Statistical Methods in Education Research
Vanderbilt University
Prof. Sean P. Corcoran

September 23, 2021
59 total points

Problem Set 4 *Solutions*

---

1. (**16 points—2 each**) In a population of students, the number of absences during the school year ranges from 3 to 7. The probabilities of a randomly drawn student from this population having 3, 4, 5, 6, or 7 absences are shown in the table below. Define the event A as the student being absent *more than* 4 days, and the event B as the student being absent *fewer than* 6 days.

# of Days	3	4	5	6	7
Probability	0.08	0.24	0.41	0.20	0.07

   (a) What is the probability of event A? $\mathbf{P(A) = P(5) + P(6) + P(7) = 0.41 + 0.20 + 0.07 = 0.68}$

   (b) What is the probability of event B? $\mathbf{P(B) = P(3) + P(4) + P(5) = 0.08 + 0.24 + 0.41 = 0.73}$

   (c) What is the probability of $\sim A$? $\mathbf{P(\sim A) = P(3) + P(4) = 0.08 + 0.24 = 0.32}$ or alternatively $\mathbf{1 - P(A) = 1 - 0.68 = 0.32}$

   (d) Are events A and B mutually exclusive? Explain why or why not. **No.** $\mathbf{A \cap B = 5}$ **(i.e. 5 absences appears in both events, so they are not mutually exclusive).**

   (e) What is the probability of $A \cap B$? $\mathbf{P(A \cap B) = P(5) = 0.41}$

   (f) What is the probability of $A \cup B$? $\mathbf{P(A \cup B) = 1.0}$

   (g) Show that $P((A \cap B) \cup (\sim A \cap B)) = P(B)$. **In words, the lefthand side of this equation is the probability that B and A occur *or* B and $\sim$ A occur. In other words, B occurs and either A occurs or it doesn't. This is simply B, and P(B)=0.73. You could also recognize that these are mutually exclusive events—if B and A are true, it cannot be the case that B and $\sim$ A are true. With mutually exclusive events, you can add the two probabilities together:** $\mathbf{P((A \cap B) \cup (\sim A \cap B)) = 0.41 + 0.32 = 0.73}$

   (h) Show that $P(A \cup (\sim A \cap B)) = P(A \cup B)$. **In words, the lefthand side of this equation is the probability that A occurs *or* A doesn't occur and B occurs. In this context (looking at the above table), this is the same as A or B occurring. As seen in part (f), this is 1.**

2. (**6 points—3 each**) Using the probability distribution in Question 1, find the following (and show your work):

   (a) $E(\#\text{ of absences})$:

   $\sum_{i=1}^{n} X_i * P(X_i) = (3*0.08) + (4*0.24) + (5*0.41) + (6*0.20) + (7*0.07) = 4.94$

   (b) $Var(\#\text{ of absences})$:

   $\sum_{i=1}^{n} (X_i - E(X))^2 * P(X_i) = ((3-4.94)^2 * 0.08) + ((4-4.94)^2 * 0.24) + ((5-4.94)^2 * 0.41) + ((6-4.94)^2 * 0.20) + ((7-4.94)^2 * 0.07) = 1.04$

3. (**8 points—2 each**) Shown below is a 2 x 2 table that reports the fraction of the population in each cell:

		Education level		
		HS	<HS	Totals
Current smoker:	NO	0.614	0.130	0.744
	YES	0.194	0.062	0.256
	Totals	0.808	0.192	1.000

   (a) For a randomly drawn person, what is P(smoker)? **0.256, or 25.6%**

   (b) For a randomly drawn person, what is P(smoker | <HS diploma)? **Here we can use $P(A|B) = P(A \cap B)/P(B)$, or 0.062/0.192 = 0.323, or 32.3%**

   (c) For a randomly drawn person, what is P(smoker | HS diploma+)? **In the same manner as part (b): 0.194/0.808 = 0.240, or 24.0%**

   (d) Are education and smoking status "independent?" Why or why not? **No. The probability of being a current smoker varies depending on one's education level (as shown in parts b and c). Thus they are not independent.**

4. (**5 points**) Shown below is a 2 x 2 table. In Period 1, events A or B can happen. In Period 2, outcome C or D will result. If $P(C|B) = 0.150$ and $P(D|A) = 0.7$, then fill in the missing boxes below:

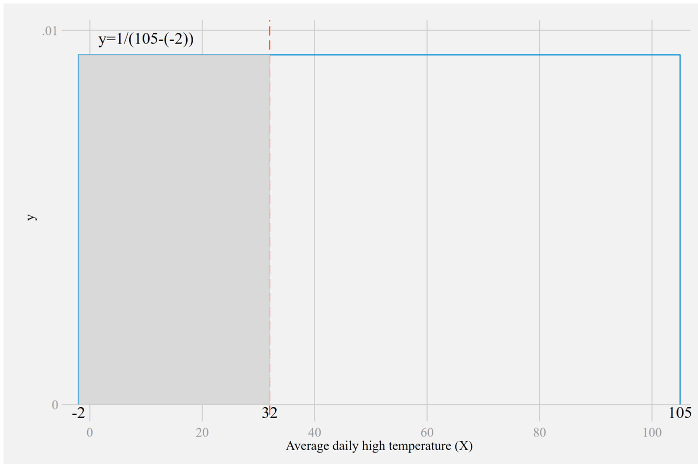
		Period 1	
		Event A	Event B
Period 2	Event C	**0.240**	0.030
	Event D	**0.560**	**0.170**
		0.800	**0.200**

- First use $P(C|B) = P(C \cap B)/P(B)$ or $0.15 = 0.030/P(B)$ which implies that $P(B) = 0.2$. This provides the first marginal probability shown in the bottom right corner.
- If $P(B \cap C) = 0.03$ and $P(B) = 0.2$ then $P(B \cap D) = 0.2 - 0.03 = 0.17$
- If $P(B) = 0.2$ then $P(A) = 1 - 0.2 = 0.8$
- Now use $P(D|A) = P(D \cap A)/P(A)$ or $0.7 = P(D \cap A)/0.8$ which implies that $P(D \cap A) = 0.56$.
- Finally $P(A \cap C) = 0.80 - 0.56 = 0.24$
- Notice that the four probabilities in the center of the table sum to 1, as they should.

5. (**6 points—3 each**) Paul and Natasha live in Los Angeles. Paul hates cold weather but Natasha has been transferred to a cold Northeastern city. Paul notes that he cannot move go to a city where more than 30% of the days have an average daily high below freezing. Suppose the average daily high temperatures (X) in a city can be described by a uniform distribution where the minimum and maximum average daily highs are -2 and 105, respectively.

   (a) What is the PDF for X, and what is $P(x \leq 32)$? Should Natasha look for a one or a two bedroom apartment? (Hint: you do not need calculus to find the requested probability).

   The PDF for a uniform distribution from $[a, b]$ is: $y = 1/(b - a)$. Or in this case: $y = 1/107$. The PDF is pictured below, and the area under the curve from -2 to 32 is shaded. The probability that this city's daily high temperature is 32 or below is this area, which is easy to calculate given the rectangular distribution: $P(X \leq 32) = 34 * (1/107) = 31.8\%$ Nathsha may want to find a one bedroom apartment! FYI the Stata code I used to produce this graph is below.




```
twoway (function y=1/107, range(-2 105) dropline(-2 105)) (function y=1/107, ///
   range(-2 32) color(gs10*0.5) recast(area)), ylabel(0(0.01)0.01) xline(32, ///
   lpattern(dash)) xtitle(Average daily high temperature (X)) legend(off) ///
   text(-0.0002 -2 "-2") text(-0.0002 32 "32") text(-0.0002 105 "105") ///
   text(0.0098 10 "y=1/(105-(-2))")
```

(b) What are $E(X)$ and $Var(X)$?

For a uniform distribution, $E(X) = \frac{a+b}{2} = \frac{-2+105}{2} = 51.5$

And $Var(X) = \frac{1}{12}(b-a)^2 = \frac{1}{12}(107^2) = 954.1$. The standard deviation would be: $\sqrt{954.1} = 30.9$

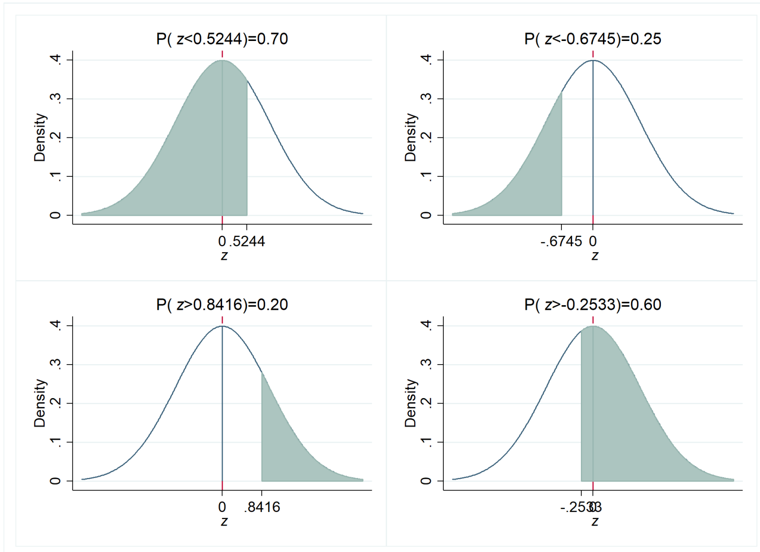
6. (**4 points**) Assume the random variable z has a standard normal distribution. Use Stata, an online calculator, or a textbook table to answer the following:

(a) The probability is 0.70 that z is less than what number? **Pr(z< 0.5244)=0.70, using** `display invnormal(0.70)`

(b) The probability is 0.25 that z is less than what number? **Pr(z< -0.6745)=0.25, using** `display invnormal(0.25)`

(c) The probability is 0.20 that z is greater than what number? **Pr(z> 0.8416)=0.20, using** `display (-1)*invnormal(0.20)`

(d) The probability is 0.60 that z is greater than what number? **Pr(z> -0.2533)=0.60, using** `display (-1)*invnormal(0.60)`



7. (**6 points**) To graduate with honors, you must be in the top 2 percent (*summa cum laude*), 3 percent (*magna cum laude*) or 5 percent (*cum laude*) of your class. Suppose GPAs are distributed normally with a mean of 2.6 and a standard deviation of 0.65. What GPA will you need in order to graduate at each of these three levels?

Under the assumption of a normal distribution, we need to find the GPA cutoff points (x_1, x_2, x_3) such that:

$P(GPA > x_1) = 0.02$ or $P\,(z > (x_1 - 2.6)/0.65) = 0.02$ (summa)

$P(GPA > x_2) = 0.03$ or $P\,(z > (x_2 - 2.6)/0.65) = 0.03$ (magna)

$P(GPA > x_3) = 0.05$ or $P\,(z > (x_3 - 2.6)/0.65) = 0.05$ (cum laude)

From the online calculator (or Stata) we find that the values of z for which 2, 3, and 5 percent of outcomes fall above are: 2.054, 1.881, and 1.645. In Stata, the command is `display (-1)*invnormal(p)`, with p=0.02, 0.03, or 0.05. The result of `invnormal` is multiplied by -1 since we are interested in the z value *above* which there is a p probability of falling. Converting z into the original units (GPA points) we find the following GPA cutoffs:

$2.054 = (x_1 - 2.6)/0.65)$ or $x_1 = 3.9351$ for summa cum laude

$1.881 = (x_2 - 2.6)/0.65)$ or $x_2 = 3.8227$ for magna cum laude

$1.645 = (x_3 - 2.6)/0.65)$ or $x_3 = 3.6693$ for cum laude

8. (**4 points**) Bob is 62 inches and he will only date women who are shorter than him. Suppose heights of females in the population follow a normal distribution with $\mu = 64$ and $\sigma = 3.9$. What fraction of women meet Bob's criteria?

This question is asking: $P(X < 62) = P(\frac{X-\mu}{\sigma} < \frac{62-64}{3.9}) = P(z < -0.51)$. Using Stata `display normal(-0.51)`, this probability is 0.305. Or, about 30.5% of women meet Bob's criteria.

9. (**4 points**) On the midterm exam in introductory statistics, an instructor always gives a grade of B to students who score between 80 and 90. One year, the scores have an approximately normal distribution with a mean $\mu = 83$ and a standard deviation $\sigma = 5$. About what fraction of the students get a B?

This question is asking: $P(80 \leq X \leq 90) = P(\frac{80-83}{5} \leq \frac{X-\mu}{\sigma} \leq \frac{90-83}{5}) = P(-0.6 \leq z \leq 1.4)$. Using Stata:

`display normal(1.4)-normal(-0.6)`, this probability is 0.645. Or, about 64.5% of students get a B.