Problem Set 1:
Probability and statistics

Mitch Downey
Econometrics I

January 21, 2024

1 Problems

1. *Useful properties of variance/covariance.* Prove that $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$.

2. *The difference between independence and mean independence.* Suppose u is uniformly distributed: $u \sim U[-1,1]$. Define $v = u^2$.

   (a) Calculate $Cov(u,v)$.

   (b) Is v independent of u? Explain why or why not using the σ-field definition of independence.

   (c) Is v independent of u? Explain why or why not using the $F_{X|Y=y}(x)$ definition of independence.

3. *The value of mean squared error for choosing an estimator.* Suppose you have an iid random sample $y_1, y_2, ..., y_n$ from some distribution. Let $\mu \equiv E(Y)$ be the mean and $\sigma^2 \equiv E\big(Y - E(Y)\big)^2$ be the variance of the distribution.

   (a) One estimator of μ is the sample mean: $\bar{y} = \frac{1}{n}\sum_i y_i$. Calculate the bias of $\bar{y}$.

   (b) One estimator of μ is the first observation: y_1. Calculate the bias of y_1.

   (c) Calculate the MSE of each estimator.

   (d) Compare them in terms of bias and MSE. Which estimator would you prefer?

   (e) One estimator of σ^2 is the sample variance: $s_n^2 = \frac{1}{n}\sum_i (y_i - \bar{y})^2$. Calculate the bias of s_n^2.
   (*Hint*: Write s_n^2 as a function of μ.)

   (f) One estimator of σ^2 is $\frac{1}{2}(y_1 - y_2)^2$. Calculate the bias of this estimator.
   (*Hint*: The formula for covariance is helpful.)

   (g) Calculate the MSE of each estimator.

   (h) Compare them in terms of bias and MSE. Which estimator would you prefer?

4. *The weak law of large numbers and its (disappointing) implications for aggregating random variation to the level of the outcomes we care about in applied research.*

   (a) In a well-known paper, Filipe Campante & David Yanigazawa-Drott show that cities with more direct flights to other major cities grow faster.[1] Because the air network is endogenous, they exploit variation caused by regulatory barriers: Because of regulations around pilot sleep schedules, two cities are far more likely to have a direct flight between them if they are just under (rather than just over) 6,000 miles apart. Table II of their paper shows that city c grows faster when $s_c \equiv \frac{N_{c,5500,6000}}{N_{c,5500,6000}+N_{c,6000,6500}}$ is larger, where $N_{c,d,d'}$ is the number of cities between d and d' miles away from cities c. In Section III.B and the appendix (not the online appendix, but the one just before the reference section), they formalize the argument that s_c is exogenous, as justified by their regression discontinuity evidence. The authors assume that conditional on a city being between 5,500 and 6,500 miles away from city c, it is random whether it is more than 6,000 miles away or less. In Online Appendix Table A.1,[2] the authors show that for the average city, there are 102 cities between 5,500 and 6,500 miles away. They also show that $\bar{s}$ (the sample mean of s_c) is .556 (row 1). Assume that each city c has 102 cities within 5,500-6,500 miles, and that the true probability that one of these cities being below 6,000 miles away is .556 (i.e., ignore that the sample mean is an estimate, and pretend that it is the true mean). Calculate the standard deviation of s_c. You may do this analytically or by simulation. Compare this with the standard deviation reported in the table.

   (b) A researcher is interested in whether math teachers exert more effort when their classes have more boys. She finds a set of schools where students are randomly assigned to classes, and thinks this is a good opportunity to test her hypothesis because the gender composition of the class will be exogenously determined. Assume that the female fraction of students in each school is 1/2 and that students are indeed randomly assigned to classes. She observes effort exerted in N different classes, each of which has k students.

      i. Assume that the researcher's hypothesis is correct, and that the data generating process for the effort of the teacher of class i (written as y_i) is determined according to $y_i = \beta m_i + \varepsilon_i$ where m_i is the fraction of students in class i who are male and $\varepsilon_i \sim N(0,1)$. Run 500 simulations: 100 for each $N \in \{50, 100, 250, 500, 1000\}$. In each case, simulate the data holding k fixed at 40 and $\beta = 1/3$, regress y_i on m_i, and save the estimate of $\hat{\beta}$. Calculate the standard deviation of $\hat{\beta}$ across each of the 100 iterations of your simulation. This gives you five different standard deviations from your five samples: $\sigma_{\hat{\beta}}$ for each $N \in \{50, 100, 250, 500, 1000\}$. Plot $\sigma_{\hat{\beta}}$ (y-axis) against N (x-axis).

      ii. Run 1500 more simulations simulations with the same data generating process. Use the same vector of five values of N, and run the simulation for $k \in \{10, 20, 60\}$. For each combination of N and k, calculate $\sigma_{\hat{\beta}}$. Add these

[1] Campante, Filipe, and David Yanagizawa-Drott. "Long-range Growth: Economic Development in the Global Network of Air Links." *The Quarterly Journal of Economics* 133.3 (2018): 1395-1458.

[2] `https://yanagizawadrott.com/wp-content/uploads/2017/10/Online-Appendix.pdf`

three additional series of $\sigma_{\hat{\beta}}$ (y-axis) against N (x-axis).

iii. One rule of thumb in applied research is that having a larger sample improves the reliability of estimates. Another rule of thumb is that having more variation improves the reliability of estimates. Relate these rules of thumb to your above answers.

iv. The researcher has currently collected data for a nationally representative sample of 200 classes which have, on average, 30 students. She has recently obtained funding to expand her sample. She can choose to use the funding in an urban area, where she could get more classes but they would be larger (another 100 40-student classes), or a rural area where she could get fewer classes but they would be smaller (another 50 15-student classes). Which would you recommend?

(c) A researcher is interested in the effect of having a female professors during college on long-run outcomes. For each course the student takes, assume that whether their professor is male or female is drawn iid at random with the probability of each being 1/2. The researcher is interested in understanding how much identifying variation she will have.

i. Write the CDF reflecting, at the student-level, the distribution of the number of female professors a student will have.
(*Hint*: Look it up; this is a well-known family of distributions, but the CDF is intractable so don't try to derive or calculate it yourself; instead, practice finding decent comprehensible documentation)

ii. Assume that students take 10 classes to get their degree. For what fraction of students will the share of professors who are female be less than 25% or more than 75%?

iii. Assume that students take 20 classes to get their degree. For what fraction of students will the share of professors who are female be less than 25% or more than 75%?

iv. Assume that students take 60 classes to get their degree. For what fraction of students will the share of professors who are female be less than 25% or more than 75%?

5. *Does normalizing variables within your sample preserve consistency?* Let x be distributed according to the exponential distribution: $x \sim exp(\lambda)$. This means that $f(x) = \lambda e^{-\lambda x}$ if $x \geq 0$ and $f(x) = 0$ otherwise.

(a) Calculate $E(x)$.

(b) Calculate $F(x)$.

(c) Define a new random variable $\tilde{x}$ such that $\tilde{x} \equiv x - E(x)$. Define a new random variable $y^{(n)}$ such that $y^{(n)} = x - \bar{x}_n$ where $\bar{x}_n$ is the sample mean from an iid random sample $x_1, x_2, ..., x_n$. Show that $y^{(n)}$ converges in distribution to $\tilde{x}$.

(d) Show that $\frac{1}{n}\sum_i 1\{x \leq 1\}$, where $1\{\cdot\}$ is the indicator function, is a consistent estimator for $Pr\big(x \leq 1\big)$. (Note: Put differently, show that the fraction of observations falling below 1 is a consistent estimator for the true probability of falling below 1.)