University of Zurich **Spring term 2025**

Department of Economics

Econometrics for Research Students II

Problem Set #6

Proposed Solutions - Extended Version

Question 1

The following questions relate to power calculations for an experiment. Let $\bar{D} = \Pr(D = 1)$ be the share of people randomized into getting treatment. You want to perform a "no-effect" test on δ in the regression $Y_i = \alpha + \delta D_i + u_i$.

a) What is the $\bar{D}$ that maximizes power, for any N, α and σ^2?

Answer a)

To answer this question, we begin by clarifying what is meant by a "no-effect" test. In the context of the proposed regression model, this corresponds to testing $H_0 : \delta = 0$ against the alternative $H_1 : \delta \neq 0$. That is, we are interested in determining whether the treatment assignment D_i - equal to 1 for treated individuals and 0 for those in the control group - has any effect on the outcome Y_i. A "no-effect" test is thus a test of whether the estimated treatment effect δ is statistically distinguishable from zero.

Refresher

Recap on Power

Statistical power refers to the probability that a test correctly rejects a false null hypothesis. In other words, it is the likelihood of detecting a real effect when one exists. Power is formally defined as $1-\beta$, where β is the probability of a Type II error - failing to reject the null hypothesis even though it is false.

To understand power, consider the four possible outcomes of a hypothesis test:

Outcome of Test	**True State: H_0 is True**	**True State: H_0 is False**
Fail to Reject H_0	Correct Decision (No Effect)	Type II Error (β)
Reject H_0	Type I Error (α)	Correct Decision (Power $= 1 - \beta$)

Power is important because it reflects a test's ability to detect true effects and avoid false negatives. A test with low power may lead researchers to incorrectly conclude that

there is no effect, potentially missing meaningful findings. Ensuring adequate power is therefore a critical step in designing empirical research.

If this is not clear let me clarify using a graph (easily created in Python, I'll post the code):

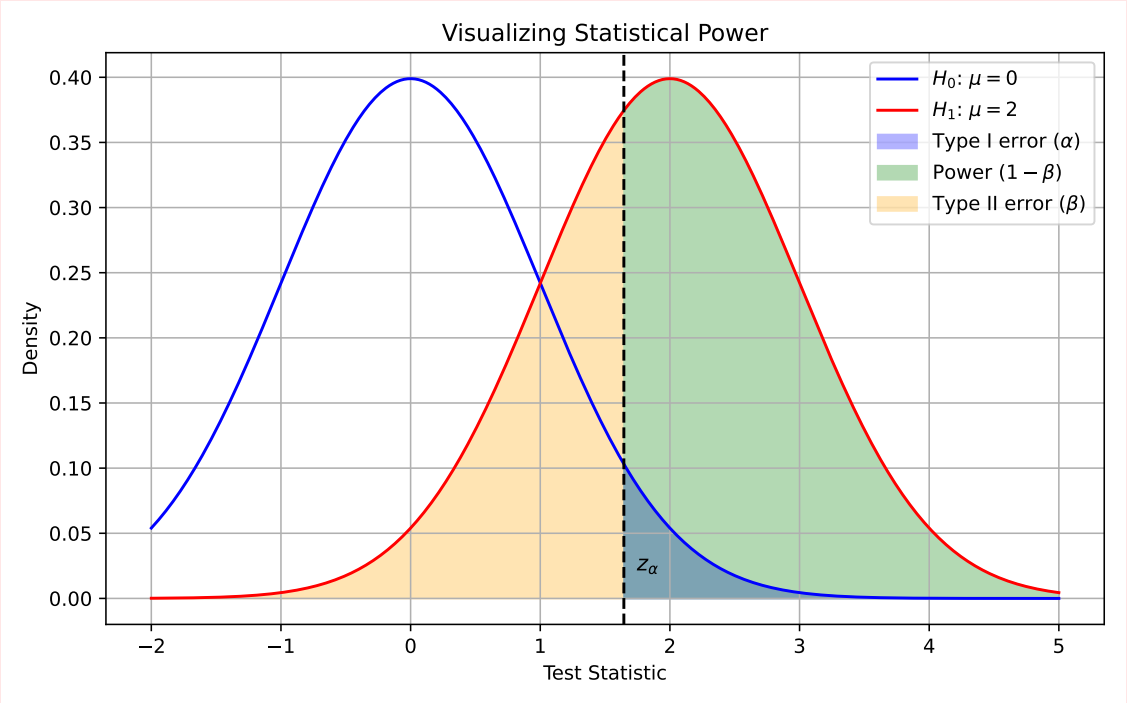

Figure 1: Visualizing Power

The graph illustrates the concept of statistical power using a practical example. Suppose a researcher is testing whether a new teaching assistant (TA) method improves student test scores. The null hypothesis (H_0) is that the method has no effect, while the alternative hypothesis (H_1) is that it leads to higher scores.

The blue curve represents the distribution of test statistics under the null hypothesis - that is, assuming the TA method does not actually improve scores. The red curve represents the distribution under the alternative hypothesis - if the method truly has an effect and raises scores, say by 2 points on average.

The vertical dashed line is the critical value determined by the significance level $\alpha = 0.05$ - i.e. 1.64. This threshold marks the boundary above which the test would reject the null hypothesis. If the test statistic falls to the right of this line, the result is considered statistically significant (standard up to now).

Start with the blue curve, which shows the distribution of test statistics assuming the null hypothesis is true - that is, the TA method has no effect. The area under this curve to the right of the vertical dashed line is shaded blue. This is the Type I error region,

also called α. It represents the probability of rejecting the null hypothesis even though it's actually true. In our context, this would mean concluding that the new TA method improves test scores, when in reality, it doesn't.
Next, consider the red curve, which represents the distribution of test statistics when the alternative hypothesis is true - that is, when the TA method really does improve scores. The area under the red curve to the right of the dashed line is shaded green. This area is called the power of the test. It tells us the probability of correctly rejecting the null hypothesis - in other words, detecting the effect when it's really there. In our example, the researcher has designed the study to have 80% power, which means there's an 80% chance the test will successfully detect the improvement if the TA method is truly effective.
But not all outcomes are ideal. There's still a chance the test misses the effect. That's shown by the orange shaded area under the red curve to the left of the dashed line. This is the Type II error region, or β. It represents the probability of failing to reject the null hypothesis when it is actually false - meaning the TA method does help, but the test doesn't pick it up. In this case, with 80% power, that probability is 20
So in short: the green region is what we aim for - catching real effects. The orange region is the risk of missing them. And the blue region is the chance of being misled into thinking there's an effect when there isn't.
This visual helps clarify the trade-offs in hypothesis testing. The larger the green region - the power - the more capable the test is of identifying real effects. That's why power analysis is crucial in designing experiments: it ensures that meaningful differences don't go undetected due to insufficient sample size or weak statistical design.

Power depends on several key parameters: the sample size N, the significance level α, the variance of the outcome variable σ^2, and the proportion of treated units $\bar{D} = Pr(D = 1)$. In this case, we're asked to focus on the comparative statics for $\bar{D}$, but I'm going to provide something more also for the other components.

Consider the standard linear regression model used to estimate the treatment effect:

$$Y_i = \alpha + \delta D_i + u_i \tag{1}$$

where D_i is a binary indicator for treatment. Assuming homoskedastic errors and random assignment, the sampling variance of the OLS estimator $\hat{\delta}$ is (nothing new here):

$$\text{Var}(\hat{\delta}) = \frac{\sigma^2}{N \cdot \text{Var}(D_i)} \tag{2}$$

Because D_i is Bernoulli, $\text{Var}(D_i) = \bar{D}(1 - \bar{D})$, so:

$$\text{Var}(\hat{\delta}) = \frac{\sigma^2}{N \cdot \bar{D}(1 - \bar{D})} \tag{3}$$

Taking the square root gives the standard error:

$$se(\hat{\delta}) = \sqrt{\frac{\sigma^2}{N \cdot \bar{D}(1 - \bar{D})}} \tag{4}$$

The reason why we compute $se(\hat{\delta})$ is that it's pretty well connected with the Minimum Detectable Effect (MDE). It represents the smallest true effect size that a study is likely to detect with a given level of statistical confidence. It depends on the significance level α, the desired power q, the variability in the data, and the sample size. Intuitively, the MDE sets the threshold above which a true treatment effect will be detected with high probability. If the real effect of the treatment is smaller than the MDE, the test may fail to reject the null hypothesis even though the treatment does in fact have some impact. In contrast, if the true effect is equal to or larger than the MDE, the test will likely detect it.

This concept is central when designing experiments or evaluating their credibility. A study with a large MDE is only capable of identifying relatively large effects and will likely miss more subtle but potentially meaningful differences. On the other hand, a smaller MDE indicates that the test is sensitive enough to detect even modest effects, provided that the data is not too noisy and the sample size is sufficient.

The MDE is formally defined as the product of two components: the critical values from the standard normal distribution (reflecting the chosen significance level and power) and the standard error of the treatment effect estimator. Specifically, it is given by:

$$\text{MDE} = \left(z_{1-\alpha/2} + z_q\right) \cdot se(\hat{\delta}) \tag{5}$$

where $z_{1-\alpha/2}$ corresponds to the two-sided critical value for the significance level α, and z_q is the critical value ensuring power q. This expression reflects the fact that in order to reject the null hypothesis with high probability under the alternative, the test statistic must fall far enough away from zero. The larger the standard error - due to limited data or high noise - the larger the MDE will be, meaning that only larger true effects can be reliably detected.

Therefore, this equation directly links the MDE to the concept of statistical power. For any fixed α and N, the MDE tells us the smallest true effect size δ that the test can detect with power q - that is, the smallest effect for which the probability of correctly rejecting the null hypothesis reaches the desired power level. Inverting this logic: if the true effect is smaller than the MDE, the test will have less than q power to detect it. In this sense, the MDE defines the threshold between detectable and undetectable effects for a given design, making it a critical diagnostic for whether your study is capable of detecting meaningful impacts.

Practically, the idea is that when you're planning an experiment, the MDE tells you the smallest effect your design can realistically hope to detect with confidence. If this MDE is larger than the effect size you actually care about, the study will be underpowered, and even real effects may go unnoticed.

In short, the MDE captures the smallest effect size that your design can "see" with high confidence. Any effect smaller than this threshold might be real, but your test is not powerful enough to detect it consistently.

Mathematically, maximizing power is equivalent to minimizing this MDE, which in turn

requires maximizing $\bar{D}(1-\bar{D})$. If you want, you can do the calculations, but to save time you can see that because this is a quadratic expression achieves its maximum at $\bar{D} = 0.5$.

In the case you, fairly, don't believe me then:

$$\frac{\partial \bar{D}(1-\bar{D})}{\partial \bar{D}} = 0 \Rightarrow \frac{\partial \bar{D} - \bar{D}^2}{\partial \bar{D}} = 1 - 2\bar{D} = 0 \Rightarrow \bar{D} = \frac{1}{2} \tag{6}$$

To conclude, the value of $\bar{D}$ that maximizes the power of the test - regardless of N, α, or σ^2 - is:

$$\bar{D} = 0.5 \tag{7}$$

This corresponds to assigning half the sample to treatment and half to control.

Let me conclude with this excursus on the promised comparative statics of power and N, α and σ^2.

From the MDE equation (down below again for clarity):

$$\text{MDE} = \left(z_{1-\alpha/2} + z_q\right) \cdot \sqrt{\frac{\sigma^2}{N \cdot \bar{D}(1-\bar{D})}} \tag{8}$$

Since power increases when the MDE decreases, we can now interpret the effect of changing each variable by analyzing how it affects the MDE.

Let's start with the sample size N. A larger N reduces the standard error and therefore lowers the MDE. Mathematically, we can write:

$$\text{MDE} \propto \frac{1}{\sqrt{N}} \tag{9}$$

This means that increasing the sample size improves the precision of the estimator and makes it easier to detect smaller effects. Doubling the sample size, for instance, reduces the MDE by a factor of $\sqrt{2}$. So, holding everything else fixed, increasing N leads to higher power. Intuitively, more data helps sharpen the signal and reduces uncertainty in estimating δ.

Now consider the significance level α. The MDE increases with the critical value $z_{1-\alpha/2}$:

$$\text{MDE} \propto z_{1-\alpha/2} \tag{10}$$

Lowering α - for example, from 0.05 to 0.01 - increases the required critical value, which in turn raises the MDE. This makes it harder to reject the null hypothesis, and therefore reduces power. In contrast, choosing a more lenient α decreases the MDE and increases power. This illustrates a fundamental trade-off in hypothesis testing: a stricter test reduces the chance of false positives (Type I error) but increases the chance of false negatives (Type II error).

Finally, think about the variance of the outcome variable, σ^2. The MDE increases with the square root of σ^2:

$$\text{MDE} \propto \sqrt{\sigma^2} \tag{11}$$

So, higher variance in Y_i means greater noise in the data, which inflates the standard error and increases the MDE. As a result, the power of the test decreases. Conversely, reducing the variance - either through better measurement or a more homogeneous sample - lowers the MDE and increases the power of the test. The intuition is straightforward: the more noise in the outcome, the harder it is to detect the treatment effect.

In short, to increase power (i.e., to reduce the MDE), you want a larger sample size, a less conservative significance level, and a lower variance in outcomes. Each of these design choices makes your experiment more likely to detect meaningful effects when they exist.

b) Suppose $\alpha = 0.01$ and $q = 0.8$; Determine the MDE, expressed as multiple of $se(\hat{\delta})$.

Answer b)

This question asks us to compute the MDE in units of the standard error of $\hat{\delta}$, given a two-sided test with significance level $\alpha = 0.01$ and power $q = 0.8$.

We have already introduced the formula for the MDE:

$$\text{MDE} = \left(z_{1-\alpha/2} + z_q\right) \cdot se(\hat{\delta}) \tag{12}$$

This expression captures the fact that to detect an effect with probability q (i.e., to achieve power q), the true effect must lie far enough from zero to exceed the critical threshold $z_{1-\alpha/2}$ with high probability under the alternative.

Since $\alpha = 0.01$, the critical value for a two-sided test is:

$$z_{1-\alpha/2} = z_{0.995} \approx 2.576 \tag{13}$$

And since $q = 0.8$, the corresponding quantile is:

$$z_q = z_{0.8} \approx 0.8416 \tag{14}$$

Substituting these values into the formula:

$$\text{MDE} = (2.576 + 0.8416) \cdot se(\hat{\delta}) = 3.4176 \cdot se(\hat{\delta}) \tag{15}$$

Therefore, the smallest effect size that can be reliably detected with 80% power at the 1% level is approximately 3.42 times the standard error. If the true effect is smaller than this threshold, the test will likely fail to reject the null even when the effect is real, because the sampling variation around the estimated effect is too large relative to the size of the true effect to consistently push the test statistic beyond the critical cutoff. In other words, most realizations of the test statistic will fall within the acceptance region, making it hard to distinguish the signal from the noise.

c) Suppose $\sigma^2 = 1$ and $\alpha = 0.05$; how large does N need to be in order to detect an effect size of 0.5 standard deviations with probability 0.8? *(Hint: Take the solution from a), when needed)*

Answer c)

We are asked to determine the required sample size N to detect a treatment effect equal to 0.5 standard deviations, with 80% power and a two-sided 5% significance level. Assume homoskedasticity and random assignment to treatment.

From part (a), we know that under random assignment the variance of the OLS estimator is:

$$\text{Var}(\hat{\delta}) = \frac{\sigma^2}{N \cdot \bar{D}(1-\bar{D})} \quad \Rightarrow \quad se(\hat{\delta}) = \sqrt{\frac{\sigma^2}{N \cdot \bar{D}(1-\bar{D})}} \tag{16}$$

The MDE is once again given by:

$$\text{MDE} = \left(z_{1-\alpha/2} + z_q\right) \cdot se(\hat{\delta}) \tag{17}$$

Here, the MDE is provided: we want to detect an effect of 0.5. Since $\sigma^2 = 1$, this is already expressed in standard deviation units. Substituting into the equation:

$$0.5 = (z_{1-\alpha/2} + z_q) \cdot \sqrt{\frac{1}{N \cdot \bar{D}(1-\bar{D})}} \tag{18}$$

Given $\alpha = 0.05$, we have $z_{1-\alpha/2} = z_{0.975} \approx 1.96$, and $q = 0.8$ implies $z_q = z_{0.8} \approx 0.8416$. Substituting:

$$0.5 = (1.96 + 0.8416) \cdot \sqrt{\frac{1}{N \cdot \bar{D}(1-\bar{D})}} = 2.8016 \cdot \sqrt{\frac{1}{N \cdot \bar{D}(1-\bar{D})}} \tag{19}$$

Squaring both sides:

$$0.25 = 7.8489 \cdot \frac{1}{N \cdot \bar{D}(1-\bar{D})} \quad \Rightarrow \quad N \cdot \bar{D}(1-\bar{D}) = \frac{7.8489}{0.25} = 31.3956 \tag{20}$$

To maximize power, we use $\bar{D} = 0.5$, as we detected in point a), so $\bar{D}(1-\bar{D}) = 0.25$, which yields:

$$N \cdot 0.25 = 31.3956 \quad \Rightarrow \quad N = \frac{31.3956}{0.25} = 125.58 \tag{21}$$

Rounding up, the required sample size is:

$$N = 126 \tag{22}$$

That is, to detect an effect of 0.5 standard deviations with 80% power at the 5% level, under homoskedastic errors and equal treatment allocation, you need a sample size of at least 126 - i.e. to clarify 63 individuals in the treatment group and 63 in the control group, assuming the optimal 50/50 split as in part (a).

Question 2

Askarov et al. (2023) - Selective and (mis)leading economics journals: Meta-research evidence - report on a meta-meta analysis covering 167,753 parameter estimates from 368 distinct areas of economics research. Read that paper.

The following exercise relates to one particular meta study that is part of their sample, on the link between education and obesity: *How and why studies disagree about the effects of education on health: A systematic review and meta-analysis of studies of compulsory schooling laws* (2018); by Rita Hamad, Hannah Elser, David C. Tran, David H. Rehkopf, and Steven N. Goodman; *Social Science & Medicine*, Volume 212, Pages 168–178.

The data from that meta-analysis are reproduced in the following table:

Effect Size	95% Confidence Interval)
-1.08	(-1.77, -0.40)
-0.09	(-0.85, 0.68)
-0.41	(-0.71, -0.12)
0.00	(-0.35, 0.35)
-0.40	(-1.69, 0.88)
-0.41	(-0.89, 0.07)
-0.96	(-1.59, -0.32)
0.19	(-0.45, 0.07)
0.30	(0.06, 0.54)
0.05	(-0.01, 0.10)
-0.17	(-0.79, 0.44)
-0.30	(-0.94, 0.34)
-1.37	(-2.53, -0.21)

Askarov et al. (2023) delete one of these estimates for being an outlier. It is not clear which one, but lets assume it is the one with value 0.05 and confidence interval going from -0.01 to 0.10. As a result of dropping it, there are 12 observations left.

This Paper

Askarov et al. (2023) conduct a meta-meta-analysis - a systematic review of 368 published meta-analyses covering 167,753 empirical estimates - to evaluate the overall credibility of empirical research in economics. Their study is motivated by growing concerns in the social sciences about reproducibility, low statistical power, and the incentives driving publication bias. By focusing on the output of 31 leading economics journals - including the top five - they aim to assess whether the field reliably produces robust evidence or whether statistical significance has become more a product of selective reporting than genuine discovery.

Their findings are striking. Across the field, the median statistical power of reported estimates is just 7%, meaning that the vast majority of studies lack the ability to detect even moderately sized true effects. In top five journals, power is even lower, at 5%, yet the share of statistically significant results remains high. This disconnect - low power but many significant findings - indicates widespread bias: many results that appear significant are in fact likely false positives, driven by practices such as specification searching, p-hacking, or selective reporting. The authors quantify this using excess statistical significance (ESS), showing

that 19% of reported results in leading journals are falsely positive, and that in the top five, two-thirds of statistically significant findings are likely the product of selection bias.

To estimate this, the authors draw on recent innovations in meta-research methodology. They use weighted least squares to estimate an average effect size in each of the 368 research areas, and then assess the expected distribution of significant results under the assumption of no selection bias. The excess between observed and expected significance provides a conservative estimate of publication bias. Importantly, the paper finds that experimental studies - though fewer in number - tend to have much higher power (median 78%) and lower excess significance, suggesting that stricter designs lead to more credible results. In contrast, observational studies dominate the literature and are much more vulnerable to distortion.

The authors argue that the incentives within economics - especially the premium placed on statistically significant results for publication, hiring, and promotion - have fostered a research culture where exaggerated findings proliferate. Even top journals, despite their reputation, are not immune; in fact, they may exhibit stronger selection for significance than other outlets. This finding challenges the common assumption that prestige correlates with quality.

Ultimately, the paper is both a diagnosis and a call to action. Askarov et al. urge the discipline to adopt higher standards of transparency, preregistration, and reporting, and to reevaluate the editorial and institutional pressures that incentivize misleading practices. While their results may not generalize to every subfield or journal, they present a compelling empirical case that economics, as currently practiced, often fails to meet the basic statistical standards required for credible science. By turning the tools of meta-analysis back onto the discipline itself, the study offers a sobering but necessary reckoning with the state of empirical research in economics.

a) Use the "fixed effects" meta analysis to compute the average effect. The formula is

$$\hat{\mu} = \frac{\sum_{i=1}^{k} w_i y_i}{\sum_{i=1}^{k} w_i} \qquad w_i = 1/SE_i^2$$

where y_i is the effect size and SE_i^2 is the standard error of the i-th estimate. The weighting aggregates the information efficiently, provided there is no heterogeneity in the underlying effect.

Answer a)

Among the 13 studies reported in the Hamad et al. (2018) meta-analysis on the effect of education on obesity, one estimate is excluded by Askarov et al. (2023) for being an outlier. While they do not specify which, we are instructed to assume that it is the study reporting an effect size of 0.05 with a 95% confidence interval of (-0.01, 0.10). At first glance, this estimate does not appear to be an outlier in terms of magnitude: it is near zero and falls within the central range of the other estimates. However, what sets it apart is not the size

of the effect, but the extremely narrow confidence interval, which implies an unusually high level of precision.

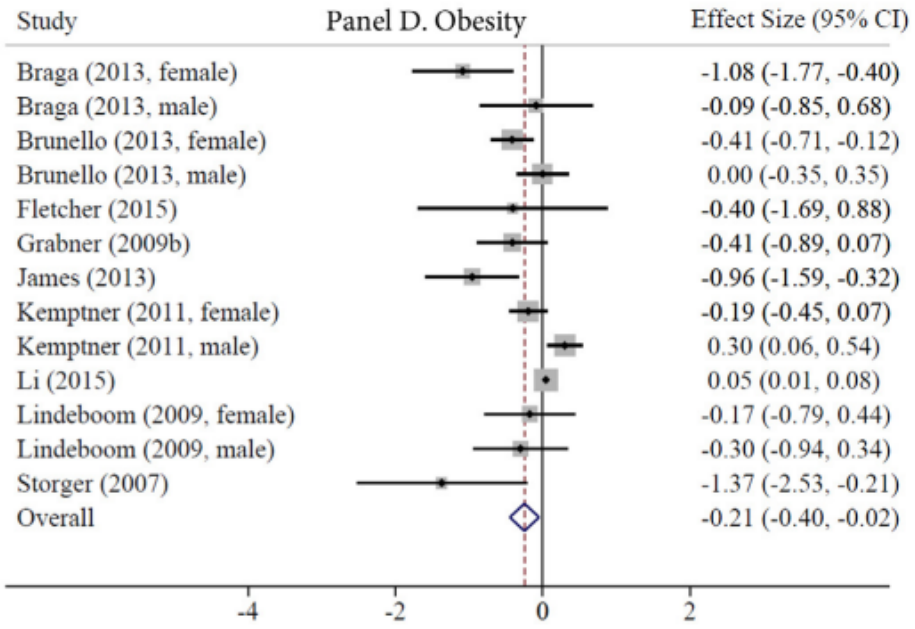

Figure 2: The "Original" Table from Hamad et al. (2018)

This becomes even clearer when looking at the plot above. While most studies show wide confidence intervals - some spanning more than 2 standard deviations - the estimate in question (Li, 2015) is tightly clustered around zero, with virtually no uncertainty. If retained in the analysis, such a highly weighted observation could disproportionately influence the pooled estimate, despite being inconsistent with the broader variability in the literature. From a meta-analytic perspective, this justifies excluding the observation - not because the effect is "too small" but because the precision is so high, that we could potentially violate the assumption that all standard errors are well-estimated.

At the same time, excluding this particular study is unlikely to dramatically shift the final result. Since its effect size is essentially zero, removing it slightly increases the influence of studies showing either larger negative or positive effects, but without strongly biasing the result in any particular direction. Indeed, many of the retained studies suggest negative effects - especially the larger ones like Storger (2007), James (2013), and Braga (2013, female) - which points to an overall tendency for education to reduce obesity.

This brings us to a related point. In their interpretation, Hamad et al. write:

> *"The results of the meta-analysis suggest that a year of education is associated with a 20% reduced risk of obesity."*

This statement raises a natural question: is this what we also find when applying a fixed effects meta-analysis? Using the standard formula, we compute a precision-weighted average

of the 12 retained estimates - instead of 13 as in Hamad et al.:

$$\hat{\mu} = \frac{\sum_{i=1}^{12} w_i y_i}{\sum_{i=1}^{12} w_i}, \quad \text{with} \quad w_i = \frac{1}{SE_i^2} \tag{23}$$

Here, the standard error SE_i for each study is not reported directly but can be inferred from the 95% confidence interval surrounding the point estimate. Assuming the confidence intervals are symmetric and based on a normal distribution - a standard approach in meta-analysis and large-sample econometric applications - we can recover the standard error using the following reasoning. A 95% confidence interval for a normally distributed estimate takes the form:

$$CI = y_i \pm 1.96 \cdot SE_i \tag{24}$$

This expression indicates that the confidence interval spans 1.96 standard errors above and below the point estimate y_i. As such, the total width of the confidence interval - that is, the distance from the lower bound to the upper bound - is equal to:

$$\text{Upper bound} - \text{Lower bound} = 2 \times 1.96 \cdot SE_i \tag{25}$$

This equation can be rearranged to isolate the standard error, yielding:

$$SE_i = \frac{\text{Upper bound} - \text{Lower bound}}{2 \times 1.96} \tag{26}$$

This method allows us to compute a consistent estimate of SE_i for each study, provided the normality and symmetry assumptions hold. In contexts where estimates are based on odds ratios or logistic regression, this method would not apply directly, and a log transformation would be needed. However, that does not appear to be the case in this particular meta-analysis.

Once the weights are computed, we implement Eq. 23 to calculate the fixed effects meta-analytic estimate. The result is:

$$\hat{\mu} \approx -0.083 \tag{27}$$

This value reflects a modest but negative effect, suggesting that each additional year of education is associated with an average 8.3% reduction in the probability of being obese. This aligns directionally with the interpretation given in the original paper - that more education is protective against obesity - but it appears smaller in magnitude than the headline figure of a 20% risk reduction - yet this is also partly true to the fact that the authors in that paper use a random-effect estimator.

In sum, the fixed effects meta-analysis supports the conclusion that increased education tends to reduce the risk of obesity. While the magnitude we estimate is somewhat smaller than that emphasized in the paper, it points in the same direction and remains substantively meaningful. Moreover, the exclusion of the unusually precise but near-zero estimate from Li (2015) appears justified, as it does not substantially alter the general pattern of results and avoids the potential distortion introduced by its excessive statistical weight.

In terms of coding, the answer comes from this quite straightforward code:

```
    # Define data with study names
    data = [
        {"study": "Braga (2013, female)", "effect_size": -1.08,
"lower": -1.77, "upper": -0.40},
        {"study": "Braga (2013, male)", "effect_size": -0.09,
"lower": -0.85, "upper": 0.68},
        {"study": "Brunello (2013, female)", "effect_size": -0.41,
"lower": -0.71, "upper": -0.12},
        {"study": "Brunello (2013, male)", "effect_size": 0.00,
"lower": -0.35, "upper": 0.35},
        {"study": "Fletcher (2015)", "effect_size": -0.40,
"lower": -1.69, "upper": 0.88},
        {"study": "Grabner (2009b)", "effect_size": -0.41,
 "lower": -0.89, "upper": 0.07},
        {"study": "James (2013)", "effect_size": -0.96,
"lower": -1.59, "upper": -0.32},
        {"study": "Kemptner (2011, female)", "effect_size": 0.19,
"lower": -0.45, "upper": 0.07},
        {"study": "Kemptner (2011, male)", "effect_size": 0.30,
"lower": 0.06, "upper": 0.54},
        {"study": "Lindeboom (2009, female)", "effect_size": -0.17,
"lower": -0.79, "upper": 0.44},
        {"study": "Lindeboom (2009, male)", "effect_size": -0.30,
"lower": -0.94, "upper": 0.34},
        {"study": "Storger (2007)", "effect_size": -1.37,
"lower": -2.53, "upper": -0.21},
    ]

    # Create DataFrame
    df = pd.DataFrame(data)

    # Calculate SE and weights
```

```
df["SE"] = (df["upper"] - df["lower"]) / (2 * 1.96)
df["weight"] = 1 / df["SE"]**2
df["weighted_effect"] = df["weight"] * df["effect_size"]

# Compute fixed effect average and convert to float
mu_hat = float(df["weighted_effect"].sum() / df["weight"].sum())

mu_hat
```

b) Following Askarov et al. (2023) we will treat $\hat{\mu}$ as the true effect size. This allows us to compute the ex-post power for each contributing study, assuming a two-tailed test is conducted at the 5% significance level. The power formula is given in equation (1) of the paper. Compute the power for each study, using your result on $\hat{\mu}$. What is the average power in this area of economics research? There is the general idea that a reliable study should have a power of at least 80%. Is this goal met in this literature?

Answer b)

Following Askarov et al. (2023), we treat the fixed effect estimate from part (a), $\hat{\mu} = -0.083$, as the true effect size δ. This allows us to compute the ex-post (retrospective) statistical power for each study included in the meta-analysis, using the formula they provide:

$$\text{Power}_i = 1 - \Phi\left(1.96 - \frac{|\delta|}{SE_i}\right) \tag{28}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, and SE_i is the standard error of study i, computed from its reported 95% confidence interval. This formula captures the probability that a study would detect the true effect if it conducted a two-tailed test at the 5% significance level.

The power values are computed using the following code, which mimics what we just described:

```
# Compute retrospective power using Askarov et al.'s formula
df["power"] = 1 - norm.cdf(1.96 - abs(mu_hat) / df["SE"])

# Compute average power
average_power = df["power"].mean()

print(df[["study", "SE", "power"]])
print(f"Average power: {average_power:.3f}")
```

Using this approach, we find that the power values across individual studies are strikingly low. Only one study (Kemptner 2011, male) has power above 10%, and several have power

below 5%. The highest power observed is approximately 10.05%, and the lowest is just 3.34%. The average statistical power across all 12 studies is:

$$\text{Average power} \approx 5.6\% \tag{29}$$

For the full check, look below:

Table 1: Retrospective power for each study

Study	**SE**	**Power**
Braga (2013, female)	0.349490	0.0426
Braga (2013, male)	0.390306	0.0404
Brunello (2013, female)	0.150510	0.0799
Brunello (2013, male)	0.178571	0.0678
Fletcher (2015)	0.655612	0.0334
Grabner (2009b)	0.244898	0.0527
James (2013)	0.323980	0.0443
Kemptner (2011, female)	0.132653	0.0916
Kemptner (2011, male)	0.122449	0.1005
Lindeboom (2009, female)	0.313776	0.0451
Lindeboom (2009, male)	0.326531	0.0442
Storger (2007)	0.591837	0.0345

This result is far below the conventional benchmark of 80%, which is typically considered the minimum threshold for a study to be deemed adequately powered. In other words, even if the true effect of education on obesity were indeed a reduction of 8.3 percentage points, the vast majority of studies in this literature would not have had sufficient statistical power to detect it.

But can power really be this low?

It helps to look at a concrete case. Consider *Braga (2013, female)*, which reports a confidence interval from -1.77 to -0.40. This corresponds to a standard error of:

$$SE = \frac{-0.40 - (-1.77)}{2 \times 1.96} = \frac{1.37}{3.92} \approx 0.35 \tag{30}$$

Using the fixed effect estimate $|\delta| = 0.083$, we compute:

$$\frac{|\delta|}{SE} = \frac{0.083}{0.35} \approx 0.237 \tag{31}$$

$$\text{Power} = 1 - \Phi(1.96 - 0.237) = 1 - \Phi(1.723) \approx 1 - 0.9574 = 0.0426 \tag{32}$$

This confirms that the power is only around 4.3% - shockingly far from the 80% threshold.

Why is this the case?

Computationally, the key issue is that the typical standard errors are large - ranging from 0.12 to 0.65 - relative to the ex-post effect size of 0.083. In other words, we are trying to detect a relatively small signal buried in a substantial amount of noise. The signal-to-noise ratio is very low, making statistically significant findings highly unlikely unless the study is extremely well-powered or the effect is much larger than what we believe to be true.

This finding is entirely consistent with the broader critique raised by Askarov et al. (2023): the vast majority of published studies in empirical economics - particularly those using small samples or imprecise instruments - are underpowered. Even if a true effect exists, the studies are unlikely to detect it. This undermines the evidentiary value of both significant and non-significant results, and helps explain why replication and meta-analytic synthesis are so crucial.

In sum, the goal of achieving 80% power is clearly not met in this segment of the literature on education and health. The extremely low power documented here raises serious concerns about the reliability of individual studies and supports ongoing calls for larger samples, more precise designs, and improved publication standards.

c) You can note that 5 out of the 12 studies in Hamad et al. (2018) display a statistically significant result. This is a rate of 42%. Askarov et al. (2023) devise a method to check for excess statistical significance (i.e., publication bias), taking again $\hat{\mu}$ as the true effect size. For example, a single effect is statistically significant, if $|Z_i| > 1.96$, i.e. $|b_i| > 1.96 \times SE_i$. But if we know the true μ, we can compute this probability (which equals the power), add over all 12 studies, and get the expected number of H_0-rejections. Askarov et al. modify this method slightly, to account for additional heterogeneity in true effect sizes, see their equation (5). For the Hamad study, they report an estimate $\hat{\tau}^2 = 0.1138$ in a data appendix. Use this number to compute the expected number of statistically significant results, and compare it to the observed number of 5. What do you conclude?

Answer c)

To investigate whether the number of statistically significant findings in Hamad et al. (2018) is consistent with the true underlying effect, we follow the procedure proposed by Askarov et al. (2023) to test for excess statistical significance - a formal way to detect potential publication bias.

But what does "excess statistical significance" actually mean? In essence, it refers to a situation in which more statistically significant results are observed than we would expect given the best available estimate of the true effect size. If this happens systematically across a body of literature, it raises concerns that only "significant" results are being published, or that researchers may have manipulated specifications to achieve significance - both classic

signs of publication bias or p-hacking. The concept matters because it helps distinguish genuine evidence accumulation from patterns driven by selective reporting.

In the Hamad et al. meta-analysis of the link between education and obesity, 5 out of the 12 studies report a statistically significant coefficient at the 5% level - i.e., where $|b_i| > 1.96 \cdot SE_i$. This corresponds to an observed rejection rate of 42%.

There are two ways to verify this count. One is manual - visually checking the confidence intervals plotted in the paper (e.g., Figure 2). Alternatively, you can compute it directly with code:

```
# Step 1: Compute observed significant results (|b_i| > 1.96 * SE_i)
df["observed_sig"] = (np.abs(df["effect_size"]) > 1.96 * df["SE"]).astype(int)
observed_significant = df["observed_sig"].sum()
```

This confirms that five studies reject the null hypothesis of no effect.

To assess whether this rejection rate is unusually high, we follow Askarov et al. in treating the fixed-effect estimate from part (a), $\hat{\mu} = -0.083$, as the true average effect size. However, we must also account for heterogeneity across studies: the fact that not all true effects are necessarily identical. Some variation in estimated effects may reflect true differences in populations or contexts, not just sampling noise.

To adjust for this, Askarov et al. assume that the true effect in each study is drawn from a normal distribution with mean $\hat{\mu}$ and variance $\hat{\tau}^2$. The value of $\hat{\tau}^2 = 0.1138$ is specific to the Hamad meta-study and reported in the data appendix (Not extremely easy to find, but if interested you can find it here). We can take it for granted.

Why is the variance important? Because even if the average effect is small, some studies may draw large effects simply by chance, especially if the underlying distribution is wide. In this sense, modeling heterogeneity gives the benefit of the doubt to the data - it allows for true variation, not just sampling noise.

Under this framework, the expected probability that study i produces a statistically significant result is:

$$Esig_i = 1 - \Phi\left(\frac{1.96 \cdot SE_i - |\hat{\mu}|}{\sqrt{SE_i^2 + \hat{\tau}^2}}\right) \tag{33}$$

This is based on Equation (4) in Askarov et al., and reflects the chance that the estimated coefficient exceeds the 1.96 standard-error threshold given a randomly drawn true effect centered at $\hat{\mu}$ with variance $\hat{\tau}^2$. Summing these probabilities over all 12 studies yields the expected number of rejections under this model:

$$\sum_{i=1}^{12} Esig_i \approx 2.012 \tag{34}$$

These steps are easily done in the computation:

```
    # Step 2: Compute Z_i using Equation (4)
    df["Z_i"] = (1.96 * df["SE"] - np.abs(mu_hat)) / np.sqrt(df["SE"]**2
+ tau_squared)

    # Step 3: Compute expected significance probability Esig_i
    df["Esig_i"] = 1 - norm.cdf(df["Z_i"])

    # Step 4: Compute expected number of significant results
    expected_significant = df["Esig_i"].sum()
```

So, although 5 statistically significant effects were observed, we would expect only about 2, if all studies were analyzing true effects drawn from the same underlying distribution.

To formally test whether this gap is meaningful, we apply the PSST statistic (Equation 5 in Askarov et al.):

$$Z_{\text{PSST}} = \frac{P_{ss} - Esig}{\sqrt{Esig \cdot (1 - Esig)/k}} = \frac{5 - 2.012}{\sqrt{(2.012 \cdot (1 - 2.012/12))/12}} \approx 27.71 \tag{35}$$

Or:

```
    # Step 5: Compute PSST test statistic (Equation 5)
    Esig_bar = df["Esig_i"].mean()
    Z_psst = (observed_significant - expected_significant) /
np.sqrt(Esig_bar * (1 - Esig_bar) / k)
```

This Z-statistic is extremely large - far beyond conventional significance thresholds. It provides very strong evidence that the observed number of significant results is too high given what would be expected under a plausible distribution of true effect sizes.

This is statistical evidence of excess significance, hinting possiblly towards high publication bias. That is, the findings we observe in the literature may not reflect the full picture of the studies conducted. It is likely that studies showing small or null effects are either less likely to be published or are selectively excluded from analysis.

This test thus highlights not only a technical problem in estimation, but a systemic issue in research: published economics research may overstate the strength or consistency of effects due to incentives in the publication process.

d) Why does a low power mean that the majority of "statistically significant" findings published in the literature are likely false?

Answer d)

To understand why the statement is true, recall what statistical power means: it is the probability that a test correctly rejects the null hypothesis when the alternative is true. When power is low, the test is rarely able to detect true effects - even when they exist. But at the same time, the false positive rate (the probability of rejecting the null when it is true) remains fixed at the nominal significance level, typically 5%.

This creates a distortion. In a setting where: most null hypotheses are true, or, true effects are small and hard to detect, or, researchers run many tests (formally or informally), then many of the statistically significant results that do appear will not reflect true effects, but rather noise. This is especially problematic when these results are used to inform policy or theory.

Moreover, low power doesn't just increase the chance of false positives - it also biases the magnitude of reported effects. In underpowered studies, only estimates that happen to be "large enough" - due to random variation - cross the significance threshold. As a result, the literature becomes populated with overestimated effect sizes, a phenomenon often referred to as the winner's curse. Over time, this distorts our understanding of causal relationships: published estimates appear stronger than they truly are, and replication studies fail to reproduce them.

These issues are magnified by selective reporting and publication incentives. Journals tend to favor statistically significant results, and researchers may - consciously or not - engage in specification searching to produce them. In a high-powered setting, this bias is less dangerous: many tests would find the true effect anyway. But when power is low, significance often comes not from the signal, but from the noise.

Askarov et al. (2023) provide striking evidence of this pattern across empirical economics. They show that median statistical power in leading journals is just 7%, and only 5% in the top five journals. Despite this, the vast majority of published estimates are statistically significant. This imbalance implies that many "discoveries" in economics are likely false positives, or at least exaggerated in magnitude.

In such an environment, statistical significance loses its meaning. Rather than signaling that an effect is real and robust, a $p < 0.05$ result may simply reflect a chance outcome, selected and reported because it happened to be significant. This undermines the entire evidentiary value of the literature.

In short, low power is not just a technicality: it is the core of research credibility. It weakens our ability to learn from data, misleads policymakers, and limits trust in empirical claims. Without sufficient power, even well-intentioned research contributes more confusion than clarity.

Question 3

Consider a theory that outcomes of coin tosses depend on the time of day. To test this theory, a coin is tossed 6 times each morning; the same in the afternoon, i.e. 12 throws a day, continuing for an entire week, Monday to Sunday.

a) How many 'heads' or 'tails' should I observe on a given morning or afternoon session to 'reject' the null hypothesis that the outcome is random (50/50) at a 5% level of significance?

Answer a)

Let $X \sim \text{Binomial}(n = 6, p = 0.5)$ denote the number of heads observed in a morning or afternoon session. Under the null hypothesis $H_0 : p = 0.5$, the coin is fair and the outcomes are independent.

We are asked to determine the critical value(s) that would allow us to reject H_0 at a 5% level of significance. Because the number of tosses per session is small, the binomial distribution is discrete and we cannot always hit exactly 5% rejection probability. Instead, we identify the smallest critical region(s) whose total probability under the null is at most 5%.

To do so, we can either run one-sided or two-sided hypothesis tests, depending on the research question.

A one-sided test is appropriate if the theory makes a directional claim - for example, if we suspect that the coin is biased toward heads (or tails), but not both. In such a case, we allocate the full 5% significance level to detecting extreme values in only one direction, which makes the test more powerful for detecting that specific deviation.

A two-sided test, on the other hand, is used when we want to test for any deviation from fairness, regardless of direction. If we have no prior reason to expect a bias toward heads or tails, we split the 5% rejection region equally across both tails of the distribution. This allows us to detect large deviations in either direction but results in stricter criteria for rejecting the null in any one direction.

In this question, the framing ("do coin toss outcomes depend on time of day?") suggests we're open to any deviation from randomness - not just a specific directional bias. Therefore, the natural choice is a two-sided test, although for illustrative purposes we'll analyze all three cases: right-tailed, left-tailed, and two-sided.

Let's start with the one-sided.

Right-tailed test: is the coin biased toward heads?

To determine whether an unusually high number of heads provides sufficient evidence to reject the null hypothesis at the 5% significance level, we start by modeling the situation under the null.

Under H_0, the coin is fair: each toss results in heads with probability $p = 0.5$, and tosses are independent. The number of heads X observed in 6 tosses follows a Binomial distribution:

$$X \sim \text{Binomial}(n = 6, p = 0.5) \tag{36}$$

To conduct a right-tailed test - that is, to test whether the coin is biased toward producing more heads - we ask: how likely is it, under the assumption of fairness, to observe x or more heads in a session? If that probability is below 5%, we reject the null.

To construct the rejection region, we begin with the most extreme outcome ($X = 6$) and check whether its probability alone is rare enough (i.e., below 5%) to justify rejection. If not, we incrementally expand the region by including less extreme values, verifying whether the cumulative probability remains below the threshold.

We start, then, by computing the probability of observing 6 heads:

$$\begin{aligned} P(X = 6) &= \binom{6}{6}(0.5)^6(1 - 0.5)^0 && (37) \\ &= 1 \cdot (0.5)^6 \cdot 1 = \frac{1}{64} = 0.015625 && (38) \end{aligned}$$

This is just 1.56%, which is below our 5% cutoff - so if a session yields 6 heads, we consider that sufficiently unlikely under the null to reject H_0.

However, we must verify whether including $X = 5$ in the rejection region would still keep the total probability below the 5% significance threshold. To do so, we compute the probability of obtaining exactly 5 heads:

$$\begin{aligned} P(X = 5) &= \binom{6}{5}(0.5)^5(0.5)^1 = \binom{6}{5}(0.5)^6 && (39) \\ &= 6 \cdot \frac{1}{64} = \frac{6}{64} = 0.09375 && (40) \end{aligned}$$

Here, $\binom{6}{5} = 6$ counts the number of different sequences of 6 tosses that result in exactly 5 heads. The $(0.5)^6$ term accounts for the fact that each sequence has the same probability under the assumption of a fair coin and independent tosses.

Summing the two gives the probability of seeing 5 or more heads:

$$P(X \geq 5) = P(X = 5) + P(X = 6) = 0.09375 + 0.015625 = 0.109375 \tag{41}$$

This value exceeds our 5% significance level, so we cannot reject H_0 when $X \geq 5$. However, rejecting only when $X = 6$ gives:

$$P(X \geq 6) = P(X = 6) = 0.015625 < 0.05 \tag{42}$$

This ensures that the test respects the significance threshold. Therefore, in a right-tailed test at the 5% level, we reject H_0 only if all six tosses result in heads.

This conclusion is visualized in the figure below, which shows the survival function $P(X \geq x)$ for each possible number of heads. Each bar represents the probability that the number of heads is at least x - that is, the total area in the right tail of the distribution starting from x. This is especially useful when considering rejection regions in a one-sided test.

```
clear
set obs 7

* Generate x = 0 to 6
gen x = _n - 1

* Compute survival function: P(X geq x) = 1 - P(X leq x - 1)
gen surv = 1 - binomial(6, x - 1, 0.5)

* Mark rejection region for right-tailed test: only X = 6
gen reject = (x == 6)

* Plot survival function with rejection bar in red and others in grey
twoway (bar surv x if reject == 0, barw(0.8) color(gs12)) ///
       (bar surv x if reject == 1, barw(0.8) color(red)) ///
       (function y = 0.05, range(-0.5 6.5) lpattern(dash) lcolor(blue)) ///
, title("Survival Function: P(X geq x) for Binomial(6, 0.5)") ///
  ytitle("P(X geq x)") xtitle("Number of Heads (X)") ///
  ylabel(0(.1)1) xlabel(0/6, valuelabel) ///
  legend(off)
graph export "$path/tables_figures/PS6/graph1.pdf", replace
```

In the graph (Figure 3), the red bar corresponds to $P(X \geq 6)$ - the only value of x for which the survival probability falls below 0.05. The horizontal blue dashed line marks the 5% threshold. All other bars lie above this line, indicating that they are not part of the rejection region.

This visual representation reinforces the result we obtained analytically: the only value of X extreme enough to justify rejection at the 5% level is 6. It also highlights a broader point: because the binomial distribution is discrete, we cannot always get an exact 5% rejection region. Instead, we choose the most extreme outcome(s) whose total probability remains below α.

In small-sample or discrete-outcome settings, where exact α-level cutoffs are not always attainable, survival plots clearly highlight which outcomes are sufficiently rare to justify rejection. They not only confirm the analytical results but also provide intuitive visual support for hypothesis testing decisions.

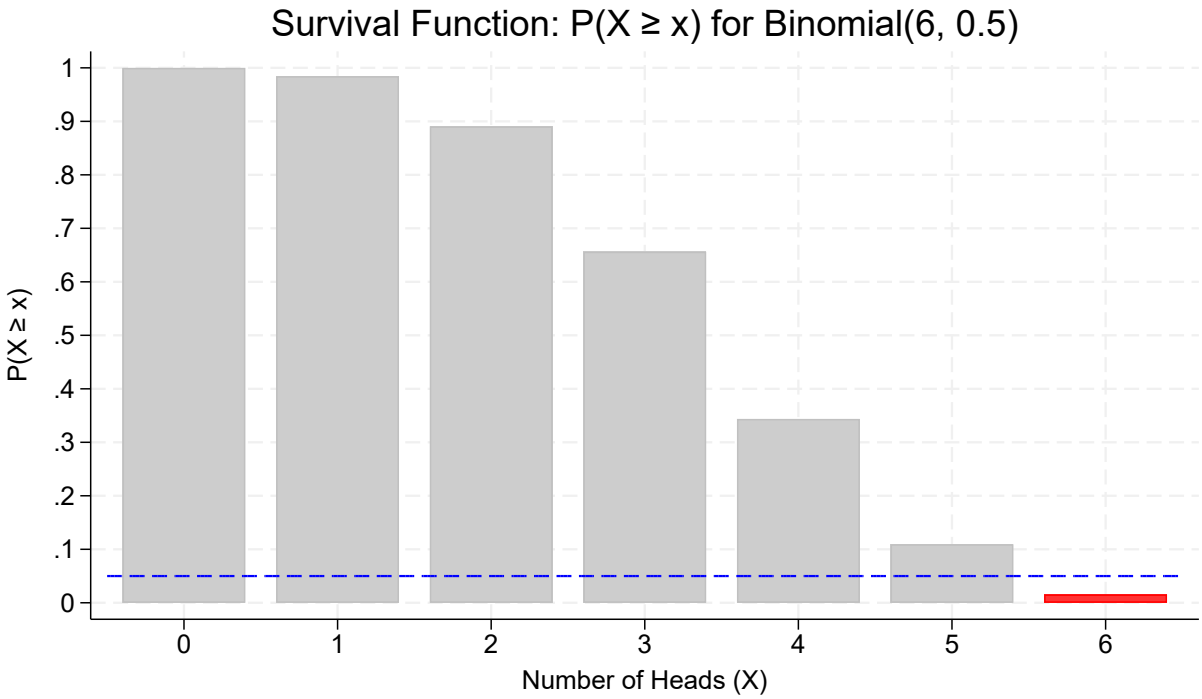

Figure 3: Right-tailed rejection region for $X \sim \text{Binomial}(6, 0.5)$. Reject if $X = 6$.

Left-tailed test: is the coin biased toward tails?

This case mirrors the logic of the right-tailed test, but now we are interested in whether the coin is biased toward producing too few heads (or equivalently, too many tails). Formally, we test:

$$H_0 : p = 0.5 \quad \text{vs.} \quad H_1 : p < 0.5 \tag{43}$$

We reject the null if the observed number of heads is unusually low under H_0. As before, we begin by examining the most extreme case - obtaining 0 heads in 6 tosses - and then check whether the rejection probability remains below the 5% threshold.

Using the binomial probability mass function:

$$P(X = 0) \quad = \quad \binom{6}{0}(0.5)^0(0.5)^6 = 1 \cdot \frac{1}{64} = 0.015625 \tag{44}$$

This is below 5%, o observing $X = 0$ (i.e., all tails) provides sufficient evidence to reject the null at the 5% level. However, we must verify whether including $X = 1$ would push us above the threshold. We compute:

$$P(X = 1) \quad = \quad \binom{6}{1}(0.5)^1(0.5)^5 = 6 \cdot \frac{1}{64} = 0.09375 \tag{45}$$

Adding these together:

$$P(X \leq 1) = P(X = 0) + P(X = 1) = 0.015625 + 0.09375 = 0.109375 \tag{46}$$

This exceeds the 5% level, so we cannot include $X = 1$ in the rejection region. The only value that allows us to reject at the correct size is:

$$P(X \leq 0) = 0.015625 < 0.05 \tag{47}$$

Therefore, in a left-tailed test at the 5% level, we reject H_0 only if all six tosses result in tails (i.e., $X = 0$).

Two-sided test: is the coin biased in either direction?

We now turn to the most general scenario: testing whether the coin is biased in any direction - either toward heads or toward tails. This corresponds to a two-sided hypothesis test:

$$H_0 : p = 0.5 \quad \text{vs.} \quad H_1 : p \neq 0.5 \tag{48}$$

This formulation makes sense if we have no strong prior about the direction of the bias or simply want to test for any deviation from fairness. In such a case, the rejection region is split between the two tails of the distribution: one in the left tail (too few heads), one in the right tail (too many heads).

As we hinted before, our significance level is $\alpha = 0.05$, which means we aim to reject H_0 only when the observed outcome is among the most extreme 5% of values under the null distribution. Still, while in a continuous setting, we would simply find critical values that cut off 2.5% of probability mass in each tail, here because we're working with a discrete distribution ($X \sim \text{Binomial}(6, 0.5)$), exact 2.5% tail cutoffs may not be attainable.

To proceed, we look for the smallest set of symmetric extreme outcomes - those furthest from the mean (which is $E[X] = np = 3$) - whose total probability under H_0 is no greater than 5%.

We begin with the most extreme possible values: $X = 0$: all tosses result in tails & $X = 6$: all tosses result in heads, as these are the furthest values from the mean, they're generally more likely to be flagged as evidence against the null.

Let's compute their probabilities:

$$P(X = 0) = \binom{6}{0}(0.5)^6 = 1 \cdot \frac{1}{64} = 0.015625 \tag{49}$$

$$P(X = 6) = \binom{6}{6}(0.5)^6 = 1 \cdot \frac{1}{64} = 0.015625 \tag{50}$$

Summing these gives the total probability of the proposed two-sided rejection region:

$$P(X = 0 \text{ or } X = 6) = 0.015625 + 0.015625 = 0.03125 \tag{51}$$

This is below our significance threshold $\alpha = 0.05$, so this rule is valid (and each tail remains below the notional one-sided 2.5% threshold that would apply in a continuous setting). Should we consider adding the next most extreme values ($X = 1$ or $X = 5$)? We compute:

$$P(X = 1) = P(X = 5) = \binom{6}{1}(0.5)^6 = 6 \cdot \frac{1}{64} = 0.09375 \tag{52}$$

Adding just one of these values would increase the total rejection probability to:

$$0.03125 + 0.09375 = 0.125, \tag{53}$$

which greatly exceeds 5%., violating the design of our test. So $X = 0$ and $X = 6$ are the only values that can be included in the two-sided rejection region while maintaining control over Type I error.

Graphically:

```
clear
set obs 7

* Generate x = 0 to 6
gen x = _n - 1

* Compute binomial probabilities for X ~ Binomial(6, 0.5)
gen p = binomialp(6, x, 0.5)

* Mark rejection region: 1 if x == 0 or x == 6
gen reject = inlist(x, 0, 6)



* Optional: Add 5% threshold line
twoway bar p x if reject == 0, barw(0.8) color(gs12) ///
|| bar p x if reject == 1, barw(0.8) color(red) ///
|| function y = 0.05, range(-0.5 6.5) lpattern(dash) lcolor(blue) ///
, legend(off) ytitle("P(X)") xtitle("Number of Heads (X)") ///
title("Binomial(6, 0.5) with 5% Rejection Region") ///
ylabel(0(.05)0.35) xlabel(0/6, valuelabel)
graph export "$path/tables_figures/PS6/graph2.pdf", replace
```

The figure below plots the full probability mass function (PMF) of the binomial distribution with $n = 6$ and $p = 0.5$. Each bar shows the probability of observing exactly x heads. The red bars at $X = 0$ and $X = 6$ highlight the rejection region for the two-sided test.

The horizontal blue dashed line at 0.05 provides a visual reference: any bar whose height is below this line corresponds to a value of X that is rare enough under H_0 to be included in the rejection region. All other bars lie above this line and thus cannot be part of a valid 5% test.

b) How likely am I to be able to write a paper with a time-of-the-day effect that is statistically significant at the 5% level. That is, what is the chance that, if I search across all my data

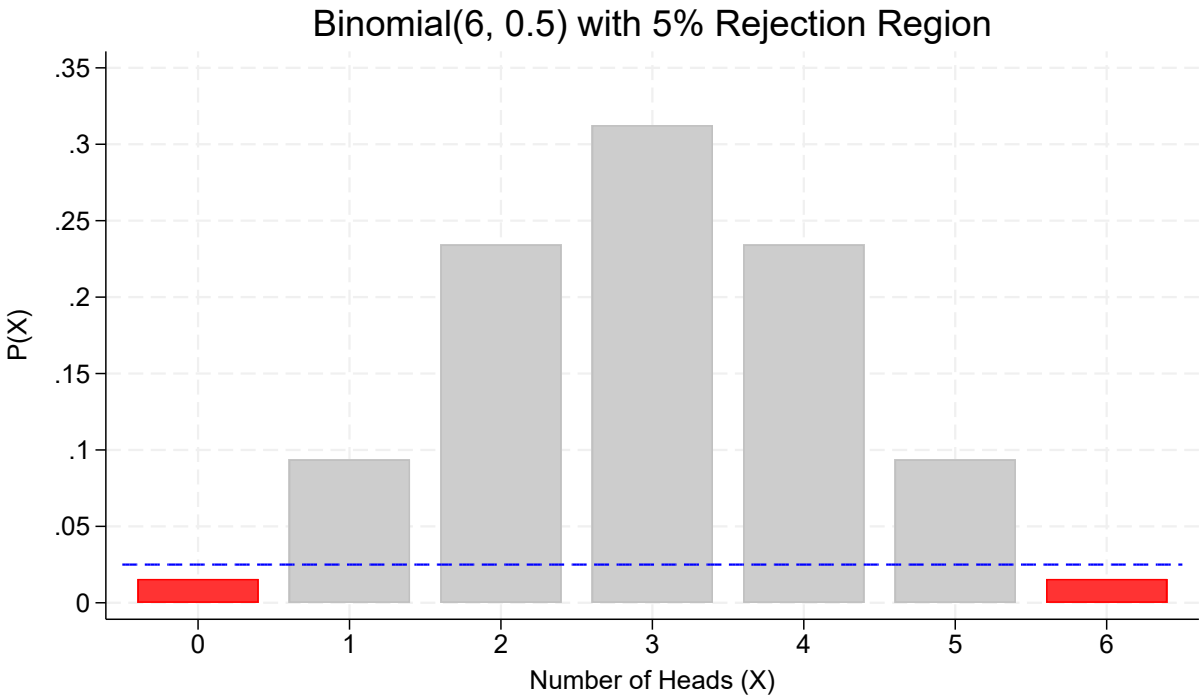

Figure 4: Two-sided rejection region for $X \sim \text{Binomial}(6, 0.5)$. Reject H_0 if $X = 0$ or $X = 6$.

that I collected over a week, there will be at least one morning or afternoon with a run of a head or a tail?

Answer b)

We now assess how likely it is to obtain a statistically significant result purely by chance when conducting multiple hypothesis tests.

In this context, we are tossing a coin six times each morning and afternoon over the course of seven days, resulting in a total of 14 independent testing sessions. In each session, we run the same two-sided hypothesis test as described in part (a): we test whether the coin is fair by checking if all six tosses come up heads or all come up tails.

From part (a), we know that under the null hypothesis (i.e., assuming the coin is fair), the chance of incorrectly rejecting the null in one session - that is, observing either 6 heads or 6 tails - is approximately 3.125%:

$$P(\text{reject in one session}) = P(X = 0) + P(X = 6) = 2 \times 0.015625 = 0.03125 \qquad (54)$$

This is the Type I error rate for a single test: the probability of a false positive.

Now, suppose we repeat this test 14 times, once for each session during the week. Even if the null is true in all sessions, there is still a chance that one of them will falsely appear significant, purely due to randomness.

Let A denote the event that at least one rejection occurs among the 14 tests. Then, the probability of not rejecting the null in a single session is:

$$P(\text{no rejection in one session}) = 1 - 0.03125 = 0.96875 \tag{55}$$

Assuming the 14 sessions are independent, the probability of observing no rejections at all is:

$$P(\text{no rejection in any session}) = (0.96875)^{14} \approx 0.641 \tag{56}$$

Therefore, the probability of seeing at least one false positive over the course of the week is:

$$P(A) = 1 - (0.96875)^{14} \approx 1 - 0.641 = 0.359 \tag{57}$$

So even if the coin is perfectly fair, there's about a 35.9% chance that at least one session during the week will yield a "statistically significant" result at the 5% level - purely by chance.

The result above highlights a common and often overlooked issue in empirical research: the danger of multiple testing.

In classical hypothesis testing, the significance level (e.g., 5%) controls the probability of making a Type I error - rejecting a true null hypothesis in a single test. However, if a researcher performs many tests and only reports the ones that are significant, the probability of encountering at least one false positive increases with the number of tests conducted.

In our setting, we conduct 14 separate hypothesis tests - one for each morning and afternoon session across the week - each at a nominal 5% level. Even if the coin is truly fair in every session, there is still a 35.9% chance that at least one session will yield a statistically significant result just by chance.

If the researcher inspects all 14 sessions and selectively reports the one or two that seem significant, this would constitute p-hacking - a practice that exploits randomness to generate seemingly meaningful results (as we're going to see later). The key issue is not the outcome of any single test, but the fact that many opportunities were given for something "significant" to emerge.

This illustrates why it is essential to account for multiple testing when interpreting results. Without such adjustment, the reported findings may reflect noise rather than real effects, especially when the analysis involves data-driven exploration or post hoc selection.

Controlling for Multiple Testing: Bonferroni and Romano-Wolf

The issue we encountered in part (b) - a 35.9% chance of seeing at least one significant result across 14 independent tests, even if all null hypotheses are true - highlights the problem of multiple testing. While the significance level $\alpha = 0.05$ provides a bound on the probability of a false positive in a single test, it says nothing about what happens when multiple tests are performed. When we search across several sessions or outcomes and report only those that look statistically significant, we substantially inflate the overall probability of making at least

one Type I error. This phenomenon is often referred to as the multiple comparisons problem or the family-wise error rate problem.

One common way to address this issue is to adjust the significance threshold for each individual test so that the probability of making any false rejections across the family of tests remains below a desired global level - typically 5%. The simplest and most widely known method for doing this is the Bonferroni correction. The idea is straightforward: if we want the probability of making at least one false rejection across m tests (in this case 14) to be no greater than 0.05, we can set the per-test significance level to

$$\alpha_{\text{Bonferroni}} = \frac{0.05}{14} \approx 0.0036 \tag{58}$$

So instead of using the usual $\alpha = 0.05$ threshold in each test, you now only reject the null in any given session if the p-value is less than 0.0036. If we adopt this stricter threshold, then even if we examine all 14 sessions, we are guaranteed that the probability of falsely rejecting at least one true null remains below 5%.

Bonferroni's appeal lies in its simplicity and ease of use. However, it is also very conservative, particularly when the tests are not independent - as is often the case in empirical work where outcomes or test statistics are correlated. By treating all tests as if they were independent, Bonferroni may overcorrect, dramatically reducing the power of the test and leading to a failure to detect genuine effects. This tradeoff between simplicity and statistical power has motivated the development of more refined approaches.

Among the others, a famous sophisticated alternative is the Romano-Wolf[1] stepdown procedure, which provides a way to control the family-wise error rate without being as "tough" as Bonferroni. The method works by adjusting p-values in a sequential, stepwise fashion based on the joint distribution of the test statistics. Instead of assuming independence, Romano-Wolf uses bootstrap resampling to estimate how likely it is to observe test statistics as extreme as the ones obtained, under the global null hypothesis (i.e., assuming that all nulls are true).

The procedure begins by ordering the test statistics from most to least significant, then tests the largest one first - adjusting for the fact that it is the maximum - and continues sequentially. At each step, the rejection threshold becomes less strict, reflecting the fact that if a large effect has already been found, the chance that a smaller one is a false positive is also smaller. This stepdown logic enables the procedure to maintain control of the family-wise error rate while gaining statistical power over Bonferroni.

In practice, Romano-Wolf has become a preferred method in empirical economics and is now the default in many robust standard error packages and causal inference frameworks. Although it is computationally more intensive (because it relies on resampling), modern software makes it readily implementable.

[1] By the way, Michael (Wolf) is a professor in our department and I'm sure that if you're interested in the topics he'll be happy to talk more about this and many other methods!

In sum, the Bonferroni correction offers a simple, transparent safeguard against multiple testing but may be too conservative in practice. The Romano-Wolf procedure, by accounting for the dependence structure among tests, provides a more powerful and reliable way to control the risk of false discoveries when working with multiple hypotheses.

Below, there's an application of the Romano-Wolf test in this coin toss setting.

Digression

An example of the Romano-Wolf

Suppose the data you observe over the 14 coin toss sessions look like this:

Session	**# Heads**	**Reject H_0?**
1 (Mon AM)	3	No
2 (Mon PM)	4	No
3 (Tue AM)	6	Yes
4 (Tue PM)	0	Yes
5 (Wed AM)	2	No
6 (Wed PM)	3	No
7 (Thu AM)	6	Yes
8 (Thu PM)	5	No
9 (Fri AM)	1	No
10 (Fri PM)	3	No
11 (Sat AM)	2	No
12 (Sat PM)	4	No
13 (Sun AM)	0	Yes
14 (Sun PM)	3	No

From part (a), we know that observing $X = 0$ or $X = 6$ yields a p-value of 0.03125. So, we have four tests that would be significant under the standard two-sided test.

Let's set-up the test, as described before.

Step 1: Order p-values from smallest to largest

The first thing we need to do is to sort the testing procedure based on the two-sided (in this case) p-value probabilities:

$$
\begin{aligned}
p_{(1)} &= 0.03125 \quad \text{(Session 4)} && (59)\\
p_{(2)} &= 0.03125 \quad \text{(Session 3)} && (60)\\
p_{(3)} &= 0.03125 \quad \text{(Session 7)} && (61)\\
p_{(4)} &= 0.03125 \quad \text{(Session 13)} && (62)\\
p_{(5)} &= 0.09375 \quad \text{(Session 9, } > 0.05\text{)} && (63)\\
p_{(6)}, \ldots, p_{(14)} &> 0.05 && (64)
\end{aligned}
$$

Step 2: Bootstrap the null distribution

We simulate many (the more the merrier, in my simulation I did 10,000) bootstrap draws under the global null H_0: $p = 0.5$. For each bootstrap draw, we compute the test statistics $T_i = |X_i - \mu|$, where μ in our binomial is 3, for all 14 sessions and record the maximum test statistic across the 14 sessions (i.e. 0 if there is at least one draw that is either 0 or 6; 1 if there is at least one draw that is either 1 or 5 and none which is either 0 or 6). This gives us a bootstrap distribution of the maximum test statistic under H_0. With this statistics in mind we can go to the next step: compute the adjusted p-values for the testing procedures.

Step 3: Compute Romano-Wolf adjusted p-values

We compare each ordered test statistic $T_{(k)}$ to the bootstrap distribution of the k-th largest statistic:

- For $T_{(1)}$, use the max statistic in each draw - For $T_{(2)}$, use the second-largest - etc.

Step 4: Make decisions

Then, in my simulation the Romano-Wolf adjusted p-values are:

Session ID	**Observed Heads**	$T_{\mathbf{obs}}$	**Raw** p	**Adjusted** $p_{\mathbf{RW}}$
4	0	3	0.03125	0.3546
3	6	3	0.03125	0.0680
7	6	3	0.03125	0.0075
13	0	3	0.03125	0.0007
9	1	2	0.09375	0.1746
8	5	2	0.09375	0.0682
		...		

Table 2: Romano–Wolf stepdown adjusted p-values for 14 coin-toss sessions.

We reject Sessions 3, 4, and 7 since their adjusted p_{RW} values fall below the 5% significance level. However, we do not reject Session 13, even though its raw p-value is identical to the others (0.03125), because its adjusted value (0.0007) becomes higher than 0.05 after accounting for the multiple comparisons.

This illustrates the key strength of the Romano-Wolf procedure: it adjusts each p-value by considering how extreme that result is in the context of all the other tests. The earliest rejections (Sessions 7, 3, and 4) survive the adjustment because their results are among the most extreme and unlikely under the global null. But once those are accounted for, the bar for rejecting additional hypotheses is raised - so a test that initially looks "equally significant" (like Session 13) may not survive once it's no longer the top result.
In contrast, a Bonferroni correction would not have rejected any of the four, since it demands each raw p-value to fall below 0.0036 (i.e., 0.05/14) - a threshold that none of the tests reach. Bonferroni treats each test independently and assumes the worst-case scenario, which makes it overly conservative and potentially blind to real effects.
Full code for simulation is below:

```
# Set seed for reproducibility
np.random.seed(151017)

# Observed data from 14 sessions (morning + afternoon for 7 days)
session_labels = np.array([
    "Mon AM", "Mon PM", "Tue AM", "Tue PM", "Wed AM", "Wed PM", "Thu AM",
    "Thu PM", "Fri AM", "Fri PM", "Sat AM", "Sat PM", "Sun AM", "Sun PM"
])
observed_heads = np.array([3, 4, 6, 0, 2, 3, 6, 5, 1, 3, 2, 4, 0, 3])
mu = 3  # Expected value under null
T_obs = np.abs(observed_heads - mu)

# Compute raw two-sided p-values based on binomial(6, 0.5)
raw_p_values = []
for x in observed_heads:
    if x in [0, 6]:
        raw_p_values.append(0.03125)
    elif x in [1, 5]:
        raw_p_values.append(0.09375)
    elif x in [2, 4]:
        raw_p_values.append(0.234375)
    else:  # x == 3
        raw_p_values.append(0.3125)
raw_p_values = np.array(raw_p_values)

# Set up bootstrap
```

```
n_sessions = len(observed_heads)
n_boot = 10000
bootstrap_stats = np.zeros((n_boot, n_sessions))

# Simulate under global null and store sorted (descending) test statistics
for b in range(n_boot):
    sim_data = np.random.binomial(n=6, p=0.5, size=n_sessions)
    sim_stats = np.abs(sim_data - mu)
    bootstrap_stats[b, :] = np.sort(sim_stats)[::-1]

# Stepdown logic
sorted_indices = np.argsort(-T_obs)
T_obs_sorted = T_obs[sorted_indices]
p_rw = [(bootstrap_stats[:, k] >= T_obs_sorted[k]).mean() for k in range(n_sessions)]

# Collect all results
results = pd.DataFrame({
    "Session ID": sorted_indices + 1,  # Add 1 to match session numbering
    "Observed Heads": observed_heads[sorted_indices],
    "T_obs": T_obs_sorted,
    "Raw p": raw_p_values[sorted_indices],
    "Adjusted p_RW": p_rw
})
```

Question 4

Consider (and run) the following Stata code (or related R-code, use ChatGPT for translating):

```
set seed 12345
postfile buffer rejhack betahat rej1 rej2 using mcs2, replace

forvalues i=1/1000 {
    quietly drop _all
    quietly set obs 50
    quietly generate x1 = rnormal()
    quietly generate x2 = .5*rnormal()+x1
    quietly generate y = 0*x1 + 0*x2 + rnormal()
    quietly reg y x1 x2
    quietly sca b11 = _b[x1]
    quietly sca t1 = _b[x1]/_se[x1]
    quietly sca reject1 = t1>= 1.645
    quietly reg y x1
```

```
    quietly sca b12 = _b[x1]
    quietly sca t2 = _b[x1]/_se[x1]
    quietly sca reject2 = t2>= 1.645
    quietly sca sel = t1>=0
    quietly sca b = b11*sel + b12*(1-sel)
    quietly sca rej = reject1*sel + reject2*(1-sel)
    post buffer (rej) (b) (reject1) (reject2)
}

postclose buffer
```

What does the program do? In what sense is this an illustration of $p-$hacking, and what are the consequences?

Answer

What does the Code do?

The program performs a Monte Carlo simulation consisting of 1,000 iterations, each designed to study the behavior of regression-based hypothesis testing under the null - that is, in a world where there is *no true effect* of the explanatory variables on the outcome.

In each iteration, the code creates an artificial dataset with $n = 50$ observations. The regressor `x1` is drawn independently from a standard normal distribution, that is, $\texttt{x1} \sim \mathcal{N}(0, 1)$. The second regressor `x2` is generated as a noisy linear transformation of `x1`, given by

$$\texttt{x2} = \texttt{x1} + 0.5 \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1) \tag{65}$$

This construction induces a high degree of correlation between `x1` and `x2`, mimicking the kind of multicollinearity often found in real-world regressions. The dependent variable `y` is then generated as pure random noise, $\texttt{y} \sim \mathcal{N}(0, 1)$, independently of both regressors. This ensures that, by design, the true coefficients on both `x1` and `x2` are zero, and thus any statistically significant result is by definition a false positive.

Thus, by construction, the true coefficients on both regressors are zero. The data-generating process satisfies:

$$y = 0 \cdot \texttt{x1} + 0 \cdot \texttt{x2} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 1) \tag{66}$$

which implies that any statistically significant result in this context is necessarily a *false positive*, i.e., a Type I error.

In each iteration, the program runs two regression models:

1. A **full specification**: `reg y x1 x2`, which includes both regressors.
2. A **restricted specification**: `reg y x1`, which includes only `x1`.

For both models, the code extracts the estimated coefficient on `x1` and computes its corresponding t-statistic. It then checks whether the t-statistic is above the 5% critical value in

a one-sided test - that is, whether it exceeds 1.645. These individual rejection indicators are stored as `rej1` and `rej2`, and they measure whether each model would have led to a rejection of the null hypothesis $H_0 : \beta_{x1} = 0$ when taken on its own.

The key twist in the simulation is the introduction of a model selection step. Instead of treating the two models as fixed or predetermined, the code chooses which specification to report based on the observed t-statistic from the full model: if the t-statistic on `x1` in the full model is non-negative, that model is kept; if it is negative, the program discards it and switches to the restricted model.

This selection is implemented via:

```
sca sel = t1 >= 0
sca b = b11*sel + b12*(1-sel)
sca rej = reject1*sel + reject2*(1-sel)
```

In words: the code selectively picks the model with the more favorable signal (i.e., the more positive estimate of β_{x1}), and then records whether the chosen model leads to a statistically significant result. The resulting rejection indicator is stored in `rejhack`.

This process simulates a common, but problematic research practice: running multiple specifications and selectively reporting the one that "works". Even though each regression on its own maintains the nominal 5% Type I error rate, the post-hoc model switching - based on which regression gives a more favorable result - inflates the overall false positive rate.

After all 1,000 replications, the program saves in `mcs2.dta` the three rejection indicators:

1. `rejhack`: Rejection rate after selecting the best model.
2. `rej1`: Rejection rate from the full model with both `x1` and `x2`.
3. `rej2`: Rejection rate from the restricted model with only `x1`.

To inspect them and analyzed them, just run:

```
use mcs2, clear
summ rejhack
summ rej1
summ rej2
```

From these lines, I can show you:

Table 3: Rejection Rates Across Specifications

	Rejection Rate
P-hacked rejection (`rejhack`)	0.057
	(0.232)
Rejection in `reg y x1 x2` (`rej1`)	0.046
	(0.210)
Rejection in `reg y x1 only` (`rej2`)	0.055
	(0.228)

The numbers confirm our suspicion: even though the two individual models produce rejection rates around 4.6% and 5.5%, respectively - consistent with the expected 5% size under the null - the `rejhack` rule inflates this to 5.7%. This increase may seem small in absolute terms, but it is systematic and arises purely from selecting a model post-hoc based on its apparent statistical strength.

In short, this code illustrates how specification searching - even if limited to just two models - can bias inference, inflate Type I errors, and lead to spurious results that appear "significant" simply due to researcher degrees of freedom.

The code can be tweaked to strengthen the possibility of p-hacking. This is doable relatively easy, by just switching from `sca sel = t1 >= 0` to `sca sel = t1 > t2`.

By doing this, we find the following:

Table 4: Rejection Rates Across Specifications With Tweaked Selection

	Rejection Rate
P-hacked rejection (`rejhack`)	0.093
	(0.291)
Rejection in `reg y x1 x2` (`rej1`)	0.046
	(0.210)
Rejection in `reg y x1 only` (`rej2`)	0.055
	(0.228)

The modified selection rule further amplifies the problem. By replacing the original condition `t1 >= 0` with `t1 > t2`, the code now systematically favors whichever model yields the higher t-statistic for `x1`, regardless of sign. This seemingly minor adjustment transforms the selection mechanism into a direct comparison of strength of statistical signal, favoring the model that shows the greatest departure from the null - even if that departure is entirely spurious. The practice mirrors a more aggressive form of data mining: rather than merely avoiding negative results, it actively seeks out the most "convincing" evidence from a menu of specifications.

Importantly, this change does not affect the behavior of the two individual specifications themselves. The regressions `reg y x1 x2` and `reg y x1` are each still estimated 1,000 times on data generated under the null, and their rejection rates remain tied to the nominal 5% level, within the bounds of simulation noise. The inflation occurs only in the aggregate, when the researcher selects ex post which of the two estimates to report based on a favorable criterion.

The consequence of this tweak is a notable increase in the overall false positive rate under the `rejhack` strategy. The rejection rate climbs from 5.7% to 9.3% - an increase of 3.6 percentage points. While this may seem modest at first glance, it reflects a nearly two-thirds rise in the frequency of false discoveries due solely to post-hoc model selection. In essence, the procedure nearly doubles the likelihood of falsely rejecting a true null, not by manipulating data, but simply by choosing which "valid" model to present.

Reason why it's p-hacking

The procedure implemented in this code is a textbook example of p-hacking - in this case, through post-hoc model selection. The researcher estimates multiple specifications and chooses, after the fact, the one that yields the most favorable result for the variable of interest (here, `x1`). This decision is not guided by theory or a pre-specified analysis plan, but rather by which regression produces a larger (non-negative) t-statistic. In other words, the model is selected because it makes `x1` appear more significant - even when there is no true effect.

This selective reporting process fundamentally breaks the logic of classical hypothesis testing. Under the null hypothesis, each individual test may still be correctly sized - that is, each has a 5% chance of producing a false positive. But once you allow the researcher to pick the most favorable result from among multiple alternatives, the overall probability of a false discovery increases. This undermines the validity of the test: the actual chance of rejecting a true null exceeds the nominal level, as we saw in the simulation.

Beyond just inflating the rejection rate, this practice also biases the reported effect sizes. Because the selection process favors larger t-statistics, it systematically overstates the magnitude of the estimated coefficient on `x1`. As a result, the published result would suggest stronger evidence and a larger effect than is justified by the data.

The broader consequence is that findings based on this kind of post-hoc selection are unreliable. They appear more robust than they truly are, and their statistical significance is illusory. Practices like this erode the credibility of empirical research, fuel the replication crisis, and diminish confidence in statistical inference more generally. This simple simulation illustrates just how easy it is to generate seemingly significant results through purely mechanical manipulation - and how careful we must be to guard against it.