What Are the Properties of a Good Empirical Study?

- Short answer: The same characteristics that contribute to any high quality work of scholarship.
- Characteristics include carefully stated premises and hypotheses, thinking deeply about one's data or arguments, assessing counter-arguments, not overstating conclusions, and many more.
- It is easier to identify characteristics that frequently are weaknesses in an empirical study.

Five Common Weaknesses in Empirical studies

- Failure to assess power (or type II error) when key results are not significant
- Failure to account for non-independence of observations
- Failure to graph data
- Failure to account for sample design
- Failure to test assumptions on which statistical analyses rely

The Importance of Power: Can be a matter of life and death

- The VIGOR study (New England Journal of Medicine 2000) compared gastrointestinal toxicity of Vioxx (rofecoxib) and naproxen. The study was supported by a grant from Merck, the maker of Vioxx.
- The study also compared the relative safety of the two drugs.
- A troubling result emerged with respect to heart attack risk. "Myocardial infarctions were less common in the naproxen group than in the rofecoxib group (0.1 percent vs.0.4%; 95% confidence interval for the difference, 0.1 to 0.6%; relative risk, 0.2; 95% confidence interval, 0.1 to 0.7)."
- Thus, **the risk of heart attacks on Vioxx was 5 times greater than the risk on naproxen.**

Explaining Away the Vioxx/Heart Attack Effect

- The New England Journal article continues:
- "Four percent of the study subjects met the criteria of the Food and Drug Administration . . . for the use of aspirin for secondary cardiovascular prophylaxis . . . but were not taking low-dose aspirin therapy. These patients accounted for 38 percent of the patients in the study who had myocardial infarctions."
- It is troubling that those at high risk (should have been on aspirin therapy) had a much greater heart attack risk on Vioxx than on naproxen.

The Power-Related Story

- But the power-related story is about the group that should not have been on aspirin therapy. With respect to that group, the article states: "the difference in the rate of myocardial infarction between groups was not significant (0.2 percent in the rofecoxib group and 0.1 percent in the naproxen group)."
- The heart attack risk doubled on Vioxx but was dismissed as statistically insignificant.

- Before viewing the statistically insignificant result as comforting, a power calculation should be done.
- A power calculation assesses the likelihood of a study detecting a statistically significant effect if one exists. The intuition is simple: one needs enough observations to reach a defensible conclusion.

Power in VIGOR re Heart Attacks

- Based on the VIGOR article, one can estimate that there were 3885 lower-risk subjects treated with Vioxx and 3868 were treated with naproxen.
- Assuming that the observed naproxen heart-attack rate (.001) is the baseline with which to compare the Vioxx heart-attack rate, how large a sample does one need to detect a statistically significant doubling of that rate in the Vioxx group?

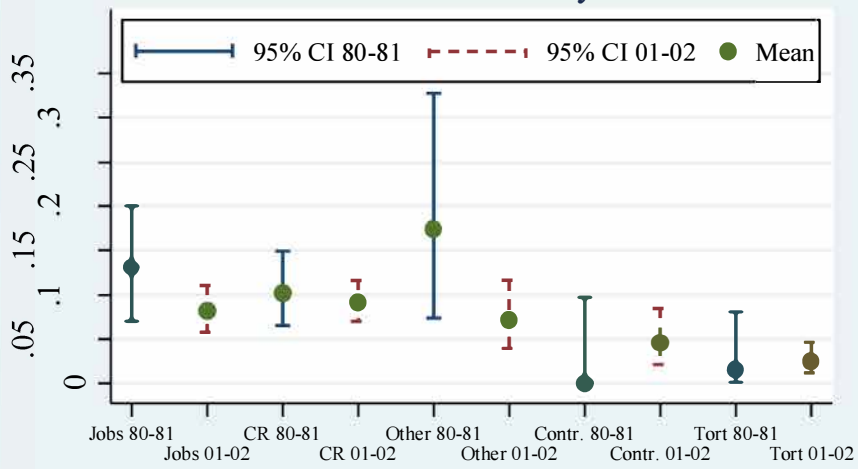
- Estimated sample size for two-sample comparison of proportions
- Test Ho: p1 = p2, where p1 is the proportion in population 1
- and p2 is the proportion in population 2
- Assumptions:
- alpha = 0.0500 (two-sided)
- power = 0.8000
- p1 = 0.0020
- p2 = 0.0010
- n2/n1 = 1.00
- **Estimated required sample sizes:**
- **n1 = 25471**
- **n2 = 25471**
- **The insignificance of the Vioxx-naproxen difference should have provided no comfort. Despite the 5 times elevated rate, Vioxx remained on the market for 4 more years; many people died. (NEJM concern)**

- Thus, to draw reasonable comfort about heart attack safety for Vioxx compared to naproxen, the study would have required over 50,000 subjects, not the 8,000 or so actually used.
- The inference that Vioxx was not more dangerous than naproxen was not scientifically supportable because the question had not been truly explored.

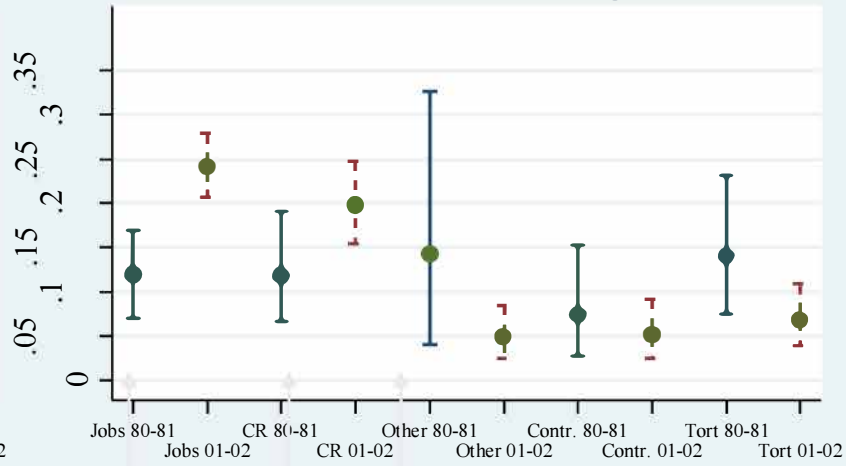
- How does this concern about power relate to our paper on summary judgment?

Figure 1. Summary Judgment Rates, Three Federal Districts, 1975-76, 1980-81 & 2001-02

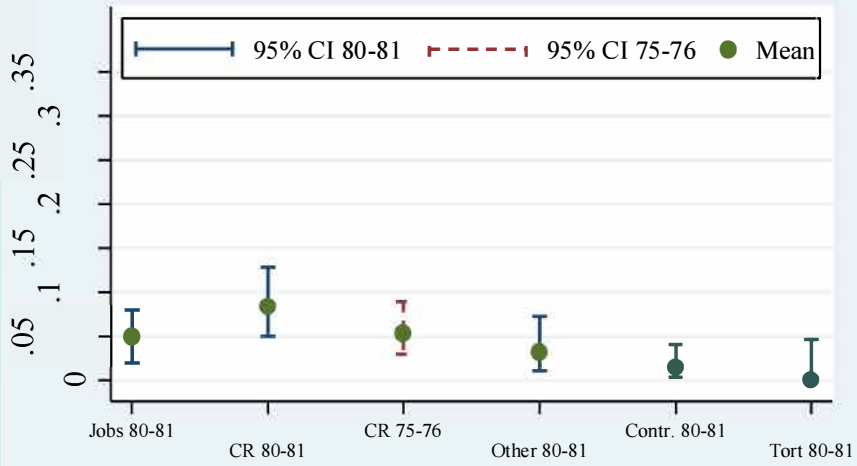
Eastern District Pennsylvania



Northern District Georgia



Central District California



Accounting for Non-independence

- Important scholars published a study of U.S. Senate votes on U.S. Supreme Court nominees
- The data cover about 3700 senators' votes on about 40 nominees (about 100 votes per nominee, though fewer in the earlier years of the study)
- The senators' votes on a nominee are not independent; nominees have characteristics that lead to the votes being associated with one another. For example, consider the "perfect" nominee, who would be expected to receive

Unadjusted Results

	Confirmation vote	
	No	Yes
Republican senator	334 (17.6%)	1559 (82.4%)
Democratic senator	110 (6.1%)	1706 (93.9%)

Chi-squared(1) = 118.08; $p<0.0001$

Adjusted Results

	Confirmation vote	
	No	Yes
Republican senator	334 (17.6%)	1559 (82.4%)
Democratic senator	110 (6.1%)	1706 (93.9%)

$F(1,39) = 4.91$; $p=.0326$

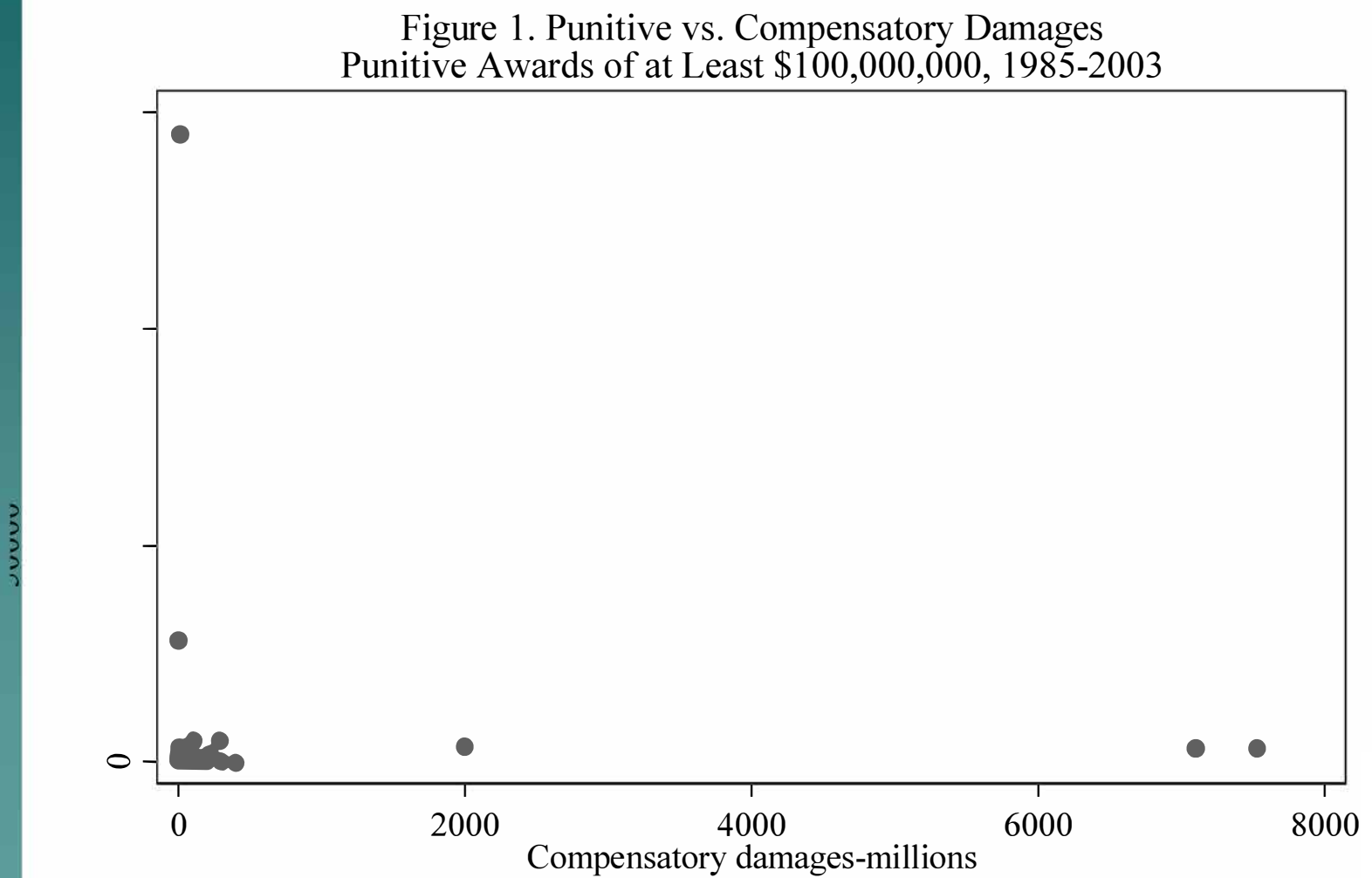
(other adjustments possible)

- How does the concern about non-independence relate to our paper on summary judgment?

Non-independence

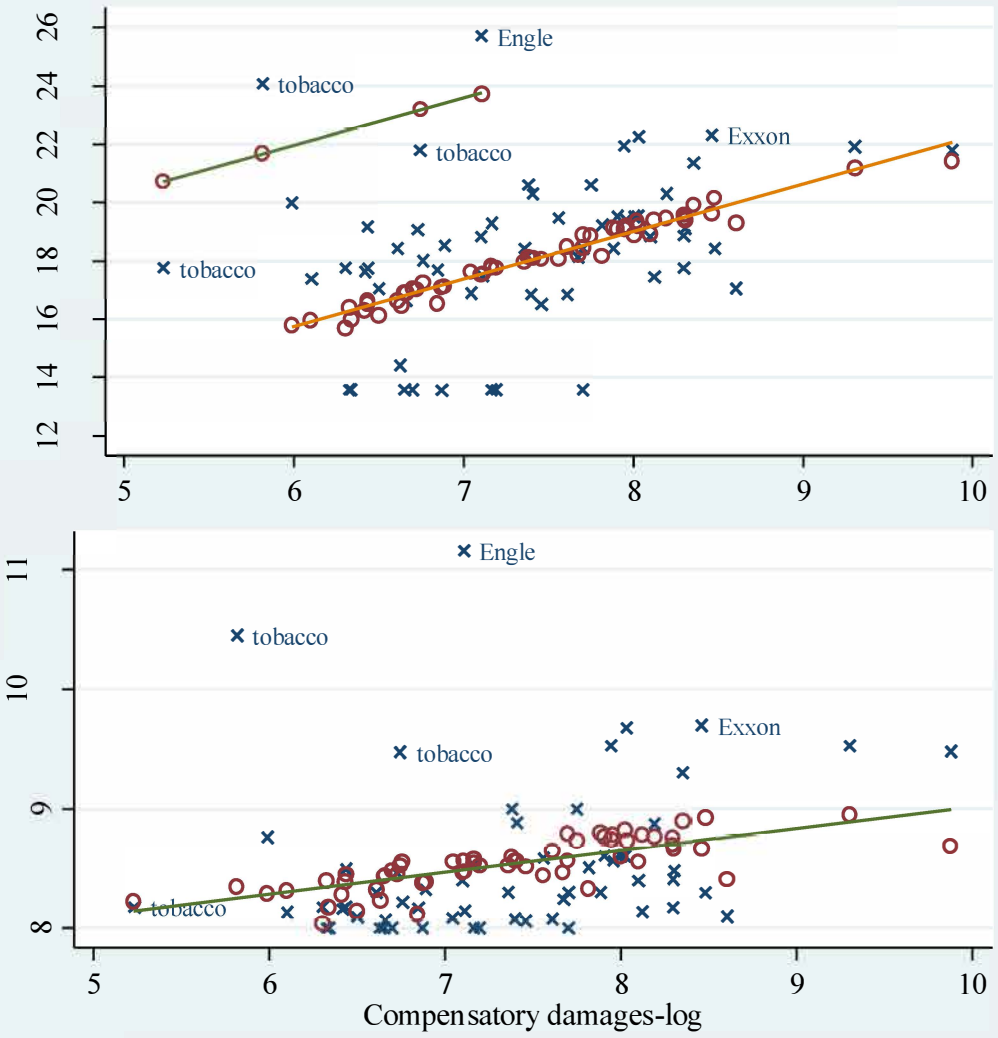
- Same judges decide many cases; judges' decisions are not independent of one another
- Cases are decided within districts, which may have characteristics that render case outcomes not independent of one another
- Other?

Graphing Data: The Relation Between Punitive & Compensatory Awards



Graphing Data: The Relation Between Punitive & Compensatory Awards

Figure 8. Comparison of Fit of Models
Eisenberg-Wells (top figure) & Hersch-Viscusi (bottom figure)

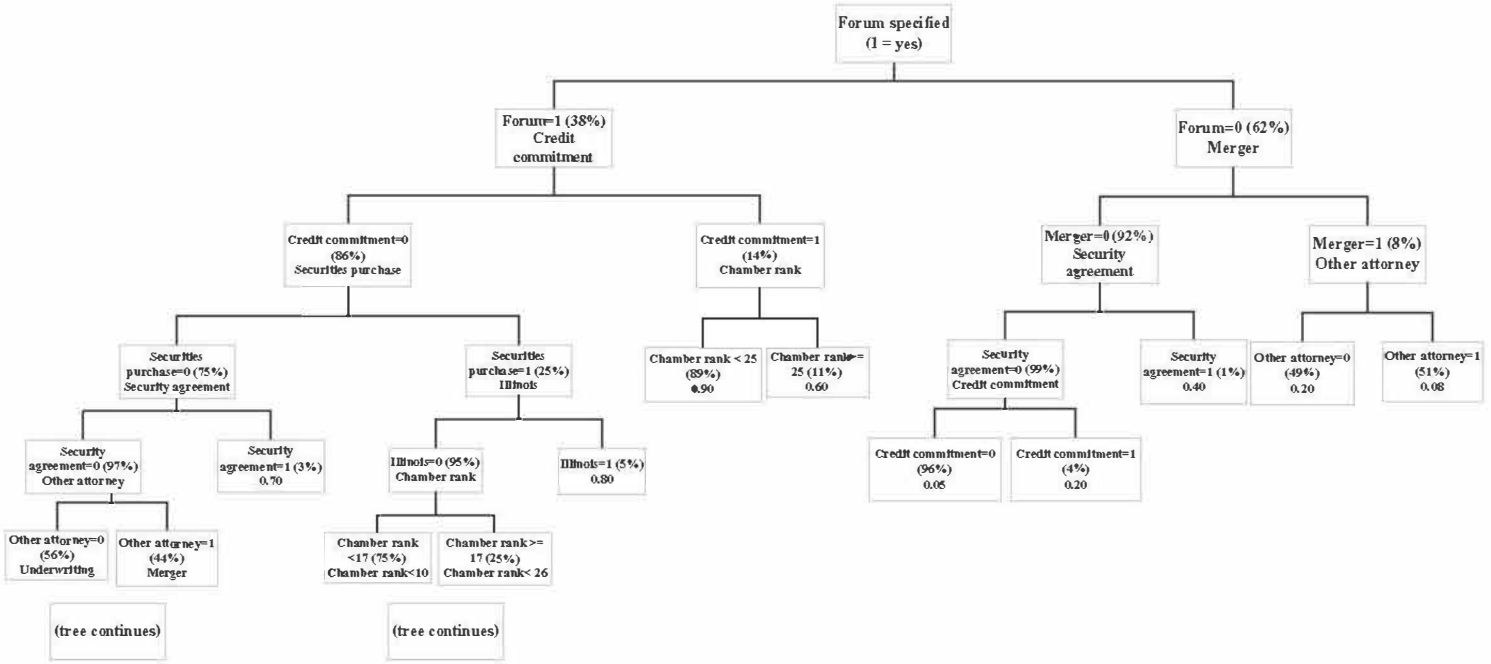


- How does the concern about graphing data relate to our paper on summary judgment?

Graphical representation of qualitative data

- See slide above, use of confidence intervals and side-by-side comparisons to assess rates
- But there is a difficulty in graphically portraying qualitative outcomes
- CART—see, e.g., Eisenberg-Miller JELS (2007) article on jury trial choice

Figure 5: Classification tree for jury trial waiver.



NOTE: This classification tree provides a nonparametric analysis of the relation between hypothesized factors and the dependent variable, jury trial waiver, in 2,749 contracts. The prominence (first node) of the forum-specified variable in this tree supports the importance of the relation between a forum being specified and jury trial being waived. The proportions reported in the terminal nodes at the end of each branch are the proportion of contracts predicted to contain jury trial waivers. For example, of the 62 percent of contracts that did not specify a forum, 8 percent involved mergers, and 51 percent of that 8 percent had attorneys other than from California, Illinois, Massachusetts, New York, Pennsylvania, or Texas: 0.08 of those contracts are expected to have jury trial waivers. Of the 38 percent of contracts that did specify a forum, 14 percent were credit-commitment contracts. The proportion of those contracts that contained jury trial waivers is expected to be associated with the Chamber of Commerce fairness ranking.

Accounting for Sample Design

- Imagine that one does a survey and oversamples a group to assure enough observations
- In the U.S., this is sometimes done with racial minorities
- E.g., sample 400 whites and 400 blacks. Suppose 75% of whites and 25% of blacks report high confidence in police fairness. It one wanted a national estimate, one could not simply combine the two results to yield 50%. The whites in the sample are "representing" many more people than the blacks. A national rate must be computed with the sample design in mind.

- How does the concern about sample design relate to our paper on summary judgment?
- In one or more districts, samples were used rather than full populations. Statistical analysis must account for this design issue.

Testing Assumptions

- Correlation coefficients
- T-tests
- Regression
- Large sample assumptions