Lecture 5: Further Issues in Statistical Inference

Jaakko Meriläinen

5304 Econometrics @ Stockholm School of Economics

Recap

- In Lecture 3, we started discussing statistical inference and hypothesis testing
  - We assumed that $u \sim Normal\left(0, \sigma^2\right)$
  - With this assumption, we could pin down the distribution of OLS estimators (conditional on sample value of regressors)

$$\widehat{\beta}_j \sim Normal\left[\beta_j, Var\left(\widehat{\beta}_j\right)\right]$$

  - In this lecture, we are going to expand this discussion

- Lecture 4 was about potential problems with getting $\hat{\beta}$s right—but this is not where the problems end...

Plan for This Lecture

1. Homoskedasticity vs. heteroskedasticity
2. Clustering
3. Randomization inference
4. Null hypothesis significance testing and misinterpreting p-values

Asymptotic Normality (Wooldridge 2013)

THEOREM 5.2

ASYMPTOTIC NORMALITY OF OLS

Under the Gauss-Markov Assumptions MLR.1 through MLR.5,

(i) $\sqrt{n}(\hat{\beta}_j - \beta_j) \overset{a}{\sim} \text{Normal}(0, \sigma^2/a_j^2)$, where $\sigma^2/a_j^2 > 0$ is the **asymptotic variance** of $\sqrt{n}(\hat{\beta}_j - \beta_j)$; for the slope coefficients, $a_j^2 = \text{plim}\left(n^{-1}\sum_{i=1}^{n} \hat{r}_{ij}^2\right)$, where the $\hat{r}_{ij}$ are the residuals from regressing x_j on the other independent variables. We say that $\hat{\beta}_j$ is *asymptotically normally distributed* (see Appendix C);

(ii) $\hat{\sigma}^2$ is a consistent estimator of $\sigma^2 = \text{Var}(u)$;

(iii) For each j,

$$(\hat{\beta}_j - \beta_j)/\text{sd}(\hat{\beta}_j) \overset{a}{\sim} \text{Normal}(0,1)$$

and

$$(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j) \overset{a}{\sim} \text{Normal}(0,1), \qquad [5.7]$$

where $\text{se}(\hat{\beta}_j)$ is the usual OLS standard error.

Homoskedasticity

$$E[(\boldsymbol{\epsilon}\boldsymbol{\epsilon}')|\boldsymbol{X}] = \begin{bmatrix} \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} = \sigma^2 \boldsymbol{I}$$

- So far we have assumed homoskedasticity
- But in real life, homoskedasticity is often going to be violated
- Instead of a common σ^2, you have σ_i^2

Heteroskedasticity

- Fortunately, a valid estimator of $Var(\widehat{\beta}_j)$ under heteroskedasticity is given by:

$$\widehat{Var}(\widehat{\beta}_j) = \frac{\sum_i^n \widehat{r}_{ij}^2 \widehat{u}_i^2}{SSR_j^2}$$

  where $\widehat{r}_{ij}$ is the i^{th} residual from regressing x_j on all other regressors and SSR_j is the sum of squared residuals from this regression

- The (positive) square root of this term is known as the heteroskedasticity-robust standard error (sometimes called Huber-White or Eicker-Huber-White errors)

- Luckily, once you compute these heteroskedasticity-robust standard errors, hypothesis testing follows as before

- The errors are (asymptotically) t-distributed

Robust Standard Errors

- In Stata, robust standard errors are computed very easily by typing `robust` as an option after your `regress` command
- We should typically use robust standard errors, but not always
- Robust standard errors have asymptotic justification
  - As $n \to \infty$, we can account for heteroskedasticity of unknown form
  - But samples are finite, and when heteroskedasticity is modest, robust SEs can be more biased than standard SEs
- Angrist and Pischke (2009) recommend that you should use the larger of the two standard errors (that is to say, be conservative!)
- Usually, robust standard errors are larger than conventional ones

Haiku

Keisuke Hirano (1998)

T-stat looks too good.

Use robust standard errors--

significance gone.

Clustering of Standard Errors

- A different problem arises when independence across observations breaks down
- In cross-sections, a common issue is the correlation of standard errors between individuals in the same cluster
  - For example, wages in a firm or test scores in a class
  - An easy way to think of the problem is correlated shocks
- Ignoring clustering ("the Moulton problem") where it is important, will typically lead to downwards-biased standard errors
  - When your standard errors are too small, you over-reject the null even with robust standard errors
- With enough clusters (again, asymptotic justification!), you can adjust for clustering using the `cluster` option in Stata

How Should We Cluster Our Standard Errors?

- There are some "rules" regarding when to cluster and at what level
- Cluster if **the sampling is clustered**
  - E.g., you first sample certain schools, then sample students in the school $\Rightarrow$ Cluster SEs at the school level
- Cluster if **the treatment/intervention/policy you study is clustered**
  - E.g., a policy is rolled out at county-level, but you have data on individuals $\Rightarrow$ Cluster SEs at the county level
- Recommendations from a recent paper by Abadie et al. (2017) are very helpfully summarized by David McKenzie in the World Bank's Development Impact blog
- However...

When Should We Not Cluster?

- Clustered standard errors will be biased if there are too few clusters
- Angrist and Pischke (2009) propose 42 clusters as a rule-of-thumb cutoff value
- What should we do if there are not enough clusters? Several options, but which one you should choose depends a bit on the situation...
  - Cluster your standard errors anyway but keep in mind that they may be biased
  - Think of another reasonable level of clustering
  - Use robust standard errors
  - Use bootstrap methods to estimate the standard errors ("wild bootstrap")
  - Collapse your data to a different level and use the adequate standard errors

The Lady Tasting Tea (Chapter II in Fisher's "The Design of Experiments")

A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup.

- The Lady was biologist Muriel Bristol, who worked with Fisher at the Rothamsted Experimental Station in Harpenden, United Kingdom
- H_0: Fisher believes that Dr. Bristol cannot taste the difference
- A test of the hypothesis: Experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgment

Rule 1: Do Not Confound the Treatment

- Critical assumption: if Dr. Bristol is unable to detect whether the milk was poured in first, then she will choose 4 cups at random
  - Fisher points out that the experimenter could mess this up
  - If all those cups made with the milk first had sugar added, while those made with the tea first had none, this might well ensure that all those made with sugar should be classed alike
  - This is sometimes referred to as excludability (or treatment uncorrelated with potential outcomes)

Rule 2: Do Not Accidentally Confound Your Own Treatment

It is not sufficient remedy to insist that "all the cups must be exactly alike" in every respect except that to be tested. For this is a totally impossible requirement.

- To minimize the likelihood of accidentally confounding your treatment, the best approach is to constrain yourself by randomizing
- Randomization minimizes the likelihood of unfortunate coincidences
- Highly controversial position at the time, and it is still debated in some circles
- The alternative is to force balance on observables, and hope that unobservables do not matter

The Lady Tasting Tea: A Hypothesis Test

- How should we interpret data from this experiment?

- Suppose Dr. Bristol correctly identified all 4 "treated" cups
  - How likely is it that this outcome could have occurred by chance?
  - There are $\binom{8}{4} = \frac{8!}{4!4!} = 70$ possible ways to choose 4 cups
  - Only one is correct; a subject with no ability to discriminate between treated and untreated cups would have a $1/70$ chance of success
  - The p-value associated with this outcome is $1/70 \approx 0.014$

The Lady Tasting Tea: A Hypothesis Test

- How should we interpret data from this experiment?
- Suppose Dr. Bristol correctly identified 3 "treated" cups
  - How likely is it that this outcome could have occurred by chance?
  - There are $\binom{4}{3} \times \binom{4}{1} = 16$ possible ways to choose 3 correct cups
  - The p-value associated with this outcome is $16/70 \approx 0.22$
- The only experimental result that would lead to the rejection of the null hypothesis was correct identification of all 4 treated cups!
- In the actual experiment, the null hypothesis was rejected

The Lady Tasting Tea: A Hypothesis Test

- The size of a test is the likelihood of rejecting a true null
- Fisher asserts that tests of size 0.05 are typical
- Alternative experiment: what if we had treated 3 out of 6 cups of tea?
  - There are $\binom{3}{6} = 20$ possible ways to choose 3 of 6 cups
  - Best possible p-value is therefore $\frac{1}{2} = 0.05$ (selecting the one correct combination)
- Alternative experiment: what if we had treated 3 out of 8 cups of tea?
  - There are $\binom{3}{8} = 56$
  - Best possible p-value is therefore $\frac{1}{56} = 0.017$

Randomization Inference (RI)

- Elegant precursor to OLS approach to experiments by Neyman (and later by Fisher) in the 1920s

- By repeating the exact routine by which the original randomization was conducted many times (e.g., 5,000), we can make inference about variance
  - For each placebo treatment assignments calculate the treatment-control difference
  - The distribution gives the variation in the treatment/control difference given the way the experiment was conducted
  - Place observed treatment/control differential into this distribution
  - If observed difference lies in bottom or top 2.5% of the distribution, reject the two-sided null at 5% level

- Purest form of RI: Calculate every permutation of the treatment assignment
  - Confidence statements become a statement of the relative probability of different counterfactual outcomes occurring

- RI allows for calculation of tight error bounds without resorting to asymptotics—works well in small samples—and makes no parametric assumptions about error distributions

Searching for the Stars?

- Null Hypothesis Significance Testing (NHST), and the use of p-values, is ubiquitous in empirical work

- There are good reasons for this
  - We need a measure of uncertainty as a guide for the strength of evidence
  - Statistical theory leads us to the NHST framework
- But it is very easy to misinterpret or over-interpret p-values
  - Especially in dichotomizing evidence based on whether a p-value crosses the relevant threshold for statistical significance
  - This is enough of a problem that some researchers advocate doing away with p-values altogether! (not economists, usually)

- So we will go through a few common issues in interpretation

Statistical Significance $\neq$ Economic Significance

- Statistical significance (embodied in p-values or CIs) only indicates the precision by which a parameter is estimated
  - A p-value is the probability of observing a value as extreme if H_0 is true
- This is **not** the same as economic significance—is the difference big enough for me to care?
- Imagine, for instance, that I estimated that Group A (say, blue-eyed people) earn more than Group B (gray-eyed people)
  - The magnitude of this coefficient is 0.01 % of annual income
  - But because I estimate it on all the population of Scandinavia from 1950-2017, this is estimated very precisely (significant with $p < 0.01$)
  - Do I care more than the difference of 10% of annual income between two groups but estimated less precisely (say with $p = 0.07$)?

The Difference Between Significant and Not Significant May Be Insignificant

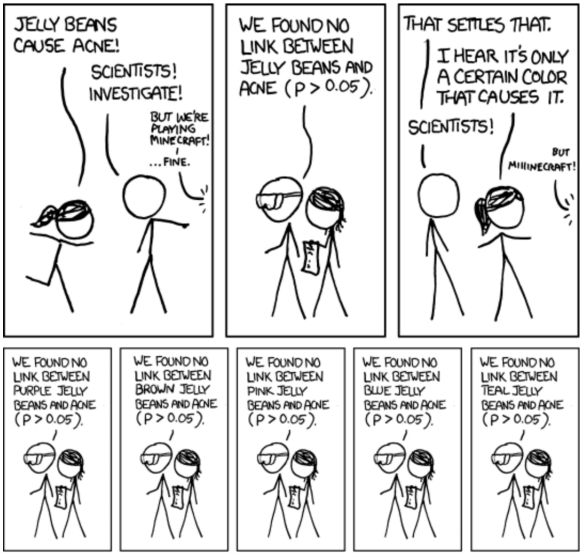
- Suppose we are interested in comparing the effectiveness of two different treatments or interventions:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

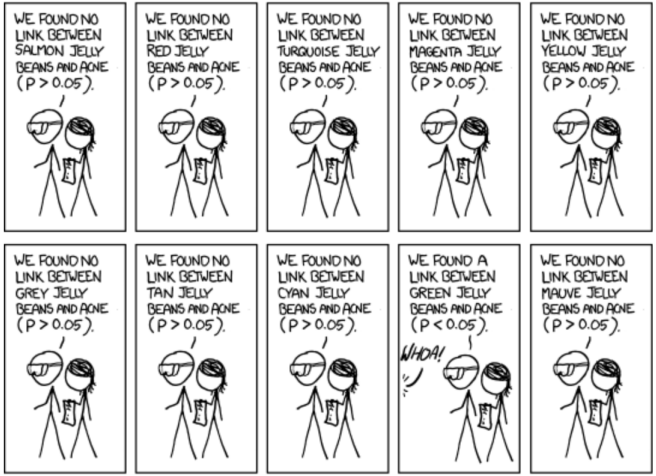
- Examples:
  - Effect of aid grants vs. loans on growth
  - Effect of two different malaria drugs on malaria prevalence

- **The difference between "significant" and "not significant" may itself not be statistically significant**
  - Need to look at the statistical significance the difference between the two effects
  - Even if $\widehat{\beta}_1$ is statistically significant and $\widehat{\beta}_2$ is not, $\widehat{\beta}_1$ and $\widehat{\beta}_2$ might not be significantly different from each other
  - See Gelman and Stern (2006) for details

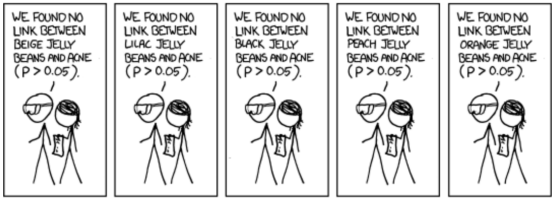
Multiple Hypotheses Testing: A Tale in Three Parts (Source: xkcd.com)

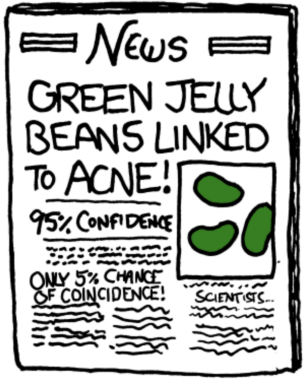


Multiple Hypotheses Testing: A Tale in Three Parts (Source: xkcd.com)



Multiple Hypotheses Testing: A Tale in Three Parts (Source: xkcd.com)





Researcher Degrees of Freedom

The introduction of norms—confidence at 95 percent or 90 percent—and the use of eye-catchers—stars—have led the academic community to accept more easily starry stories with marginally significant coefficients than starless ones with marginally insignificant coefficients.[...]

A consequence of such selection is that researchers may anticipate and consider that it is a stumbling block for their ideas to be considered. For instance, they may censor their papers with too high p-values. They may also search for specifications delivering just-significant results and ignore specifications giving just-insignificant results in order to increase their chances of being published.

Brodeur et al. (2016)

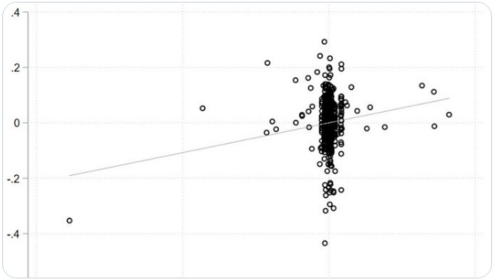
Researcher Degrees of Freedom

Our analysis suggests that the pattern of this misallocation is consistent with what we dubbed an inflation bias: researchers might be tempted to inflate the value of those almost-rejected tests by choosing a slightly more "significant" specification. [...] among the tests that are marginally significant, 10 percent to 20 percent are misreported. On the one hand, our results provide some evidence consistent with the existence of p-hacking thereby justifying the increasing concerns on data replicability and the implementation of pre-analysis plans. On the other hand, [...], the bias remains circumscribed (z-statistics from 1.4 to 2.2).

Brodeur et al. (2016)

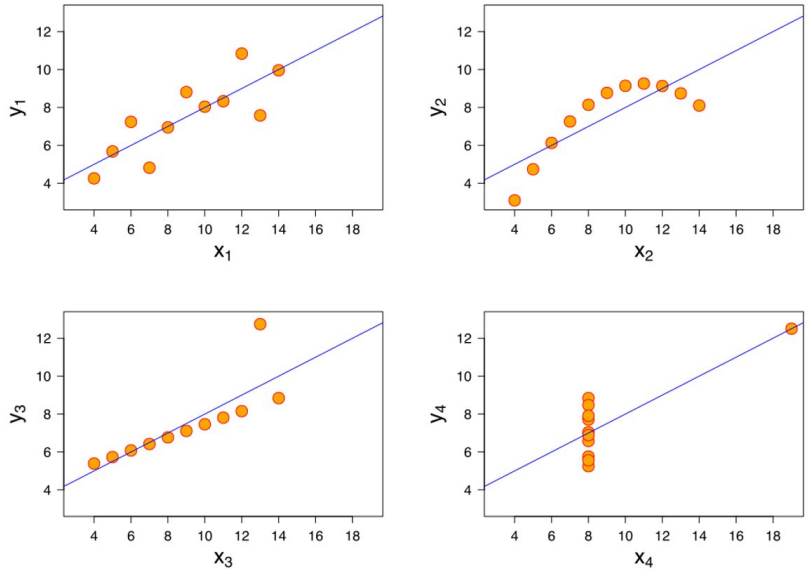
Do Not Stare at Stars Only—Inspect Your Data!

I recently fell in love with the significance stars and spent hours (100+) on a piece of research. Only recently I looked at the plot behind the stars. Remove one observation and it all went away. Down the drain with the entire thing. If I had only looked at the plot to start😩



19 52 345

Anscombe's Quartet



Guidelines from the American Statistical Association

- **P1.** p-values can indicate how incompatible the data are with a specified statistical model
- **P2.** p-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone
- **P3.** Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold
- **P4.** Proper inference requires full reporting and transparency
- **P5.** A p-value, or statistical significance, does not measure the size of an effect or the importance of a result
- **P6.** By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis

McShane and Gal (2017)

- Knowing how to interpret statistical findings and their significance is important...
- ...but also very difficult!
- McShane and Gal (2017) conducted experiments with researchers (statisticians and others) who were presented hypothetical research findings and asked to interpret them
- Consider first their Study 1 where $p = 0.01$ was alternated with $p = 0.27$

The study aimed to test how different interventions might affect terminal cancer patients' survival. Subjects were randomly assigned to one of two groups. Group A was instructed to write daily about positive things they were blessed with while Group B was instructed to write daily about misfortunes that others had to endure. Subjects were then tracked until all had died. Subjects in Group A lived, on average, 8.2 months post-diagnosis whereas subjects in Group B lived, on average, 7.5 months post-diagnosis ($p = 0.01$).

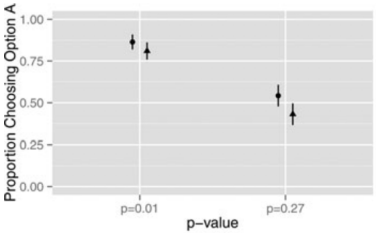
What Is the Correct Interpretation?

Speaking only of the subjects who took part in this particular study, ...

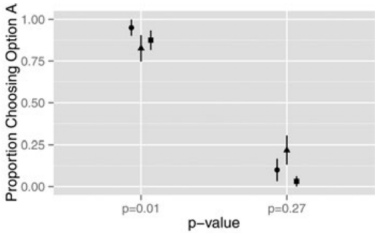
- A. the average number of post-diagnosis months lived by the subjects who were in Group A was greater than that lived by the subjects who were in Group B
- B. the average number of post-diagnosis months lived by the subjects who were in Group A was less than that lived by the subjects who were in Group B
- C. the average number of post-diagnosis months lived by the subjects who were in Group A was no different than that lived by the subjects who were in Group B
- D. it cannot be determined whether the average number of post-diagnosis months lived by the subjects who were in Group A was greater/no different/less than that lived by the subjects who were in Group B

Do you know what is the correct answer?

Interpretations of Statisticians and Medical Doctors



(a) *JASA*



(b) *NEJM*

Study 2 in McShane and Gal (2017)

- Does the pattern of results observed in Study 1 extend from the interpretation of data to likelihood judgments (i.e., predictions) and decisions (i.e., choices) made based on data?

- How does varying the degree to which the *p*-value is above the threshold for statistical significance affect likelihood judgments and decisions?

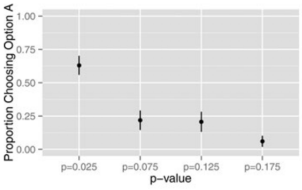
The study aimed to test how two different drugs impact whether a patient recovers from a certain disease. Subjects were randomly drawn from a fixed population and then randomly assigned to Drug A or Drug B. Fifty-two percent (52%) of subjects who took Drug A recovered from the disease while forty-four percent (44%) of subjects who took Drug B recovered from the disease. A test of the null hypothesis that there is no difference between Drug A and Drug B in terms of probability of recovery from the disease yields a p-value of 0.025.

Interpreting Study 2

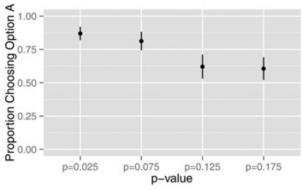
Assuming no prior studies have been conducted with these drugs, which of the following statements is most accurate? A person drawn randomly from the same population as the subjects in the study is...

- A. more likely to recover from the disease if given Drug A than if given Drug B
- B. less likely to recover from the disease if given Drug A than if given Drug B
- C. equally likely to recover from the disease if given Drug A than if given Drug B
- D. It cannot be determined whether a person drawn randomly from the same population as the subjects in the study is more/less/equally likely to recover from the disease if given Drug A or if given Drug B

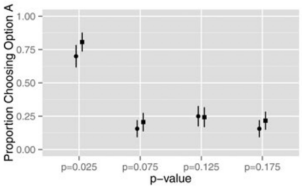
Judgements and Choices of Statisticians and Epidemiologists



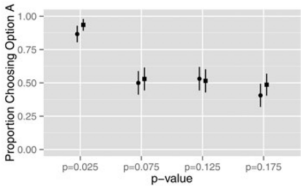
(a) *JASA*: Judgment



(b) *JASA*: Choice



(c) *AJE*: Judgment



(d) *AJE*: Choice

Readings

- **Randomization inference:** Section 4.2 in Cunningham (2021)
- **Clustering:** Read the blog post by David McKenzie summarizing a paper on clustering by Athey, Abadie, Imbens and Wooldridge (2017) available at `https://blogs.worldbank.org/impactevaluations/when-should-you-cluster-standard-errors-new-wisdom-econometrics-oracle`
- McShane, B. B., & Gal, D. (2017). Statistical significance and the dichotomization of evidence. Journal of the American Statistical Association, 112(519), 885-895.

Optional:

- Brodeur, A., Le, M., Sangnier, M., & Zylberberg, Y. (2016). Star wars: The empirics strike back. American Economic Journal: Applied Economics, 8(1), 1-32.
- Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. The American Statistician, 60(4), 328-331.