Prelim 1 for BTRY6010/ILRST6100

September 28, 2017: 7:30pm-9:30pm

Name:________________________________

NetID:________________________________

Lab: (circle one)

Lab 402: Tues 1:25PM - 2:40PM

Lab 403: Tues 2:55PM - 4:10PM

Lab 404: Wed 2:55PM - 4:10PM

Lab 405: Tues 7:30PM - 8:45PM

Score: ______ / 75

[This exam is for 78 points. The maximum you can score is 75.]

Instructions

1. Please **do not turn to next page** until instructed to do so.
2. You have 120 minutes to complete this exam.
3. The last page of this exam has some useful formulas.
4. No textbook, calculators, phone, computer, notes, etc. allowed (please keep your phones off or do not bring them to the exam).
5. Please answer questions in the spaces provided. Feel free to use the blank sides for additional calculations.
6. When asked to calculate a number, it is sufficient to write out the full expression in numbers without actually calculating the value. E.g., $\frac{1+3\times\frac{4}{7}}{3+0.7}$ is a valid answer.
7. Please sign the following statement before beginning the exam.

Academic Integrity

I, ______________________, certify that this work is entirely my own. I will not look at any of my peers' answers or communicate in any way with my peers. I will not use any resource other than a pen/pencil. I will behave honorably in all ways and in accordance with Cornell's Code of Academic Integrity.

Signature: _________________ **Date:**_________

For each problem in this section, please circle one of the following answers. No justification is required.

1. [2 points]

The City Council of Ithaca wants to estimate the percentage of Ithacans who believe having Uber in Ithaca will benefit the city. They surveyed a sample of 200 Cornell students using e-mails and found that 121 believe having Uber will benefit the city. In this study, the population of interest is

a) the 200 students who were surveyed

b) All Ithacans who believe having Uber will benefit the city

c) ***All Ithacans***

d) the 121 students who believe having Uber will benefit the city

2. [2 points]

Let X be a random variable such that $Var(X) = 4$. What is the standard deviation of the random variable $2X + 3$?

a) ***4***

b) 7

c) 11

d) 16

3. [2 points]

Which of the following type of plots can be used to detect if the BTRY6010 Prelim 1 scores have a bimodal distribution?

a) scatterplot

b) boxplot

c) ***histogram***

d) none of the above

4. [2 points]

If a unimodal histogram is skewed right, we will expect that

(a) ***mean is larger than median***

(b) mean is smaller than median

(c) median is larger than mode

(d) median is smaller than mode

True or False? For each of the following questions, please answer either *True* or *False.* While justification is not required, it is encouraged and may allow in some cases for partial credit to be awarded.

5. [2 points]

Standard deviation describes the average distance of values from their mean.

FALSE. This is approximately the average distance, not the exact average distance.

6. [2 points]

For two independent random variables X and Y, $Var(X - Y) = Var(X) + Var(Y)$.

TRUE.

$Var(X - Y) = Var(X + (-Y)) = Var(X) + Var(-Y) = Var(X) + (-1)^2 \cdot Var(Y) = Var(X) + Var(Y)$

7. [2 points]

If the distribution of a data set is approximately bell-shaped with mean μ and standard deviation σ, we expect 95% of data to be smaller than $\mu + 2\sigma$.

FALSE. We would expect 97.5%.

8-10. [5 points]

Consider the Ithaca City Council survey described in Problem 1. Assume the 200 Cornell students in the sample were selected randomly using the list of all students in the Registrar's office. Then this is an example of a

8. Simple Random Sampling. (Write TRUE or FALSE)

FALSE. The population is all Ithacans.

9. Probability Sampling. (Write TRUE or FALSE)

FALSE.

10. Non-probability sampling. (Write TRUE or FALSE)

TRUE. Ithacans who are not Cornell students have no chance of being included in the sample.

[Hint: Consider the population of interest in this problem.]

11. [5 points]

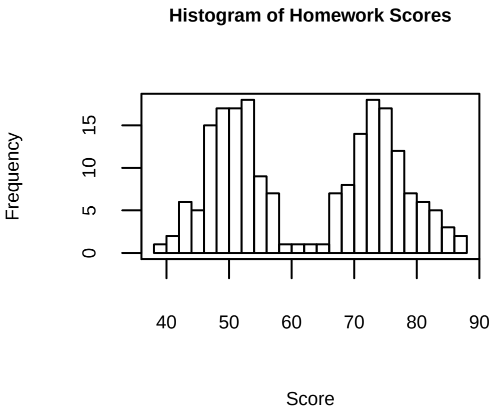
Fill in (A)-(E) with numerical values or variable names in the following R code chunk for calculating the mean and standard deviation of a set of numbers.

```
# Calculate sample mean
x = c(0, 1, 2, 4, -2, 5, -7)
n = length(x)
tot = (A) # <--- (A)=?

for (i in 1:n)
  tot = tot+x[i]

xbar = tot/(B) # <--- (B)=?

# Calculate sample standard deviation8
totb = 0
for (i in 1:(C)){ # <--- (C)=?
  totb=totb+(x[i]-(D))^2 # <--- (D)=?
}
s2 = totb/(E) # <--- (E)=?
s = sqrt(s2)
```

(A) = ***0*** (B) = ***n*** (C) = ***n*** (D) = ***xbar*** (E) = ***n-1***

Please answer the following questions. An answer without justification will not receive full credit.

12. [3 points] Below is a histogram of 200 homework scores in a class. Based on this histogram, explain if the mean score (in this case, 62.5) is a good summary statistic of the overall score distribution. If not, comment on how you will summarize the data.



The mean score is not a good summary statistic of the overall score distribution, since the histogram shows that the data has bimodal distribution with two modes around 50 and 75.

I will divide the data into two parts (i. more than 62 and ii. less than or equal to 62), and report summary statistics of each part separately.

13.

a) [3 points] Two events A and B are independent, with $P(A) = 0.5$ and $P(B) = 0.6$. Find $P(A \textit{ or } B)$.

Since A ***and*** B ***are independent,*** $P(A \cap B) = P(A)P(B) = 0.6 \times 0.5 = 0.3$***.***

Then $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.5 + 0.6 - 0.3 = 0.8$***.***

b) [2 points] Assume $0 < P(A) < 1$. Are the events A and A^c independent?

$A \cap A^c = \emptyset$***, thus*** A ***and*** A^c ***are disjoint. Now we know that non empty disjoint events could not be independent.***

c) Alice lives in a village where every family has two kids. Assume genders of kids are independent of each other, and each one is equally likely to be a boy or a girl.

i. [3 points] Find out the probability that a randomly selected family has 1 boy and 1 girl. Make sure to list all the outcomes **(a)** in the sample space and **(b)** in the event of interest.

The sample space is $\{B, B\}, \{B, G\}, \{G, B\}, \{G, G\}$***. The event contains two sample points*** $\{G, B\}, \{B, G\}$***. Thus probability of the event is*** $\frac{2}{4} = 0.5$***.***

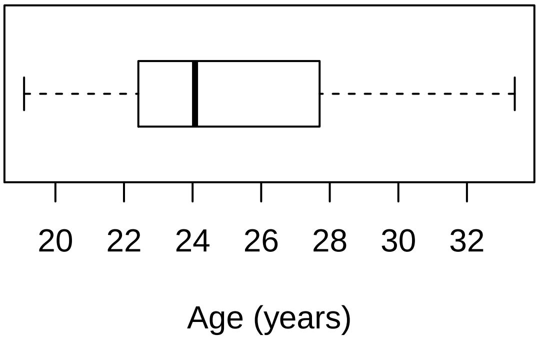
ii. [2 points] Calculate the probability that Alice's mom has a brother.

The event that there is at least one girl in the family (Alice's mom) has three sample points $\{B, G\}, \{G, B\}, \{G, G\}$***. The event that there is one boy and one girl (Alice's mom has a brother) consists of 2 sample points*** $\{B, G\}, \{G, B\}$***. Thus the required probability is*** $\frac{2}{3}$***.***

iii. [2 points] Calculate the probability that Alice's dad has a brother.

The event that there is at least one boy in the family (Alice's dad) contains three sample points $\{B, G\}, \{G, B\}, \{B, B\}$***. The event that both children are boys (Alice's dad has a brother) has only one possible sample point*** $\{B, B\}$***. Thus the required probability is*** $\frac{1}{3}$***.***

14. Here is a box plot of age of students in a graduate class.



a) [2 points] Approximately 75% of students are at most ***28*** years old.

The boxplot shows that the 75^{th} ***percentile or*** 3^{rd} ***quartile (***Q_3***) of age distribution is approximately*** 28 ***years.***

b) [2 points] Approximately ***50***% of students are between 22 and 28 years old.

50% ***of the observations lie between*** Q_3 ***and*** Q_1***, which are approximately*** 22 ***and*** 28 ***years.***

c) [2 points] Based on this boxplot, can you conclude that the distribution of student ages is skewed?

No, we need to look at the histogram of the data to make a conclusion about its shape. Although the two quartiles are at different distances from the median (suggesting skewness and lack of symmetry), this could be due to a bimodal data distribution.

d) [2 points] Comment on why the upper and lower whisker extensions have different lengths.

The upper and lower whiskers reach beyond Q_3 ***and*** Q_1 ***by a common length of*** $1.5 * IQR$***. However, if there is no data point present at*** $Q_3 + 1.5 * IQR$ ***or*** $Q_1 - 1.5 * IQR$***, the whiskers are shortened to the closest data points within these limits. Since these two closest data points may be at different distances from the quartiles, the extensions of the whiskers can be different.***

15. A finance company does credit check before approving loan applications. Their analytics team recently found that 30% of their approved loans went to applicants with low credit scores ($ < 600$) and the rest went to applicants with good or excellent credit scores (600 or more). Historically, it is known that loan accounts with poor credit scores have a 10% chance of being defaulted (not being completely repaid) while loan accounts with good or excellent credit scores have only a 2% chance of being defaulted.

a) [1 point] What is the chance that a randomly selected loan account is associated with good or excellent credit score?

We know $P(poor) = 0.3$ ***and hence*** $P(\textit{good or excellent}) = P(poor^C) = 1 - P(poor) = 0.7$

b) [2 points] What is the chance that a randomly selected loan account will be defaulted?

We know $P(default|poor) = 0.1$ ***and*** $P(default|poor^C) = 0.02$***. So by the law of total probability,***

$$\begin{aligned} P(default) &= P(poor)P(default|poor) + P(poor^C)P(default|poor^C) \\ &= 0.3 \times 0.1 + 0.7 \times 0.02 \\ &= 0.03 + 0.014 = 0.044 \end{aligned}$$

c) [2 points] Suppose a loan account is being defaulted. What is the chance that it was associated with good or excellent credit score?

Using Bayes theorem,

$$\begin{aligned} P(poor^C|default) &= \frac{P(default \cap poor^C)}{P(default)} \\ &= \frac{P(default|poor^C)P(poor^C)}{P(default)} \\ &= \frac{0.7 \times 0.02}{0.044} \end{aligned}$$

d) [5 points] John, an employee in the same finance company, manages 2 loan accounts. Every time an account is defaulted, he has to work 1 additional hour for following up with the applicant. Let X denote the number of additional hours John will have to spend on these 2 accounts. Write down the probability mass function of X [List all the values X can take and their probabilities].

Let $p = P(default) =$ ***0.044. Then X can take the values 0,1,2 and its pmf is given by***

$$\begin{aligned} P(X = 0) &= (1 - p)^2 \\ P(X = 1) &= 2p(1 - p) \\ P(X = 2) &= p^2 \end{aligned}$$

In other words, $X \sim Binomial(2, p)$

16. Bob is the manager at a Gimme! coffee shop near downtown Ithaca, where a box of spinach feta scones is a popular order among Cornell seminar committees. Every morning Bob starts with two boxes in the shelves, and orders more from their bakery if needed. Based on previous sales records, Bob knows that on average 1 box is sold on every weekday. Let X denote the number of boxes sold on a Monday.

a) [3 points] What sort of probability distribution would be appropriate for modeling X? Justify your answer. In addition to naming the type of distribution, also specify the values of all parameters.

- ***The most appropriate distribution for modeling X is the Poisson distribution.***
- ***This is because the event is this case is "one order of spinach feta scones". This event can occur anytime throughout the day and there is no upper bound on the number of times it can occur.***
- ***Since the average number of times this event occurs is 1, so the parameter of the Poisson distribution is also 1, i.e., $X \sim Poisson(1)$.***

b) [2 points] What is the expected value and standard deviation of X?

Expected value of X is 1.
Standard deviation of X is $\sqrt{1} = 1$.

c) [2 points] What is the chance that no box of scones will be sold on a Monday?

The chance that no box of scones will be sold is $P(X = 0) = \frac{e^{-1}1^0}{0!} = 1/e$.

d) [2 points] What is the chance that Bob will have to order additional boxes of scones from their bakery on a Monday?

Bob orders additional scones when he runs out of the 2 scones on his shelf and there is demand for more, i.e., he sells 3 or more scones. Hence the probability of this event is

$$\begin{aligned} P(X \geq 3) &= 1 - P(X < 3) \\ &= 1 - P(X = 0) - P(X = 1) - P(X = 2) \\ &= 1 - \frac{e^{-1}1^0}{0!} - \frac{e^{-1}1^1}{1!} - \frac{e^{-1}1^2}{2!} \\ &= 1 - e^{-1}\left(1 + 1 + \frac{1}{2}\right) = 1 - \frac{5}{2e} \end{aligned}$$

e) [2 points] What is the chance that Bob will have to order additional boxes of scones on both Monday and Tuesday next week?

If Y is the number of scones sold on a Tuesday then X and Y are independent random variables but have the same distribution. So the probability that Bob has to order additional scones on Monday and Tuesday is $P(X \geq 3 \text{ and } Y \geq 3) = P(X \geq 3) \cdot P(Y \geq 3) = \left(1 - \frac{5}{2e}\right)^2$.

17. For weekend flights to Vegas, an airline sells more tickets than the flight capacity anticipating some passengers to not show up in time (a ‘no-show’). Suppose a randomly selected passenger has a 1% chance of being a no-show. One weekend the airline has sold 120 tickets for a flight with 110 seats. Assume every passenger who bought a ticket is traveling alone (e.g., no family has bought a ticket for this flight). Let X denote the number of passengers who shows up for the flight this weekend.

a) [3 points] What kind of probability distribution is appropriate to model the random variable X? (justify your answer, be sure to give the name of the distribution and the values of its parameters).

Each person has a $P('no-show'^C) = 1-0.01 = 0.99$ ***probability of showing up for the flight. Thus a binomial distribution with*** $n = 120$ ***and*** $p = 0.99$ ***is appropriate because:***

- ***The presence of one person is independent of the presence of another***
- ***The maximum value of*** X ***is fixed in advance***
- ***The probability that one person shows up is fixed***

b) [1 point] If there were any families traveling, would it still be appropriate to use the same distribution? Explain why.

No, because then the “trials” (each person showing up) are no longer independent.

c) [2 points] What is the probability that all the passengers will show up (i.e., zero ‘no-show’)?

There are zero ‘no-shows’ if all 120 passengers turn up with probability

$$P(X = 120) = \binom{120}{120}(0.99)^{120}(0.01)^{0}$$

d) [2 points] How many passengers are expected to show up for the flight?

Expected number of people to show up for the flight is $\mathrm{E}(X) = np = 120 \times 0.99 = 118.8$

e) [2 points] What is the probability that the flight will be overbooked (i.e., more than 110 passengers shows up)?

More than 110 people show up with probability

$$P(\boldsymbol{oerbooked}) = P(X > 110) = \sum_{k=111}^{120} \mathrm{P}(X = k) = \sum_{k=111}^{120} \binom{120}{k}(0.99)^{k}(0.01)^{120-k}$$

Formula Sheet

The law of total probability:

$$P(B) = P(A)P(B|A) + P(A^C)P(B|A^C)$$

Bayes' theorem:

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(A^C)P(B|A^C)}.$$

Discrete Random Variables: X can take n distinct values $x_1, x_2, x_3, \ldots, x_n$

- probability mass function (pmf): $P(X = x_j) = p_j$, $j = 1, 2, 3, \ldots, n$
- expected value $E(X) = \mu = \sum_{j=1}^{n} x_j p_j$
- variance $\text{Var}(X) = \sum_{j=1}^{n} (x_j - \mu)^2 p_j$

Bernoulli distribution: $X \sim \text{Bernoulli}(p)$

- probability mass function (pmf):

$$P(X = x) = p^x (1-p)^{1-x} \text{ for } x = 0, 1$$

- expected value $E(X) = p$
- variance $\text{Var}(X) = np(1-p)$

Binomial distribution: $X \sim \text{Binomial}(n, p)$

- probability mass function (pmf):

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, 1, 2, \ldots, n$$

- expected value $E(X) = np$
- variance $\text{Var}(X) = np(1-p)$

Poisson distribution: $X \sim \text{Poisson}(\lambda)$

- probability mass function (pmf):

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!} \text{ for } x = 0, 1, 2, \ldots$$

- expected value $E(X) = \lambda$
- variance $\text{Var}(X) = \lambda$