Problem Set 1 Solutions:
Probability and statistics

Thomas Mikaelsen[*]
Econometrics I

January 26, 2024

Problem 1

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

By definition

$$\begin{aligned}
Var(X+Y) &:= E[((X+Y) - (\mu_X + \mu_Y))((X+Y) - (\mu_X + \mu_Y))] \\
&= E[(X - \mu_X + Y - \mu_Y)(X - \mu_X + Y - \mu_Y)] \\
&= E[(X - \mu_X)^2 + (Y - \mu_Y)^2 + 2(X - \mu_X)(Y - \mu_Y)] \\
&= E[(X - \mu_X)^2] + E[(Y - \mu_Y)^2] + 2E[(X - \mu_X)(Y - \mu_Y)] \\
&= Var(X) + Var(Y) + 2Cov(X, Y).
\end{aligned}$$

Problem 2

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

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

[*] thomas.mikaelsen@ne.su.se

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

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

2.1: First note the following useful result that holds for any two random variables X, Y (with finite second moments)

$$\begin{aligned}
Cov(X,Y) &:= E[(X-\mu_X)(Y-\mu_Y)] \\
&= E[XY] - E[X\mu_Y] - E[\mu_X Y] + E[\mu_X\mu_Y] \\
&= E[XY] - \mu_Y E[X] - \mu_X E[Y] + \mu_X\mu_Y \\
&= E[XY] - \mu_Y\mu_X - \mu_X\mu_Y + \mu_X\mu_Y \\
&= E[XY] - \mu_X\mu_Y \\
&= E[XY] - E[X]E[Y] \qquad (1)
\end{aligned}$$

Plugging in u, v for X, Y, we get

$$Cov(u,v) = E[uv] - E[u]E[v] = E[u^3] - E[u]E[v] \qquad (2)$$

Let's calculate the moments from equation (2):

$$E[u^3] = \int_{-1}^{1} x^3 f_u(x)dx = \frac{1}{2}\int_{-1}^{1} x^3 dx = \frac{1}{2}[\frac{x^4}{4}]_{-1}^{1} = 0 \qquad (3)$$

$$E[u] = \int_{-1}^{1} x f_u(x)dx = \frac{1}{2}\int_{-1}^{1} x dx = \frac{1}{2}[\frac{x^2}{2}]_{-1}^{1} = 0 \qquad (4)$$

We see that it doesn't matter what $E[v]$ is, so plugging (3) and (4) into (2), we get $Cov(u,v) = 0$. Thus u and v are *mean independent* or *uncorrelated*.

2.2: We will show that u and v are *not* independent. Intuitively, this should make sense since v is entirely a function of u. Let's show it formally.

By definition, u, v are independent if their induced sigma-algebras $\sigma(u), \sigma(v)$ are independent. Two sigma-algebras are independent if for any two sets $A_1 \in \sigma(u), A_2 \in \sigma(v)$, A_1 and A_2 are independent. Two sets A_1, A_2 are independent if $P_{u,v}(A_1 \cap A_2) = P_u(A_1)P_v(A_2)$. Since we want to show that

u, v are not independent, it therefore suffices to show that there exists two sets $A_1 \in \sigma(u), A_2 \in \sigma(v)$ such that $P_{u,v}(A_1 \cap A_2) \neq P_u(A_1)P_v(A_2)$.

Consider the two sets

$$\begin{aligned} A_1 &= \{\omega \in \Omega : u(\omega) \in [0, 0.5]\} \\ A_2 &= \{\omega \in \Omega : v(\omega) \in [0.75, 1]\} \end{aligned}$$

We have $P_u(A_1) = \frac{1}{4}, P(A_2) = \frac{1}{4}$ but $P_{u,v}(A_1 \cap A_2) = 0$ because $v = u^2 \leq u$ for all $u \in [0, 0.5]$ and so

$$\begin{aligned} & A_1 \cap A_2 = \{\omega \in \Omega : u(\omega) \in [0, 0.5], v(\omega) \in [0.75, 1]\} = \emptyset \\ \Rightarrow \quad & P_{u,v}(A_1 \cap A_2) = P(\emptyset) = 0 \end{aligned}$$

Thus $P_u(A_1)P_v(A_2) = \frac{1}{16} \neq 0 = P_{u,v}(A_1 \cap A_2)$, as we wanted.

2.3: An equivalent way of defining independence is by the use of distribution functions. In particular, two random variables X, Y are independent if and only if $F_{X|Y=y}(x) = F_X(x)$ for all x, y. That is, the conditional distribution of X given $Y = y$ is just the distribution of X. Intuitively, information about Y is uninformative for the distribution of X.

Once again, we provide a counter-example. Consider

$$F_{V|U=u}(\frac{1}{2}) = P(V \leq \frac{1}{2} \mid U = u) = \begin{cases} 0, & \text{if } u > \frac{1}{\sqrt{2}} \\ 1, & \text{if } u = \frac{1}{\sqrt{2}} \end{cases}$$

Why is the above true? First, if $u > \frac{1}{\sqrt{2}}$ then $v = u^2 > \frac{1}{2}$ and so $v \leq \frac{1}{2}$ cannot happen. Conversely, if $u = \frac{1}{\sqrt{2}}$ then $v = u^2 = \frac{1}{2}$ and so, in particular, $v \leq \frac{1}{2}$ with certainty. We have shown that the conditional distribution function of v depends on u and so u, v are not independent.

Problem 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.

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

2. One estimator of μ is the first observation: y_1. Calculate the bias of y_1.

3. Calculate the MSE of each estimator.

4. Compare them in terms of bias and MSE. Which estimator would you prefer?

5. 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.)

6. 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.)

7. Calculate the MSE of each estimator.

8. Compare them in terms of bias and MSE. Which estimator would you prefer?

3.1: The bias of an estimator $\hat{\theta}$ is $E[\hat{\theta}] - \theta$, so let's calculate it for $\bar{y}$:

$$\begin{aligned} E[\bar{y}] - E[Y] &= E[\frac{1}{n}\sum_{i=1}^{n} y_i] - E[Y] = \frac{1}{n}\sum_{i=1}^{n} E[y_i] - E[Y] \\ &= \frac{n}{n}E[Y] - E[Y] = 0 \end{aligned}$$

Where we use linearity of expectations and that the observations are iid along the way. So the sample average is an unbiased estimator of the mean.

3.2: We have $E[y_1] - E[Y] = E[Y] - E[Y] = 0$ so the first observation is also an unbiased estimator of the mean.

3.3: The mean squared error of an estimator is defined as $MSE[\hat{\theta} \mid \theta] = E[(\hat{\theta} - \theta)^2]$. From the slides we use the result that $MSE[\hat{\theta} \mid \theta] = Var(\hat{\theta}) +$

(bias$[\hat{\theta} \mid \theta])^2$. Since both estimators are unbiased, we just need to compare their variance. We have

$$\begin{aligned} Var(\bar{y}) &= Var(\frac{1}{n}\sum_{i=1}^{n} y_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(y_i) = \frac{\sigma^2}{n} \\ Var(y_1) &= \sigma^2 \end{aligned}$$

Where we use $Var(kZ) = z^2 Var(Z)$ and $Var(Z_1 + Z_2) = Var(Z_1) + Var(Z_2)$ if Z_1, Z_2 are independent.

3.4: Comparing the two expressions, we notice that

$$\begin{aligned} MSE[\bar{y} \mid \mu] &= \frac{\sigma^2}{n} \leq \sigma^2 = MSE[y_1 \mid \mu] \quad \forall n \in \mathbb{N} \quad \text{and} \\ \lim_{n\to\infty} MSE[\bar{y} \mid \mu] &= 0 \end{aligned}$$

Since both are unbiased, we would (weakly) prefer the sample average to y_1 for all $n \geq 1$ if we only care about bias and MSE.

3.5: Bias is still the difference between the mean of the estimator and the estimand, so let's calculate the mean of s_n^2 using Mitch's hint in the second equality:

$$\begin{aligned} E[s_n^2] &= \frac{1}{n}\sum_{i=1}^{n} E[(y_i - \bar{y})^2] = \frac{1}{n}\sum_{i=1}^{n} E[((y_i - \mu) - (\bar{y} - \mu))^2] \\ &= \frac{1}{n}\sum_{i=1}^{n} E[(y_i - \mu)^2] + \frac{1}{n}\sum_{i=1} E[(\bar{y} - \mu)^2] - 2\frac{1}{n}\sum_{i=1} E[(y_i - \mu)(\bar{y} - \mu)] \\ &= \frac{n}{n} Var(Y) + \frac{n}{n} Var(\bar{y}) - \frac{2n}{n} Cov(y_1, \bar{y}) \\ &= \sigma^2 + \frac{n}{n^2}\sigma^2 - 2\frac{1}{n} Cov(y_1, y_1) \\ &= \sigma^2 + \frac{1}{n}\sigma^2 - \frac{2}{n}\sigma^2 \\ &= \sigma^2 - \frac{\sigma^2}{n} \\ \Rightarrow \quad &\text{bias}[s_n^2 \mid \sigma^2] := E[s_n^2] - \sigma^2 = -\frac{\sigma^2}{n}. \end{aligned}$$

We use that $Cov(y_1, y_j) = 0$ for all $j \neq i$ because they are iid to get the fourth line. Thus, the sample variance is biased but consistent as $\frac{-\sigma^2}{n} \to 0$.

3.6: Using the same idea as in (3.5), we have

$$\begin{aligned}
E[\frac{1}{2}(y_1 - y_2)^2] &= \frac{1}{2}E[((y_1 - \mu) - (y_2 - \mu))^2] \\
&= \frac{1}{2}(E[(y_1 - \mu)] + E[(y_2 - \mu)] - 2E[(y_1 - \mu)(y_2 - \mu)]) \\
&= \frac{1}{2}(\sigma^2 + \sigma^2 - 2Cov(y_1, y_2)) \\
&= \frac{1}{2}(2\sigma^2 - 0) = \sigma^2 \\
\Rightarrow \quad & \text{bias}[\frac{1}{2}(y_1 - y_2)^2 \mid \sigma^2] = \sigma^2 - \sigma^2 = 0
\end{aligned}$$

So this is an unbiased estimator or σ^2.

3.7: We use once again that MSE is variance plus bias. We have already calculated bias in (3.5) and (3.6) so let's focus on variances. First the sample variance which has a non-trivial derivation (taken from here: `https://math.stackexchange.com/questions/2476527/variance-of-sample-variance?noredirect=1&lq=1`). It gives us

$$Var(\frac{1}{n-1}\sum_i (y_i - \bar{y})^2) = \frac{\mu}{n} - \frac{\sigma^2(n-3)}{n(n-1)}$$

Second the other estimator

$$
\begin{aligned}
Var(\frac{1}{2}(y_1 - y_2)^2) &= \frac{1}{4}Var(y_1^2 + y_2^2 - 2y_1y_2) \\
&= \frac{1}{4}[Var(y_1^2) + Var(y_2^2 - 2y_1y_2) + 2Cov(y_1^2, y_2^2 - 2y_1y_2)] \\
&= \frac{1}{4}[Var(y_1^2) + Var(y_2^2) + 4Var(y_1y_2) + 2Cov(y_2^2, -2y_1y_2) + 2Cov(y_1^2, y_2^2 - 2y_1y_2)] \\
&= \frac{1}{4}[2Var(y^2) + 4Var(y_1y_2) - 4Cov(y_2^2, y_1y_2) + 2Cov(y_1^2, y_2^2) - 4Cov(y_1^2, y_1y_2)] \\
&= \frac{1}{4}[2Var(y^2) + 4Var(y_1y_2) - 8Cov(y^2, y_1y_2) + 2Cov(y_1^2, y_2^2)] \\
&= \frac{1}{4}[2Var(y^2) + 4Var(y_1y_2) - 8Cov(y^2, y_1y_2)] \\
&= \frac{1}{4}[2Var(y^2) + 4[Var(y)^2 + 2Var(y)E[y]^2] - 8Cov(y^2, y_1y_2)] \\
&= \frac{1}{4}[2Var(y^2) + 4[Var(y)^2 + 2Var(y)E[y]^2] - 8Cov(y^2, y_1y_2)] \\
&= \frac{1}{4}[2Var(y^2) + 4[Var(y)^2 + 2Var(y)E[y]^2] - 8(E[y^3]E[y] - E[y^2]E[y]^2)] \\
&= \frac{1}{2}Var(y^2) + Var(y)^2 + 2Var(y)E(y)^2 - 2E(y^3)E(y) - 2E(y^2)E(y)^2
\end{aligned}
$$

Where we use some implicit calculations in the background. The following is used in line 5:

$$
\begin{aligned}
Cov(y_1^2, y_1y_2) &= E[y_1^2y_1y_2] - E[y_1^2]E[y_1y_2] \\
&= E[y_1^3]E[y_2] - E[y_1^2]E[y_1]E[y_2] \\
&= E[y_2^3]E[y_1] - E[y_2^2]E[y_1y_2] \\
&= E[y_2^2y_2y_1] - E[y_2^2]E[y_1y_2] \\
&= Cov(y_2^2, y_1y_2)
\end{aligned}
$$

We then used that for independent variables (see here: `https://math.stackexchange.com/questions/4271183/variance-of-product-of-two-random-variables-f` to get line 6

$$
\begin{aligned}
Var(y_1y_2) &= Var(y_1)Var(y_2) + Var(y_1)E[y_2]^2 + Var(y_2)E[y_1]^2 \\
&= Var(y^2) + 2Var(y)E[y]^2
\end{aligned}
$$

and we use that y_1, y_2 are identically distributed. Finally, to get line 7 we used

$$\begin{aligned} Cov(y_1^2, y_1 y_2) &= E[y_1^2 y_1 y_2] - E[y_1^2]E[y_1 y_2] \\ &= E[y_1^3]E[y_2] - E[y_1^2]E[y_1]E[y_2] \\ &= E[y^3]E[y] - E[y^2]E[y]^2 \end{aligned}$$

since they are identically distributed.

Conclusion: The sample variance is biased but consistent and its MSE is much lower since it uses much more data. In particular, it converges to 0. Hence I prefer the sample variance.

Problem 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.

1. 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

[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`

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.

2. 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.

   (a) 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).

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

   (c) 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.

(d) 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?

3. 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.

   (a) 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)

   (b) 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%?

   (c) 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%?

   (d) 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%?

4.1: Given the assumptions that there are 102 cities within $5,500 - 6,500$ miles of each city and $P(c \in [5,500,6,000]) = 0.556$ we can think of $N_{c,5500,6000}$ as the random variable that asks the question "How many of the 102 cities lie within 5,500-6,000 miles of c?". Such a random variable follows a Binomial distribution $N_{c,5500,6000} \sim Bi(n = 102, p = 0.556)$ and since we are given that

$N_{c,5500,6000} + N_{c,6000,6500} = 102$, it follows that s_c is a scaled version of said Binomial, $s_c = \frac{N_{c,5500,6000}}{102}$. To calculate the standard deviation, we use that a Binomial distribution is the sum of iid Bernoulli distributions[3]

$$Bi(n,p) = \sum_{i=1}^{n} Ber(p)$$

Which is nice because it is not too difficult to calculate mean and variances of Bernoullis, since they are binary random variables

$$\begin{aligned} E[Ber(p)] &= P(B=1)\cdot 1 + P(B=0)\cdot 0 = P(B=1) =: p \\ Var[Ber(p)] &= E[B^2] - E[B]^2 = P(B^2=1)\cdot 1^2 + P(B^2=0)\cdot 0^2 - p^2 \\ &= p - p^2 = p(1-p) \end{aligned}$$

$$[independence] \quad \Rightarrow \quad Var(Bi(n,p)) = Var(\sum_{i=1}^{n} Ber(p)) = np(1-p) \tag{1}$$

Thus, we can use equation (1) to calculate the standard deviation $\sigma(s_c)$ of s_c

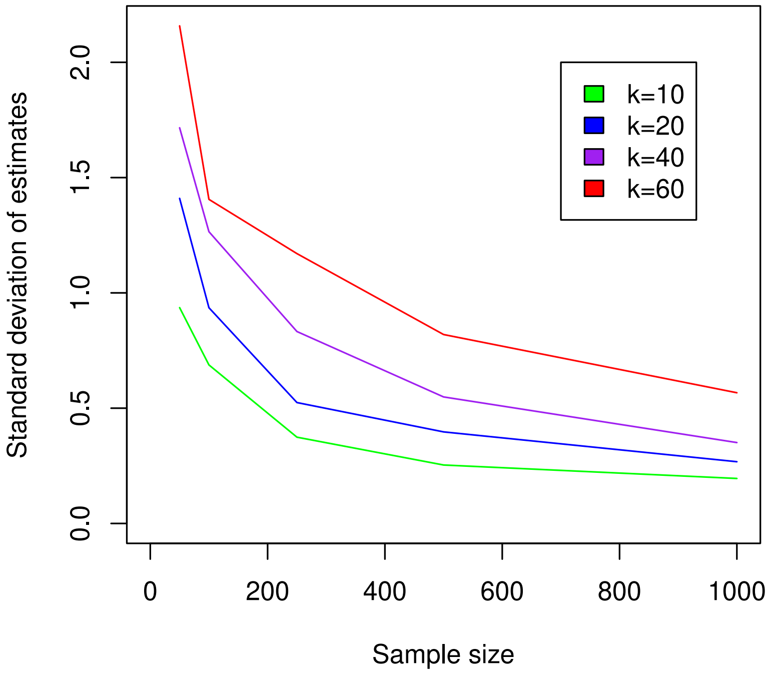
$$\begin{aligned} \sigma(s_c) &= \sqrt{Var(s_c)} = \sqrt{Var(\frac{N_{c,5500,6000}}{102})} \\ &= \frac{\sqrt{Var(Bi(n=102, p=0.556))}}{102} \\ &= \frac{\sqrt{102 \cdot 0.556(1-0.556)}}{102} \\ &\approx 0.05 \end{aligned}$$

This standard deviation is roughly 42% of the standard deviation they report in the paper, or equivalently, they report having almost two and a half times as much identifying variation than they should theoretically have. In general, the message here is that if treatment occurs at the individual level then, if we aggregate up, some (potentially much, if not all) variation gets lost.

4.2(a+b):

[3] Because "If the probability of a city lying in [5500,6000] is p, how many of the n cities are within [5500,6000]?" is the same as asking "If the probability of flipping heads on a coin is p and I flip it n times, how many heads will I get?". The first experiment follows a Binomial distribution, the second experiment repeats an iid Bernoulli experiment 102 times.

Figure 1: Sample standard deviation of $\hat{\beta}$: $\sigma_{\hat{\beta}}$



4.2(c): We see that increasing sample size does reduce the standard deviation of the estimates, but the effects are non-linear. For instance, if classes are 40 students, then going from 250 to 1000 classes (i.e. increasing the sample size by four) decreases the standard deviation by roughly as much as going from 40 to 20 students per class, without changing N. That's because decreasing the size of classrooms generates much more variation in the female share of the class because the law of large numbers means that as the size of classes grows, the share of males in the class converges to the true mean of 0.5 – so in the limit as, k increases, all classes will have the same female share and hence no variation at all. Why does this increase variation on the estimate of beta? Because the variation in y_i is increasingly coming from noise and not from β because all the m_is converge to the same value.

4.2(d): If we look at the graph, we see that is we start with $N = 200$ classes, then adding 100 vs 50 doesn't seem to make that much of a difference. However, the difference between the standard deviation of classes with 40 students vs 15 seems rather larger. I would recommend the data on rural schools.

4.3(i): Suppose the student takes n classes. Then we want to model the random variable that asks "how many of the n classes has a female professor?", or equivalently, "out of the n trials, how many of them will be a success?". This is modeled by a Binomial distribution $Bi(n, 1/2)$ and it has probability mass function

$$f(k, n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

which measures "what is the probability of getting exactly k successes out of the n trials?". The intuition for f is that the k successes happen with probability p^k, the remaining failures occur with probability $(1-p)^{n-k}$, but each of the k successes can happen anywhere among the n trials and there are $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ (pronounced *n-choose-k*) ways of distributing these k successes among the n draws.

4.3(ii): If you take 10 classes, then you have less than 25% female teachers if 0,1, or 2 of the classes have a female teacher. Similarly, more than 75% translates into 8,9 or 10 classes with female teachers. Since each of these draws are mutually exclusive, the probability of their union is the sum of the individual probabilities

$$\begin{aligned}
P(k \in \{0,1,2,8,9,10\}) &= P(k = 0 \cup k = 1 \cup ... \cup k = 10) \\
&= \sum_{i \in \{0,1,2,8,9,10\}} P(k = i) \\
&= \sum_{i \in \{0,1,2,8,9,10\}} \binom{10}{i} p^i (1-p)^{(10-i)} \\
&\approx 0.11
\end{aligned}$$

4.3(iii): This amounts to setting $n = 20$ and then the relevant outcomes are $k \in I = \{0,1,2,3,4,16,17,18,19,20\}$

$$\begin{aligned}
P(k \in I) &= \sum_{i \in I} \binom{20}{i} p^i (1-p)^{(20-i)} \\
&\approx 0.01
\end{aligned}$$

4.3(iv): Then the relevant ks are $K = \{0,1,2,...,14,46,47,...,60\}$ and we get

$$\begin{aligned}
P(k \in I) &= \sum_{i \in I} \binom{60}{i} p^i (1-p)^{(60-i)} \\
&\approx 0.00005
\end{aligned}$$

The message from these three sub-questions is two-foled: First we observe that for any number of reasonable number of courses, the share of students who experienced a very high or very low female share of professors is small. Hence, nearly all of the identifying variation – variation in the share of female professors a student experienced during their studies – will be close to the 50-50. In other words, few students will have experienced very few or very many female professors during their studies and as such the data is less well suited to speak to certain theories, such as role model effects, and at least should be addressed by the researcher. Second, this again illustrates the point of question 4.2: The identifying variation also shrinks as the size of the groups – in this case the number of courses taken during studies – increases, due to the law of large numbers. In this case, finding an individual with 25% female professors is much more likely when people take 10 classes than when people take 60 classes. This effect exacerbates the previous point.

Problem 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.

1. Calculate $E(x)$.
2. Calculate $F(x)$.
3. 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 $\tilde{x}$.
4. 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.)

5.1: Recall integration by parts

$$\int uv' = uv - \int u'v$$

and the nice way to derive it (I can never remember the formula): Use the product rule and take an integral while appealing to the Fundamental Theorem of Calculus, like so

$$\begin{aligned} (uv)' &= u'v + uv' \\ \Rightarrow \quad uv &= \int vu' + \int uv' \\ \Rightarrow \quad \int uv' &= uv - \int u'v. \end{aligned}$$

The expectation then is

$$
\begin{aligned}
E(x) &:= \int_{-\infty}^{\infty} xf(x)dx = \lambda \int_{0}^{\infty} xe^{-\lambda x}dx \\
&= \lambda[-\frac{x}{\lambda}e^{-\lambda x}]_0^\infty + \frac{\lambda}{\lambda}\int_0^\infty e^{-\lambda x}dx \\
&= [-\frac{e^{-\lambda x}}{\lambda}]_0^\infty \\
&= \frac{1}{\lambda}
\end{aligned}
$$

as we wanted.

5.2: By definition

$$
\begin{aligned}
F(x) = P(X \leq x) &= \int_{-\infty}^{x} f(z)dz \\
&= \lambda \int_0^x e^{-\lambda z}dz = -\frac{\lambda}{\lambda}[e^{-\lambda z}]_0^x \\
&= -(e^{-\lambda x} - 1) \\
&= 1 - e^{-\lambda x}
\end{aligned}
$$

as we wanted.

5.3: By Weak Law of Large Numbers $\frac{1}{n}\sum_{i=1}^n x_i \xrightarrow{p} E(x) = \frac{1}{\lambda}$ which is a constant. x from $\tilde{x}$ and x from $y^{(n)}$ are identical, so they in particular converge in distribution. Thus, by the second property of convergence in distribution we have

$$
\begin{aligned}
y^{(n)} &= x - \frac{1}{n}\sum_{i=1}^n x_i \xrightarrow{d} x - E(x) = x - \frac{1}{\lambda} \\
&= \tilde{x}
\end{aligned}
$$

as we wanted.

5.4: By WLLN and definitions

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x\leq 1\}} &\xrightarrow{p} E[\mathbb{1}_{\{x\leq 1\}}] \\
&:= \int_{supp(X)} \mathbb{1}_{\{x\leq 1\}} f_x(x) dx \\
&= \int_{\{x\leq 1\}} f_x(x) dx \\
&=: P(x \leq 1)
\end{aligned}$$

as we wanted.