Practice Exam 2

Name: Solutions

Section: ________________

You may use a calculator and two 8.5" by 11" sheets of notes (front and back), which you will turn in with your exam.

Please show all your work, including all calculations, and explain your answers. Whenever needed, please round numbers (including intermediate calculations) to the nearest 0.001.

Cell phones and any other electronic devices (aside from your calculator) are not permitted. No interaction of any sort is allowed with your classmates.

I Conceptual Questions

Please answer the following in no more than 1-2 sentences each.

1. (4 points) Describe stratified sampling and give an example of when it might be used.

In stratified sampling, we divide the population up into groups and take a random sample from within each group. Usually, the groups are of people or items that are similar to each other in some way, and we want to be sure our sample is representative of those different groups. For example, in political polls, pollsters often stratify by race in order to ensure that a representative sample of minority groups is included in the sample.

2. (4 points) Describe the difference between a population and a sample.

The population is the entire group of people or items we want to learn about. A sample is a smaller set of people or items that we select from the population. We only record data about the people in the sample, but we hope that what we learn about the sample will apply to the whole population.

3. (4 points) Describe the difference between disjoint events and independent events.

Disjoint events are events that share no outcomes in common, and so they could not happen at the same time.
Events are independent if knowing that one of them has occurred does not affect the probability that the other has occurred.

II Word Problems

1. Police often set up sobriety checkpoints (roadblocks where drivers are asked a few brief questions to allow the officer to judge whether or not the person may have been drinking). If the officer does not suspect a problem, drivers are released to go on their way. Otherwise, drivers are detained for a Breathalyzer test that will determine whether or not they will be arrested. The police say that based on the brief initial stop, trained officers can make the right decision 80% of the time. Suppose the police operate a sobriety checkpoint after 9:00 p.m. on a Saturday night, a time when national traffic safety experts suspect that about 12% of drivers have been drinking.

   Define A to be the event that a driver who is stopped has been drinking.

   Define B to be the event that the police decide a driver they stop has been drinking.

   a) Write a sentence describing what the event A^c means, in a similar style to the definition of the event A above.

   A^c is the event that a driver who is stopped has <u>not</u> been drinking.

   b) Write a sentence describing what the event B^c means, in a similar style to the definition of the event B above.

   B^c is the event that the police decide a driver they stop has <u>not</u> been drinking.

   c) For this part of the problem, you're just matching up numbers in the problem statement to the probabilities of events. No calculations are necessary.

   Find $P(A)$: 0.12

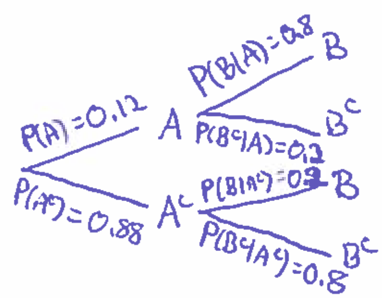
   Find $P(B|A)$: 0.8

   Find $P(B^c|A^c)$: 0.8

   d) For this part of the problem, you may find it helpful to draw a probability tree diagram.

   Find $P(B^c|A)$: 0.2

   Find $P(B|A^c)$: 0.2



   e) What's the probability that a driver who is stopped has been drinking, and the police decide (correctly) that he has been drinking?

   $P(A \text{ and } B) = P(A) \cdot P(B|A) = 0.12 \cdot 0.8 = 0.096$

   (in the probability tree diagram, multiply the probabilities along the branches leading to A and B)

f) If I know that the police decided a particular driver has been drinking, what is the probability that the driver actually had been drinking? That is, find $P(A|B)$

Use Bayes' Rule since we need to reverse the order of conditioning to get $P(A|B)$ from the things we know like $P(B|A)$.

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(A) \cdot P(B|A) + P(A^c) \cdot P(B|A^c)}$$

$$= \frac{0.12 \cdot 0.8}{0.12 \cdot 0.8 + 0.88 \cdot 0.2}$$

$$= 0.353$$

g) Are the events A and B independent? Justify your answer with a comparison of numeric probabilities that you have already calculated in previous parts of the question.

The events A and B are <u>not</u> independent, since

$$P(A|B) \neq P(A)$$
$$0.353 \neq 0.12$$

(we could also have checked whether $P(B|A) = P(B)$ or $P(A \text{ and } B) = P(A) \cdot P(B)$)

2. After menopause, some women take supplemental estrogen. There is some concern that if these women also drink alcohol, their estrogen levels will rise too high. Twelve volunteers who were receiving supplemental estrogen were randomly divided into two groups, as were 12 other volunteers not on estrogen. In each case, one group drank an alcoholic beverage, the other a nonalcoholic beverage. An hour later, everyone's estrogen level was checked. Only those on supplemental estrogen who drank alcohol showed a marked increase.

a) Was this study an experiment or an observational study?

Experiment.

b) Was blocking used? If so, describe how and why.

Blocking was used, since the study participants were divided into two groups (those receiving supplemental estrogen and those not receiving supplemental estrogen), and assignment to either consume alcohol or not consume alcohol was made within those groups. This ensures that all four combinations of alcohol consumption ~~for~~ or not and taking estrogen or not are represented in the data.

c) Why was it important to randomize the assignment of the study participants to the groups that either drank or did not drink an alcoholic beverage? What problem is randomization trying to prevent?

We want to ensure that there are not any systematic differences between those who are in the group that drinks an alcoholic beverage and those who are not in the group that drinks an alcoholic beverage. Randomization helps with this reducing the chances that any differences in estrogen levels between the groups are due to a lurking variable and allowing us to be more comfortable in making statements about causality based on the experimental results.

d) What is blinding? From the description above, is there any evidence that blinding was used in this study?

Blinding is when people involved in the study, either participants or the people running the study, do not know which study participants have been assigned to which treatments. There is no evidence of blinding in the study description above.

e) Is it possible to make a conclusion about a causal connection between estrogen supplements, alcoholic beverages, and estrogen levels based on this study? Why or why not?

Because of the experimental design which randomized the study participants to the treatment groups, I feel more comfortable making a conclusion that there is a causal relationship than I would in an observational study. However it is worth noting that the study participants were all volunteers and the assignment to take estrogen or not was not randomized. This leaves the possibility that there may be differences between the people who were taking estrogen and those not taking estrogen that were not accounted for in this study design.

3. In the 1980s, it was generally believed that congenital disorders (that is, a health condition present at or before birth) affected about 5% of the nation's children. Some people believe that the increase in the number of chemicals in the environment has led to an increase in the incidence of congenital disorders. A recent study examined randomly selected sample of 392 children and counted the number who showed signs of a congenital disorder.

   a) In this study, what is the population parameter of interest? (We don't have a number for this; just describe the parameter in a sentence.)

   The proportion of children born recently who have congenital disorders.

   b) Let X be a random variable equal to the number of children in the sample who show signs of a congenital disorder. What would the sampling distribution of X be, assuming the proportion of children born with congenital disorders has not changed since the 1980's? Check all assumptions.

   $X \sim \text{Binomial}(392, 0.05)$
   - 2 outcomes: child has congenital disorder or not
   - same probability of congenital disorder for each randomly selected child
   - independence
     - the children included in the study were selected randomly, so there is not likely to be any connection between them (e.g. they are probably not related, etc.)
     - the sample size, 392 children, is much less than 10% of all children born recently.

   All assumptions for the binomial model are met.

   c) Let $\widehat{p}$ be the proportion of children in the sample who show signs of a congenital disorder. What would the sampling distribution of $\widehat{p}$ be, assuming the proportion of children born with congenital disorders has not changed since the 1980's? Check all assumptions (if you already checked any in part b, you do not need to check them again).

   $$\hat{p} \sim \text{Normal}\left(0.05, \sqrt{\frac{0.05(1-0.05)}{392}}\right)$$
   $$\hat{p} \sim \text{Normal}(0.05, 0.011)$$

   We need to check all the same assumptions as in part b), plus one more:

   $$np \geq 10 \quad \text{and} \quad n(1-p) \geq 10$$
   $$392 \cdot 0.05 = 19.6 \geq 10 \quad \text{and} \quad 392(.95) = 372.4 \geq 10$$

   All assumptions for the normal model for the sampling distribution of $\hat{p}$ are met.

d) In the sample, 24 of the children showed signs of a congenital disorder. Using your choice of the distribution from part b) or part c) and the R output below, find the probability of getting at least 24 children with congenital disorders in this sample, assuming that the proportion of children born with congenital disorders has not changed since the 1980's. You will only need one of the numbers below to calculate your answer.

```
> pbinom(q = 23, size = 392, prob = 0.05)

[1] 0.8189341

> pbinom(q = 24, size = 392, prob = 0.05)

[1] 0.8704152

> pnorm(q = 0.0609, mean = 0.05, sd = 0.218)

[1] 0.5199388

> pnorm(q = 0.0609, mean = 0.05, sd = 0.044)

[1] 0.5978273

> pnorm(q = 0.0609, mean = 0.05, sd = 0.011)

[1] 0.839135
```

I should have put $24/392 = 0.612$ here when I wrote the problem.

Using binomial: $P(X \geq 24) = 1 - P(X \leq 23)$

$= 1 - 0.819$

$= 0.181$

Using normal: $P(\hat{p} \geq 24/392) = 1 - P(\hat{p} \leq 24/392)$

$= 1 - P(\hat{p} \leq 0.061)$

$\approx 1 - 0.839$

$= 0.161$

4. Twix Minis have, on average, 2.00 grams of saturated fat per bar, with a standard deviation of 0.17 grams. The amount of saturated fat per bar follows a normal distribution. The day after Halloween I ate 3 Twix Minis.

a) What is the mean and standard deviation of the **total amount** of saturated fat I consumed?

Define X_1 = amount of saturated fat in first bar, $E(X_1)=2.00$, $SD(X_1)=0.17$

X_2 = " " " " " second ", $E(X_2)=2.00$, $SD(X_2)=0.17$

X_3 = " " " " " third ", $E(X_3)=2.00$, $SD(X_3)=0.17$

Total saturated fat across all 3 bars is

$$E(X_1+X_2+X_3)=E(X_1)+E(X_2)+E(X_3)=3\cdot 2.00=6.00$$

$$SD(X_1+X_2+X_3)=\sqrt{Var(X_1+X_2+X_3)}=\sqrt{Var(X_1)+Var(X_2)+Var(X_3)}$$

$$=\sqrt{SD(X_1)^2+SD(X_2)^2+SD(X_3)^2}$$

$$=\sqrt{3\cdot 0.17^2}$$

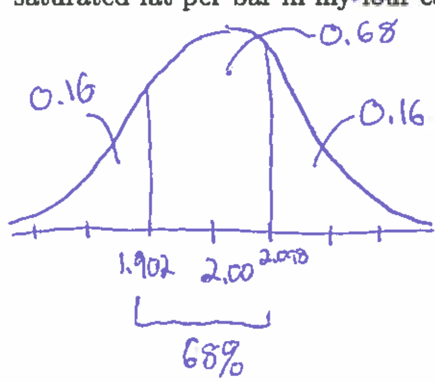
$$=0.294$$

b) What distribution would you use to model the **sample mean** of the number of grams of saturated fat per bar? Be as specific as possible.

$$\bar{X} \sim Normal(\underset{\mu}{2.00}, \underset{\sigma/\sqrt{n}}{0.17/\sqrt{3}})$$

$$\bar{X} \sim Normal(2.00, 0.098)$$

c) Using the 68/95/99.7 Rule, find the probability that the average number of grams of saturated fat per bar in my ~~four~~ three candy bars was greater than 1.902 grams.



Since the area between 1.902 grams and 2.098 grams is 0.68 and the normal density curve is symmetric, the area to the left of 1.902 is 0.16 and the area to the right of 2.098 is 0.16.

The total area to the right of 2.098 is 0.84.

$P(\bar{X} \geq 1.902) = 0.84$