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

STAT 131: Take-Home Test 1 *[300 total points]*

(please watch email and `Canvas` for the final due date)

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 is the only way to get no credit; to maximize your score, please ensure that you attempt to answer every part of every problem.

This test is to be entirely your own efforts; do not collaborate with anyone or get help from anyone but me or our TAs. 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 (out of 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.

> **Uploading your solutions constitutes agreement by you with these rules about cheating.**

You're free to use `Wolfram Alpha` (or some other symbolic computing environment) for any calculations you make, if you wish; if you do, your answers should look something like this: "The relevant integral is $\int_{-1}^{+1} x^2 \, dx$; I gave this integral to `Wolfram Alpha` and the answer was $\frac{2}{3}$"; but (of course) be sure that you approach `Wolfram Alpha` with the correct query.

1. *[100 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.

Tables 1–3 summarize some of the data collected in this sample survey: Table 1 *cross-tabulates* X (smoking) against Y (mortality) for the 1,072 women whose ages Z were between 18 and 64 (inclusive) in 1972; Table 2 performs the same cross-tabulation but for the 242 women whose ages were 65 and over in 1972; and Table 3 *aggregates* (combines) Tables 1 and 2 for all n =1,314 women (for example, the counts of women in the upper left-hand cells in the Tables are related additively: (93 [Table 1] + 46 [Table 2]) = 139 [Table 3]).

Table 1: Age Group 18–64

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

Table 2: Age Group 65+

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

Table 3: Overall

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

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 CFs: an apparent relationship between X and Y may have been caused, in whole or in part, by a

The best way to remove the pos anding is to *hold it constant*: to examine t a possible value of Z — if you see something isons, the association between X and Y ca constant. This holding-constant process is

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. *[10 points]* = 0.44 29.

in total we 1314 women, there are 582 smokers. P(smoker) = 582/1314

(b) Compute $P(\text{smoker})$ for a randomly chosen woman from Table 3, and compare this with your computation of $P(\text{smoker} \mid 18\text{–}64)$ for a woman picked at random from Table 1 and $P(\text{smoker} \mid 65+)$ 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. *[25 points]*

Conditional probability

(c) For a woman chosen at random from the 1,314 in Table 3, compute $P(\text{dead})$, $P(\text{dead} \mid \text{smoker})$, and $P(\text{dead} \mid \text{nonsmoker})$. 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. *[30 points]*

X. Observational studies can only obtain association not causality.

Confounding factors

(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 $P(\text{dead})$, $P(\text{dead} \mid \text{smoker})$, and $P(\text{dead} \mid \text{nonsmoker})$. 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 (c)? *[25 points]*

(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 at least since the 1890s). 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. *[10 points]*

① as smoking changes from no to yes, how mortality changes?

2. *[90 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

② as age goes up (changes from 18–64 to 65+), how mortality changes?

③ as age goes up, how smoking

balls in an urn, or Democrats and Republicans among people who stick to the major political parties in the U.S.); in other words, this classification *partitions* the population into two non-overlapping and exhaustive subsets. 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 are $\binom{N_1}{n_1}$ ways to choose the n_1 type–1 individuals and $\binom{N_2}{n_2}$ ways to end up with exactly n_2 individuals of type 2. Thus

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

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 (i.e., each ticket bought specifying your choice of 5 white numbers and 1 red number) costs \$2, and you can play as many times as you like; note that any single ticket cannot win more than one prize.

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]*

(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]*

Table 4: *The nine ways to win in Powerball and the associated "odds," as stated on the Powerball website* `www.powerball.com/games/home` .

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

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

correct red balls

1 choose 1 = 1

25 wrong red balls

25 choose 0 = 1

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

k correct whites

26 choose 1 = 26

the red one is also correct

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

Use these formulas to verify the rest of the "odds" entries in Table 4. *[50 points]*

(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)" could be misinterpreted but can be made unambiguous with the insertion of a single word. *[10 points]*

(e) Suppose that T tickets were bought across the entire U.S. in a given drawing, 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]*

3. *[30 points]* (logic and Bayes's Theorem) Here's a small fictitious drama with five actors: three

$$P(A \mid \text{Warden says } B) = \frac{P(\text{Warden says } B \mid A) \cdot P(A)}{P(\text{Warden says } B)}$$

1/2 ✓ ✓ 1/3

1/2

$P(A) = P(B) = P(C) = \frac{1}{3}$

$P(A \mid \text{Warden says } B) = \frac{1}{3}$

A, B, C on death row

Warden knows which person is picked up.

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.

if B pardoned, Warden say C.

if A pardoned, Warden say B or C. 50/50

Warden says B.

> 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.

if C pardoned, Warden say B.

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. *[30 points]*

$\frac{1}{3}$.

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.

lose.

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

r=5

candidates

- 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: call these r candidates the *quality pool*, and suppose that you've assigned a numerical quality score $0 \leq Q_i \leq 100$ to each of them.

n=10

r candidates

- Beginning with the next candidate $(r+1)$, you continue interviewing until the current candidate is the best you've seen so far (including the people in the quality pool), 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.

i = 10, r = 4

first 10 people. 4 candidate pool.

$\frac{4}{10} = \frac{\binom{4}{1}\binom{6}{0}}{\binom{10}{1}}$ → best candidate

$= \frac{4}{10} = \frac{r}{i}$

$\frac{\binom{r}{1}\cdot\binom{i-r}{0}}{\binom{i}{1}} = \frac{r}{i}$

6 candidate people out of pool.

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

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

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

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

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

But you still want to find the best candidate. $P_0 = \frac{1}{n}$.

(i) $p_0 = \frac{1}{n}$, and quality pool is 0. You don't want to build a quality pool.

(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] $P_r = P(A)$ Law of total prob $\sum_{i=1}^{n} P(A \mid B_i)\cdot P(B_i) = \sum_{i=1}^{r} P(A \mid B_i)P(B_i) + \sum_{i=r+1}^{n} P(A \mid B_i)P(B_i)$

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

(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]*

$P_r - P_{r-1}$

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

(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]*

$= \frac{1}{n}\left[r\sum_{i=r+1}^{n}\frac{1}{i-1} - (r-1)\sum_{i=r}^{n}\frac{1}{i-1}\right] = \frac{1}{n}\left[r\cdot\sum_{i=r+1}^{n}\frac{1}{i-1} - r\sum_{i=r}^{n}\frac{1}{i-1} + \sum_{i=r}^{n}\frac{1}{i-1}\right] = q_r$

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

(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.)

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

$P(A \mid B_i) = \frac{P(A \cap B_i)}{P(B_i)}$

A: You find the best candidate

B_i: the best candidate is person i.

C_i: keep until candidate i.

Law of total probability:

$P(A \cap B_i) = P(A \cap B_i \cap C_i) + P(A \cap B_i \cap C_i^c)$ (the second term $= 0$)

$P(A \cap B_i \cap C_i)$:

A: You find the best

B_i: the best is i-th person

C_i: keep interview until i-th

$P(A \cap B_i \cap C_i^c)$:

A: You find the best candidate.

B_i: the best candidate is person i

C_i^c: but did not interview until person i

For $i \le r$ → interview person → candidate pool,

$P(A \mid B_i) = 0$

↓ ↳ the best candidate is i-th person.

You find the best candidate.

① $i \le r$, all people you interviewed goes into the candidate pool.

you can't the best candidate.

② $i > r$. we have r many candidate pods.

C_i: keep interview until i-th person,

B_i: You find the best i-th person.

A: You find the best

$$P(A \mid B_i) = \frac{P(A \cap B_i)}{P(B_i)} = \frac{P(A \cap B_i \cap C_i)}{P(B_i)} = P(A \cap C_i \mid B_i)$$

The $i-1$ people are not the best.

All $i-1$ people should go to the quality pool.

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

(c)

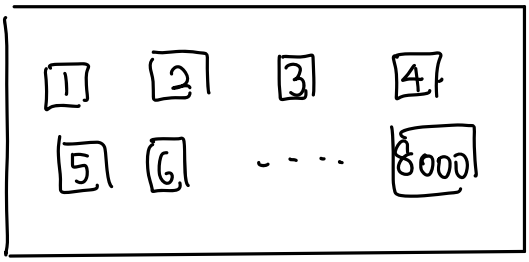
SRS: simple random sampling.

iid: identically → same distribution
independently → independent

Population: All UCSC undergraduates.

distribution

a SRS of this population:



randomly draw without replacement.

Sample size = 200.

① 200 random draws without replacement from this box

② computer programs