EXAMINATION:
QUANTITATIVE EMPIRICAL METHODS

Yale University
Department of Political Science
August 2015

You have seven hours (and fifteen minutes) to complete the exam. You can use the points assigned to each question as a (rough) guide to allocation of your time. The exam is graded out of 40 points, which corresponds approximately to 1 point per 10 minutes. This exam consists of two parts.

Part One (9:00am - 2:15pm; 28 points)
Part One will be distributed at 9am. Your answers for Part One will be collected at 2:15pm. The only aids permitted for Part One are (1) one page of notes (double-sided permitted), (2) a calculator, and (3) a word processor on one of the Statlab computers to write up your answers (you may also write up answers using pencil/pen and paper). At 2:15 all your answers for Part One must be submitted to the proctor. Hold on to the instructions for Part One of the exam.

Advice: Use math to back up your assertions wherever appropriate, and remember to show your work. Do not leave sub-parts of questions unanswered. Your answers should be succinct and to the point.

Part Two (2:30pm - 4:30pm; 12 points)
Part Two of the exam will be distributed at 2:30pm. Your answers will be collected at 4:30pm. Part Two may involve using statistical software. A complete answer to Part Two will include the code and output, as well as your written answers. You should provide digital copies of all your files to the proctor. Ask the proctor how you should provide them, and make sure that the proctor receives them (for example if you use email). Your files should all be in a folder labelled "Exam#", where # is your your exam number. Avoid including identifying information in your code and files. If part of your answers were done on paper, you may hand that in to the proctor as well.

For Part Two you are permitted (1) unrestricted use of your own computer with access to the internet or (2) use of a Statlab computer with access to the internet. You may also use any written reference material. The only restriction for Part Two is that you may not interact with anyone, online or otherwise. In addition, you must credit in your answer any sources (code or references) that help you.

Advice: Explain what you are trying to do in comments. Even if you are not able to execute your program correctly, you can receive partial grades for explaining clearly what you wanted to do and why. We recommend you use a program such as *knitr* (see here) for R or *log* for Stata to easily record the input and output from your analysis. You should not expect graders to compile your code; all materials (code, figures, tables, results) should be directly provided to the proctor.

Part I

9:00-2:15. (5 hours; 28 points)

Rules: No references or aids permitted except one page of notes and a calculator. Answers may be written up by hand or word processor.

1 Probability (50 minutes; 5 points)

Suppose you have four urns. In Urn A, there are three balls, numbered 1 to 3. In Urn B, there is one white ball and two black balls; Urn C contains two white balls and one black ball; Urn D has a white ball, a black ball, and a red ball. You draw one ball from Urn A, and one ball from one of the other urns. Let X, a random variable, represent the number of the ball drawn from Urn A. Let Y, a random variable, represent the color of the ball drawn from the other urns (B, C, or D).

(a) Describe a process for drawing two balls that makes X and Y statistically dependent. For the rest of the questions, assume that you will use this process for drawing the balls.

(b) What is the joint p.m.f. of X and Y?

(c) Demonstrate that X and Y are statistically dependent.

(d) What is the conditional p.m.f. of X given $Y = black$? Given $Y = white$? Given $Y = red$?

(e) What is the conditional expectation function of X given Y?

(f) Is there a way to rearrange the balls in Urns B, C, and D to make X and Y statistically independent? Explain why or why not.

2 Causal Effects in Factorial Designs (50 min; 5 points)

Smedley has run two experiments. In the first (E1), he randomly manipulated factor X1. In the second (E2), he (randomly and independently) manipulated three factors: X1, X2, and X3. Each of these factors has two levels, denoted 0 or 1 (so X1=0 or X1=1). Smedley assigns each of these factors to be 1 with 50% probability. E2 is called a full factorial design.

(a)

Consider two quantities for E2. First, the difference in means: $E(Y|X1 = 1) - E(Y|X1 = 0)$. Call this Q1E2. Second, the difference in means when conditioning on the values of the other variables, for example: $E(Y|X1 = 1, X2 = x2, X3 = x3) - E(Y|X1 = 0, X2 = x2, X3 =$

$x3$). Call this Q2E2(x2,x3). Note that there are four different versions of this second quantity Q2E2(x2,x3).

What is the relationship between these two quantities? Can you write one in terms of the other?

(b)

What is the relationship between Q1E1 (the difference in means from the first experiment) and Q1E2? Which one would you prefer as an estimate of the causal effect of $X1$? You may want to discuss specific (hypothetical) experiments to illustrate your argument.

(c)

Suppose Smedley just ran experiment two (E2). What would he need to assume in order to calculate Q1E1 from the results from E2? Be specific.

3 OLS (40 minutes; 4 points)

(a) Suppose that data are generated according to the following regression equation: $Y_i = \alpha + \beta X_i + \epsilon_i$ for each unit i. However, a researcher measures $X_i^* = X_i + \nu_i$. Here, ϵ_i and ν_i are both i.i.d. random variables and are independent of X_i. Now, suppose the researcher regresses Y_i on a constant and X_i^*. What are the consequences for the unbiasedness and variance of the OLS estimator of β?
(b) Does your answer change if the true data-generating process is $Y_i = \alpha + \beta X_i + \gamma Z_i + \epsilon_i$ and the researcher regresses Y_i on a constant, X_i^*, and Z_i?

4 p-values (50 minutes; 5 points)

The following question was asked on the Polmeth listserve.

"The Journal of Basic and Applied Social Psychology (BASP) recently banned statistical significance testing. While political science journals have not done this (yet), what is your opinion on the use of p-value? What are the alternatives that we can use to do "better" science and avoid issues that using statistical significance testing introduces?"

Answer this question. Make sure that you discuss in detail (1) the problems that can arise with the use of p-values, and (2) how these problems can best be addressed. Answer as if every graduate student in political science would read your answer in their methods training.

5 Linear Probability Model vs Probit (3 points; 30 minutes)

Smedley has a dichotomous dependent variable. (Say the onset of war.) Smedley wants to estimate a maximum likelihood model of determinants (X and D) of this outcome, with emphasis on his treatment variable D. Smedley isn't sure whether to use a linear probability model or a Probit model. (1) Write out each of these models using math. (2) Explain in

words the different assumptions behind them. (3) Explain which model is preferred under what circumstances.

6 Fuzzy RD (30 minutes; 3 points)

Suppose we have a potential regression discontinuity design. We have an outcome variable Y, a treatment variable D, and a variable Z such that it is known that $E(D_i) = 0.8$ if $Z_i \geq 0$, and $E(D_i) = 0.2$ if $Z_i < 0$. What are some conditions for identification of an average effect of D_i on Y_i. How would you characterize the population for which this effect applies? What are the testable implications, if any, of these assumptions?

7 FE, RE, Pooled (30 minutes; 3 points)

For the analysis of longitudinal data (i.e., TSCS or panel data), there is debate in the social sciences about the use of fixed effects, random effects, and pooled regression for estimating causal effects. What are fixed effects regression, random effects regression, and pooled regression? Summarize and critically assess the arguments made in favor and against each of these types of estimators.

Part II

2:30-4:30. (2 hours; 12 points)

8 The Electoral College; 12 points

You are applying to work for Nate Silver. He wants to test your simulation skills. He asks you to consider the following simplistic electoral college. There are seven states $i \in \{A, B, C, D, E, F, G\}$. These states have different electoral votes (η_i) as summarized in the table below. Silver tells you to assume that the true proportion p_i of voters from each state that will vote Democratic on election day can be thought of as drawn from a normal distribution with known mean μ_i and known variance σ_i^2: $p_i \sim N(\mu_i, \sigma_i^2)$.

The Democrats will win a state if they get over 50% of the vote, otherwise the Republicans win. Winning a state gives that party the number of electoral votes held by that state. The party that wins the most electoral votes wins the election. Formally, the Democrats win if:

$$\sum_i \eta_i \mathbf{1}_{p_i > 0.5} > \frac{\sum_i \eta_i}{2} = 23.5$$

where $\mathbf{1}_{p_i>0.5}$ is a count function such that $\mathbf{1}_{p_i>0.5} = 1$ if $p_i > 0.5$ and $\mathbf{1}_{p_i>0.5} = 0$ if $p_i \leq 0.5$.

Name	Electoral Votes (η_i)	μ_i	σ_i
A	5	0.51	0.03
B	8	0.45	0.03
C	3	0.63	0.05
D	12	0.51	0.025
E	4	0.42	0.05
F	6	0.45	0.03
G	9	0.54	0.025
Total	47		

For ease of entry into R:

```
eta <- c(5,8,3,12,4,6,9)
mu <- c(0.51, 0.45, 0.63, 0.51, 0.42, 0.45, 0.54)
sigma <- c(0.03, 0.03, 0.05, 0.025, 0.05, 0.03, 0.025)
```

Silver wants you to tell him (1) what is the probability that the Democrats will win the election, and (2) what is the probability mass distribution for how many electoral votes the Democrats will get. In ordinary language, summarize the Democrats' prospects for winning the election, and what it depends on.

(3) Instead of the "Winner Takes All" function used above, allocate electoral votes for each state in proportion to the proportion of the state that votes Democrat. Thus, the Democrats win if and only if $\sum_i \eta_i p_i > 23.5$. (You can think about this election system as representing the National Popular Vote, so long as the population of each state is proportionate to the number of electoral votes.) Answer the following questions:

1. What is the probability that the Democrats will win the election under this election rule?

2. Plot the probability distribution for the proportion of the total vote that the Democrats will get.

3. Calculate how likely it is that the Democrats would win under the first election rule but lose under the second. How likely is it that they would lose under the first, but win under the second?

4. Explain in ordinary language why these different election rules give rise to different outcomes. Which electoral rule do you think that the Democrats would prefer?