Math 241 Fall 2016
Final Examination

Solutions

Professor Adam Kapelner

December 15, 2016

Full Name ______________________________ Section (A or B) ____

Code of Academic Integrity

Since the college is an academic community, its fundamental purpose is the pursuit of knowledge. Essential to the success of this educational mission is a commitment to the principles of academic integrity. Every member of the college community is responsible for upholding the highest standards of honesty at all times. Students, as members of the community, are also responsible for adhering to the principles and spirit of the following Code of Academic Integrity.

Activities that have the effect or intention of interfering with education, pursuit of knowledge, or fair evaluation of a student's performance are prohibited. Examples of such activities include but are not limited to the following definitions:

Cheating Using or attempting to use unauthorized assistance, material, or study aids in examinations or other academic work or preventing, or attempting to prevent, another from using authorized assistance, material, or study aids. Example: using a cheat sheet in a quiz or exam, altering a graded exam and resubmitting it for a better grade, etc.

I acknowledge and agree to uphold this Code of Academic Integrity.

______________________________ ______________

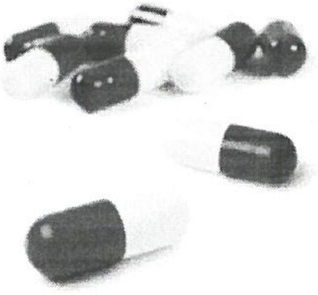
signature date

Instructions

This exam is 120 minutes and closed-book. You are allowed three pages (front and back) of a "cheat sheet." You may use a graphing calculator of your choice but *no cell phones.* Please read the questions carefully. If the question reads "compute," this means the solution will be a number otherwise you can leave the answer in choose, permutation, factorial, summation or any other notation which could be resolved to a number with a computer. I advise you to skip problems marked "[Extra Credit]" until you have finished the other questions on the exam, then loop back and plug in all the holes. I also advise you to use pencil. The exam is 100 points total plus extra credit. Partial credit will be granted for incomplete answers on most of the questions. $\boxed{\text{Box}}$ in your final answers. Good luck!

Problem 1

You are the biostatician for a drug company that is trying to develop a drug to treat skin cancer which has spread to lymph nodes (we will call this the "disease"). There are no drugs currently available for the disease so the company is developing a drug and wants to run a trial in the future.



The recovery rate right now for patients with the disease is 63%. Assume this is the truth known with certainty.

(a) [2 pt / 2 pts] Bob who has the disease wants to be part of the drug trial. Model Bob's recovery as of now as 1 if he recovers and 0 if he does not recover as a r.v. below. Be sure to indicate parameter value(s).

$X \sim$ Bernoulli $(p = 0.63)$

(b) [3 pt / 5 pts] 100 people just like Bob with the disease want to be part of the trial. Model the total number of people of those 100 who will recover as of now below as the r.v. X. Be sure to indicate parameter value(s).

$X \sim$ Binom $(n = 100, p = 0.63)$

(c) [4 pt / 9 pts] What two assumptions did you make about the 100 people in order to build the model in part (b)? Explain each in this context.

Assumption Name: Identically Distributed	Assumption Name: Independence
Discussion: Each subject has the same probability of recovering from the disease. This is assumed in the question header.	Discussion: The probability of one subject recovering is unaffected by whether or not others recovered / did not recover. This is plausible.

(d) [2 pt / 11 pts] Calculate the expected number of people of the 100 who will recover.

$E(X) = n \cdot p = 100 \cdot 0.63 = 63$

(e) [3 pt / 14 pts] Compute explicitly the probability as % percentage that exactly 63 people will recover. Round to two decimal places.

$P(X=63) = \binom{100}{63} .63^{63} .37^{37}$ [illegible] $= \boxed{8.24\%}$

(f) [5 pt / 19 pts] For those 100 people just like Bob with the disease, model the *proportion* of people who will recover below as the r.v. $\hat{P}$ and provide its *approximate* distribution. Be sure to indicate parameter values.

$\hat{P} \sim N\left(p, \left(\sqrt{\frac{p(1-p)}{n}}\right)^2\right) = N(0.63, .04828^2)$

(g) [2 pt / 21 pts] What theorem did you use to answer (h)? No need to discuss.

Central Limit Theorem

(h) [2 pt / 23 pts] Find $\mathbb{P}\left(\hat{P} = 0.63\right)$ below assuming the approximate distribution (and not the distribution of the r.v. in part b).

0 ($\hat{P}$ is a continuous r.v.)

(i) [2 pt / 25 pts] We will now start testing our drug. Denote the proportion of people who truly recover as the population parameter p. We are unsure if the drug works and want to be scientifically honest. Thus our null hypothesis will be that the proportion of people who recover when taking the drug will be *at most* the same as the population overall. Write this in our notation we used in class below.

$H_0: p \leq 0.63$

(j) [2 pt / 27 pts] Write the alternative hypothesis using the notation we learned in class.

$H_a: p > 0.63$

(k) [1 pt / 28 pts] What is the official name of this statistical test?

One-sided test of one proportion (or one-tailed or right-tailed)

(l) [2 pt / 30 pts] We will set $\alpha = 2.5\%$ since the scientific community is very conservative when it comes to new drugs. What is the probability of a Type I error?

$\alpha := P(\text{Type I err}) = 2.5\%$

(m) [3 pt / 33 pts] Explain what a Type I error would be in this context.

Concluding that the drug is effective in helping people recover from the disease when it's not.

(n) [3 pt / 36 pts] Explain what a Type II error would be in this context.

Concluding that there is not enough evidence to say the drug is effective but in reality it is.

(o) [3 pt / 39 pts] Discuss what you believe the cost of a type II error would be in this case.

Those who could be helped by the drug will not be helped. ~~Given that the prob. of recovery is 63% (unless it's a miracle drug~~, ~~the~~

(p) [5 pt / 44 pts] For the case where we have $n = 100$ people who take the drug, find the retainment region for H_0 in this hypothesis test. Round to three digits.

$$\text{Ret Region} = \left(-\infty, p + z_\alpha \sqrt{\frac{pq}{n}}\right] = \left(-\infty, .63 + 2 \cdot .04828\right]$$

$$= \left(-\infty, .727\right]$$

(q) [5 pt / 49 pts] Imagine 72 people of the 100 in the drug trial recovered. Run the hypothesis test, state the conclusion and then write a sentence which explains your conclusion in the context of drug trial.

$\hat{p} = \frac{72}{100}$ ∈ Retainment Region ⇒ Fail to reject H_0. There is not enough evidence to suggest this drug is effective in increasing the chance of recovery above the baseline recovery rate.

(r) [1 pt / 50 pts] Regardless of the conclusion of the hypothesis test, you *feel* deep down inside that the drug is effective in increasing the recovery rate of the disease. What type of error do you *feel* you made here? No explanation necessary.

Type II

(s) [3 pt / 53 pts] If you were to design another experiment to prove that the drug is effective, what would you do differently and why? Hint: increasing α is *not* the answer.

Increase n (the sample size). This will lower the probability of the Type II error in the future.

(t) [2 pt / 55 pts] Regardless of the conclusion of the hypothesis test, we will pretend that the point estimate you computed, i.e. $\hat{p} = 0.72$, represents the real probability of recovery under the drug. Denote R as the event of recovery and D as the event that the drug was taken. Thus R^C denotes the person did not recover and D^C denotes the person did not take the drug. What is $\mathbb{P}(R \mid D)$?

$P(R|D) = 0.72$

(u) [1 pt / 56 pts] Find $\mathbb{P}(R^C \mid D)$.

$P(R^c|D) = 1 - P(R|D) = 0.28$

(v) [6 pt / 62 pts] Let's say the drug is on the market and you estimate 20% of the people who have the disease take the drug. You see someone who recovers, what is the probability they took the drug?

$$P(D) = 0.2$$

$$P(D|R) = \frac{P(RD)}{P(R)} = \frac{P(R|D)\,P(D)}{P(RD) + P(RD^c)} = \frac{\overset{.72}{\overset{\shortparallel}{P(R|D)}}\ \overset{.2}{\overset{\shortparallel}{P(D)}}}{\underset{.72}{\underset{\shortparallel}{P(R|D)}}\ \underset{.2}{\underset{\shortparallel}{P(D)}} + \underset{\uparrow}{P(R|D^c)}\,P(D^c)_{=.8}}$$

$$= \boxed{.222}$$

0.63 (from question header) this is the recovery rate without the drug

Problem 2

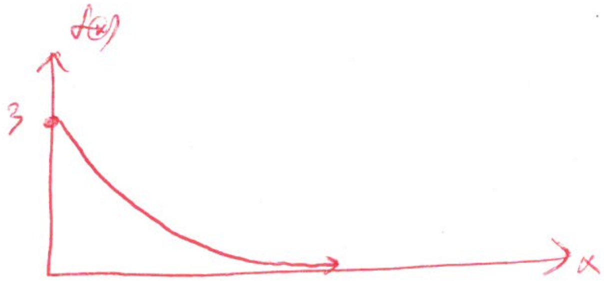
You are part of a call center for a very large company.

You have millions of clients but the probability each of them call in at any moment is very small. If there are 3,000,0000 clients and the probability any of them call in any second is 1 in 1,000,000 then you can say $\lambda = 3$ and model time in seconds until the next call as an *exponential* distribution.

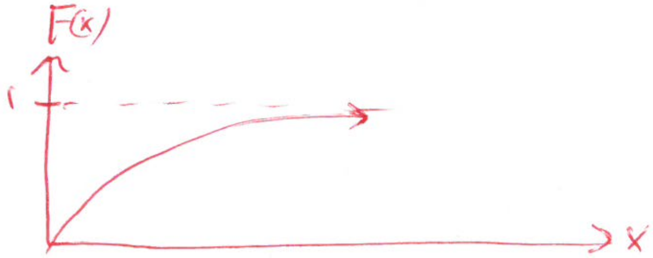
(a) [2 pt / 64 pts] Based on the description above, denote the time until the next call as r.v. X. Notate its distribution below as $X \sim$ something.

$$X \sim \text{Exp}(\lambda = 3)$$

(b) [4 pt / 68 pts] Provide a rough illustration of $f(x)$. Label the axes and label one point on the y-axis that is not 0. Do not worry about scale. 4 points



(c) [4 pt / 72 pts] Provide a rough illustration of $F(x)$. Label the axes and label one point on the y-axis that is not 0. Do not worry about scale. 4 points



(d) [4 pt / 76 pts] What is the probability we will wait more than 5 seconds for the next call that comes in? No need to compute explicitly.

$$P(X > 5) = 1 - F(x) = e^{-3 \cdot 5} = \boxed{e^{-15}}$$

(e) [4 pt / 80 pts] 4 seconds just elapsed. Given this information, what is the probability we will wait more than 5 seconds for the next call that comes in? No need to compute explicitly.

e^{-15} the exponential r.v has the memoryless property

Problem 3

Some theoretical exercises are below.

(a) [3 pt / 83 pts] On the homework you were introduced to the r.v. $X \sim \text{Poisson}(\lambda)$. The PMF is $p(x) = \lambda^x e^{-\lambda}/x!$. If $\text{Supp}[X] = \mathbb{N} \cup \{0\}$, write an expression for $\mathbb{E}[X]$ but *do not solve.*

$$E(X) = \sum_{x=0}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x!}$$

(b) [5 pt / 88 pts] On the homework you were given that if $X \sim \text{Poisson}(\lambda)$ then $M_X(t) = \underline{e^{\lambda(e^t-1)}}$. Find $\mathbb{E}[X]$ by any way you believe to be best.

$$e^{\lambda e^t} e^{-\lambda}$$

$$M'_X(t) = e^{-\lambda} e^{\lambda e^t} \lambda e^t$$

$$M'_X(0) = e^{-\lambda} e^{\lambda e^0} \lambda e^0 = \cancel{e^{-\lambda}} e^{\lambda} \lambda = \boxed{\lambda}$$

(c) [3 pt / 91 pts] Calculate

$$\int_2^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-2)^2} dx = \frac{1}{2}$$

↑

this is $f_X(x)$ for $X \sim N(2,1)$

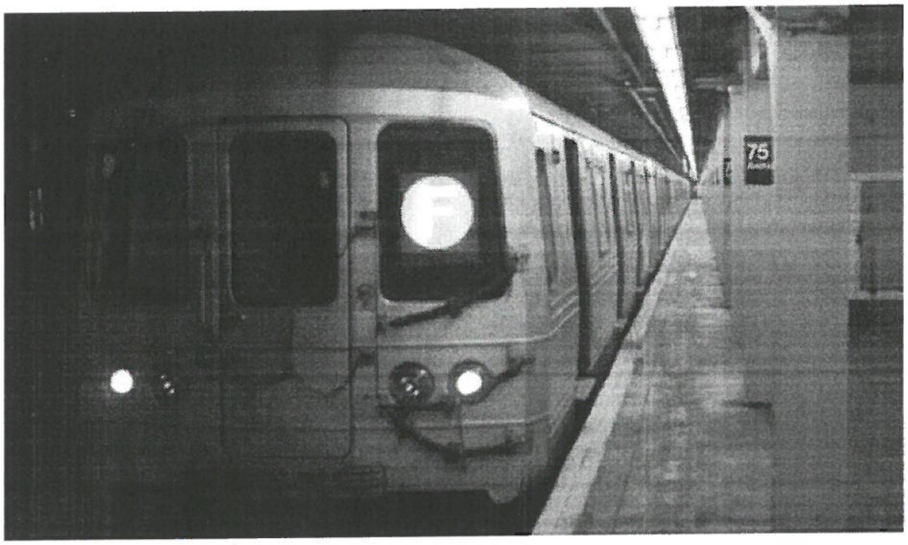
(d) [2 pt / 93 pts] Let $X \sim U(0, 1)$. Compute $\mathbb{SE}[X]$ to two decimal places.

$$SE(X) = \frac{1}{\sqrt{12}} = \boxed{.29}$$

(e) [2 pt / 95 pts] Assume $\mathbb{P}(A) > 0$. Compute $\mathbb{P}(A \mid A)$.

1

(f) [5 pt / 100 pts] You are riding the F train from Roosevelt Ave Jackson Heights to Forest Hills 71Av/Continental.



That means you are traveling for 6 stops. The traveling time of each stop can be modeled as a r.v. with distribution $\mathcal{N}(1,\ 0.2042^2)$ minutes and we will assume independent travel time for each stop. Find the probability it takes you more than 7 minutes to get from Roosevelt Ave Jackson Heights to Forest Hills 71Av/Continental.

$$X_1, X_2, X_3, X_4, X_5, X_6 \overset{iid}{\sim} N(1,\ .2042^2)$$

$$T = X_1 + X_2 + X_3 + X_4 + X_5 + X_6 \sim N\left(6 \cdot 1 + \left(\sqrt{6 \cdot .2042^2}\right)^2\right)$$

$$= N(6,\ .500^2)$$

$$P(T > 7) = P\left(\frac{T-6}{.500} > \frac{7-6}{.500}\right) = P(Z > 2) = \boxed{2.5\%}$$

(g) [5 pt / 105 pts] [Extra credit] Prove or disprove: $X \sim \mathrm{U}(a, b)$ has the memorylessness property.

(h) [5 pt / 110 pts] [Extra credit] If $X \sim \mathrm{Poisson}(\lambda)$, find $\mathbb{V}\mathrm{ar}[X]$.

(i) [5 pt / 115 pts] [Extra credit] Find the skewness of the standard normal.