Midterm (100 points possible)
ECNS 561
Fall 2018
ANSWER KEY

_____Name

1.) Proofs of variance and covariance propositions.

a.) (5 points) Property Var. 2 states that, for any constants a and b

$$Var(aX + b) = a^2Var(X).$$

Prove this property.

$$\begin{aligned} Var(aX + b) &= E[((aX + b) – E(aX + b))^2] \\ &= E[(aX + b – aE(X) - b)^2] \\ &= E[(aX – aE(X))^2] \\ &= E[a^2(X – E(X))^2] \\ &= a^2E[(X – E(X))^2] \\ &= a^2Var(X) \end{aligned}$$

b.) (5 points) Property Cov. 2 states that, for any constants a, b, c, and d

$$Cov(aX + b, cY + d) = acCov(X, Y).$$

Using the fact that $Cov(X,Y) = E[(X – E(X))(Y – E(Y))]$, prove this property.

$$\begin{aligned} Cov(aX + b, cY + d) &= E[(aX + b – E(aX + b))(cY + d – E(cY + d))] \\ &= E[(aX + b – aE(X) - b)(cY + d – cE(Y) - d)] \\ &= E[a(X – E(X))c(Y – E(Y))] \\ &= acE[(X – E(X))(Y – E(Y))] \\ &= acCov(X, Y) \end{aligned}$$

2.) The uniform distribution is characterized as follows:

Let X be a continuous random variable and its support be a closed interval of real numbers:

$$R_X = [l, u].$$

We say that X has a uniform distribution on the interval $[l, u]$ if its probability density function is:

$$f_X(x) = \begin{cases} \frac{1}{u-l} & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

a.) (7 points) Show that the expected value of a uniform random variable X is $\mathrm{E}[X] = \frac{u+l}{2}$.

$$\mathrm{E}[X] = \int_l^u x f(x)dx = \int_l^u x \frac{1}{u-l} dx = \frac{1}{u-l}\int_l^u x dx = \frac{1}{u-l}\left[\frac{1}{2}x^2\right]\Big|_l^u = \frac{1}{u-l}\frac{1}{2}[u^2 - l^2] = \frac{(u-l)(u+l)}{2(u-l)} = \frac{u+l}{2}$$

b.) (8 points) Show that the variance of a uniform random variable X is $\mathrm{Var}[X] = \frac{(u-l)^2}{12}$.

$$\mathrm{E}[X^2] = \int_l^u x^2 f(x)dx = \int_l^u x^2 \frac{1}{u-l} dx = \frac{1}{u-l}\int_l^u x^2 dx = \frac{1}{u-l}\left[\frac{1}{3}x^3\right]\Big|_l^u = \frac{u^3-l^3}{3(u-l)}$$

$$\mathrm{E}[X]^2 = \frac{(u+l)^2}{4}$$

$$\mathrm{E}[X^2] - \mathrm{E}[X]^2 = \frac{u^3-l^3}{3(u-l)} - \frac{(u^2+2ul+l^2)(u-l)}{4(u-l)} = \frac{u^3-l^3-3u^2l+3ul^2}{12(u-l)} = \frac{(u-l)^3}{12(u-l)} = \frac{(u-l)^2}{12}$$

c.) (5 points) Let X be a uniform random variable with support:

$$R_X = [4, 12].$$

Calculate the following probabilities:

$$P(5 \leq X \leq 10)$$

and

$$P(X > 6)$$

$P(5 \leq X \leq 10) = \int_5^{10} f(x)dx = \int_5^{10} \frac{1}{12-4} dx = \frac{x}{8}|_5^{10} = 10/8 - 5/8 = 5/8$

$P(X > 6) = 1 - P(X \leq 6) = 1 - \int_4^6 \frac{1}{12-4} dx = 1 - \frac{x}{8}|_4^6 = 1 - (6/8 - 4/8) = 1 - 1/4 = 3/4$

3.) (10 points) Given a positive constant k > 0, the exponential density function is

$$f_X(x) = \begin{cases} ke^{-kx} \text{ if } x \geq 0 \\ 0 \text{ if } x < 0 \end{cases}$$

Let X be a continuous random variable with an exponential density function with parameter k. Solve for E[X].

E[X] = $\int_0^\infty kxe^{-kx}dx$

Integrating by parts and setting u = kx and dv = e^{-kx}dx, we obtain

= $-xe^{-kx}|_0^\infty + \int_0^\infty e^{-kx}dx$

= $-xe^{-kx}|_0^\infty - \frac{1}{k}e^{-kx}|_0^\infty$

Taking limits and evaluating,

= (0 + 0) – (0 – 1/k)
= 1/k

4.) Let $\bar{Y}$ denote the sample average from a random sample with mean μ and variance σ^2. Consider two alternative estimators of μ:

$$G_1 = \frac{(n-3)}{n}\bar{Y}$$

and

$$G_2 = \frac{\bar{Y}}{3}$$

a.) (7 points) Find the probability limits of G_1 and G_2, which estimator is consistent?

$plim[G_1] = plim[\frac{(n-3)}{n}\bar{Y}] = plim[\frac{(n-3)}{n}]*plim[Y] = 1*\mu = \mu$

$plim[G_2] = plim[\frac{\bar{Y}}{3}] = \mu/3$

G_1 is consistent

b.) (8 points) Using the concept of mean squared error, argue whether G_1 is a better estimator than $\bar{Y}$ if μ is arbitrarily close to zero.

$MSE[G_1] = Var[G_1] + [Bias(G_1)]^2 = Var[\frac{(n-3)}{n}\bar{Y}] + \{E[\frac{(n-3)}{n}\bar{Y}] - \mu\}^2 = [\frac{(n-3)}{n}]^2 Var[\bar{Y}] + \{\frac{(n-3)}{n}\mu - \mu\}^2 = [\frac{(n-3)}{n}]^2 * \frac{\sigma^2}{n} + \frac{3\mu^2}{n^2}$

At $\mu = 0$, $MSE[G_1] = \frac{(n-3)^2}{n^3}\sigma^2$.

We know that $MSE[\bar{Y}] = Var[\bar{Y}] = \frac{\sigma^2}{n}$.

So, because $\frac{(n-3)}{n} < 1$, we also know that $MSE[G_1] < Var[\bar{Y}]$. Consequently, when μ is arbitrarily close to zero, we prefer G_1 over $\bar{Y}$.

5.) For each of the following assertions, state whether it is a legitimate statistical hypothesis.

a.) (3 points) H_0: $\sigma < 100$

Yes. It is an assertion about the value of a parameter.

b.) (3 points) H_0: $\bar{y} = 45$

No. The sample average is not a parameter.

c.) (3 points) H_0: $\frac{\sigma_1}{\sigma_2} < 1$

Yes. Again, it is an assertion about parameters of a particular problem.

6.) Adult males are taller, on average, than adult females. Visiting two recent American Youth Soccer Organization (AYSO) under 12 year old (U12) soccer matches on a Saturday, you do not observe an obvious difference in the height of boys and girls of that age. You suggest to your little sister that she collect data on height and gender of children in 4th to 6th grade as part of her science project. The accompanying table shows her findings (Y=mean, sd = standard deviation, N = sample size).

Height of Young Boys and Girls, Grades 4-6, in inches

Boys			Girls		
$\bar{Y}_{boys}$	sd_{boys}	N_{boys}	$\bar{Y}_{girls}$	sd_{girls}	N_{girls}
57.8	3.9	55	58.4	4.2	57

a.) (4 points) Let your null hypothesis be that there is no difference in the height of females and males at this age level. Write down the null and alternative hypotheses.

$H_0 : \mu_{boys} - \mu_{girls} = 0$ vs. $H_1 : \mu_{boys} - \mu_{girls} \neq 0$

b.) (3 points) Find the difference in height and the standard error of the difference. Note that $se(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{sd_1^2}{N_1} + \frac{sd_2^2}{N_2}}$.

$\bar{Y}_{boys} - \bar{Y}_{girls} = -0.6$

$se(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{3.9^2}{55} + \frac{4.2^2}{57}} = 0.77$.

c.) (4 points) Generate a 95% confidence interval for the difference in height.
$-0.6 \pm 1.96 \times 0.77 = (-2.11, 0.91)$

Critical values	
p-value	two-tailed test
.050	1.96
.025	2.24
.010	2.58
.005	2.81
.001	3.3

d.) (5 points) Calculate the *t*-stat for comparing the two means. Using a two-tailed test, is the difference stat. significant? Would the critical value be smaller if you had assumed a one-sided alternative hypothesis? Why? Give no more than a two-sentence answer on the intuition here.

$t = -0.6/0.77 = -0.78$, so $|t| < 2.58$, which is the critical value at the 1% level. Hence you cannot reject the null hypothesis. The critical value for the one-sided hypothesis would have been 2.33. Assuming a one-sided hypothesis implies that you have some information about the problem at hand, and, as a result, can be more easily convinced than if you had no prior expectation.

7.) (10 points) Consider the following archetypal regression in the style of the "Equality of Educational Opportunity" study (EEOS):

$$Y_{isn} = \alpha + \boldsymbol{X1_{isn}'\beta_1} + \boldsymbol{X2_{sn}'\beta_2} + \boldsymbol{X3_{n}'\beta_3} + \varepsilon_{isn},$$

where *i* indexes individuals, *s* indexes schools, and *n* indexes neighborhoods. The dependent variable, *Y*, represents some measure of student achievement and the vectors ***X1***, ***X2***, and ***X3***, represent individual (e.g., race, parents' education, number of siblings, peer characteristics), school (e.g., funding per student, class size, teacher qualifications), and neighborhood characteristics (e.g., share of households who rent vs. own), respectively.

Based on this type of analysis, Dr. James Coleman, a pioneer in the social scientific analysis of education, concluded that family and peer characteristics explained a statistically and consequently significant amount of variation in the measure of achievement. School inputs and neighborhood characteristics did so to a much lesser extent. Given these results, Coleman concluded that families and peers had an effect on achievement that schools and neighborhoods did not.

In the space provided below, briefly describe Dr. Caroline Hoxby's critique of Coleman's conclusions.

Dr. Hoxby's critique is based off the fact that none of the variables in Coleman's analyses were causally identified. To do so would have required exploiting well-defined natural experiments that induced plausibly exogenous variation in the covariates of interest. Because Coleman did no such thing in his research, interpreting his estimated coefficients as causal effects was incorrect and came with serious policy implications.

8.) (10 points) "The Marshmallow Test" is a famous psychological experiment in the delay of self-gratification. It goes something like this…A researcher gives a child a marshmallow and tells her that she can either eat the marshmallow now or wait and receive another marshmallow later. Before leaving the room, the researcher instructs the child that she will receive the second marshmallow upon the researchers return if the original marshmallow is uneaten.

It turns out that research finds that the children who are able to delay self-gratification actually do better in terms of long-run outcomes (e.g., educational attainment, employment, likelihood of incarceration, etc.). Based on these findings, many schools have adopted lessons where they stress teaching children the delay of self-gratification and emphasize to parents that this is a vitally important early-childhood skill. This research, however, is generally based on raw correlations between marshmallow-test performance and a long-run outcome of interest.

Do you think these raw correlations are sufficient evidence to support allocating resources to teaching delay in self-gratification to school-aged children? Why or why not? As an aspiring econometrician, what would be some of the first things you would want to do to test the robustness of these correlations?

No, definitely not. Even in the most basic analysis, one would want to control for factors such as parental education, family socioeconomic status, etc. It is entirely possible that, when controlling for these potential confounds, the observed relationship between performance on the marshmallow test and subsequent outcomes goes away. In fact, a recent study by Watts et al. (2018) in *Psychological Science* found this exact result…the marshmallow test does a much poorer job predicting adult outcomes after holding constant factors such as family background, early cognitive ability, and home environment.