple reason that most Americans don't walk around thinking about this. All sorts of issues only becomes salient when a pollster asks people about them which is one of the reasons that "opinions" are so sensitive to question wording and ordering on many issues – the opinions don't really exist as pre-existing cognitive facts to be discovered by the poll, rather the pollster *invokes* or *conjures* these opinions through asking questions – many of which have things like "forced choice" or allowing "don't know" only an implicit choice in order to ensure that most people will state an opinion. Most of the questions in the GSS ought to be prefaced by "when asked by a pollster ..."

One of the best understood instances of this problem is crime statistics.[2] As shown at great length in the tv show *The Wire*, police departments are under pressure to reduce crime, which really means they are under pressure to reduce the measured amount of crime. This can be accomplished by actually reducing crime but it's much easier accomplished in the short-run by juking the stats through such means as downgrading felonies to misdemeanors. This can lead to wide divergence between the Uniform Crime Report (which is based on local police reports) and Victimization (which is based on an opinion poll). For similar reasons, homicide is the gold standard of criminology because a dead body with a bullet hole in it is rather objectively a murder whereas rape is notoriously variable and fuzzy because over time there have been huge changes in what the police will recognize as a rape and in when victims feel comfortable reporting the crime. In all sorts of ways statistics do not document but socially construct their phenomena of focus. For instance, presidential approval ratings are probably more important than presidential approval per se in that if polls did not measure this and the press did not report it, it wouldn't matter very much.

This can lead to nihilism until you realize that some statistics are more problematic than others. By thinking seriously about where data come from you can be discerning rather than naive (or cynical). Even variables that lack face validity can still be useful if you treat them cautiously. Self-reported church attendance is a thoroughly mediocre variable if you're interested in how often people go to church.[3] However it's an increasingly useful predictor of politics, fertility, and gender attitudes – a much better predictor than denominational affiliation and most demographic traits. So go ahead and throw church attendance in your model, just interpret it carefully.

Overcontrolling

Quantitative social scientists worry a lot about "spuriousness," which is when we think something is caused by its association with something else. At talks or in peer review the easiest comment is "but did you control for X"? In part to avoid such inevitable criticisms, we tend to throw as many control variables as we can get into the model and what's left is our effect.

However this can lead to myopia. First, often we are not interested in the variable itself in a literal and direct sense but see the variable as indicative of something bigger. If two variables are both indicating the same latent tendency of interest it's inappropriate to consider one a "control" for the other. Second, often "controls" are actually mechanisms through which the variable exerts its impact. When Blau and Duncan showed that most father-son occupational reproduction occurs through education, no sensible person would say they showed occupational reproduction to be "spurious," rather they found a mechanism to

[2]Kitsuse, John I. and Aaron V. Cicorel. 1963. "A Note on the Uses of Official Statistics" *Social Problems* 11:131-139.

[3]Hadaway, C. Kirk, Penny Long Marler, and Mark Chaves. 1993. "What the Polls Don't Show: A Closer Look at U.S. Church Attendance." *American Sociological Review* 58:741-752.

explain how it works. Likewise in their study "occupation" stood for a whole host of related variables such as income (which they did not observe).

The issue of "overcontrolling" is at the core of the structural discrimination literature. If you simply compare whites and blacks on various dimensions (income, wealth, imprisonment) you quickly see that whites have it much better off than blacks. Take another step and throw in all the control variables you can think of – education, labor force experience, social networks, etc – and most of the gap goes away. There are two ways to interpret this. One is that there is no discrimination. The other is that we have identified the mechanisms through which inequality is perpetuated.

This is not to say that the "throw controls at it" approach is silly as identifying mechanisms is enormously important for policy intervention. For instance, the evidence suggests that "racial profiling" is basically a red herring which is why more sophisticated scholars who worry about mass incarceration (Bruce Western, Glenn Loury) don't give it any attention but focus on structural conditions like unemployment and mandatory minimums. This implies a completely different set of policy implications than looking for racist cops. Likewise, with our early example of gender gap versus mommy gap, it's very difficult to say whether it is appropriate to control for occupation. If you think that in some meaningful sense occupations are not freely chosen but are imposed on people in gendered ways then selection to occupation is merely the mechanism through which gender inequality is achieved. (This is the premise behind things like the legal theory of "comparable worth").

A closely related problem in regression is colinearity. This when you put two highly correlated predictors in the same model. Often this will lead to one variable simply dropping out or the model going haywire – for instance it might make one of the two variables have large positive effects and the other large negative effects.

The Assumption of Independence

Statistics are premised on the notion that cases are independent of each other. This is a pretty good assumption if you're doing carefully controlled experiments or you're analyzing a telephone survey. On the other hand, independence is often a horrible assumption. Part of the reason the economy is in trouble is that Wall Street thought it could hedge risk by pooling mortgages. If a mortgage has risk p of default and mortgage defaults are independent then the risk of two mortgages both defaulting is p^2. Except that they aren't independent as when one mortgage defaults it lowers the price which puts other mortgages underwater and greatly increases their risk of default.

Fortunately there has been a lot of work in statistics lately aimed at relaxing this assumption. The most basic is Huber-White or "robust" standard error (which you'll learn about in 210B) but more complex techniques (210C) allow you to build error structure into the model. Random effects(which in stata are part of the "xt" syntax) allow you to identify clusters of observations with shared error. So independence is an important assumption but you can relax it if you

know what you're doing and can identify the structure of the error.

Assymetric Causation

Pretty much all quantitative models are based on comparing some variables and seeing what values on one tend to correspond to those on another. Particularly with regression models, we tend to say things like "gaining a year of education is worth another $1500 of annual income" and (if the model is linear) to treat this as equivalent to saying "forgoing a year of education means losing $1500 in annual income." Under some circumstances this symmetry may be justified, but in other cases it is not.

For instance, cognitive psychology and behavioral economics show that we discount gains relative to equivalent losses ("the endowment effect"). So this implies that there's a difference between earning $50,000 having previously earned $30,000 and earning $50,000 having previously earned $70,000. So some causal processes have what we might call a memory, it matters not just where you are but how you got there.

In other cases, we might imagine a ratchet mechanism, where applying a condition might create a change, but removing the condition will not reverse it. For instance, the state tends to increase during war time but at peace it does not revert back to pre-war levels. Likewise, Weber thought Calvinism created capitalism but the decline of the reform church did not lead to the decline of capitalism (Daniel Bell argued something similar to this, but even he didn't think we'd go back to feudalism).

This is largely an issue of model specification. Is it the thing itself, delta, or cumulative? It depends but by default we tend to assume it is the thing itself.