Prelim 1 for BTRY6010/ILRST6100

October 3, 2019: 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

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 or code without actually calculating the value. E.g., $\frac{1+3\times\frac{4}{7}}{3+0.7}$ or `1 - pnorm(3)` are valid answers.
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 library advisory committee in Cornell University wants to know what percentage of current Cornell students would like to borrow physical material (books, dissertations, microfilms etc.) from other non-Cornell libraries and universities. They surveyed a sample of 200 Cornell students using e-mails and found that 180 would like to borrow from other non-Cornell libraries and universities. In this study, the population of interest is

a) the 200 students who were surveyed

b) *All current Cornell students*

c) All students in Ithaca who would like to borrow from non-Cornell libraries and universities

d) the 180 students who would like to borrow from other non-Cornell libraries and universities

2. [2 points]

Which is NOT one of the four fundamental principles in designing experiments?

a) Controlling

b) Randomization

c) *Treatment*

d) Replication

3. [2 points] Let X be a Poisson random variable with mean $\lambda = 4$. What is the standard deviation of the random variable $3X$?

a) 36

b) 18

c) *6*

d) 12

4. [2 points] Which of the following plots is the most effective to look for potential skewness in the income distribution of US citizens?

a) scatterplot

b) boxplot

c) *histogram*

d) barplot

5. [2 points] Let A and B be independent events. Which of the following must be true?

a) $P(A \cap B) = 0$

b) $P(A \cup B) = P(A) + P(B)$

c) $\underline{P(A) = P(A|B)}$

d) $P(A) = P(B|A)$

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.

6. [2 points] A sampling strategy falls under the category of probability sampling if every unit in the population has a positive probability of getting selected in the sample.

FALSE: every unit in the population must have a known and positive probability of getting selected in the sample.

7. [2 points] Eye color is an example of ordinal variable.

FALSE: Eye color is a nominal variable.

8. [2 points] Two events are considered independent if their intersection is empty.

FALSE: Two events are independent if probability of one event conditional on the other event is same as the probability of the first event.

9. [2 points] In the presence of extreme outliers in data, interquartile range is often considered a more appropriate measure of dispersion than standard deviation.

TRUE: interquartile range is robust to presence of outliers, standard deviation is not.

10. [2 points] The calculation $Var(X + X) = Var(X) + Var(X) = 2Var(X)$ is NOT correct.

TRUE: $Var(X + X) = Var(2X) = 4Var(X)$.

11. [2 points] In a boxplot, the upper whisker is always at $Q_3 + 1.5 * IQR$.

FALSE: upper whisker is set at the largest data point less than or equal to $Q_3 + 1.5 * IQR$.

12. [2 points] In a bell-shaped distribution, a z-score of 2 corresponds to 95^{th} percentile.

FALSE: In a bell-shaped distribution, a z-score of 2 corresponds to 97.5^{th} *percentile.*

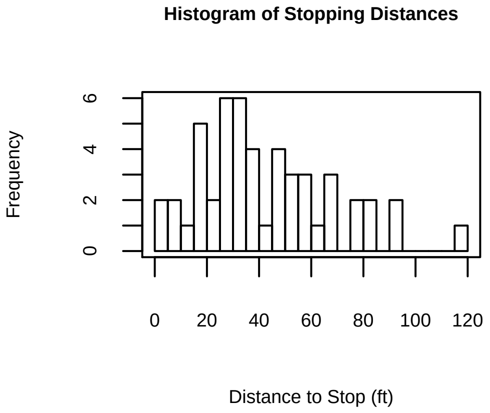
13. [5 points] Fill in (A)-(E) with numerical values or variable names in the following R code chunk for simulating a 1000 tosses of a fair coin and calculating the proportions of heads and tails.

```r
# simulate 1000 tosses of a fair coin
coin = c('head', 'tail')
n = (A) # <--- (A)=?
tosses = sample(coin, n, replace = TRUE)

# Calculate proportion of heads
count_head <- (B) # <--- (B)=?
for (i in 1:n)
  if (tosses[i] == 'head')
    count_head = count_head + (C) # <--- (C)=?

prop_head = count_head / (D) # <--- (D)=?

# Calculate proportion of tails
prop_tail = (E) - prop_head # <--- (E)=?
```

(A) = _ 1000 _ (B) = _ 0 _ (C) = _ 1 _ (D) = _ n _ (E) = _ 1 _

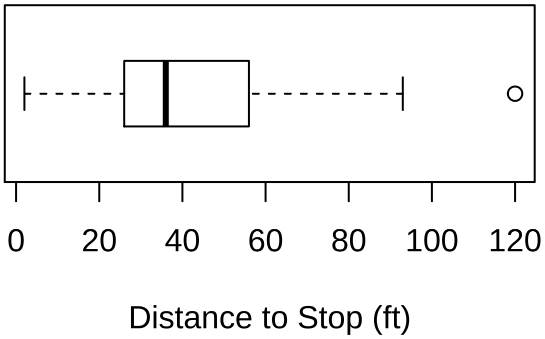
Please answer the following questions. An answer without justification will not receive full credit.

14. [4 points] Below is a histogram of stopping distances (in ft.) of 50 cars. Comment on the distribution of stopping distances based on this histogram.



Distribution of stopping times is unimodal, slightly skewed towards right. Almost all the stopping times of cars in the data are between 0 *and* 100 *feet, with a majority of them centered around* $25 - 35$ *feet. There is a potential outlier at* $115 \sim 120$ *feet.*

15. Here is a box plot of stopping distances of 50 cars from the last question. Complete the following sentences using information from this plot (reasonable approximations are acceptable).



a) [2 points] Approximately 25% of cars take at least _ 55 _ feet to stop upon applying brake.

The boxplot shows that the 75^{th} *percentile or* 3^{rd} *quartile* (Q_3) *of the distribution of stopping distance is approximately* 55 *ft. [Note: any answer between* 50 *and* 60 *ft is acceptable.]*

b) [2 points] The range of stopping distances of these 50 cars is _ 120 _ feet.

minimum: 0 *foot, maximum:* 120*feet. Range = maximum - minimum =* 120 *feet.*

c) [2 points] The inter quartile range of stopping distances of these 50 cars is *30* feet.

IQR = Q_3 - Q_1*. Acceptable answers for* Q_1*: between 20 and 30 feet, acceptable answers for* Q_3*: between 50 and 60 feet*

16. Heights of 10 year olds, regardless of gender, closely follow a normal distribution with mean 55 inches and standard deviation 6 inches.

a) [2 points] What is the probability that a randomly chosen 10 year old is shorter than 48 inches?

```
pnorm(48, mean = 55, sd = 6)
```

b) [2 points] If the tallest 10% of the class is considered "very tall", what is the height cutoff for "very tall"?

```
qnorm(0.9, mean = 55, sd = 6)
```

17. Foobaritis is a strange disease with no obvious symptoms that randomly develops in 1 in 1000 people across the world. Hello World Inc. has recently developed a highly accurate diagnostic test that correctly identifies 99% of all positive cases (in other words, sensitivity = 99%), and correctly identifies all unaffected individuals (in other words, specificity = 100%).

Let $T_\pm$ denote the test result and $D_\pm$ the disease state for a given individual.

a) [1 point] What is the probability that you have foobaritis?

$$P(D_+) = \frac{1}{1000} = 0.001$$

b) [2 points] What is the probability that you will develop foobaritis and test positive?

$$P(T_+ \cap D_+) = P(T_+|D_+)P(D_+) = 0.99 \times 0.001 = 0.00099$$

c) [3 points] What is the chance that you will test positive for foobaritis?

We apply the law of total probability to get

$$P(T_+) = P(T_+|D_+)P(D_+) + P(T_+|D_-)P(D_-) = 0.00099 + 0 \times 0.999 = 0.00099.$$

d) [4 points] Suppose that you have tested positive for foobaritis. What is the probability that you have foobaritis? Your doctor tells you that this means that you have a 99% chance of developing foobaritis, as this is the proportion of foobaritis patients that the test correctly diagnoses. Is her reasoning correct?

We apply Bayes' theorem, and simplify to get

$$P(D_+|T_+) = \frac{P(T_+|D_+)P(D_+)}{P(T_+|D_+)P(D_+) + P(T_+|D_-)P(D_-)} = \frac{0.00099}{0.00099} = 1.$$

(Alternatively, observe that the specificity is 1–there is no way a person without the disease could test positive, so $P(D_+|T_+) = 1$) The doctor's reasoning is incorrect. $0.99 = P(T_+|D_+)$, not $P(D_+|T_+) = 1$. A positive test must mean that you have a 100% chance of having foobaritis.

18. In the fictional land of middle earth, a small army of humans, elves, dwarves and hobbits is marching on its way to a battlefield. The table below is a distribution of favorite hobbies of these soldiers when they are on a break. For example, the first row indicates that out of 76 human soldiers in the army, 20 like to sharpen their weapons, 24 like to smoke pipe and the other 32 like to make wood craft when they are on a break.

	sharpen	smoke	make wood	Total
	weapon	pipe	craft	
Humans	20	24	32	76
Elves	2	5	10	17
Dwarves	14	7	4	25
Hobbits	0	3	0	3
Total	36	39	46	121

a) [2 points] What is the probability that a randomly chosen soldier is a dwarf?

$$P(a\ dwarf) = \frac{\#\ dwarves}{\#\ soldiers} = \frac{25}{121}$$

b) [2 points] What is the probability that a randomly chosen soldier is a dwarf and likes to make wood craft?

$$P(dwarf\ and\ make\ wood\ craft) = \frac{\#\ dwarves\ and\ make\ wood\ craft}{\#\ soldiers} = \frac{4}{121}$$

c) [2 points] Given that a soldier likes to make wood craft, what is the probability that he is an elf?

$$P(elf|make\ wood\ craft) = \frac{\#\ elves\ and\ make\ wood\ craft}{\#\ soldiers\ and\ make\ wood\ craft} = \frac{10}{46}$$

d) [2 points] Given that a soldier likes to smoke pipe and is not a human, what is the probability that he is a hobbit?

$$P(hobbit|smoke\ and\ not\ human) = \frac{\#\ hobbit\ AND\ smoke\ and\ not\ human}{\#\ smoke\ and\ not\ human} = \frac{3}{39-24} = \frac{3}{15}$$

e) [2 points] Define events A: a randomly chosen soldier is a hobbit, and B: a randomly chosen soldier likes smoking pipe. Are A and B independent?

$$P(A) = \frac{3}{121}, \quad P(B) = \frac{39}{121}, \quad P(A\ AND\ B) = \frac{3}{121}$$

$A\ and\ B\ are\ not\ independent\ because\ P(A)P(B) = \frac{3\times 39}{121\times 121} \neq P(A\ AND\ B)$

19. A tree ecologist is studying oak trees in a forest in upstate New York. She knows from prior research that oaks occur at 8.2 trees per square mile in this forest. She marks off a random one-square-mile region in the forest, and a research assistant will count the number of trees in that region. Let Y be the random variable representing the number of trees counted.

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

Because we are dealing with a number of events (trees) occurring in a fixed region of space, a Poisson model is appropriate. We are given a rate of 8.2 trees per square mile, so $Y \sim \text{Poisson}(\lambda = 8.2)$.

b) [2 point] Calculate the expected value and standard deviation of Y.

A Poisson random variable has $\lambda = E[Y] = \text{Var}(Y)$, so $E[Y] = 8.2$ and $\text{SD}(Y) = \sqrt{\text{Var}(Y)} = \sqrt{8.2}$.

c) [2 points] What is the probability that she counts 10 trees?

$P(Y = 10) = \frac{e^{-8.2} 8.2^{10}}{10!} =$ `dpois(10,lambda=8.2)`.

d) [3 points] Her research assistant stops to take a break and reports that she's counted four trees so far, so $Y \geq 4$. Knowing this, what is the probability that after she finishes counting, she's counted $Y = 10$ trees? Feel free to leave this in terms of probability statments, i.e. P(A).

Since she's counted four trees already, we know that $Y \geq 4$, but we don't what Y is yet (she hasn't finished counting). We are computing the conditional probability that she counts ten trees, given that there four or more trees. This is given by

$$P(Y = 10 | Y \geq 4) = \frac{P(Y = 10 \text{ and } Y \geq 4)}{P(Y \geq 4)} = \frac{P(Y = 10)}{P(Y \geq 4)}.$$

The simplification comes from noticing that "$Y = 10$ and $Y \geq 4$" is the same as "$Y = 10$". Drawing a Venn diagram can help with this if it's not clear. The result can be computed in R as `dpois(10,lambda=8.2)/(1 - ppois(3,lambda=8.2))`, where the `ppois` has a three because $P(Y \geq 4) = 1 - P(Y < 4) = 1 - P(Y \leq 3)$.

e) [2 points] In the course of the experiment, four tags are left on each tree. Let $X = 4Y$ denote the total number of tags used in the experiment. What is $\text{Var}(X)$?

Remembering that $\text{Var}(Y) = \lambda = 8.2$, we can plug in X and get

$$\text{Var}(X) = \text{Var}(4Y) = 4^2 \text{Var}(Y) = 16 \times 8.2 = 131.2$$

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 $\text{Var}(X) = \lambda$

Uniform distribution: $X \sim Unif(a, b)$

- probability density function (pdf):

$$f(x) = \frac{1}{b-a} \text{ for } a \leq x \leq b$$

- expected value $E(X) = (a+b)/2$
- variance $\text{Var}(X) = (b-a)^2/12$

Normal distribution: $X \sim \text{Normal}(\mu, \sigma)$

- probability density function (pdf):

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ for } -\infty < x < \infty$$

- expected value $E(X) = \mu$
- variance $\mathrm{Var}(X) = \sigma^2$
- If $X \sim N(\mu, \sigma)$, then $Z = \frac{X-\mu}{\sigma}$ has a $N(0,1)$ distribution.