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

STAT 131:

Take-Home Test 1 Target due date: at canvas.ucsc.edu by 11.59pm Sun 26 Apr 2020 [330 total points]

Name: Christy Yuen
Note: Attended office hours and got lots of inspiration

Here's a style guide for all of the written work in this class. In figuring out how to write up answers to quizzes and take-home tests, pretend that 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.

This test is to be entirely your own efforts; do not collaborate with anyone or get help from anyone but me or our TAs (Isabelle Grenier, Hyotae Kim, DK = Daniel Kirsner, and Arthur Lui). The intent is that the course lecture notes and readings should be sufficient to provide you with all the guidance you need to solve the problems posed below, but you may use other written materials (e.g., the web, journal articles, and books other than those already mentioned in the readings), provided that you cite your sources thoroughly and accurately; you will lose (substantial) credit for, e.g., lifting blocks of text directly from wikipedia and inserting them into your paper without full attribution.

If it's clear that (for example) two people have worked together on a part of a problem that's worth 20 points, and each answer would have earned 16 points if it had not arisen from a collaboration, then each person will receive 8 of the 16 points collectively earned (for a total score of 8 out of 20), and I reserve the right to impose additional penalties at my discretion. If you solve a problem on your own and then share your solution with anyone else (because people from your cultural background routinely do this, or out of pity, or kindness, or whatever motive you may believe you have; it doesn't matter), you're just as guilty of illegal collaboration as the person who took your solution from you, and both of you will receive the same penalty. This sort of thing is necessary on behalf of the many people who do not cheat, to ensure that their scores are meaningfully earned. In an AMS graduate class I taught in 2012, five people failed the class because of illegal collaboration; don't let that happen to you.

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 socalled 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]

This is an observational study as the researchers don't have control over what the women did over the course of the 20 years.

(b) Compute P(smoker) for a randomly chosen woman from Table 3, and compare this with your computation of P(smoker| 18–64) for a woman picked at random from Table 1 and P(smoker| 65+) for 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]

Table 3: P(smoker) = Smoker total/ total = $\frac{582}{1314} = 0.4429 = 44.3\%$

Table 1: P(smoker| 18–64) = Smoker total/ total = $\frac{533}{1072} = 0.4972 = 49.7\%$

Table 2: P(smoker| 65+) = Smoker total/ total = $\frac{49}{242} = 0.2024 = 20.2\%$

There is an association between age and smoking habits in this sample of 1,314 women. As age goes up, the probability of smoking goes down. **Age ↑ P(Smoker) ↓**

(c) For a woman chosen at random from the 1,314 in Table 3, compute P(dead), P(dead| smoker), and P(dead|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]

Table 3:

P(dead) = $\frac{369}{1314} = 0.2808 = 28.1\%$

P(dead| smoker) = $\frac{139}{582} = 0.2388 = 23.9\%$

P(dead|nonsmoker) = $\frac{230}{732} = 0.3142 = 31.4\%$

P(dead| smoker) < P(dead|nonsmoker) This is shocking because smoking lowers mortality rate.

This does establish an association between smoking and mortality for these women as smoking lowers mortality rate. And as we move from nonsmoker to smoker, the probability of dying goes down.

(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(dead), P(dead|smoker), and P(dead|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 (c)? [15 points]

Table 1: (18–64)

P(dead) = $\frac{162}{1072} = 0.1577 = 15.77\%$

P(dead|smoker) = $\frac{93}{533} = 0.1745 = 17.45\%$

P(dead|nonsmoker) = $\frac{69}{539} = 0.1280 = 12.80\%$

Note: P(dead|smoker) > P(dead|nonsmoker)

Table 2: (65+)

P(dead) = $\frac{207}{242} = 0.8554 = 85.54\%$

P(dead|smoker) = $\frac{46}{49} = 0.9389 = 93.89\%$

P(dead|nonsmoker) = $\frac{161}{193} = 0.8341 = 83.41\%$

Note: P(dead|smoker) > P(dead|nonsmoker)

Age is a CF in studying the relationship between smoking and mortality for these 1,314 women.
When age is taken into consideration, as we move from nonsmoker to smoker, the probability of dying goes up, which is in the opposite direction than in part ©.

Testing to see if age is a CF, Z and Y are associated as Z goes from 0(65+) to 1(18-64), the probability of being dead goes down. The other variables, Z and X are associated with as Z goes from 0(65+) to 1(18-64), the probability of being a smoker goes up.

(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 (1922–2019), 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]

This is a Simpson's Paradox between tables 1&2 and table 3 because if we didn't take age into consideration, the data would be that smoking lowers mortality rate. But when we take age into consideration, we learned that women who are 65+ when the study began in 1972 are less likely to live another 20 years, making age a factor in this study. Tables 1&2 are more accurate than table 3, this is Simplson's Paradox because the direction of the association from smoker to nonsmoker, the probability of being dead went up in tables 1&2 but went down in table 3.
Smoking proxy for being young.

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 N1 are of type 1 and N2 of type 2 (with N1 + N2 = 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 n1 individuals of type 1 and n2 of type 2? Evidently there are some restrictions here: $0 \leq n1 \leq N1$, and $0 \leq n2 \leq N2$, and n1 + n2 = 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 are $\binom{N1}{n1}$ ways to choose the n1 type–1 individuals and $\binom{N2}{n2}$ ways to end up with exactly n2 individuals of type 2. Thus

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

$$P(n_1 \text{ type–1 individuals and } n_2 \text{ type–2 individuals}) = \frac{\binom{N_1}{n_1}\binom{N_2}{n_2}}{\binom{N}{n}}. \tag{1}$$

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 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]*.

$$o_A = \frac{P(A)}{1-P(A)}$$

$$\text{P(no whites, 1 red)} = \frac{1}{38.32} \longrightarrow o_A = \frac{\frac{1}{38.32}}{\frac{37.32}{38.32}} = \frac{1}{37.32} = 0.02 = 2\%$$

(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]

These aren't "odds" but probabilities due to the lottery people used permutations instead of combinations.

$$\frac{1}{35{,}064{,}160{,}560} = 2.85 * 10^{-11}$$

5! As there are 5 balls

$$\frac{35{,}064{,}160{,}560}{5!} = 292201338 \longleftarrow \text{That's the right number}$$

The stated "odds" value was essentially correct but it should have been called a probability.

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

The second prize forgot to specify that it's all 5 white and no reds. Adding in no reds, gets us the right answer.

P(\$1M) = P(All 5 white balls and no red balls)

$$= \frac{\binom{5}{5}\binom{64}{0}}{\binom{69}{5}} \bullet \frac{\binom{1}{0}\binom{25}{1}}{\binom{26}{1}} = \left(\frac{1*1}{11238513}\right) \bullet \left(\frac{25}{26}\right) = \frac{1}{11238513}$$

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

$$P(k \text{ whites and the red}) = \frac{\binom{5}{k}\binom{64}{5-k}\binom{1}{1}\binom{25}{0}}{\binom{69}{5}\binom{26}{1}} \text{ and}$$

$$P(k \text{ whites (and not the red)}) = \frac{\binom{5}{k}\binom{64}{5-k}\binom{1}{0}\binom{25}{1}}{\binom{69}{5}\binom{26}{1}}. \qquad (2)$$

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

Red and white balls are independent meaning it's an unconditional probability.

P(k whites and the red) = $\frac{1}{26}$

P(k whites (and not the red)) = $\frac{1}{25}$

P(0 whites and the red) $= \frac{\binom{5}{0}\binom{64}{5-0}\binom{1}{1}\binom{25}{0}}{\binom{69}{5}\binom{26}{1}} = \frac{7624512}{292201338} = 0.026$

The 0.026 equals to $\frac{1}{38.32}$ *for the probability of* "the red"

P(0 whites (and not the red)) $= \frac{\binom{5}{0}\binom{64}{5-0}\binom{1}{0}\binom{25}{1}}{\binom{69}{5}\binom{26}{1}} = \frac{190612800}{292201338} = 6.52$

$\neq \frac{1}{38.32}$

P(1 whites and the red) $= \frac{\binom{5}{1}\binom{64}{5-1}\binom{1}{1}\binom{25}{0}}{\binom{69}{5}\binom{26}{1}} = \frac{3176880}{292201338} = 0.01 = \frac{1}{91.98}$

P(1 whites (and not the red)) $= \frac{\binom{5}{1}\binom{64}{5-k}\binom{1}{0}\binom{25}{1}}{\binom{69}{5}\binom{26}{1}} = \frac{79422000}{292201338} = 0.272 \neq \frac{1}{91.98}$

$$\text{P(2 whites and the red)} = \frac{\binom{5}{2}\binom{64}{5-2}\binom{1}{1}\binom{25}{0}}{\binom{69}{5}\binom{26}{1}} = \frac{416640}{292201338} = 0.0014 = \frac{1}{701.33}$$

$$\text{P(2 whites (and not the red))} = \frac{\binom{5}{2}\binom{64}{5-2}\binom{1}{1}\binom{25}{1}}{\binom{69}{5}\binom{26}{1}} = \frac{10416000}{292201338} \neq \frac{1}{701.33}$$

$$\text{P(3 whites and the red)} = \frac{\binom{5}{k}\binom{64}{5-k}\binom{1}{1}\binom{25}{0}}{\binom{69}{5}\binom{26}{1}} = \frac{20160}{292201338} = 6.89 * 10^{-5} = \frac{1}{14494.11}$$

$$\text{P(3 whites (and not the red))} = \frac{\binom{5}{k}\binom{64}{5-k}\binom{1}{0}\binom{25}{1}}{\binom{69}{5}\binom{26}{1}} = \frac{504000}{292201338} = 0.00173 \neq \frac{1}{14494.11}$$

$$\text{P(4 whites and the red)} = \frac{\binom{5}{k}\binom{64}{5-k}\binom{1}{1}\binom{25}{0}}{\binom{69}{5}\binom{26}{1}} = \frac{320}{292201338} = 1.09 * 10^{-6} = \frac{1}{913129.18}$$

$$\text{P(4 whites (and not the red))} = \frac{\binom{5}{k}\binom{64}{5-k}\binom{1}{0}\binom{25}{1}}{\binom{69}{5}\binom{26}{1}} = \frac{8000}{292201338} \neq \frac{1}{913129.18}$$

$$\text{P(5 whites and the red)} = \frac{\binom{5}{k}\binom{64}{5-k}\binom{1}{1}\binom{25}{0}}{\binom{69}{5}\binom{26}{1}} = \frac{1}{292201338} = 3.422 * 10^{-9} = \frac{1}{292201338}$$

$$\text{P(5 whites (and not the red))} = \frac{\binom{5}{k}\binom{64}{5-k}\binom{1}{0}\binom{25}{1}}{\binom{69}{5}\binom{26}{1}} = \frac{25}{292201338} \neq \frac{1}{292201338}$$

The probability did not match table 4 when it was missing in the probability about the red ball. The formulas are correct as the include the red ball or not.

(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]

These aren't "odds", they are probabilities, changing the word "odds" to probabilities fixes the table.

$P(Win)$ = P(Grand Prize OR $1M OR $50K OR $100 OR $100 OR $7 OR $7 OR $4 OR $4)

$= P(Grand\,Prize) + P(\$1M) + P(\$50K) + P(\$100) + P(\$100) + P(\$7) + P(\$7) + P(\$4) + P(\$4)$

$= \frac{1}{292201338} + \frac{1}{11688053.52} + \frac{1}{913129.18} + \frac{1}{36525.17} + \frac{1}{14.494.11} + \frac{1}{579.76} + \frac{1}{701.33} + \frac{1}{91.98} + \frac{1}{38.32}$

$= \frac{1}{24.87} = 0.04 = 4\%$

(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]

This problem is similar to the Tay-Sachs case study.

T = 182,900,000 p = $\frac{1}{292,201,338}$ ← Grand prize winner

P(1 or more winners in n IID trials) = $1 - (1 - p)^T$

$= 1 - (1 - \frac{1}{292,201,338})^{182,900,000} = 0.465 = 46.5\%$

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.

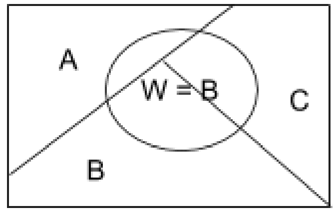
> Please tell me the name of one of the other prisoners who's *not* going to be pardoned — no harm done, since you won't be identifying the lucky person. Let's agree on these rules: if B will be pardoned, you say C; if C will get the pardon, you say B; and if I'm the lucky person, you toss a 50/50 coin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

This problem is similar to the Monty Hall problem.
Variables for who will be pardon
A = (A gets pardon)
B = (B gets pardon)
C = (C gets pardon)
Variable for the Warden
W = x (meaning x will not be pardon)
What warden says is data, looking for who gets the pardon

Math:
$P(A) = \frac{1}{3}$
$P(W = A \mid A) = 0$
$P(W = B \mid B) = 0$
$P(W = C \mid C) = 0$

$P(W = B \mid A) = \frac{1}{2}$
$P(W = C \mid A) = \frac{1}{2}$
$P(A \mid W{=}B) = ???$
$P(A \mid W{=}B) = \frac{P(A) \bullet P(W{=}B|A)}{P(W{=}B)} = \frac{\frac{1}{3} \bullet \frac{1}{2}}{P(W{=}B)}$
Hard to find the truth as we need to know the warden's behavior so need to partition over the truth



$$\text{P(W=B)} = P(W = B \text{ and } A) + P(W = B \text{ and } B) + P(W = B \text{ and } C)$$

$$= P(A) \bullet P(W = B | A) + P(B) \bullet P(W = B | B) + P(C) \bullet P(W = B | C)$$

$$= \frac{1}{3} \bullet \frac{1}{2} + \frac{1}{3} \bullet 0 + \frac{1}{3} \bullet 1$$

$$= \frac{1}{6} + 0 + \frac{1}{3}$$

$$= \frac{1}{2}$$

Therefore,

$$\text{P(A | W=B)} = \frac{\frac{1}{3} \bullet \frac{1}{2}}{\frac{1}{2}} = \frac{1}{3}$$

There is no information in what the warden said that changes A's chance of being pardoned. Because the probability before and after what the warden said, did not change.

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 $0 \le r < n$, 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]*

Applicants arrive in a quality-random order, the ELM applies to any single slot having the best person in it, $\frac{\textit{number of quality pool}}{\textit{total slots}} = \frac{r}{i}$

Using a special case to calculate,

i = 5

P(best candidate among the first 5 in one of the first 2) = $\frac{2}{5} = \frac{r}{i}$

(b) Explain why $P(A \mid B_i) = 0$ for $i \le r$, 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]*

A = the best person to hire

B_i = best person in slot i

n = 8

r = 3

_ _ | _ _ _ _

Qp | interview pool

C_i = we are still interviewing the candidates

$P(A|B_i)$ = ???

Partitioning C_i VS. (not C_i)

$P(A|B_i) = P(A \text{ and } C_i|B_i) + \cancel{P(A \text{ and not } C_i|B_i)} \leftarrow 0$

$P(A \text{ and } C_i|B_i) = P(C_i|B_i)\underline{P(A|C_iB_i)}$

=1

$P(A|B_i) = P(A \text{ and } C_i|B_i)$

$= P(C_i|B_i)$ = P(interviewing slot i | best person on slot i)

Using a special case,

N = 7

<u>40</u> <u>36</u> <u>50</u> | _ <u>98</u> | _ _

Qp | ^i=5

^r =3

= P(2nd best person so far | best person is in slot i)

$$= \frac{r}{i-1}$$

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

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

[15 points]

(i)

A = hiring the best candidate

$p_r = P(A\ having\ pre-specified\ r\ before\ the\ interviews)\ 0 \leq r \leq n$

$p_o = \frac{1}{n}$

If we have no quality pool, we hire the first person we see.

(ii)

P(A) = ??? ← looking for

$P(A|B_i) = \begin{cases} 0, & i \leq r \\ \frac{r}{i-1}, & i > r \end{cases}$

$P(B_i) = \frac{1}{n}$

Partitioning over the truth

$P(A) = P[(A \text{ and } B_1) \text{ OR } (A \text{ and } B_2) \text{ OR } (...) \text{ OR } (A \text{ and } B_n)]$

B_i is mutually exclusive (partition)

$= P(A \text{ and } B_1) + P(A \text{ and } B_2) + \dots + P(A \text{ and } B_n)$

$= P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \dots + P(B_n)P(A|B_n)$

$= (\frac{1}{1})(0) + (\frac{1}{2})(0)$ … most of these values are zeros so

$= P(B_{r+1})P(A|B_{r+1}) + \dots + P(B_n)P(A|B_n)$

P(A and B) = $P(A)P(B_1|A) = P(B_1)P(A|B_1)$

$= \frac{1}{n} P(A|B_1)$

From 4a, we know $P(A|B_i)$ as $\frac{r}{i-1}$

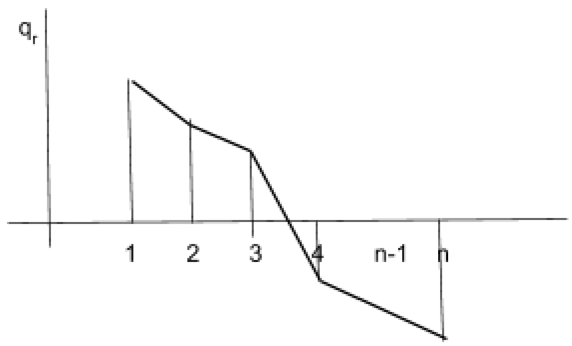
So, P(A) is

$P(A) = \frac{1}{n} P(A|B_{r+1}) + \ldots + \frac{1}{n} P(A|B_n)$

$P(A) = p_r$

$p_r = \frac{1}{n} \sum_{i=r+1}^{n} P(A|B_i) = \frac{1}{n} \sum_{i=r+1}^{n} \frac{r}{i-1} = \frac{r}{n} \sum_{i=r+1}^{n} \frac{1}{i-1}$ ← which is what we are solving for

(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]*



$q_r = p_r - p_{r-1}$ for $1 \le r \le n$

$q_1 = p_1 - p_0 \Rightarrow (\frac{1}{n} \sum_{i=2}^{n} \frac{1}{i-1}) - \frac{1}{n}$

$q_2 = p_2 - p_1 \Rightarrow (\frac{2}{n} \sum_{i=3}^{n} \frac{1}{i-1}) - (\frac{1}{n} \sum_{i=2}^{n} \frac{1}{i-1})$

The $\sum_{i=3}^{n} \frac{1}{i-1}$ and $\sum_{i=2}^{n} \frac{1}{i-1}$ cancel each other out almost.

$\sum_{i=3}^{n} \frac{1}{i-1} = \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n-1}$

$\sum_{i=2}^{n} \frac{1}{i-1} = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n-1}$

$\sum_{i=1}^{\infty} \frac{1}{i} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} = +\infty$

This series will diverge, meaning the function is strictly decreasing.

Hope the graph helps to explain what I was thinking.

(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]*

$$p_r = p_o + \sum_{i=1}^{r} q_i$$

$$p_\infty = p_o + \sum_{i=1}^{\infty} q_i$$

$n \geq 1$

$0 \leq r \leq n$

for $r \geq 1$, $q_r = p_r - p_{r-1}$

$p_1 = p_o + q_1 = p_o + (p_1 - p_o)$

$p_2 = p_o + q_1 + q_2 = p_o + (p_1 - p_o) + (p_2 - p_1)$

$p_2 - p_{2-1} = p_2 - p_1 = [p_o + (p_1 - p_o) + (p_2 - p_1)] - [p_o + (p_1 - p_o)] = p_2 - p_1$

^canceled out to make $p_2 - p_1$

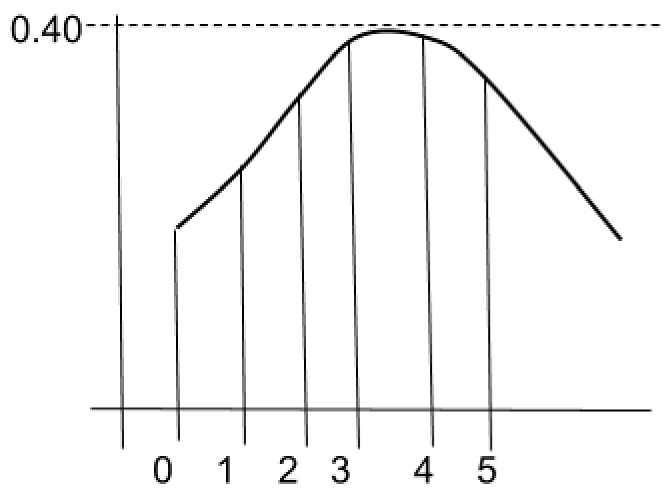
(f) Use (e) to find the best value of r when $n = 10$ and the resulting optimal value of p_r. Does the adaptive hiring strategy examined in this problem look good to you? Explain briefly. *[15 points]*

n = 10

$p_0 = \frac{1}{10}$

$$p_r = \frac{r}{n} \sum_{i=r+1}^{n} \frac{1}{i-1}$$

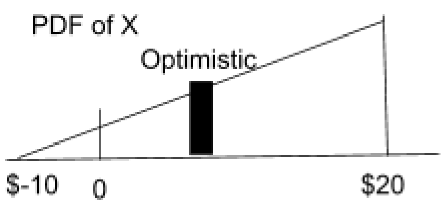
r	p_r
0	0.1
1	0.283
2	0.366
3	0.399
4	0.398
5	0.372



When n = 10, the quality pool has about 3 or 4 people. The optimal value of p_r is about 40%. This hiring strategy could be higher but 40% is better than nothing.

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]*

Trying to solve for VaR = v to $F_Y(v) = (1-\alpha)$

$$F_Y^{-1}[F_Y(v)] = F_Y^{-1}(1-\alpha) = v$$

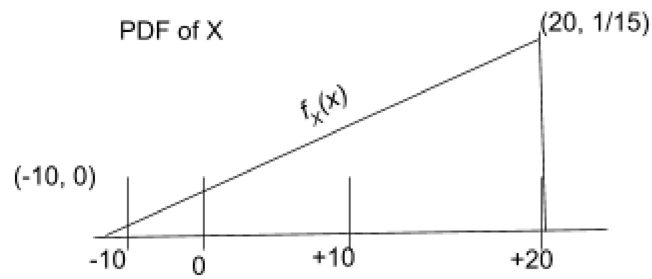
If both v = $F_Y^{-1}(1-\alpha)$ and VaR, then VaR = $F_Y^{-1}(1-\alpha)$

(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 $x = -10$).



(i)

Steps: PDF of X, CDF of X, then the inverse of CDF of X

1) $f_x(x) \geq 0$

2) $$\int_{-10}^{+20} f_x(x) = 1$$

Area of the triangle is 0.5*base*height

0.5(20-(-10)) *c = 1

$c = \frac{1}{15}$

$$\frac{y - y_1}{x - x_1} = \frac{y_2 - y_1}{x_2 - x_1}$$

$$\frac{y - 0}{x - (-10)} = \frac{\frac{1}{15} - 0}{20 - (-10)}$$

$$y - 0 = \frac{\frac{1}{15} - 0}{20 - (-10)}(x + 10)$$

$$y = \frac{1}{450}(x + 10)$$

$f_x(x) = \{\frac{x+10}{450}\ , for\ -10\ <\ x\ <\ +20$

$\{\ 0,\ else$

var = v = - $F_x^{-1}(\alpha)$

Looking for the CDF:

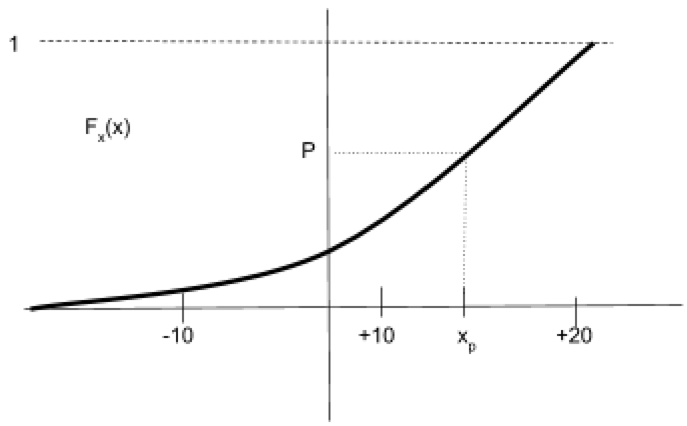
for -10 < x < +20

$F_x(x) = \int_{-10}^{x} \frac{a+10}{450} da\ = \left[\frac{(x+10)^2}{2*450}\right] from\ x\ to\ -10\ = \frac{1}{900}(x+10)^2 - 0$

$F_x(x)\ = \{0 \qquad x <\ -10$

$\{\frac{(x+10)^2}{900} \qquad -10 < x < 20$

$\{1 \qquad x\ <+20$



$F_x(x_p) = p = \frac{(x_p+10)^2}{900}$

$(x_p + 10)^2 = 900p$

$x_p + 10\ =\ 30\sqrt{p}$

$x_p = 30\sqrt{p} - 10\ = F_x^{-1}(p)$

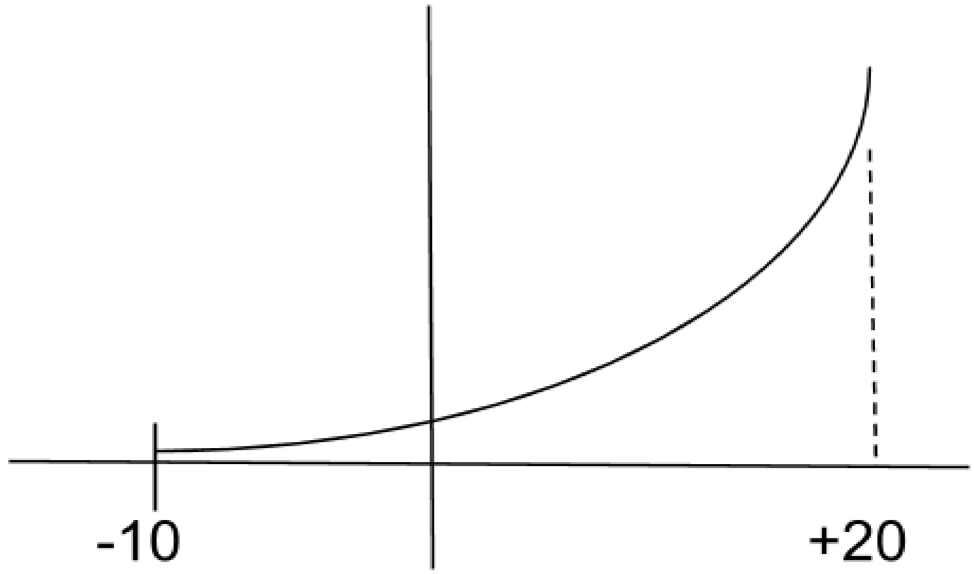
$= 10 - 30\sqrt{\alpha}$

$\alpha = 0.01$

=$7M

var with this PDF for Z = $-F_x^{-1}(\alpha) =-(30\sqrt{\alpha} - 10)$

(ii)



Steps: PDF of X, CDF of X, then the inverse of CDF of X

$y = (x + 10)^2$

$f_x(x) \geq 0$

$f_x(x) = \{ c \bullet (x + 10)^2$, for $-10 \leq x \leq +20$

$\{ 0 \; else$

c = $\frac{1}{9000}$ because $(20+10)^2$

$1 = \int_{-10}^{+20} c\,(x + 10)^2\, dx$

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

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

$F_x(-v) = F[F_x^{-1}(\alpha)] = \alpha = 0.01$

Looking for the CDF:

$F_x(x) = \int_{-10}^{+20} \frac{1}{9000}(t + 10)^2 dt = \frac{(x+10)^3}{27000}$ from Wolfram Alpha

$F_x(x) = \{0 \qquad x < -10$

$\{\frac{(x+10)^3}{27000} \qquad -10 < x < 20$

$\{1 \qquad x <+ 20$

Inverse of CDF:

$F_x(x_p) = p = \frac{(x_p+10)^3}{27000}$

$(x_p + 10)^3 = 30p$

$x_p + 10 = 30\sqrt[3]{p}$

$x_p = 30\sqrt[3]{p} - 10 = F_x^{-1}(p)$

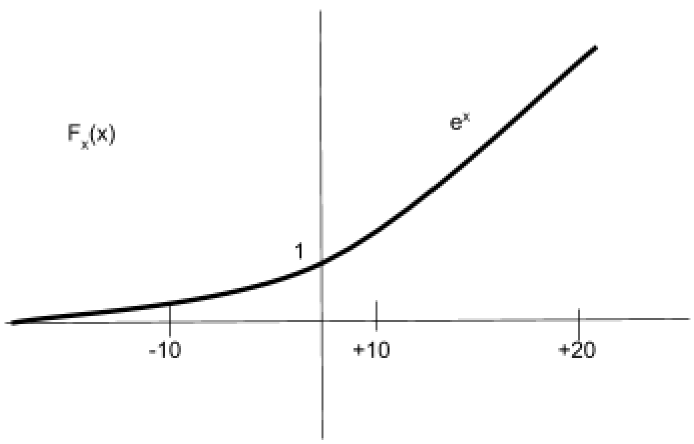
$= 10 - 30\sqrt[3]{\alpha}$

$\alpha = 0.01$

=\$3.537

var with this PDF for Z = $-F_x^{-1}(\alpha) = -(30\sqrt[3]{\alpha} - 10)$

(iii)



$f_x(x) = \{c * exp(\frac{\lambda x}{10}) for\ -10 \le x \le +20 \qquad \{0$ else

Finding the CDF

$\int_{-10}^{x} f_x(t)dt,\ for\ -10 \le x \le +20$ (one part of the CDF)

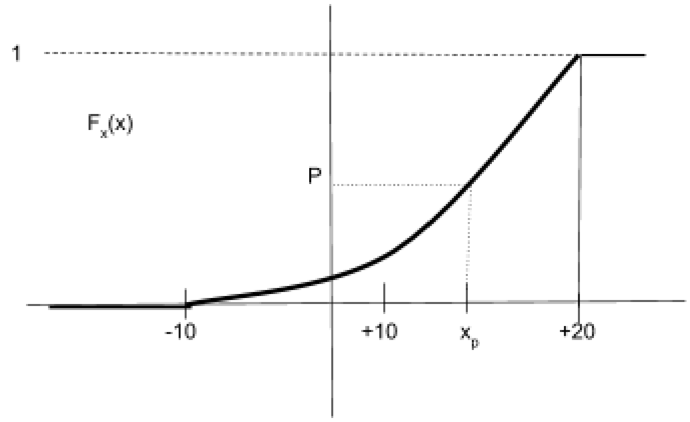
$F_x(x) = \{0,\ for\ x \le -10$

$\{ \int_{-10}^{x} f_x(t)dt,\ for\ -10 \le x \le +20$

$\{ 1,\ x \ge +20$

From wolfram Alpha,

$0.0324306 * e^{0.171723x_p} - 0.00582333 = \text{p}$



$F_x^{-1}(x) = 5.8233 \log (30.8351 + 0.1796)$

$v = -F_x^{-1}(0{,}01) = \$4.18$ M

(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]

(i)

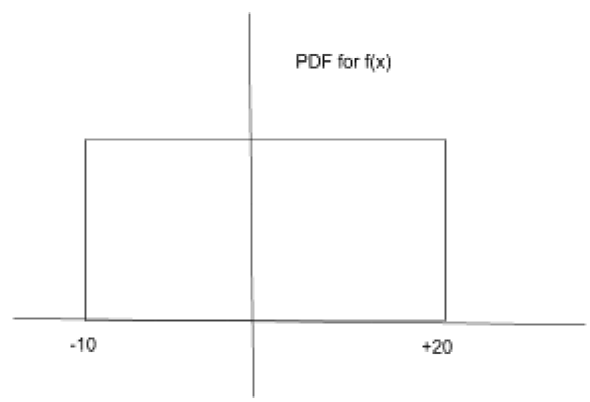
1 = base * height

1 = (30) * height

Height = $\frac{1}{30}$

CDF of

$F_x(x) = \{ 0 \, for \, x \leq -10$



$\{ \int_{-10}^{x} \frac{1}{30} dt, \, for \, -10 \leq x \leq +20$

$= \frac{x+10}{30}$

$\{ 1, \, for \, x \geq +20$

$F_x(x) = P(X \leq x)$

$F_x(x_p) = p = \frac{x+10}{30}$

$(x_p + 10) = 30p$

$x_p + 10 = 30p$

$x_p = 30p - 10 = F_x^{-1}(p)$

$= 10 - 30p$

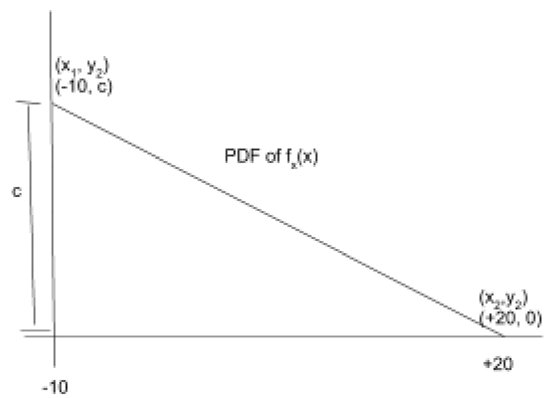
$\alpha = 0.01$

$=\$9.7M$

var with this PDF for Z = $-F_x^{-1}(\alpha) =- (90\sqrt{\alpha} - 10)$

(ii)

$f_x(x)$ linear with a negative slope going through (-10,c) and (+20, 0)



$c = \frac{30}{2} \bullet 1 = \frac{1}{15}$

$\frac{y - y_1}{x - x_1} = \frac{y_2 - y_1}{x_2 - x_1}$

$y - y_1 = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1)$

$y = y_1 + \frac{y_2 - y_1}{x_2 - x_1}(x - x_1)$

$= \frac{1}{15} + \frac{0 - \frac{1}{15}}{20 - (-10)}[x - (-10)]$

$= \frac{1}{15} - \frac{1}{450}(x + 10)$

$= (\frac{1}{15} - \frac{1}{45}) - \frac{x}{450}$

PDF

$$f_x(x) \; = \{ \; \frac{2}{45} - \frac{x}{450}, \; for \; -10 \; \leq x \leq +20$$
$$\{0 \; , \qquad else$$

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

$\alpha = 0.01$

CDF

$$F_x(x) \; = \{0, \quad for \; x \leq -10$$
$$\{\frac{1}{900}(500 + 40x - x^2), \quad -10 \leq x \leq +20$$
$$\{1, \qquad x \geq +20$$

For -10$\leq x \leq +20$

$$F_x(x) \; = \int_{-10}^{x} (\frac{2}{45} - \frac{t}{450})dt = \frac{1}{900}(500 + 40x - x^2)$$

Inverse of CDF:

For $-10 \leq x_p \leq +20$

$$F_x(x_p) \; = \; \frac{1}{900}(500 + 40x_p - x_p^2\,) = p$$

$$x_p = 20 - 30\sqrt{1-p} = F_x^{-1}(p)$$

$$= \; 30\sqrt{1-\alpha} - 20$$

$\alpha = 0.01$

=\$9.849M

var with this PDF for Z $= -F_x^{-1}(\alpha) = \; 30\sqrt{1-\alpha} - 20$