# Quantitative Empirical Methods Exam

## Yale Department of Political Science, Fall 2021

You have 24 hours to complete the exam. The exam consists of three parts.

Back up your assertions with mathematics where appropriate and show your work. Good answers will provide a direct answer that illustrates an understanding of the question, and calculations or statistical arguments to validate the answer. Where applicable, exceptional answers will include all of these *as well as* proofs that are technically complete, including formally articulating sufficient assumptions and regularity conditions. Questions will not be weighted equally. A holistic score will be assigned to the exam. Therefore, it is important to demonstrate your understanding of the material to the best of your ability.

**Part 1** (Theoretical section) consists of eight shorter questions that can be answered with pen and paper. *Advice*: There may be multiple correct answers to some questions. We encourage you to give the most complete (but still succinct) solution possible. Do not leave sub-parts of questions unanswered.

**Part 2** (Essay section) contains a recent empirical article. We will ask you to offer an evaluation of its methodological approach and presentation of results. In particular, we will advise you to pay particular attention to the identification conditions (either explicit or implicit), the associated estimation strategy, and possible threats to inference. Your response may be anywhere from 500 to 1500 words.

**Part 3** (Computer assisted section) will involve using statistical software to answer one longer exercise with several associated questions. A complete answer to Part 3 will include code and output, as well as your written answers. Most students will need to consult textbooks and other references to complete this part. *Advice*: We recommend that you explain what you are trying to do in comments in your code. Even if you are not able to execute your program correctly, you can receive partial credit for explaining clearly what you wanted to do and why. Make your code readable and succinct; it should not be an issue to complete this exercise with less than 40 lines of code (excluding comments).

For the whole exam, you are permitted access to any and all written materials, as well as unrestricted use of your own computer with access to the internet. The only restriction is that you may not interact with any person, online or otherwise. A warning: problems may closely resemble problems from previous years, but may have small differences.

Please turn in your answers as an email to colleen.amaro@yale.edu.

# 1 Theoretical section

1. Suppose there is a population of 1000 subjects, where 400 of the subjects are Democrats, 150 are Independents, and 450 of the subjects are Republicans. You randomly sample one person from this population, and record whether or not this person was a Democrat, Independent or Republican.

   (a) Formally represent this random generative process as a probability space.

   (b) Formally define a random variable $X$ that takes on the value $0$ if the sampled person is a Democrat, $1$ if the sampled person is an Independent, and $2$ if the sampled person is a Republican.

   (c) Find the CDF of $X$.

   (d) Find the PMF of $X$.

   (e) Find $\mathrm{E}[X]$.

   (f) Find $\mathrm{E}[X^2]$.

   (g) Find $\mathrm{Var}[X]$.

   (h) Find $\Pr[X = 1 | X > 0.5]$.

   (i) Find $\mathrm{E}[X^2 | X > 0.5]$

2. Assume that the conditional expectation function of $Y$ given $X$ is given by

$$\mathrm{E}[Y|X = x] = x^3.$$

   Further assume that $X$ is distributed according to the standard uniform distribution.

   (a) What is the marginal effect of $X$ on $Y$ when $X = 0$?

   (b) What is the marginal effect of $X$ on $Y$ when $X$ is at its mean?

   (c) What is the average marginal effect of $X$ on $Y$?

   (d) Find the best linear predictor of $Y$ given $X$.

   You are encouraged to use a computer to verify your answers numerically (and help verify intuitions, derive integrals, etc.), but the answer should be justified analytically.

3. A paper includes the following regression table, computed using ordinary least squares and robust standard errors. It reports the results from a regression conducted on 1000 survey respondents. The outcome is $Donations$, or respondent's donations to a senator's reelection campaign in US Dollars. We have two predictors:

   - $Ideology$: self-reported ideology, on a scale from -2 (Very Liberal) to 2 (Very Conservative).

- $Income$: income, scaled as quantile in the US income distribution, on a scale from 0 to 1.

| Dependent Variable: $Donations$ | | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| $Ideology$ | 1.725 | 0.835 | 3.401 |
| | (0.456) | (0.640) | (1.494) |
| $Ideology^2$ | | | 1.317 |
| | | | (0.517) |
| $Income$ | 0.140 | 1.131 | 0.587 |
| | (0.363) | (0.854) | (0.913) |
| $Ideology \times$ | | 1.067 | 0.364 |
| $Income$ | | (0.569) | (0.635) |
| Intercept | 2.539 | 1.466 | 1.851 |
| | (0.775) | (1.054) | (1.145) |
| $n$ | 1000 | 1000 | 1000 |

The paper also includes the following summary statistics:

- $\overline{Ideology} = -1.144$.
- $\overline{Income} = 0.695$.

Consider the following inferential targets:

- $\theta_1 = \mathrm{E}[Donations|Ideology = 2, Income = 0.5]$
- $\theta_2 = \left.\frac{\partial\, \mathrm{E}[Donations|Ideology,Income]}{\partial Ideology}\right|_{Ideology=2,Income=0.5}$
- $\theta_3 = \mathrm{E}\left[\frac{\partial\, \mathrm{E}[Donations|Ideology,Income]}{\partial Ideology}\right]$

(a) In words, what are $\theta_1$, $\theta_2$ and $\theta_3$?

(b) Under specification (1), what are the estimates of $\theta_1$, $\theta_2$ and $\theta_3$?

(c) Under specification (1), compute a 95% normal approximation-based confidence interval for $\theta_2$.

(d) Under specification (2), what are the estimates of $\theta_1$, $\theta_2$ and $\theta_3$?

(e) Under specification (3), what are the estimates of $\theta_1$, $\theta_2$ and $\theta_3$?

(f) Under what assumptions can $\theta_2$ and $\theta_3$ be interpreted causally? Discuss their causal interpretation under your chosen assumptions.

4. Suppose that you want to study whether the electorate in a particular country favors men over women as members of parliament. You collect election results from $250$ districts where the main two candidates were a man and a woman. In $146$ of these elections, the winner was a man. You can assume that the $250$ observed election results were drawn i.i.d. from the population of interest.

   To formalize this setting, let $X$ be a binary random variable describing the population distribution of the election results, where $X = 1$ denotes that a man won and $X = 0$ denotes that a woman won. Define the parameter $\theta = \Pr(X = 1)$. Let the random variables $X_i$ for $i \in \{1, \ldots, 250\}$ denote the sample observations.

   (a) You want to test the null hypothesis that men and women are equally likely to be elected as members of parliament. What value $\theta_0$ of the parameter corresponds to this hypothesis?

   (b) Construct a test statistic suitable to test the null hypothesis. Motivate your choice. Derive the $p$-value of the test statistic for the observed data under the null hypothesis. Hint: `pbinom()` in R may be helpful.

   (c) What conclusions, if any, can be drawn from the test? Does the data provide support for the proposition that women are discriminated in the country under study? Explain why or why not.

   (d) A likelihood-ratio test is an alternative way to test the null hypothesis. To construct this test, follow these steps:

      i. Provide an expression of the log-likelihood function of $\theta$, denoted $\ell(\theta)$.

      ii. What is the maximum likelihood estimate of $\theta$? Denote this estimate with $\widehat{\theta}_{\text{ML}}$.

      iii. Derive the test statistic $\lambda = 2\ell(\widehat{\theta}_{\text{ML}}) - 2\ell(\theta_0)$.

      iv. Wilks' theorem implies that $\lambda$ asymptotically follows a (central) chi-squared distribution with one degree of freedom under the null hypothesis. Using an asymptotic approximation, what is the $p$-value with $\lambda$ as test statistic? Hint: `pchisq()` in R.

      v. Explain why the $p$-values of the two tests are similar but not exactly the same. Which test do you prefer and why?

5. Assume that SUTVA holds, so that $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$. Consider

$$\theta = \mathrm{E}[Y_i(1)|D_i = 0].$$

   (a) In words, what is $\theta$?

   (b) Give an example of an assumption that would allow you to identify $\theta$ from the distribution $(Y_i, D_i)$.

   (c) Give an example of a research design that would allow you justify the assumption you invoked in (b).

6. Suppose you have three variables: $Z$, $D$, and $Y$. Suppose that $Z$ is a binary instrument, $D$ is a binary treatment, and $Y$ is a real-valued outcome.

   (a) Articulate the assumptions behind the LATE Theorem: when does $\mathrm{Cov}(Y, Z)/\mathrm{Cov}(D, Z)$ characterize the "local average treatment effect" (LATE) of $D$ on $Y$?

   (b) How do empirical scholars seek to evaluate the plausibility of the assumptions behind the LATE Theorem? How convincing do you find these efforts? [Recommended length: 100-300 words]

   (c) Why is the LATE a controversial inferential target? How might you defend the LATE as an inferential target? [Recommended length: 100-300 words]

7. What are the problems associated with a "weak instrument"? What solutions might be available to the researcher faced with the problems posed by a weak instrument? [Recommended length: 100-300 words]

8. Consider a sharp regression discontinuity design with outcome $Y$, treatment $D$ and continuous forcing variable $X$. Assume the cutpoint is at zero, so that

$$D = \begin{cases} 1: & X \geq 0 \\ 0: & X < 0 \end{cases}.$$

   Suppose that you have a consistent estimator of

$$\theta = \lim_{x \to 0^+} \mathrm{E}[Y|X = x] - \lim_{x \to 0^-} \mathrm{E}[Y|X = x].$$

   Denote this estimator $\hat{\theta}$.

   (a) Articulate a set of (nontrivial) conditions under which $\hat{\theta}$ is consistent for a causal effect of $D$ on $Y$. Under these conditions, what population of units does this causal effect apply to?

   (b) In what ways do scholars seek to validate the presence of these conditions in empirical research? How persuasive are these efforts? [Recommended length: 100-300 words]

# 2 Essay section

Read the article attached to your exam. Offer a critical evaluation of its methodological approach and presentation of results. Note: "critical" does not imply that you must only criticize – it is recommended that you give credit to the authors if and when their arguments are convincing and/or novel with respect to standard practice. Your response may be anywhere from 500 to 1500 words.

We advise you to pay particular attention to the identification conditions (either explicit or implicit), the associated estimation strategy, and possible threats to inference. Justify each of your claims and, where applicable, suggest ways in which this line of research might be improved. As appropriate, we encourage you to reference related work in your answer.

Article: Peterson, Erik (2021). Paper Cuts: How Reporting Resources Affect Political News Coverage. *American Journal of Political Science* 65(2): 443–459.

# 3 Computer assisted section

You are studying political scandals among American state governors. You have collected data on the number of scandals that a sample of $n = 200$ governors have been subject to during their terms in office. Let $X$ be a random variable with support on the non-negative integers $\{0, 1, 2, \dots\}$, describing the population distribution. You posit the following statistical model:

$$\Pr(X = x) = f(x; \theta) = \frac{\theta^x}{\exp(\theta)x!},$$

where $\theta$ is a positive real-valued parameter, $\exp(\cdot)$ denotes the exponential function, and $x!$ denotes the factorial of $x$. In R, these are implemented as `exp()` and `factorial()`. The sample observations are denoted $X_i$, and you can assume that they were drawn i.i.d. from the population of interest. The data set can be accessed here:

`https://www.dropbox.com/s/r0cwr62cju84wcg/exam-part3.csv?dl=1`.

1. Confirm that $f(x; \theta)$ is a proper probability mass function. Hint: $\exp(z) = \sum_{k=0}^{\infty} z^k/k!$
2. Plot the empirical log-likelihood function of $\theta$ given the observed data.
3. Find the maximum likelihood estimate of $\theta$ numerically by searching all values between $0.05$ and $10$ with a step size of $0.05$. Hint: `seq(0.05, 10, 0.05)` in R.
4. Show analytically that the maximum likelihood estimator of $\theta$ is $\sum_{i=1}^{n} X_i/n$. Confirm that your numerically derived estimate is close to the analytically derived one. Hint: $\ln f(x; \theta) = x \ln(\theta) - \theta - \ln(x!)$
5. Pick either the numerical or analytical estimator of $\theta$ for the remaining exercises. Briefly motivate your choice.
6. What is the estimated probability that a new observation from the population is not subject to any scandal in their time in office?
7. What is the estimated probability that a new observation from the population is subject to four or more scandals?
8. Estimate the standard error of the maximum likelihood estimator of $\theta$ using bootstrap.
9. Construct a $99\%$ confidence interval for $\theta$ using a normal approximation and the estimated standard error from the previous exercise.
10. Explain in words how the confidence interval should be interpreted.
11. Construct a $99\%$ confidence interval for $\theta$ using the percentile bootstrap approach (see, e.g., Aronow & Miller, p. 132). Briefly explain why this confidence interval is similar but not exactly the same as the previous one.