Stat 88 Fall 2019 FINAL EXAM **A. Adhikari**

NAME (FIRST LAST): ______________________________ **SID:** ________________

BUILDING (circle one): Stanley/LeConte **SEAT NUMBER:** ________________

TIME AND CONDITIONS: You have three hours. A two-sided reference sheet is provided. No other materials are allowed; nor are calculators, computers, or the internet.

QUESTIONS AND ANSWERS

- There are 10 questions.
- **Give brief explanations or show calculations in each question** unless the question says this is not required. You may use, without proof, any result proved or used in lecture, the textbook, and homework, **unless the question asks for a proof**.
- Unless the question says otherwise, you may leave answers as unsimplified arithmetic or algebraic expressions including finite sums and the standard normal cdf Φ.

GRADING

- The exam is worth 100 points: 10 points for each of the 10 questions. Points for parts are indicated in the question.
- Please commit yourself to a single answer for each question. If you give multiple answers (such as both True and False) then please don't expect credit, even if the right answer is among those that you gave.
- Please stop writing immediately when proctors announce that time is up, and please make no delay in following instructions to turn in your test. If you delay, you will be penalized 20% of your score in fairness to students who stop writing when instructed. See Honor Code below.

FORMAT

- There is a space for your name and SID number on one side of each page. Please fill this in. It will ensure that we can identify your work during the scanning process.
- There is space for your answer below each question. **Please do not write outside the black boundary**; the scanner and Gradescope won't read it.
- If you need scratch paper please use the backs of the exam pages. **But be aware that these pages will neither be scanned nor graded.**
- Please turn in only your exam, not the reference sheet.

HONOR CODE

Data Science and the entire academic enterprise are based on one quality – integrity. We are all part of a community that doesn't fabricate evidence, doesn't fudge data, doesn't present other people's work as our own, doesn't lie and cheat. You trust that we will treat you fairly and with respect. We trust that you will treat us and your fellow students fairly and with respect. **Please abide by UC Berkeley's Honor Code:**

"As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others."

Your signature: ____

Name: ________________________________ **SID:** ____________________

1. A fair coin is tossed 400 times. Let X be the number of heads.

a) [**3 points**] What is the distribution of X?

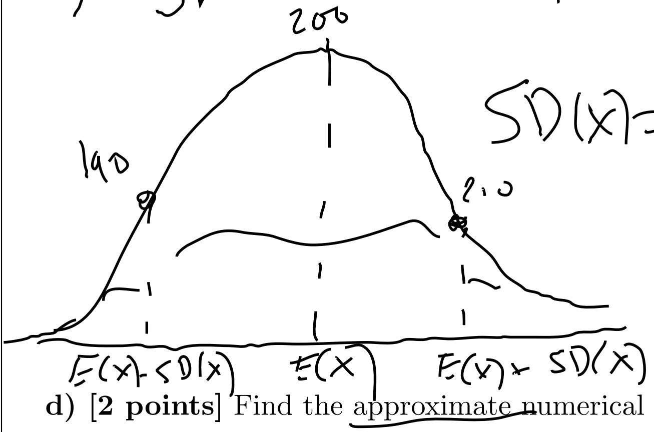
$\mathrm{Bin}(N = 400, p = 1/2)$

b) [**3 points**] Write a numerical formula for $P(X > 210)$.

$$P(X > 210) = P(X \geq 211) = \sum_{k=211}^{400} \binom{400}{k} (1/2)^k (1/2)^{400-k}$$

c) [**2 points**] Sketch the rough shape of the distribution of X, and mark the numerical values of $E(X)$ and $SD(X)$ appropriately on your sketch.

$X \sim \mathrm{Bin}$



$E(X) = np = 400(1/2) = 200$

$SD(X) = \sqrt{Var(X)} = \sqrt{npq}$

$= \sqrt{(400)(1/2)(1/2)}$

$= 10$

d) [**2 points**] Find the approximate numerical value of $P(X > 210)$.

$X \sim \mathrm{Normal}(\mu = 200, \sigma^2 = 100)$

$\Rightarrow P(X > 210) = 1 - P(X \leq 209) = \boxed{1 - \Phi\left(\frac{209 - 200}{10}\right)}$

Also $\boxed{1 - \Phi\left(\frac{210 - 200}{10}\right)}$

$P(X \geq 210)$
$\sim P(X > 210)$

Name: ________________________________ **SID:** ________________

2. In a simple random sample of 500 students taken from among all community college students in a large state, the following data are recorded:

- The annual tuition paid by the sampled students has an average of $1200 and an SD of $450.
- Among the sampled students, 37% are less than 21 years old.

a) [5 points] Construct an approximate 95% confidence interval for the average tuition paid by community college students in the state.

Sample: $\bar{x} = 1200$ $\hat{\sigma} = 450$

CI: $\mu \pm 2\sigma$ by 68-95-99 rule

Apply CLT. $\mu = \bar{x} = 1200$ $\sigma = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{450}{\sqrt{500}}$

$\therefore \boxed{1200 \pm 2 \cdot \frac{450}{\sqrt{500}}}$

b) [5 points] Construct an approximate 99% confidence interval for the percent of the state's community college students who are less than 21 years old.

95% CI: $\left(\bar{x} - 2\frac{\sigma}{\sqrt{n}}, \bar{x} + 2\frac{\sigma}{\sqrt{n}}\right)$

obs-value of $\bar{x} = 0.37$ $n = 500$

$\sigma = \sqrt{p(1-p)} \approx \sqrt{(0.37)((0.63)} = 0.4828$

$\Rightarrow \boxed{\left(0.37 \pm 2 \cdot \frac{0.48}{\sqrt{500}}\right)}$

Name: ______________________________ **SID:** ______________

3. A class consists of 50 students of whom 30 are sophomores and 20 are juniors. For the midterm exam, a simple random sample of 25 of the 50 students are assigned to Room A and the remaining 25 to Room B.

a) [**5 points**] Find the chance that exactly 15 sophomores are assigned to Room A.

50 students 30 soph 20 jr $\Rightarrow$ HyperGeo

$$\frac{\binom{30}{15}\binom{20}{10}}{\binom{50}{25}}$$

$N = 50$

$G = 30$

$n = 25$

b) [**5 points**] Find the chance that all the juniors are assigned to the same room.

$$\frac{\overset{\text{All in first}}{\binom{20}{20}\binom{30}{5}}}{\binom{50}{25}} + \frac{\overset{\text{None in first}}{\binom{20}{0}\binom{30}{25}}}{\binom{50}{25}}$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

All = None

$$\frac{20!}{0!\,20!} = 1 \qquad \frac{20!}{20!\,0!} = 1$$

$$\frac{30!}{25!\,5!} = \frac{30!}{5!\,25!}$$

Name: ______________________________ **SID:** ______________

4. On the first day of the term, students at a college sign up for brief individual appointments with an undergraduate advisor. She has two advising sessions: one in the morning and one in the afternoon.

- Each appointment in the morning session is for 10 minutes. The number of students who sign up for morning appointments is a random variable N_m that has the Poisson (6) distribution.
- Each appointment in the afternoon session is for 15 minutes. The number of students who sign up for afternoon appointments is a random variable N_a that has the Poisson (3) distribution and is independent of N_m.

Let T be the total time (in minutes) of the appointments for which students sign up that day.

a) [**2 points**] Find $E(T)$.

$T = M + A$ $\qquad E(T) = E(M) + E(A) = E(10N_m) + E(15N_a)$

$= 10E(N_m) + 15E(N_a)$

$M = \underset{\substack{\downarrow \\ \text{duration}}}{10} \cdot \underset{\substack{\downarrow \\ \text{\# of appointments}}}{N_m}$ $\qquad A = 15 \cdot N_a$

$= 10(6) + 15(3) = \boxed{105}$

b) [**2 points**] Find $Var(T)$.

$Var(T) = Var(M + A) = Var(M) + Var(A)$ b/c M & A indep.

$= Var(10N_m) + Var(15N_A)$

$= 100\,Var(N_m) + 225\,Var(N_A) = \boxed{100(6) + 225(3)} \rightarrow 1275$

c) [**3 points**] Find $P(N_m + N_a \leq 20)$. $\qquad N_m + N_a = N_T \sim Pois(9)$

$$P(N_T \leq 20) = \boxed{\sum_{i=0}^{20} \frac{e^{-9}9^i}{i!}} \qquad \frac{e^{-\lambda}\lambda^k}{k!}$$

d) [**3 points**] Select **all** the correct options below and justify your choices: The answer to Part (c) is

[✓ (i) at least $\frac{1}{2}$] [✓ (ii) at least $\frac{3}{4}$] [✓ (iii) at least $\frac{8}{9}$] [✓ (iv) at least $\frac{15}{16}$]

$P(N_T \leq 20) \Rightarrow$ Upper bound $P(N_T > 20) \Rightarrow P(N_T \geq 21)$

By Markov: $P(N_T \geq 21) \leq \frac{E(N_T)}{21} = \frac{9}{21} = \frac{3}{7} \Rightarrow$ complement $= \boxed{\frac{4}{7}}$

Chebyshev's: $P(N_T \geq 21) \leq P(|N_T - 9| \geq 12) = \frac{1}{k^2} = \frac{1}{16} \Rightarrow$ complement $\boxed{\frac{15}{16}}$

$\sigma^2 = 9$ $\qquad P(|N_T - \mu| \geq k\sigma) = \frac{1}{k^2}$

$\sigma = 3$ $\qquad (4)\,3$

$12 = k\sigma \quad 12 = 3k \quad k = 4$

Name: ______________________________ **SID:** ______________

5. A student whose roommate claims to be able to predict the future builds a machine that works as follows: There are four lightbulbs. Each time the machine is run, it picks a bulb at random and that bulb lights up. The roommate's job is to predict which bulb will light up. (Such a machine was used for a similar purpose by a professor at UC Davis.)

The machine is run 48 times and the roommate makes correct predictions 15 times. The student who built the machine says it looks as though the roommate was just guessing at random.

To see whether or not the results resemble guessing at random, perform a test of hypotheses in the following steps.

a) [**2 points**] State a precise null hypothesis in terms of random variables.

H_0: Student guesses at random, iid Ber(1/4) gets 12 correct

Bin($N=48, p=1/4$)

b) [**2 points**] State an appropriate alternative hypothesis.

H_A: Student does not guess at random ie correct $\neq 12$

c) [**2 points**] Select an appropriate test statistic, justify your choice, and provide the observed value of the statistic.

Two sided prue correct $\neq 12$

$T = |X-12|$ $\quad X \sim \text{Bin}(N=48, p=1/4)$ $\quad E(X)=np=12$

$T_{\text{observed}} = 3$

d) [**2 points**] Write a numerical expression for the exact p-value of the test.

Large T support the alternative

$$P(T \geq 3) = P(|X-12| \geq 3) = P(X \geq 15) + P(X \leq 9)$$

$$\sum_{0}^{9} \binom{48}{k} (1/4)^k (3/4)^{48-k} + \sum_{15}^{48} \binom{48}{k} (1/4)^k (3/4)^{48-k}$$

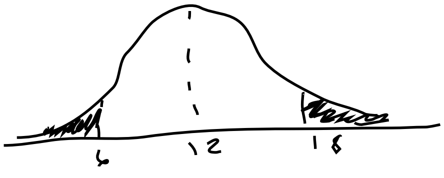
e) [**2 points**] Which of the two hypotheses is better supported by the data, and why? [You can answer this by using simple arithmetic; no long calculations are needed. Examine properties of the distribution you used in Part (d).]

95% $\sim \mu \pm 2\sigma$ $\qquad 12+3=15$

$n=48$ is large enough for CLT

$SD(x) = \sqrt{npq} = \sqrt{48(1/4)(3/4)} = 3$

$\mu = np = 12$ $\qquad$ CI: $(6, 18)$ $\qquad$ 1SE$(6,18)$



$\boxed{\Rightarrow \text{Support } H_0}$

Name: ______________________________________ **SID:** ____________________

6. Let X have the probability density defined by

pdf

$$f(x) \;=\; \begin{cases} 12x^2(1-x), & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

In the parts below, **please work out all integrals** as arithmetic or algebraic expresssions. Don't leave integrals in your answers.

a) [**3 points**] Find $E(X)$.

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_0^1 x \cdot 12x^2(1-x)dx = \int_0^1 12x^3 - 12x^4 dx$$

$$= \left(3x^4 + \frac{12}{5}x^5\right)\Bigg|_0^1 = 3(1) + \frac{12}{5}(1) = \boxed{\frac{3}{5}}$$

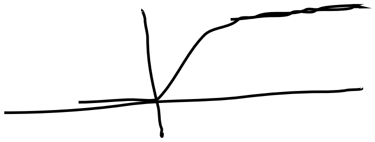
$$\frac{15}{5} - \frac{12}{5}$$

b) [**3 points**] Let F be the cumulative distribution function (cdf) of X. Find $F(x)$ for $0 < x < 1$.

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_s(s) ds = \int_0^x 12s^2(1-s)ds$$

$$= \int_0^s 12s^2 - 12s^3 ds = \left(4s^3 - 3s^4\right)\Big|_0^x = 4x^3 - 3x^4$$

$$F(x) = \begin{cases} = 0 & 0 \leq x \\ 4x^3 - 3x^4 & 0 < x < 1 \\ 1 & x \geq 1 \end{cases}$$



c) [**4 points**] Find $P(X > 0.6 \mid |X - 0.5| < 0.3)$ in terms of $\underline{F}$.

$$P(A|B) = \frac{P(A,B)}{P(B)} \searrow \quad |X - 0.5| < 0.3 = X < 0.8, X > 0.2$$

$$\frac{P(X > 0.6, X < 0.8, X > 0.2)}{P(X < 0.8, X > 0.2)} = \frac{P(0.6 < X < 0.8)}{P(0.2 < X < 0.8)}$$

in terms of F : $\boxed{\dfrac{F(0.8) - F(0.6)}{F(0.8) - F(0.2)}}$

Name: ____ **SID:** ____________________

7. A random number generator draws repeatedly at random with replacement from the 10 digits 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9.

$E(geo) = \frac{1}{p}$

a) [**3 points**] Find the expected number of draws till all 10 digits have appeared.

X = # of draws until all 10 appear

$X = 1 + geo(9/10) + geo(8/10) + \cdots + geo(1/10)$

$= 1 + \frac{10}{9} + \frac{10}{8} + \cdots + 10$

$= 1 + \frac{10}{9} + \frac{10}{8} + \frac{10}{7} + \frac{10}{6} + \frac{10}{5} + \frac{10}{4} + \frac{10}{3} + \frac{10}{2} + 10$

b) [**4 points**] Find the variance of the number of draws till all 10 digits have appeared.

$Var(X) = Var(1) + Var(geo(9/10)) + \cdots + Var(geo(1/10))$

$0 + \frac{1/10}{(9/10)^2} + \frac{2/10}{(8/10)^2} + \cdots + \frac{9/10}{(1/10)^2}$

$Var(geo) = \frac{1-p}{p^2}$

c) [**3 points**] Find the chance that all 10 digits appear in the first 10 draws.

$P(X=10) = \left(\frac{10}{10}\right)\left(\frac{9}{10}\right)\left(\frac{8}{10}\right) \cdots \left(\frac{1}{10}\right)$

$= \frac{10!}{10^{10}}$

Name: ____ **SID:** ____________________

8. Let $N > 25$ be a positive integer, and let $X_1, X_2, \ldots, X_{25}$ represent the results of 25 draws made uniformly at random **without** replacement from the integers $\{1, 2, 3, \ldots, N\}$.

a) [**2 points**] What is the distribution of X_7? Explain briefly.

$X_7 \sim \text{Unif}(1 \ldots N)$ by symmetry each of the X_i ~ are identical and previous draws give no info (not unconditional) of previous draws.

b) [**3 points**] Let $\bar{X} = \frac{1}{25}\sum_{i=1}^{25} X_i$. Find $E(\bar{X})$ and hence show that $2\bar{X} - 1$ is an unbiased estimator of N.

if $2\bar{x} - 1$ is unbiased estimator of N, $E(2\bar{x} - 1) = N$

$$E(2\bar{x} - 1) = 2E(\bar{x}) - 1$$
$$= 2E\left(\frac{1}{25}\sum_{i=1}^{25} x_i\right) - 1$$
$$= 2E(X_1) - 1$$
$$= 2\left[\frac{N+1}{2}\right] - 1$$
$$= N + 1 - 1 = N \checkmark$$

c) [**5 points**] Let H be the sample median; we are using H for "halfway". That is, let H be the 13th value in the sorted sample. It will be the same value whether you sort in increasing or decreasing order, but it is more helpful to imagine the sample sorted in increasing order.

Is $2H - 1$ an unbiased estimator of N? If not, can you construct an unbiased estimator of N based on H?

[Hint: A good way is to draw a row of spaces for the integers 1 through N, mark the sampled values, and count gaps.]

Does $E[2H - 1] = N$? What is $E(H)$?

+1

_ _ _ _ _ . . . . _ _ _
1 2 3 4 5 N-2 N-1 N

There are 26 gaps, find the expected length of each gap.

$$E(\text{length of gap}) = \frac{\text{total length}}{\#\text{ of gaps}} = \frac{N-25}{26} \Rightarrow E[2H-1]$$

$$E[H] = \underset{\#\text{ of gaps}}{13}\left(\frac{N-25}{26}\right) + 12 + 1 = \frac{N-25}{2} + 13$$

12 x's before median

median itself

$$2E\left[\frac{N-25}{2} + 13\right] - 1$$
$$N - 25 + 26 - 1 = \boxed{N}$$

2H-1 is unbiased!

Name: ______________________________ **SID:** ______________

9. An individual picked randomly from a population has educational level V and income M. Here educational level is measured in years and income in dollars. Define the following notation:

- r is the correlation between V and M — $r(V, M)$
- $E(V) = \mu_V$, $Var(V) = \sigma_V^2$
- $E(M) = \mu_M$, $Var(M) = \sigma_M^2$
- $\hat{M}$ is the least squares linear predictor of M based on V

In what follows, you can use the equation of the least squares line provided on the reference sheet. Please derive all other results.

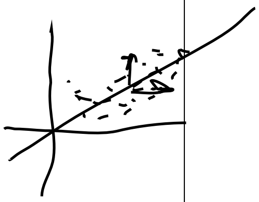
a) [3 points] Let $W = M - \hat{M}$ be the residual. Express W in terms of the deviations $\underline{D_V = V - \mu_V}$, $\underline{D_M = M - \mu_M}$, and the quantities defined above.

$$W = M - \hat{m} = M - (\hat{a}V + \hat{b}) = M - (\hat{a}V + \mu_m - \hat{a}\mu_v)$$

$$\hat{a} = r\frac{\sigma_m}{\sigma_v} \qquad = M - \hat{a}V - \mu_m + \hat{a}\mu_V$$

$$\hat{b} = \mu_m - \hat{a}\mu_v \qquad = M - \mu_m - \hat{a}(V - \mu_v)$$

$$\boxed{W = D_m - r\frac{\sigma_m}{\sigma_v}D_V}$$



b) [2 points] Use Part (a) to find $E(W)$. Show your work.

$$E(W) = E(D_m) - E\left(\underset{\substack{\uparrow \\ \text{constants}}}{r\frac{\sigma_m}{\sigma_v}}D_v\right) = E(D_m) - r\frac{\sigma_m}{\sigma_v}E(D_v)$$

$$= E(M - \mu_m) - r\frac{\sigma_m}{\sigma_v}E(V - \mu_v) = 0 - r\frac{\sigma_m}{\sigma_v}(0) = 0$$

c) [5 points] Use Parts (a) and (b) to derive the formula $Var(W) = (1 - r^2)\sigma_M^2$. Show all the details of your work; false or unjustified arguments will not receive credit.

$$Var(W) = E(W^2) - (E(W))^2 \qquad \text{Since } E(W) = 0,\ E(W)^2 = 0$$

$$Var(W) = E(W^2) \qquad E(W^2) = E\left[\left(D_m - r\frac{\sigma_m}{\sigma_v}D_v\right)^2\right] = E\left[D_m^2 + 2r\frac{\sigma_m}{\sigma_v}D_mD_v + r^2\left(\frac{\sigma_m}{\sigma_v}\right)^2D_v^2\right]$$

$$E(D_m^2) + 2r\frac{\sigma_m}{\sigma_v}E(D_mD_v) + r^2\left(\frac{\sigma_m}{\sigma_v}\right)^2E(D_v^2)$$

$$\downarrow$$

$$E((M - \hat{m})^2)$$

$$\downarrow$$

$$Var(M) - 2r\frac{\sigma_m}{\sigma_v}E(D_mD_v) + r^2\left(\frac{\sigma_m}{\sigma_v}\right)^2Var(V)$$

$$\sigma_m^2 - 2r\frac{\sigma_m}{\sigma_v}r\sigma_m\sigma_v + r^2\left(\frac{\sigma_m}{\cancel{\sigma_v}}\right)^2\cancel{\sigma_v^2} = \sigma_m^2 - 2r^2\sigma_m + r^2\sigma_m$$

$$= \sigma_m^2 - r^2\sigma_m$$

$$r(M, V) = \frac{E(D_mD_v)}{\sigma_m\sigma_v} \Rightarrow E(D_mD_v) = r\sigma_m\sigma_v \qquad \boxed{(1 - r^2)\sigma_m^2}$$

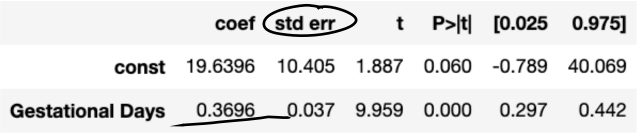
Name: ____ **SID:** ____________________

10. A dataset you have analyzed in Data 8 consists of observations on pairs of mothers and their newborns. The data were gathered to study the differences between the outcomes of mothers who smoked and those who didn't smoke. The sample of smokers was independent of the sample of nonsmokers.

The response variable was the birth weight of the newborn, measured in ounces. We will use only one predictor variable: the number of gestational days.

There were 715 nonsmokers. Assume the simple linear regression model for the relation between the birth weight and gestational days of the babies of these women. Let β_{ns} denote the slope of the true line in the model. Here is a summary of the results of the regression. It is a portion of the Python regression output familiar to you from class and assignments.

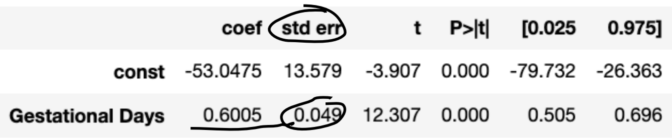
$\beta_{NS} = 0.3696$



	coef	std err	t	P>|t|	[0.025	0.975]
const	19.6396	10.405	1.887	0.060	-0.789	40.069
Gestational Days	0.3696	0.037	9.959	0.000	0.297	0.442

There were 459 smokers. Assume the simple linear regression model for the relation between the birth weight and gestational days of the babies of these women, and let β_s denote the slope of the true line in the model. Here is a summary of the results of the regression.

$\beta_S = 0.6005$



	coef	std err	t	P>|t|	[0.025	0.975]
const	-53.0475	13.579	-3.907	0.000	-79.732	-26.363
Gestational Days	0.6005	0.049	12.307	0.000	0.505	0.696

To compare the slopes of the two true lines, a data scientist wants to test the hypothesis $H_0: \beta_{ns} = \beta_s$ versus $H_A : \beta_{ns} < \beta_s$. ← One sided' two sided $\beta_{ns} \neq \beta_s$

a) [**3 points**] Construct a test statistic to test these hypotheses, and provide the observed value of the statistic.

TS: $\beta_s - \beta_{ns}$ ⇒ Large values of TS support the alternate H_a, β_s is larger than β_{NA}

b) [**7 points**] What is the exact or approximate distribution of your test statistic under H_0, and why? Provide numerical expressions for the exact or approximate parameters of the distribution.

Under H_0: $\beta_s = \beta_{NS}$ $E[\beta_s - \beta_{NS}] = 0$ (under H_0)

For large enough samples, invoke CLT, β_s, $\beta_{NS} \sim$ approx normal

— Treat SE as the SD for approx normal.

$\beta_s - \beta_{NS} \sim$ Approx Normal with mean $= 0$, var $= (0.049)^2 + (0.037)^2$

$Var(TS) = Var(\beta_s - \beta_{NS}) = Var(\beta_s) + Var(\beta_{NS})$