Applied Microeconometrics – Fall 2022 – Helsinki School of Economics

Problem set 1

Some suggested answers

The problem set is inspired by the following two papers:

- Gelman, Andrew and David Weakliem (2009), "Of beauty, sex, and power", American Scientist 97(4), 310-316 .
- Gelman, Andrew and Eric Loken (2013), "The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time", mimeo.

A. The garden of forking paths

A.1. Chocolate helps to lose weight!

Please read the following article by John Bohannon:

"I Fooled Millions Into Thinking Chocolate Helps Weight Loss. Here's How."
which discusses the following "academic" paper:

Bohannon, Johannes, Diana Koch, Peter Homm and Alexander Driehaus (2015) "Chocolate with high cocoa content as a weight-loss accelerator," International Archives of Medicine, Vol. 8(55).

A.1.1. Discuss the potential existence of a problem of multiple-testing in Bohannon et al. (2015)

The author considered purposefully 18 different outcome variables, anticipating that due to random sampling, each regression has a 5% probability of generating a false positive. Assuming independence between the different outcome variables, the possibility of obtaining at least one false positive was around 60% (1-(0.95)^18).

A.1.2. Propose a method to deal with multiple-testing in this context

First, it would have helped to pre-register the hypotheses that were going to be tested. Second, the author should have adjusted the standard errors

appropriately. The simplest way to do it would be using Bonferroni (e.g., multiplying p-values by the number of tests). However, given that different outcome variables are not independent, Bonferroni is generally too conservative. Instead, it would be more appropriate to use some of the methods that have been proposed in the literature, such us Anderson (2008) or the procedure proposed by procedure set out by John List, Azeem Shaikh and Yang Xu (2016, command mhtexp in stata). A useful discussion is provided in this blog: https://blogs.worldbank.org/impactevaluations/overview-multiple-hypothesis-testing-commands-stata

Alternatively, he could have created a single index that summarizes all the different outcome variables (e.g., the sum of the standardized individual measures).
Finally, the results would be more convincing if they were replicated by a new independent study.

A.2.3. Does the small sample size of the experiment increase the probability of a false positive?

The probability of a false positive (type I error) is set by the researcher and does not in itself vary with sample size.

A.2.4. The authors use a very small sample size but they find significant estimates of very small magnitude. Does it sound plausible?

It sounds "fishy". A small sample size implies that the standard errors will be large, and it should not possible to estimate effects of a very small magnitude. I would suspect that there is some error in the calculations.

A.2.5. Despite the questionable quality of the paper it was accepted for publication. Why?

The paper was published in a journal that has been classified by the librarian Jeff Beall as predatory. These journals are typically run by individuals who do not belong to the academic community, and who are willing to publish any article, without conducting any scientific evaluation of its content, in exchange for a payment.

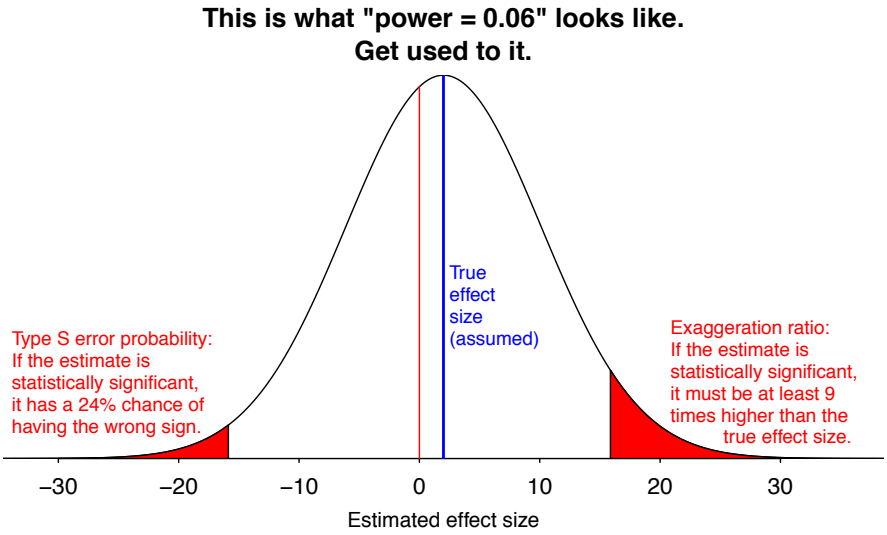
B. Of beauty, sex, and power

B.1. Maternal stress and gender ratio

B.1.1. Imagine that you are asked to conduct a study about the impact of maternal stress on the gender of children. Due to budget constraints, you would be able to measure the stress level of 338 women who are trying to conceive, and you expect approximately 130 of them to give birth during the period of study.

Discuss whether it is a good idea to conduct such study. In your discussion please provide a quantitative estimate of the power of the study (how likely you are to find any significant results, given the sample size and some reasonable assumption about the expected magnitude of the effect of stress), and discuss verbally the potential existence of a type M (magnitude) error and a type S (sign) error.

The sample size is expected to be around 130. Let us also consider for simplicity that we will classify women in two equally sized groups: women above and below the median level of stress. The standard error for the difference in the frequency of boys between the two groups is equal to 8.8 ((1/130)^.5).
Based on the existing literature, let us consider that, if there is an effect, its magnitude might be at most around 2 percentage points (other assumptions are equally valid)
A simple power calculation indicates that the authors only have a 5.6% probability of detecting such an effect (`power twoproportions 0.51 0.49, n(130)').
And if the authors find a significant difference, the magnitude error will be large.
*If we use a 5% significance level, they can only detect differences above or below 17.6 (=2*8.8). If the estimate is statistically significant, it must be at least 9 times higher than any plausible true effect size. Moreover, it has a non-negligible probability of having the wrong sign (in this case approximately 25%).*
In order to provide any informative results, the study should have a much larger sample size. For instance, if the goal is to detect a 0.01 effect with a 80% probability, a conservative sample size would be around 235,000, assuming a one third probability of giving birth during the study period (n =78,400 = [2.8/(p1-p2)]^2=[2.8/0.01]^2))



B.1.2. Please read the following press article:
"Stressed women more likely to have baby girls"
Available at http://www.telegraph.co.uk/news/health/news/8830036/Stressed-women-more-likely-to-have-baby-girls.html
which reports on the findings the following academic paper:

Chason et al. (2012), "Preconception stress and the secondary sex ratio: a prospective cohort study", Fertility and Sterility Vol. 98, No. 4, 937-941.

Imagine that the newspaper contacts you before publishing the article and requests your expert opinion. Write a short letter explaining the journalist how should we think about the findings of this scientific article

Here you were expected to explain in a non-technical way that (i) the sample size is too small to be informative, (ii) there might be a multiple testing problem (the authors make only one of several possible comparisons: they compare the top and the bottom quartiles) and, even if we give face value to the findings, (iii) readers should keep in mind that the observed correlation may not be causal.

B.2. The effect of minimum wage on employment

Card and Krueger (1995)[1] perform a meta-analysis of published studies on the effect of the minimum wage on employment. The following graph describes the relationship between the estimates found in these studies (i.e. *absolute value* of the elasticity of substitution between minimum wage and employment) and the accuracy of these estimates (i.e., standard errors).
The graph displays two interesting patterns: (i) point estimates tend to be twice as large as the standard error and (ii) more precise estimates (typically due to larger sample size) tend to yield lower point estimates.

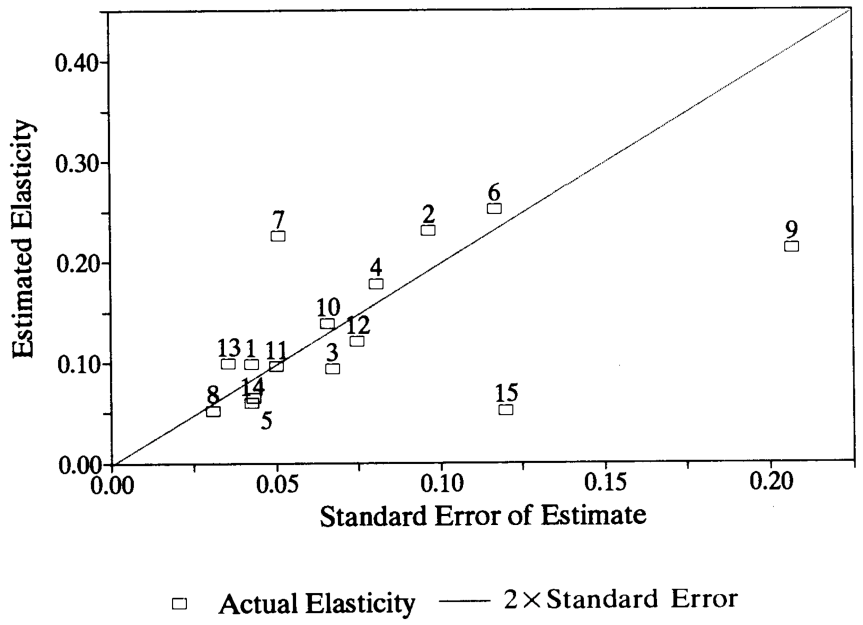


FIGURE 2. RELATION OF ESTIMATED EMPLOYMENT ELASTICITY TO STANDARD ERROR

B.2.1. Please, provide an explanation for why we observe these two patterns.

A likely explanation is publication (or author) bias. In order to be statistically significant at the 5% level the point estimate should be approximately twice as large as the standard error. Interestingly, note that the two outliers correspond to papers that have not been published in academic journals.
The fact that more precise estimates tend to find smaller magnitudes is consistent with the true effect being relatively small.

B.2.2. Based on this graph (assuming that we give face value to the empirical

strategies of these papers), what would be your educated guess about the relationship between minimum wages and employment in terms of its magnitude? (is it around 5%, 10%, 15%, 20%, 25%?) Justify your answer.

If we discount for the possibility that publication bias affects the possibility that studies with small samples are only published in the magnitude of the point estimate is large enough, we may want to give more credibility to the findings of results with relatively larger sample size (and smaller standard errors), and conclude that the effect cannot be larger than 5-10%.

---

[1] Card, David and Alan B. Krueger 1995, "Time-Series Minimum-Wage Studies: A Meta-analysis" The American Economic Review Papers and Proceedings, Vol. 85(2), pp. 238-243.

C. Multiple choice questions

1. Assuming that the treatment has no effect, the probability of a false positive tends to:
    a. increase with sample size
    b. decrease with sample size
    c. it is unaffected by sample size

2. The power (1-beta) tends to:
    a. increase with sample size
    b. decrease with sample size
    c. it is unaffected by sample size
    d. it is unaffected by sample size

3. Assuming that the treatment has an effect, the probability of obtaining a statistically significant result in general tends to:
    a. increase with sample size
    b. decrease with sample size
    c. it is unaffected by sample size

4. The magnitude of a statistically significant coefficient tends to
    a. increase with sample size
    b. decrease with sample size
    c. it is unaffected by sample size

5. The probability of a false positive tends to
    a. increase as the number of potential independent variables increases
    b. decrease as the number of potential independent variables increases
    c. it is unaffected by the number of potential independent variables

6. Conditional on obtaining a statistically significant result, the probability that the magnitude of this estimate is too large:
    a. increases with sample size
    b. decreases with sample size
    c. it is unaffected by sample size

7. Conditional on obtaining a statistically significant result, the probability that this estimate has the `wrong' sign:
    a. increases with sample size
    b. decreases with sample size
    c. it is unaffected by sample size

D. Hepatitis B and the Case of the Missing Women

In the paper `Hepatitis B and the Case of the Missing Women', Emily Oster presents evidence that, she argues, would be consistent with "an existing scientific literature, that carriers of the hepatitis B virus have offspring sex ratios around 1.50 boys for each girl."

The following table provides information on these studies:

TABLE 3
HEPATITIS B AND SEX RATIO: INDIVIDUAL-LEVEL ESTIMATES

Location and HBV Status	Sons	Daughters	Sex Ratio
Greenland:			
Positive	64	60	1.07
Negative	174	194	.90
Kar Kar Island:			
Positive	63	54	1.17
Negative	163	206	.79
Greece 1:			
Positive	90	51	1.77
Negative	287	255	1.13
Philippines:			
Positive	66	41	1.61
Negative	304	301	1.01
Greece 2:			
Positive	52	30	1.73
Negative	1,006	955	1.05
France:			
Positive	20	12	1.66
Negative	149	122	1.22

SOURCE.—Greenland: Drew (1986); Kar Kar Island: Drew et al. (1982); Greece 1: Hesser et al. (1975); Philippines: Chahnazarian et al. (1988); Greece 2: Livadas et al. (1979); France: Cazal et al. (1976).

NOTE.—This table shows sex ratios among the children of carrier and noncarrier parents in four regions. Data were collected by testing married women and, in all cases except for Greenland, their husbands for HBV. Detailed reproductive histories were also collected. The table represents all births to women in these samples, with generally more than one birth to each woman. The last two studies (Greece 2 and France) were designed specifically to test the hypothesis that HBV affects the offspring sex ratio and were run after the original theory was published.

1. First, let us consider one of these studies: "Greece 2". It includes information on the gender of 82 children from HBV positive individuals and 1961 children from HBV negative individuals. The authors compare the share of daughters in each group. Calculate the power of this study assuming that, if HBV affects the probability of having a daughter, it would decrease the

share of daughters by three percentage points (e.g. from 0.488 in the HBV negative group to 0.458 in the HBV positive group). (hint: you may use in Stata the command *power twoproportions*) Or in other words, how likely were the authors to obtain a significant result, assuming that HBV decreases the probability of having a daughter by 3 p.p.?

Using Stata we see that the power of the study is a measly 8%.

2. If you conduct a study with this sample size, what would be approximately the size of your standard errors? (please apply the formula from the slides, lecture 2, slide 27)

Calculating the standard errors using sqrt((0.488^2/1961) + (0.458^2/82)) we find that they will be approximately 5.2 percentage points.

3. Imagine that you had to design the study. What would be the sample size required in order to be able to detect with a 80% probability an effect of magnitude 3%, assuming that the size of the HBV group is expected to be 7 times smaller? (hint: you may use in Stata the command *power twoproportions,* with the options power(.) and nratio(.))

Calculating the required power in tells us that for a study with 80% power and 3% estimated true effect we would require almost 20,000 observations with 2,500 being affected by the disease.

4. Let us now move to the results of this study. The share of daughters is equal to 0.366 (30/82) in the HBV positive group and 0.487 (955/1961) in the HBV negative group. Is this difference statistically significant? At which level? (you can use for instance the Stata command *prtesti*)

Yes, the difference is statistically significant at 5% level (p-value is 0.032). However, note that the effect is implausible large (12% > 3%).

5. Next, let us consider the six studies reported in Table 3 jointly. Overall, they include information on the gender of 603 children from HBV positive individuals and 4116 children from HBV negative individuals. Given this sample size, calculate the power of this (six-sample) study assuming that, if HBV affects the probability of having a daughter, it would decrease the share of daughters by three percentage points, and that α is equal to 0.05.

Now the power would be equal to 28%.

6. Let us now consider jointly the results of these six studies. The share of daughters is equal to 0.41 (248/603) in the HBV positive group and 0.49 (2033/4116) in the HBV negative group. Is this difference statistically significant? At which level?

The difference is significant at 0.01% level.

7. Taking α =0.05 and 1-β from answer 5, please calculate the post-study probability (PSP). For instance, you can consider 10% as your prior (π) for the possibility that the HBV in fact decreases the share of daughters by 3 p.p.
   (note: do not be surprised if your PSP happens to be quite large)

   The probability of a true positive is 1-β * π = 0.28*0.10=0.028
   The probability of a false positive is α * (1- π) = 0.05 * 0.90= 0.045
   Therefore the post-study probability is equal to =0.028/(0.028+0.045)=38%

In a more recent study, Lin and Luoh (2008) use a large dataset from Taiwan and they find that, among first borns from a HBV positive mother, the share of daughters is equal to 0.48288 (N= 122,561) and, within the group of HBV negative mothers, the share of daughters is equal to 0.4856 (N=598,629).

8. Calculate the power of this study assuming again a potential effect of 3 p.p.

The power is almost equal to 1 (100.00%).

9. Let us now look at the results. Is the share of daughters in each group significantly different? What is the maximum difference that we can reject?

The difference between shares is significantly different only at the 10% level (p value is 0.08). The maximum difference we can reject is 0.58%.

10. If we give face value to the findings of Lin and Luoh (2008), how would you explain the evidence provided by the six studies reported by Oster (2005), which tend to find a large significant correlation between HBV and the gender of children? Should Emily Oster have realized that the previous evidence was not reliable? How?

There are several red flags that should have raised suspicion. First, according to the cited evidence, the effect of HBV would be dramatically larger than the effect of any other factor that has been documented before in the medical literature. Second, the six studies that Oster considered rely on very small samples and, therefore, have limited power. Therefore, even if the six studies taken together seem to lead to a high PSP, the lack of any evidence that relies on larger samples is worrying. The author should have realized that the lack of studies with small samples and non-significant results is likely to reflect the impact of publication bias. Finally, it would have been useful to explore some of the testable implications of the different theories (e.g., according to Oster's theory birth order should not be correlated with gender.)