Prof. David Draper
Department of
Applied Mathematics and Statistics
University of California, Santa Cruz

Name: Jeremy Lafond

AMS 131: Take-Home Test 1

Target due date: Fri 6 Jul 2018 *[330 total points]*

Here's a style guide for all of the written work in this class. In figuring out how to write up answers to homework (and quiz, and midterm, and final) problems, pretend the grader is sitting there with you and you're having a brief discussion with her/him on each question — that is, write down in a few sentences what you would say to someone to support your position. It's never enough in this class to just say "yes" or "10.3," even if the right answer is "yes" or "10.3"; you need to say "yes (or 10.3), because ... ." The right answer with no reasoning to support it, or the wrong reasoning, will get about half credit in this course, as will the wrong answer arrived at with a good effort. Leaving a problem or a part of a problem blank will get no credit.

1. *[55 points]* (public health) In 1972 a one-in-six random survey of the electoral roll — largely concerned with studying heart disease and smoking — was carried out in Whickham, a mixed urban and rural district near Newcastle upon Tyne in England. Twenty years later a follow-up study was conducted, with the results published in the journal *Clinical Endocrinology* in 1995.

The dataset summarized below in this problem pertains to the subsample of 1,314 women in the study who were classified in the original survey either as current smokers or as never having smoked. There were relatively few women (162) who had smoked but stopped, and only 18 whose smoking habits were not recorded; these women are not included in the data here. The 20-year survival status was determined for *all* the women in the original survey.

The *outcome* variable Y of interest here was mortality, recorded as dead or alive in 1992; the researchers regarded X, smoking behavior in 1972 (current smoker or never smoked), as the *supposedly causal factor (SCF)*, and they also measured the variable Z, age (18–64 or 65+) in 1972.

Several definitions and conclusions from the field of *experimental design* are relevant here:

- A *controlled experiment* is a study in which the investigators have control over X, in the sense that they assign participants to different groups defined by X (in this case, smoker (the so-called *treatment* group T) versus never-smoked (the *control* group C)); controlled experiments become *randomized controlled trials (RCTs)* when the investigators assign the participants to T and C at random. Investigations in which the researchers have no control over who gets into T and C — typically because the participants themselves choose which group they're in — are called *observational studies*.

- Two variables V and W are *associated* if as V increases W tends on average to increase or decrease, and vice versa; two variables that are not associated are *independent*. If both of the variables are *binary* — i.e., if they each have only two possible values, which may without loss of generality be taken as 0 and 1 — then $\{V$ and W are associated$\} \longleftrightarrow \{$as V moves from 0 to 1, $P(W = 1)$ increases or decreases$\}$.

- A *confounding factor (CF)* is a third variable Z, distinct from Y and X, that satisfies two properties:
  - Z and X are associated, and
  - Z and Y are associated.

  The conclusion that changes in X *cause* changes in Y (at least probabilistically) may validly be drawn from RCTs, but not necessarily from observational studies, because of CFs: an apparent relationship between X and Y in an observational setting may in fact have been caused, in whole or in part, by a CF Z.

  The best way to remove the possibility of a CF Z confounding your causal understanding is to *hold it constant*: to examine the relationship between X and Y separately for each possible value of Z — if you see something going on between Y and X in each of these comparisons, the association between X and Y cannot have been caused by Z, because it's been held constant. This holding-constant process is called *controlling for the CF Z*.

Table 1: Age Group 18–64

Mortality	Smoker? Yes	No	Total
Dead	93	69	162
Alive	440	470	910
Total	533	539	1072

Table 2: Age Group 65+

Mortality	Smoker? Yes	No	Total
Dead	46	161	207
Alive	3	32	35
Total	49	193	242

Table 3: Overall

Mortality	Smoker? Yes	No	Total
Dead	139	230	369
Alive	443	502	945
Total	582	732	1314

(a) Is the investigation described in this problem a controlled experiment or an observational study? If it's a controlled experiment, is it an RCT? Explain briefly. *[5 points]*

The investigation in this problem is clearly an observational study because the subjects themselves choose whether or not to smoke over the 20 year period and their mortality rates were merely compared to those who did not smoke during that period.

(b) Compute $P(\text{smoker})$ for a randomly chosen woman from Table 3, and compare this with your computation of $P(\text{smoker} \,|\, 18\text{–}64)$ for a woman picked at random from Table 1 and $P(\text{smoker} \,|\, 65+)$ for a women chosen at random from Table 2. Are age and smoking habits independent in this sample of 1,314 women, or does an association between these two variables exist in this data set (and if so, in which direction does the relationship go)? Explain briefly. *[10 points]*

$P(\text{smoker}) = \frac{582}{1314} \approx 44\%$ of those studied were smokers.

$P(\text{smoker} \,|\, 18-64) = \frac{533}{1072} \approx 50\%$ of those studied from ages 18-64 were smokers.

$P(\text{smoker} \mid 65+) = \frac{49}{242} \approx 20\%$ of those studied that were 65 or older were smokers.

Age and smoking habits are dependent as you would be $\frac{533-49}{1314} \approx 37\%$ more likely to be between the ages of 18-64 as a smoker in this data set. The results from the 18-64 subset will skew the data because they are represented at more than 4:1 to the older age group. In this data-set, this relationship trends in the negative direction because an increase in age correlates to a DECREASE in smoking.

(c) For a woman chosen at random from the 1,314 in Table 3, compute $P(\text{dead}), P(\text{dead}|\text{ smoker})$, and $P(\text{dead}|\text{nonsmoker})$. Does this establish an association between smoking and mortality for these women, and if so in which direction? Is the direction of this relationship surprising? Does this prove that smoking *causes* higher or lower mortality for these women? Explain briefly. *[10 points]*

$P(\text{dead}) = \frac{369}{1314} \approx 28\%$ of those studied died after 20 years.

$P(\text{dead}|\text{smoker}) = \frac{139}{582} \approx 24\%$ of those studied that were smokers died after 20 years.

$P(\text{dead}|\text{nonsmoker}) = \frac{230}{732} \approx 31\%$ of those studied that were NOT smokers died after 20 years.

This shows that there is a $\frac{139-230}{1314} \approx -0.069$ or a 7% less likely chance of dying as a woman smoker after 20 years. Thus this relationship trends in the slightly NEGATIVE direction because an increase in smokers indicates a slight decrease in mortality.

The direction of this relationship is surprising because one would assume smoking for 20 years would certain be more likely to kill you or at the very least be worse than NOT smoking. This does not PROVE that smoking causes higher or lower mortality for these women because there are external factors that can affect the complexity of their mortality beyond just cigarettes; such as sample-space, health habits, and health history. It merely indicates some CORRELATION or dependence.

(d) By looking at Tables 1 and 2 and computing any relevant probabilities (unconditional or conditional), explain why age is a CF in studying the relationship between smoking and mortality for these 1,314 women. Separately for each of the age groups {18–64} and {65+} (i.e., for women chosen randomly from Tables 1 and 2), compute $P(\text{dead}), P(\text{dead}|\text{smoker})$, and $P(\text{dead}|\text{nonsmoker})$. How can you explain the fact that, when age is taken into consideration, the association between smoking and mortality for these women goes in the opposite direction than in part (b)? *[15 points]*

$P(\text{dead}) = \frac{369}{1314} \approx 28\%$ of those studied died after 20 years in the entire data set.

$P(\text{dead 18-64}) = \frac{162}{1072} \approx 15\%$ of those studied between the ages of 18-64 died after 20 years.

$P(\text{dead 65+}) = \frac{207}{242} \approx 86\%$ of those studied that were 65 or older died after 20 years.

$P(\text{dead 18-64}|\text{smoker}) = \frac{93}{533} \approx 17\%$ of those studied that were smokers between the ages of 18 and 64 died after 20 years.

$P(\text{dead 65+}|\text{smoker}) = \frac{46}{49} \approx 94\%$ of those studied that were smokers 65 or older died after 20 years.

$P(\text{dead 18-64}|\text{nonsmoker}) = \frac{69}{539} \approx 13\%$ of those studied that were non-smokers between the ages of 18 and 64 died after 20 years.

$P(\text{smoker}) = \frac{582}{1314} \approx 44\%$ of those studied were smokers in the entire data set.

$P(\text{smoker} \,|\, 18-64) = \frac{533}{1072} \approx 50\%$ of those studied from ages 18-64 were smokers.

$P(\text{smoker} \,|\, 65+) = \frac{49}{242} \approx 20\%$ of those studied that were 65 or older were smokers.

This indicates there is more going on that what is being studied as it is plausible to assume that people 18-64, especially on the lower end, may easily go 20 years without dying from smoking; whereas, someone who is 65+ is statistically more likely to have died after a 20 year study regardless of whether or not they smoked. Additionally, it is plausible to assume that as age varies, the likelihood of smoking may increase which in turn may increase mortality rate, regardless of prior health. The reversal of these variable's directional relationship is likely due to various possibly unintentional biases that occur in the collected data. For example, in this data set for people of age 65+, more people died that were non-smokers than of those that were. Additionally, a very small percentage of people of age 18-64 actually ended up dying after smoking for 20 years. Neither of these metrics are necessarily factual or even statistically relevant in a larger data set; however, they do come up in THIS data set.

(e) If the relationship between X and Y changes direction when a CF Z is controlled for, the situation is referred to as a *Simpson's Paradox* (named for the British statistician Edward Simpson (1922–), who wrote about it in 1951, although the phenomenon had been known about for a long time before that). By examining the directions of the relationships between $(X, Y), (X, Z)$ and (Y, Z), explain intuitively why the Simpson's Paradox occurred here. Which conclusion about the effects of smoking on mortality is more trustworthy, the one in part (c) or its opposite in part (d)? Explain briefly. *[15 points]*

The 18-64 data is represented in the aggregate data by more than a 4:1 ratio compared to people 65 or older. This gives the data a bias which results in a somewhat surprising result when considering age. When holding age constant, we can show a Simpson's paradox occurs as the relationship between smoking and mortality reverses direction. This is merely because the aggregate data skews the results counter-intuitively because certain CF's such as age are effectively averaged when combined rather than being considered and appropriately conditioned. Thus, holding age constant in part c results in aggregate data that doesn't tell the whole story; whereas, in part d considering age shows us that smoking is still more likely to kill you after 20 years than not smoking which is far more trustworthy as it at least considers age as a CF.

2. *[70 points]* (gambling) To solve this problem I need to tell you about *hypergeometric* probabilities (we'll revisit this topic in the unit on discrete distributions). Suppose that you're considering a finite population of individuals, each of which can be classified in one of two ways (e.g., black and green balls in an urn, or Democrats and Republicans among people who stick to the major political parties). Let the total number of individuals in the population be N, of which N_1 are of type 1 and N_2 of type 2 (with $N_1 + N_2 = N$). If you now take a simple random sample (without replacement) of size n from this population, what's the probability that you'll end up with exactly n_1 individuals of type 1 and n_2 of type 2?

Evidently there are some restrictions here: $0 \le n_1 \le N_1$, and $0 \le n_2 \le N_2$, and $n_1 + n_2 = n$. From our discussion of permutations and combinations, you can immediately see that there are $\binom{N}{n}$ possible simple random samples, all of which are equally likely, and furthermore that there

Table 4: *The nine ways to win in Powerball and the associated "odds," as stated on the Powerball website.*

Match	Prize	"Odds"
All five whites and the red	Grand Prize	1 in 292,201,338.00
All five whites	\$1,000,000	1 in 11,688,053.52
Four whites and the red	\$50,000	1 in 913,129.18
Four whites	\$100	1 in 36,525.17
Three whites and the red	\$100	1 in 14,494.11
Three whites	\$7	1 in 579.76
Two whites and the red	\$7	1 in 701.33
One white and the red	\$4	1 in 91.98
The red	\$4	1 in 38.32

are $\left(\begin{array}{c} N_1 \\ n_1 \end{array}\right)$ ways to choose the n_1 type–1 individuals and $\left(\begin{array}{c} N_2 \\ n_2 \end{array}\right)$ ways to end up with exactly n_2 individuals of type 2. Thus

$$P(n_1 \text{ type–1 individuals and } n_2 \text{ type–2 individuals}) = \frac{\left(\begin{array}{c} N_1 \\ n_1 \end{array}\right)\left(\begin{array}{c} N_2 \\ n_2 \end{array}\right)}{\left(\begin{array}{c} N \\ n \end{array}\right)}\,. \tag{1}$$

OK, now we can get on with the problem, which makes extensive use of these hypergeometric probabilities.

Powerball is a national lottery in the U.S. with drawings every Wednesday and Saturday night at 10.59pm Eastern time. The money left over after paying the winners is used by each state for projects designated by the legislatures, such as helping to fund K–12 education. In the *Powerball* game, five numbered white balls are drawn — in a manner certified by the lottery to be as close as humanly possible to *at random without replacement* — from a drum containing white balls numbered from 1 to 69, and one red ball is then also drawn at random from a second smaller drum that has 26 numbered red balls in it. Table 4 below lists the nine ways you can win and the "odds" against you. Each play of the game cost \$2, and you can play as many times as you like.

There are several errors on the *Powerball* website. The first error is that when the *Powerball* people said "Odds" in Table 4 what they really meant was "the probability of occurrence, expressed as a fraction $\frac{1}{x}$." Another error was present in something the *Powerball* website further stated:

> The overall "odds" of winning a prize are 1 in 24.87. The "odds" presented here are based on a \$2 play (rounded to two decimal places) *[quotes added].*

(a) Explain why the "odds" value in the first row of Table 4 was not 1 in $(69 \cdot 68 \cdot \cdots \cdot 65 \cdot 26) =$ 35,064,160,560, and why the stated "odds" value was essentially correct. *[10 points]*

P(Grand Prize) = Getting all 5 white balls AND getting the red ball.

$$\text{P(Grand Prize)} = \frac{1}{\left(\begin{array}{c}\textbf{69 possible numbered balls}\\ \textbf{5 drawn white balls}\end{array}\right)} \text{ AND } \frac{1}{\left(\begin{array}{c}\textbf{26 possible numbered balls}\\ \textbf{1 drawn red ball}\end{array}\right)}$$

which is:

$(\left(\begin{array}{c}\textbf{69}\\ \textbf{5}\end{array}\right) * 26)^{-1} = \frac{1}{292{,}201{,}338.00}$ or 1 in 292,201,338.00

(b) Explain why the "odds" value for the Second Prize of \$1,000,000 was not $\left(\begin{array}{c}69\\ 5\end{array}\right)^{-1} = 1$ in 11,238,513, and show that the lottery people got the correct answer. *[10 points]*

We can use part a to justify this.

We know the odds of getting all 5 balls white is:

$$\left(\begin{array}{c}\textbf{69 possible numbered balls}\\ \textbf{5 drawn white balls}\end{array}\right)$$

but we must also take into account that the player DID NOT get the red ball.

Thus:

$$\text{P(Second Prize)} = (\left(\begin{array}{c}\textbf{69}\\ \textbf{5}\end{array}\right)^{-1} * \frac{\textbf{25 non-red balls}}{\textbf{26 total balls}}) \approx \frac{1}{11{,}688{,}053.52} \text{ or 1 in } 11{,}688{,}053.52$$

(c) For $(k = 0, 1, \ldots, 5)$, explain why the following formulas are correct:

$$P(k \text{ whites and the red}) = \frac{\left(\begin{array}{c}5\\ k\end{array}\right)\left(\begin{array}{c}64\\ 5-k\end{array}\right)\left(\begin{array}{c}1\\ 1\end{array}\right)\left(\begin{array}{c}25\\ 0\end{array}\right)}{\left(\begin{array}{c}69\\ 5\end{array}\right)\left(\begin{array}{c}26\\ 1\end{array}\right)} \text{ and}$$

$$P(k \text{ whites (and not the red)}) = \frac{\left(\begin{array}{c}5\\ k\end{array}\right)\left(\begin{array}{c}64\\ 5-k\end{array}\right)\left(\begin{array}{c}1\\ 0\end{array}\right)\left(\begin{array}{c}25\\ 1\end{array}\right)}{\left(\begin{array}{c}69\\ 5\end{array}\right)\left(\begin{array}{c}26\\ 1\end{array}\right)}. \quad (2)$$

This formula is essentially saying:

For the following let:

X = ways of drawing k white balls from 5 total white balls.

Y = ways of drawing 5-k non-white balls from 64 total non-white numbered balls

Z = ways to draw red ball from 26 total red balls.

Z_0 = ways to NOT draw red ball from 26 total red balls.

P(k whites and the red) = . . .

$$\frac{\text{(X)}\mathbf{AND}\text{(Y)}\mathbf{AND}\text{(1)}\mathbf{AND}\text{(Z)}}{\text{(combos drawing all 5 white balls)}\mathbf{AND}\text{(combos drawing a bonus ball)}}$$

P(k whites (and not the red)) = . . .

$$\frac{(X)\mathbf{AND}(Y)\mathbf{AND}(1)\mathbf{AND}(Z_0)}{\text{(ways of drawing all 5 white balls)}\mathbf{AND}\text{(ways of drawing a bonus ball)}}$$

The only difference between the two is the combinations of drawing the red ball in the numerator(Z and Z_0. When k = 5 all white balls are drawn and the numerator becomes 1 over the possibility space of combinations of draws; unless the red ball is not drawn in which case the numerator becomes 25. The $\binom{1}{1}$ is always 1 along with $\binom{1}{0}$ so they always cancel out.

Use these formulas to verify the rest of the "odds" entries in Table 4. *[30 points]*

P(4 whites and the red) =

$$\frac{\binom{5}{(4)}\binom{64}{1}\binom{1}{1}\binom{1}{1}}{\binom{69}{5}\binom{26}{1}} = \frac{320}{292{,}201{,}338} \approx 1 \text{ in } 913{,}129.18$$

P(4 whites (and not the red) =

$$\frac{\binom{5}{(4)}\binom{64}{1}\binom{1}{1}\binom{25}{1}}{\binom{69}{5}\binom{26}{1}} = \frac{320*25}{292{,}201{,}338} \approx 1 \text{ in } 36{,}525.17$$

P(3 whites and the red) =

$$\frac{\binom{5}{(3)}\binom{64}{2}\binom{1}{1}\binom{1}{1}}{\binom{69}{5}\binom{26}{1}} = \frac{20160}{292{,}201{,}338} \approx 1 \text{ in } 14{,}494.11$$

P(3 whites (and not the red) =

$$\frac{\binom{5}{(3)}\binom{64}{2}\binom{1}{}\binom{25}{}}{\binom{69}{5}\binom{26}{1}} = \frac{20160*25}{292{,}201{,}338} \approx 1 \text{ in } 579.76$$

P(2 whites and the red) =

$$\frac{\binom{5}{(2)}\binom{64}{3}\binom{1}{}\binom{1}{}}{\binom{69}{5}\binom{26}{1}} = \frac{416640}{292{,}201{,}338} \approx 1 \text{ in } 701.33$$

P(1 white and the red) =

$$\frac{\binom{5}{(1)}\binom{64}{4}\binom{1}{}\binom{1}{}}{\binom{69}{5}\binom{26}{1}} = \frac{5*635376}{292{,}201{,}338} \approx 1 \text{ in } 91.98$$

P(The red) =

$$\frac{\binom{5}{(0)}\binom{64}{5}\binom{1}{}\binom{1}{}}{\binom{69}{5}\binom{26}{1}} = \frac{1*7624512}{292{,}201{,}338} \approx 1 \text{ in } 38.32$$

(d) Show that the lottery people were right when they said that the overall "odds" of winning a prize are 1 in about 24.87, and explain why the statement "The "odds" presented here are based on a $2 play (rounded to two decimal places)" initially sounds ridiculous but can be made correct with the insertion of a single word. *[10 points]*

The odds of winning any prize would be the sum of the odds of winning each prize:
$\frac{1}{292291338} + \frac{1}{11688053.52} + \frac{1}{913129.18} + \frac{1}{36525.17} + \frac{1}{14494.11} + \frac{1}{579.76} + \frac{1}{701.33} + \frac{1}{91.98} + \frac{1}{38.32} \approx 0.04021623$
or ≈ 1 in 24.87.

The second statement sounds ridiculous because it should be clarified that the ODDS are rounded to two decimal places.

(e) Suppose that T tickets were bought across the entire U.S. in a given week, that no one was clairvoyant or otherwise privy to knowledge about the winning numbers, and (for simplicity) that everybody made their lottery picks independently of everybody else. In the drawing on 30 Jul 2016, for which the Grand Prize (or *jackpot*) was \$487 million, it could be estimated from historical records on numbers of tickets purchased as a function of jackpot size that T was about 182.9 million. Show that the chance of at least one Grand Prize winner on this occasion was about 46.5%. (In actuality, one winning ticket was sold in a supermarket in Raymond, New Hampshire.) *[10 points]*

If we know the odds of NOT winning are $1 - \frac{1}{292,201,338}$

and we know the odds of 182.9 million people not winning is:

$(1 - \frac{1}{292,201,338}) * (1 - \frac{1}{292,201,338}) * (1 - \frac{1}{292,201,338}) * ... * (1 - \frac{1}{292,201,338})$

multiplied 182.9 million times, we get:

$(1 - \frac{1}{292,201,338})^{182,900,000} \approx 0.535$ or a 53.5% chance of NOT winning.

We can extend this logic to show that there is then a 1 - 53.5% or 46.5% chance of winning.

3. *[30 points]* (logic and Bayes's Theorem) Here's a small fictitious drama with five actors: three people — A, B and C — on death row; the governor, who has chosen one of them at random to be pardoned; and a warden in the prison, who knows the identity of the person the governor picked but isn't allowed to tell A, B or C who the lucky person will be. Person A now speaks to the warden, as follows.

> Please tell me the name of one of the other prisoners who's *not* going to be pardoned — no harm done, since you won't be identifying the lucky person. Let's agree on these rules: if B will be pardoned, you say C; if C will get the pardon, you say B; and if I'm the lucky person, you toss a 50/50 coin to decide whether to say B or C.

The warden thinks it over and says "B won't get the pardon." This is good news to A, because he secretly didn't believe that the warden's statement contains no information relevant to him: he thinks that, given what the warden said, his chance for the pardon has gone up from $\frac{1}{3}$ to $\frac{1}{2}$. Use Bayes's Theorem to show that A's reasoning is incorrect, thereby working out whether there *was* information in what the warden said that's relevant to A's probability of being pardoned. *[25 points]*

After the warden tells A there are technically only two elements remaining but they are not "equally probable."

We have a number of possible outcomes before the warden speaks:

A = [A will be pardoned] Pr(A) = 1/3

B = [B will be pardoned] Pr(B) = 1/3

C = [C will be pardoned] Pr(C) = 1/3

But then the warden says: "B won't get the pardon."

Lets call this E = evidence or information given by the warden.

We can now look at A's probability of being pardoned, given that E is true, with Bayes theorem below:

$$Pr(A|E) = \frac{Pr(A)*Pr(E|A)}{Pr(A)Pr(E|A)+Pr(B)Pr(E|B)+Pr(C)Pr(E|C)} = \frac{\frac{1}{3}*\frac{1}{2}}{\frac{1}{3}*\frac{1}{2}+\frac{1}{3}*0+\frac{1}{3}*1} = \frac{1}{3}$$

Obviously if the warden is telling the truth, B will not be pardoned.

The above shows that A and C are not equally probable to be pardoned and prisoner A's reasoning was incorrect because if Pr(A—E) = 1/3 and Pr(B—E) = 0, then Pr(C—E) = 1 - 1/3 = 2/3.

4. *[80 points]* (optimal hiring strategy) Here's an oversimplified version of a common problem for personnel managers that nevertheless contains elements of realism. You've advertised an open position in your organization, and $n \geq 1$ candidates have put their names forward for consideration. You want to hire the best candidate, but before interviewing any of them — suppose that their resumes don't provide strong information with which to create a ranking — each of them in your judgment has equal probability $\frac{1}{n}$ of being the best. It would be great if you could just interview all n of them, because you would then know for sure who's best, but (as with the tech sector, for example) this is a fast-moving hiring environment (by the time you get to the end and figure out that (say) candidate 3 is best, that person has probably already taken another job), so you need to be adaptive. Here are the ground rules:

- Once the interviews start, you can rank the candidates you've already seen, but you'll have no information about how the remaining candidates will fit into the ranking; and
- After each interview (because of the fast-moving environment), you either immediately hire the candidate you've just seen (and stop the interviewing process) or let that candidate go, with no opportunity to call her or him back.

Here's the adaptive strategy you've decided to use:

- To get information about the quality of the applicant pool, you pick a number $0 \leq r < n$, and you (callously) interview the first r candidates without intending to hire any of them.
- Beginning with the next candidate $(r + 1)$, you continue interviewing until the current candidate is the best you've seen so far, at which point you stop the interviewing process and hire that candidate.
- If none of the candidates from $(r + 1)$ to n is best, you just throw up your hands and hire candidate n.

The goals in this problem are twofold: to compute the probability that you hire the best candidate with this strategy, and to choose r to maximize this probability. Let A = (you hire the best candidate) and B_i = (the best candidate is person i in the interviewing sequence).

(a) For any $i > r$, show that the probability that {the best candidate among the first i people interviewed is one of the first r people} is $\frac{r}{i}$. *[10 points]*

We know that $Pr(B_i) = \frac{1}{i}$ because there is an equal chance that any candidate i is the best in the sequence.

To be successful, we need to make sure that out of the first i-1 candidates that one of the first r candidates is the best amongst the first i - 1 candidates.

$Pr(B_i|A) = Pr(B_i)(Pr(A)|B_i) = \frac{1}{i} * \frac{r}{i-1} = \frac{r}{i} * \frac{1}{i-1}$

Thus, $Pr(A) = \frac{r}{i}$ for any $i > r$

(b) Explain why $P(A \mid B_i) = 0$ for $i \leq r$, and (hard) show that $P(A \mid B_i) = \frac{r}{i-1}$ for $i > r$. (*Hint:* it helps to define the events C_i = (you keep interviewing until you see candidate i).) *[15 points]*

$P(A|B_i) = 0$ for $i \leq r$ because if the best candidate is person i in the sequence, they would be interviewed before anyone was intended to be hired (i would be among the first r candidates)

$P(A|B_i) = \frac{r}{i-1}$ for $i > r$ because we know that out of the first i - 1 candidates, at least one of the first r candidates must be best in order to hire successfully.

(c) Having specified a value of r before interviewing begins, let $p_r = P(A)$ with the chosen r value, and show that

(i) $p_0 = \frac{1}{n}$,

If r is zero there is an equal chance that any of the n candidates would be the best hire so there is 1 out of n or $\frac{1}{n}$ odds of picking the best person for the job.

and

(ii) for $0 < r < n$, $p_r = \frac{r}{n} \sum_{i=r+1}^{n} \frac{1}{i-1}$. (*Hint:* Use the results from part (b).)

If we know that $r > 0$ and $r < n$ and we know that i is the best candidate then we can multiply $P(B_i)$ with $P(A|B_i)$.

We know that $P(B_i) = \frac{1}{i}$

We also know from part a and b above that $P(A|B_i) = \frac{r}{i-1}$ for $i > r$ and 0 for $i \leq r$

Combining the results gives us a general $P_r = \frac{1}{i} * \frac{r}{i-1}$

Summing these probabilities together, we are left with:
P_r for $0 < r < n = (\sum_{i=1}^{r} 0 \sum_{i=r}^{n} P(A|B_i)) * \frac{1}{i} = \frac{1}{n} * \sum_{i=r+1}^{n} \frac{r}{i-1} = \frac{r}{n} \sum_{i=r+1}^{n} \frac{1}{i-1}$

[15 points]

(d) On the way to finding the optimal value of r, define $q_r = (p_r - p_{r-1})$ for $r = 1, \ldots, (n-1)$ and show that q_r is a strictly decreasing function of r for $r > 0$. *[15 points]*

As n approaches infinity we get what resembles a Riemann integral.

Given the above result:

$P(r) = \frac{r}{n} \sum_{i=r+1}^{n} \frac{1}{i-1}$

We can approximate the integral as a collection of Riemann rectangles of width $1-$
We can also show that: q_r as n approaches infinity $\approx -\frac{r}{n} * ln(\frac{r}{n})$

With some further cleverness we can substitute $\frac{r}{n}$ with x and differentiate using the product rule:

$\frac{d}{dx} = -x * ln(x) = -ln(x) - 1$

Substituting back in we get: $-ln(\frac{r}{n}) - 1$

and since: $-ln(\frac{r}{n}) < 1$,

$-ln(\frac{r}{n}) - 1 < 0$,

Therefore, q(r) is strictly decreasing because $q\acute{(}r) < 0$.

(e) Use (d) to show that the value of r that maximizes p_r is the largest r such that $q_r > 0$. (*Hint:* For $r > 0$, from the definition of q_r, it helps to write $p_r = p_0 + \sum_{i=1}^{r} q_i$.) *[10 points]*

In order to maximize P_r, we can check the value at $P_r' = 0$:

$P_r' = -ln(x) - 1$
$-ln(x) - 1 = 0$
$e^{ln(x)} = e^{-1}$
$x = \frac{1}{e}$
$r = \frac{n}{e}$

By showing $P_r'' < 0$, we can show P_r is at a maximum.
$P_r'' = -\frac{1}{x}$
$-\frac{1}{\frac{1}{e}} = -e < 0$

We have found our maximum, now after plugging the value back into q_r:

$q_r(\frac{n}{e}) \approx -\frac{(\frac{n}{e})}{n} * ln(\frac{(\frac{n}{e})}{n}) = -\frac{1}{e} * ln(\frac{1}{e}) = \frac{1}{e} > 0$

We can see that the value of r that maximizes p_r (r = $\frac{n}{e}$) is indeed the largest r for $q_r > 0$

(f) Use (e) to find the best value of r when $n = 10$ and the resulting optimal value of p_r. Does the adaptive hiring strategy examined in this problem look good to you? Explain briefly. *[15 points]*

If n = 10, we know that the optimal value for r is $r = \frac{(10)}{e}$.

The resulting optimal value of r is $\approx$ 3.679; however, for practical purposes(because there's no such thing as .679 of a person) we would round down and begin hiring after the 3rd interview rather than the 4th. This strategy looks good if you have to interview a massive amount of people for a job with very measurable skills. At the same time, it seems a little ridiculous to arbitrarily toss a large portion of your hiring pool. I guess time is money though. This approach does seem to be very stable because as n grows, probabilistic certainty is maintained!

(Remarkable fact (not part of what you're asked to show in this problem), for those of you who like to think about math: it turns out, weirdly, that for $0 < r < n$, $\sum_{i=r+1}^{n} \frac{1}{i-1} = \Psi(n) - \Psi(r)$, where $\Psi(x) \triangleq \frac{d}{dx} \ln \Gamma(x)$ is the *digamma* function.)

5. *[95 points]* (portfolio management) You're a portfolio manager at a hedge fund, meaning that you make investment decisions about other people's money. Naturally enough, the people whose money you're investing want to know how risky your investment decisions are. To this end, a standard metric in the investment industry is the *Value at Risk (VaR)* of a portfolio. Letting the continuous random variable X represent the (unknown) change in value of the portfolio in question over a fixed time horizon, for example one month, suppose that the PDF of X — in your judgment, based on the best current information — concentrates most of its probability on the positive part of the real number line $\mathbb{R}$; in other words, in your judgment the portfolio will probably increase in value over the next month but may instead decrease. Let $Y = -X$, so that Y is the pessimistic side of the X coin (so to speak): if $X > 0$ with high probability then $Y < 0$ with the same high probability. To quantify the term "high," let α be a small positive number, so that $(1 - \alpha)$ is close to 1; then the *VaR* of the portfolio is defined to be the $(1-\alpha)$ quantile of the distribution of Y. The tough part of implementing this idea is pinning down the PDF of X; in this problem you'll examine how sensitive the *VaR* is to this PDF specification. Let's take $\alpha = 0.01$ in what follows; this is a frequent choice in calculating VaR values.

Note that a portfolio based on sensible trading of stocks on the New York Stock exchange will typically appreciate at a rate of about 7% per year, which translates to a rate of about 0.6% per month; this implies that, if the portfolio is expected to increase in value by about $10 million in the next month, which is consistent with the PDFs in part (b) below, the total value of the portfolio at the beginning of the month was about $1.7 billion.

(a) Let $F_X(x)$ and $F_Y(y)$ be the CDFs for the random variables X and Y, respectively. By definition, to say that $VaR = v$ means that $F_Y(v) = (1-\alpha)$. Work out how F_Y depends on F_X, and use this to show that $VaR = -F_X^{-1}(\alpha)$. *[10 points]*

If the VaR of the portfolio = v and $F_Y(v) = (1-\alpha)$ then v relates to the 0.99 quantile of the distribution of Y this shows that any change in $F_X(v)$ in either direction will cause F_Y to drop as Y is almost certainly less than or equal to v. This also shows that v is closely related to the 0.01 quantile of the distribution X. The 0.01 quantile α has the property that:

$Pr(X < \alpha) = 0.01$.

But $Pr(X < \alpha) = Pr(Y > -\alpha) = 1 - Pr(Y \leq -\alpha)$.

Thus $-\alpha$ is the 0.99 quantile of Y.

Since $-\alpha$ is the 0.99 quantile of Y and $1 - \alpha$ is the 0.01 quantile of the distribution of X:

$VaR = v = -F_X^{-1}(\alpha)$

(b) Suppose that in your judgment the support of X is $[-10, +20]$, where the units are in millions of dollars, and that you think that the PDF of X should be monotonically increasing on its support (in other words, if $20 \geq x_2 > x_1 \geq -10$ then in your view $P(X \doteq x_2) > P(X \doteq x_1)$, where $\doteq$ means *is approximately equal to*). For each of the un-normalized PDFs below, compute the normalizing constant, make a rough (or refined) sketch of the PDF, compute the CDF, and work out the resulting *VaR*. Given the initial portfolio value of about \$1.7 billion, does *VaR* seem highly sensitive to you across this range of PDF shapes? Explain briefly. (Below, c is a generic constant [real number], not necessarily equal as you move from (i) to (ii) to (iii), and $f_X(x)$ is nonzero only on $[-10, +20]$.)

(i) (triangular [linear]) $f_X(x)$ is linear with positive slope and passes through the points $(-10, 0)$ and $(20, c)$.

(ii) (quadratic) $f_X(x)$ is quadratic and achieves its minimum at the point $(-10, 0)$.

(iii) (exponential) $f_X(x) = c \exp\left(\frac{\lambda x}{10}\right)$, with $c = 0.005569078782$ and $\lambda = 1.717229651$ (chosen to make the PDF just slightly positive (0.001) at $x = -10$).

[45 points]

(c) What if (i), (ii) and (iii) are all unrealistically cheerful about the fate of the portfolio over the next month? Repeat (b) with the two PDFs below. How would you describe the *VaR*'s sensitivity (e.g., moderately insensitive, or highly sensitive, or ...) across the entire range of the five PDFs you examined? Explain briefly.

(i) (uniform [constant]) X Uniform$(-10, 20)$.

(ii) (triangular [linear]) $f_X(x)$ is linear with negative slope and passes through the points $(-10, c)$ and $(20, 0)$.

[40 points]