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

September 19, 2023
62 total points

Problem Set 3 *Solutions*

---

1. (**16 points—2 each**) In a population of students, the number of absences during the school year ranges from 0 to 8. The probabilities of a randomly drawn student from this population having 0, 1, 2, ..., 8 absences are shown in the table below. Define the event A as the student being absent *fewer than* 4 days, and the event B as the student being absent *more than* 3 days.

# of Days	0	1	2	3	4	5	6	7	8
Probability	0.2	0.14	0.25	0.11	0.1	0.09	0.05	0.03	0.03

   (a) What is the probability of event A? $\mathbf{P(A) = P(0) + P(1) + P(2) + P(3) = 0.2 + 0.14 + 0.25 + 0.11 = 0.7}$

   (b) What is the probability of event B? $\mathbf{P(B) = P(4) + P(5) + P(6) + P(7) + P(8) = 0.1 + 0.09 + 0.05 + 0.03 + 0.03 = 0.3}$

   (c) What is the probability of $\sim A$? $\mathbf{P(\sim A) = 1 - P(A) = 1 - 0.7 = 0.3}$

   (d) Are events A and B mutually exclusive? Explain why or why not. **Yes. If event A occurred then B didn't occur, and vice versa**.

   (e) What is the probability of $A \cap B$? $\mathbf{P(A \cap B) = \emptyset}$. **The two events do not intersect.**

   (f) What is the probability of $A \cup B$? $\mathbf{P(A \cup B) = 1.0}$. **A and B represent all possible outcomes in the sample space.**

   (g) Show using values from the table 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)=0.3. 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:** $\mathbf{P((A \cap B) \cup (\sim A \cap B)) = 0 + 0.3 = 0.3}$

   (h) Show using values from the table that 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) $E(\#\text{ of absences})$:

   $\sum_{i=1}^{n} X_i * P(X_i) = (0*0.2) + (1*0.14) + (2*0.25) + (3*0.11) + (4*0.1) + (5*0.09) + (6*0.05) + (7*0.03) + (8*0.03) = \mathbf{2.57}$

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

   $Var = \sum_{i=1}^{n}(X_i - E(X))^2 * P(X_i) = ((0-2.57)^2*0.20) + ((1-2.57)^2*0.14) + ((2-2.57)^2*0.25) + ((3-2.57)^2*0.11) + ((4-2.57)^2*0.1) + ((5-2.57)^2*0.09) + ((6-2.57)^2*0.05) + ((7-2.57)^2*0.03) + ((8-2.57)^2*0.03) = \mathbf{5.23}$
   $SD = \sqrt{Var} = \sqrt{5.23} = \mathbf{2.287}$

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. (**5 points**) After the attacks of September 11, 2001, the TSA implemented a program called SPOT (Screening of Passengers by Observation Techniques) in which passengers were flagged for suspicious behavior and given additional searching or screening. Suppose that:

   - There are 2 billion plus 100 passenger trips per year (2,000,000,100).
   - 100 of these passengers are terrorists (i.e., less than 0.00000001%).
   - Nearly all (99%) terrorists exhibit the kinds of behaviors that were flagged.
   - Some non-terrorists exhibit these suspicious behaviors, but it is rare (1%).

   The SPOT test has low false negative and false positive rates, suggesting it is an effective way to catch would-be terrorists. Use Bayes' Theorem to calculate the probability that a flagged passenger is, in fact, a terrorist.

Bayes' Theorem applied here is:

$$
\begin{aligned}
P(\text{terrorist}|\text{flagged}) &= \frac{P(\text{flagged}|\text{terrorist})P(\text{terrorist})}{P(\text{flagged})} \\
&= \frac{\frac{99}{100} * \frac{100}{2{,}000{,}000{,}100}}{\frac{20{,}000{,}099}{2{,}000{,}000{,}100}} \\
&= \frac{99}{20,000,099} \\
&= 0.00000495
\end{aligned}
$$

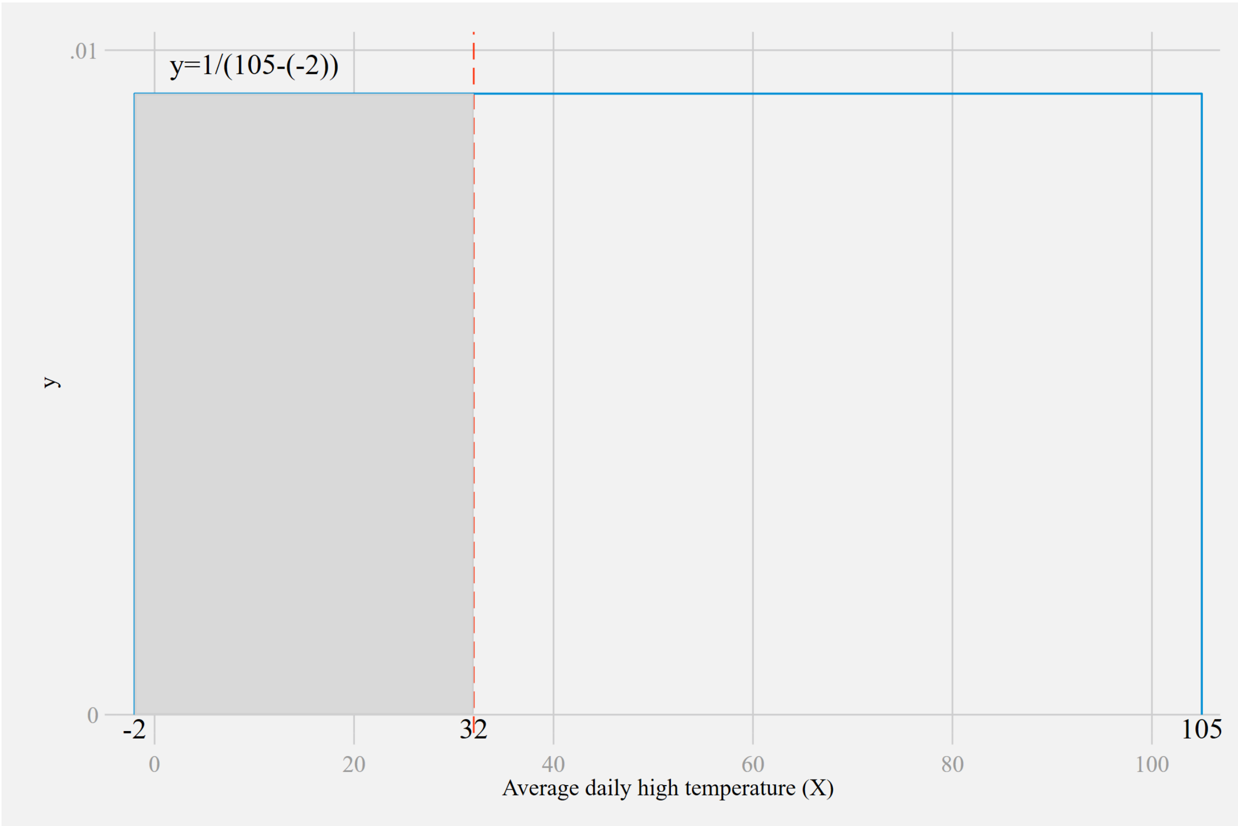
In other words, very small! While the system involves a test that will "catch" nearly all terrorists, the baseline probability of being a terrorist is very low. Even with a low false positive rate, the SPOT system flags a very large number of innocent passengers.

6. (**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: $\sqrt{954.1} = 30.9$

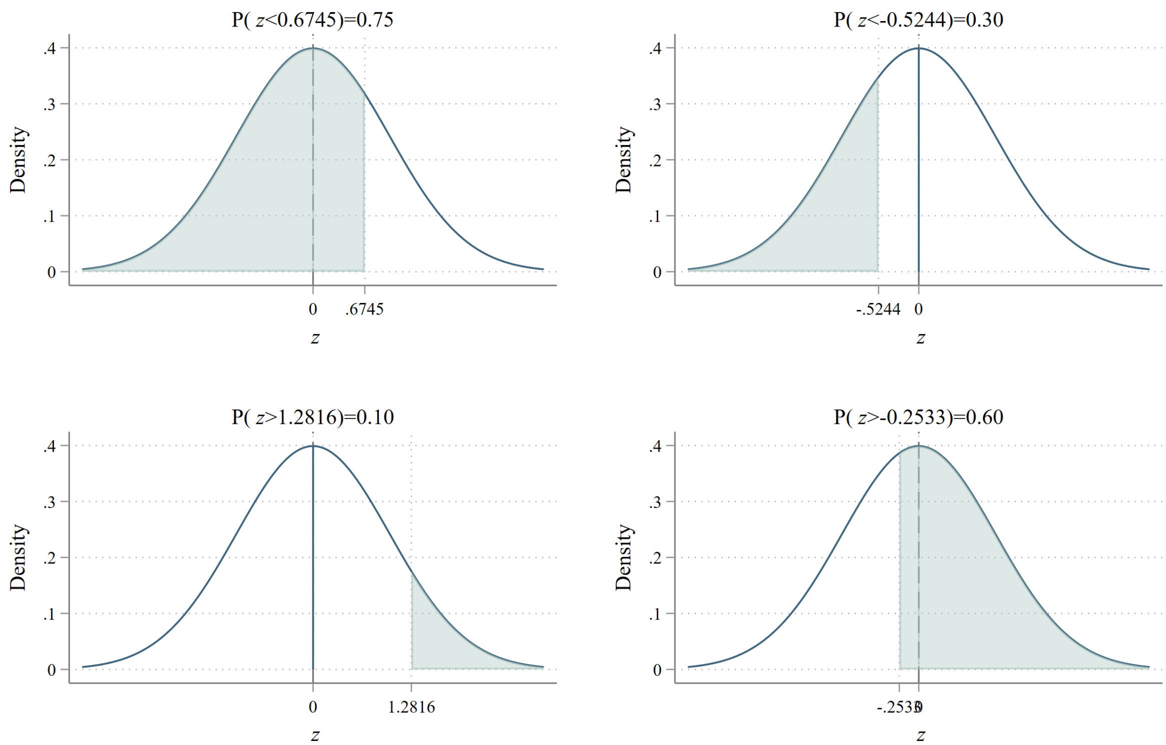
7. (**4 points**) Assume the random variable z has a standard normal distribution. Use Stata, an online calculator, or a textbook table to answer the following:

(a) The probability is 0.75 that z is less than what number? **Pr(z< 0.6745)=0.75, using `display invnormal(0.75)`**

(b) The probability is 0.30 that z is less than what number? **Pr(z< -0.5244)=0.30, using `display invnormal(0.30)`**

(c) The probability is 0.10 that z is greater than what number? **Pr(z> 1.2816)=0.10, using `display (-1)*invnormal(0.10)`**

(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)`**

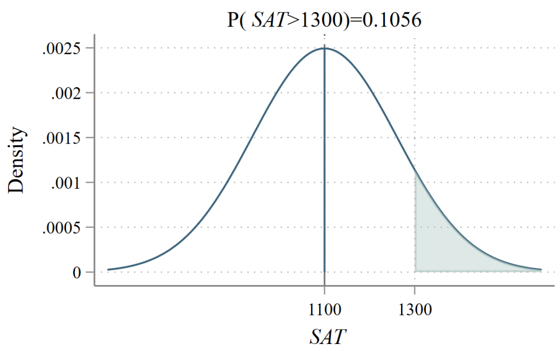


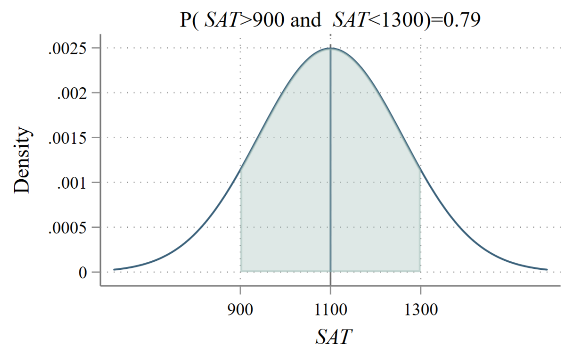
8. (**6 points**) Applicants to Local U. have SAT composite scores that follow a normal distribution with a mean of 1100 and a standard deviation of 160.

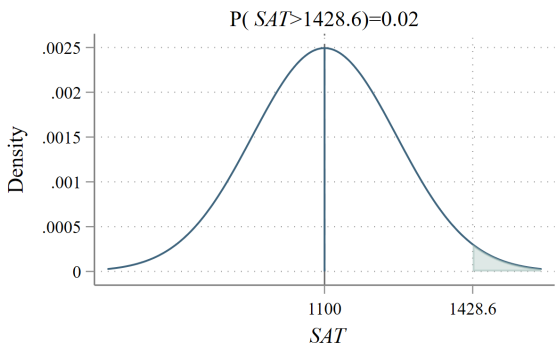
   (a) What is the probability that a Local U. applicant will have a composite SAT score of 1300 or higher? $\mathbf{Pr(SAT \geq 1300) = Pr(z \geq (1300 - 1100)/160) = 0.1056,}$ **using** `display 1 - normal((1300-1100)/160)`

   (b) What is the probability that a Local U. applicant will have a composite score above 900 but below 1300? $\mathbf{Pr(900 < SAT < 1300) = Pr(z < (1300 - 1100)/160) - Pr(z < (900 - 1100)/160) = 0.7887,}$ **using** `display normal((1300-1100)/160) - normal((900-1100)/160)`

   (c) Showing strikingly bad judgment, Local U. wishes to offer a tuition-free scholarship to applicants who score in the top 2% of applicants on the SAT. Above what score will they offer this scholarship? **The score above which 2% of applicants fall is 1428.6, found using** `display 1100 + invnormal(0.98)*160`**. Note the score above which 2% of applicants fall is also the score below which 98% of the applicants fall.**







9. (**6 points—3 each**) Suppose the probability that a teenage driver gets into an accident during a one-year period is 0.12, and assume the probability of getting into an accident is independent across drivers.

   (a) A particular family has 5 teenage drivers. What is the probability that *at least one driver* in this family will have an accident over the coming year? Show or explain how you obtained your answer.

   This is an application of the binomial distribution, with 5 identical independent trials and 1 or more "successes" (getting into an accident) when the probability of "success" is 0.12.

```
. display binomialp(5,1,0.12) + binomialp(5,2,0.12) + ///
          binomialp(5,3,0.12) + binomialp(5,4,0.12) + ///
          binomialp(5,5,0.12)
.47226808

. *** or the function binomial gives you the probability of
      k or fewer successes
```

```
.
. display 1 - binomial(5,0,0.12)
.47226808
.
. *** or the function bionmialtail gives you the probability of
      k or more successes
.
. display binomialtail(5,1,0.12)
.47226808
```

You could alternatively use an online calculator to find probabilities from the binomial distribution, or the formula below where $\pi = 0.12$:

$$P(x \geq k) = \sum_{i=k}^{n} \binom{n}{i} \pi^i (1-\pi)^{n-i}$$

(b) Now consider the population of families with 5 teenage drivers, and define X as the number of accidents that occurred among their 5 drivers. For these families in a typical year, what is $E(X)$ and $sd(X)$? Show or explain how you obtained your answer.

In a binomial distribution, $\mathbf{E(X) = n\pi}$ and $\mathbf{Var(X) = n\pi(1-\pi)}$, so $\mathbf{E(X) = 5 * 0.12 = 0.6}$ and $\mathbf{sd(X) = \sqrt{5 * 0.12 * (1 - 0.12)} = 0.727}$. In other words the mean number of accidents in a year with 5 drivers is 0.6, with a standard deviation of 0.727.