Midterm 2 Suggested Review Problems

Here are problems that are similar to the ones you might see on the exam. Be sure to also review old quiz and workshop questions too.

Experiments vs. Observational Studies

Know the difference between explanatory variables, response variables, and lurking variables. Also, make sure that you understand why randomized controlled experiments let you establish cause and effect but observational studies do not.

1. A NY Times article titled *Risks: Smokers Found More Prone to Dementia* states the following:

   > "Researchers analyzed data from 23,123 health plan members who participated in a voluntary exam and health behavior survey from 1978 to 1985, when they were 50-60 years old. 23 years later, about 25% of the group had dementia, including 1,136 with Alzheimer's disease and 416 with vascular dementia. After adjusting for other factors, the researchers concluded that pack-a-day smokers were 37% more likely than nonsmokers to develop dementia, and the risks went up with increased smoking; 44% for one to two packs a day; and twice the risk for more than two packs."

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

   Solution: It's an observation study because there was no treatment imposed on the individuals.

   (b) What are the explanatory and response variables?

   Solution: Explanatory: smoking frequency (packs per day), Response: whether they have dementia.

   (c) Based on this study, can we conclude that smoking causes dementia later in life? Explain your reasoning.

   Solution: No, you can't prove cause & effect with an observational study. There might be lurking variables like diet, exercise, income, etc. that are the real cause of the correlation.

2. In a public health study on the effects of consumption of fruits and vegetables on psychological well-being in young adults, participants were randomly assigned to three groups: (1) diet as usual, (2) an intervention involving text message reminders to increase their fruits and vegetable consumption plus a voucher to purchase them, or (3) a fruit and vegetable intervention in which participants were given two additional daily servings of fresh fruits and vegetables to consume on top of their normal diet. Participants were asked to take a nightly survey on their smartphones. Participants were student volunteers at the University of Otago, New Zealand. At the end of the 14-day study, only participants in the third group showed improvements to their psychological well-being across the 14-days relative to the other groups.

   (a) What type of study is this?

   Solution: This was an experiment.

   (b) Identify the explanatory and response variables.

   Solution: Explanatory: treatment group (or fruit and vegetable consumption). Response: psychological well-being.

   (c) Were the individuals in the study a random sample from the population?

   Solution: No, they were student volunteers at one university.

   (d) Were the individuals randomly assigned to different treatment groups?

   Solution: Yes.

   (e) Does this study support the claim that giving young adults fresh fruits and vegetables to eat can *cause* psychological benefits?

   Solution: Yes! You can establish cause and effect with a randomized controlled experiment like this one.

Probability

3. Data collected at elementary schools in DeKalb County, GA suggest that each year roughly 25% of students miss exactly one day of school, 15% miss 2 days, and 28% miss 3 or more days due to sickness.

   (a) What is the probability that a student chosen at random doesn't miss any days of school due to sickness this year?

Solution: Let X represent the number of days a kid misses school.

$$P(X = 0) = 100\% - 25\% - 15\% - 28\% = 32\%.$$

(b) What is the probability that a student chosen at random misses no more than one day?

Solution:

$$P(X \leq 1) = 32\% + 25\% = 57\%.$$

(c) What is the probability that a student chosen at random misses at least one day?

Solution:

$$P(X \geq 1) = 100\% - 32\% = 68\%.$$

Weighted Averages and Expected Value

Expected value (also known as the theoretical average) is the weighted average of the outcomes in a probability model. Make sure you understand why it is called "expected" and how to calculate it. You should know the Law of Large Numbers too.

4. Andy is always looking for ways to make money fast. Lately, he has been trying to make money by gambling. Here is the game he is considering playing: The game costs \$2 to play. He draws a card from a deck. If he gets a number card (2-10), he wins nothing. For any face card (jack, queen or king), he wins \$3. For any ace, he wins \$5, and he wins an extra \$20 if he draws the ace of clubs.

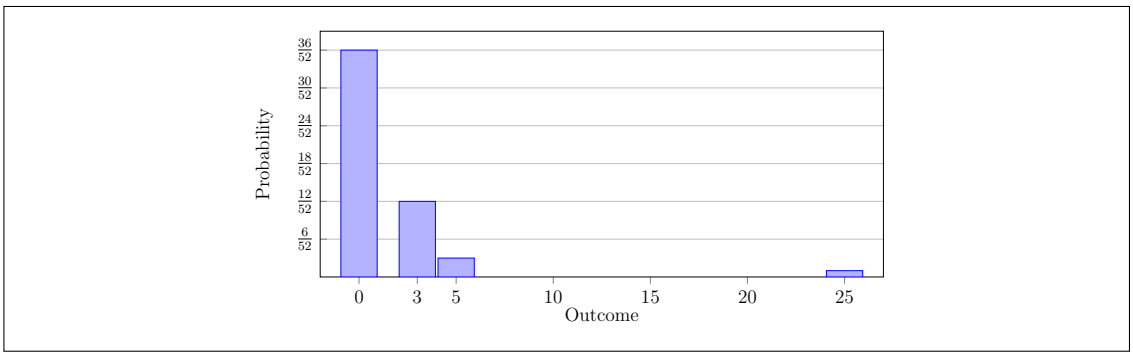
   (a) Create a probability model for this game.

   Solution:

Outcome (Amount Won)	\$0	\$3	\$5	\$25
Probability	$\frac{36}{52}$	$\frac{12}{52}$	$\frac{3}{52}$	$\frac{1}{52}$

   (b) Draw a probability histogram for the game.

   Solution:



(c) Find the expected value of the game.

Solution:

$$0\left(\tfrac{36}{52}\right) + 3\left(\tfrac{12}{52}\right) + 5\left(\tfrac{3}{52}\right) + 25\left(\tfrac{1}{52}\right) = \$1.461.$$

Random Variables

When we use a letter to represent the numerical outcome of a probability model, that letter is called a random variable. You should be comfortable with the way random variables are used in notation, like $P(X > 5)$ for example.

5. Suppose X is a $N(500, 80)$ random variable. Find the following.

(a) $P(X > 540)$

Solution: Use the normal distributions app.

$$P(X > 540) = 30.9\%.$$

(b) $P(400 < X < 540)$

Solution:

$$P(400 < X < 540) = 100 - 10.6\% - 30.9\% = 58.5\%.$$

Sampling Distributions

Make sure you know the shape, center, and spread for the sampling distributions of the sample mean $\bar{x}$ and the sample proportion $\hat{p}$. Be sure you can describe how they change as the sample size gets larger.

6. Data collected by the Substance Abuse and Mental Health Services Administration (SAMSHA) suggests that 69.7% of 18-20 year olds consumed alcoholic beverages in any given year. Suppose we consider a random sample of fifty 18-20 year olds.

   (a) Describe the sampling distribution for the proportion of students in the sample that have consumed alcoholic beverages in the past year. What is the shape, center, and spread of the distribution?

   Solution:

   Shape The shape will be approximately normal.

   Center The center is 69.7%.

   Spread The spread is the standard deviation for $\hat{p}$:

   $$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{N}} = \sqrt{\frac{0.697(1-0.697)}{50}} = 6.5\%.$$

   (b) Estimate the probability that at least 80% of the individuals in the sample have consumed alcohol in the past year.

   Solution: Use the normal distributions app to find

   $$P(X \geq 80\%) = 5.7\%.$$

7. (Exercise 5.1 from OpenIntro Statistics) For each of the following situations, state whether the parameter of interest is a mean or a proportion. It may be helpful to examine whether individual responses are numerical or categorical.

   (a) In a survey, one hundred college students are asked how many hours per week they spend on the Internet.

   Solution: Mean (response is numerical)

   (b) In a survey, one hundred college students are asked: "What percentage of the time you spend on the Internet is part of your course work?"

   Solution: Mean (response is a percent, which is numerical)

(c) In a survey, one hundred college students are asked whether or not they cited information from Wikipedia in their papers.

Solution: Proportion (response is categorical)

(d) In a survey, one hundred college students are asked what percentage of their total weekly spending is on alcoholic beverages.

Solution: Mean (response is numerical)

(e) In a sample of one hundred recent college graduates, it is found that 85 percent expect to get a job within one year of their graduation date.

Solution: Proportion (response is categorical: expect to have a job or not?)

8. As part of a quality control process for computer chips, an engineer at a factory randomly samples 212 chips during a week of production to test the current rate of chips with severe defects. She finds that 27 of the chips are defective.

(a) What population is under consideration in the data set?

Solution: All the chips produced at the facility.

(b) What parameter is being estimated?

Solution: The proportion of all chips with severe defects.

(c) What is the best guess estimate for the parameter?

Solution: The sample proportion

$$\hat{p} = \frac{27}{212} = 12.7\%.$$

(d) Calculate the standard error $\text{SE}_{\hat{p}} = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{N}}$.

Solution:

$$\text{SE}_{\hat{p}} = \sqrt{\frac{0.127(1-0.127)}{212}} = 2.29\%.$$

(e) The historical rate of defects is 10%. Should the engineer be surprised by the observed rate of defects during the current week?

Solution: The rate of defects (12.7%) is only a little more than one standard deviation (2.29%) above the historical rate of defects, so that is not a surprising result.

9. American adults have an average weight of 170 lbs. with a standard deviation of 40 lbs.

(a) Describe the sampling distribution for the average weight of a random group of 25 adults.

Solution:

Shape The shape is approximately normal.

Center $\mu = 170$ lbs.

Spread The standard deviation of $\bar{x}$:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} = \frac{40}{\sqrt{25}} = 8 \text{ lbs.}$$

(b) Estimate $P(\bar{x} > 180)$ using the normal distribution.

Solution:

$$P(\bar{x} > 180) = 10.6\%.$$

Confidence Intervals for Proportions

Make sure you understand that the confidence interval is a tool to estimate the population proportion p using the sample proportion $\hat{p}$. The confidence level is how confident we are that the true value of p is inside the interval.

10. (Exercise 6.45 from OpenIntro Statistics) We are interested in estimating the proportion of graduates at a mid-sized university who found a job within one year of completing their undergraduate degree. Suppose we conduct a survey and find out that 348 of the 400 randomly sampled graduates found jobs. The graduating class under consideration included over 4500 students.

    (a) Describe the population parameter of interest. What is the value of the point estimate of this parameter?

    Solution: The parameter of interest is the percent of all graduates at the university who found a job within one year. The best guess (point estimate) is $\hat{p} = \frac{348}{400} = 87\%$.

    (b) Check if the conditions for constructing a confidence interval based on these data are met.

    Solution: The graduates were randomly sampled, which means we can hope that there was no bias. And there were 348 successes and 52 failures, so both are large enough to assume normality.

    (c) Calculate a 95% confidence interval for the proportion of graduates who found a job within one year of completing their undergraduate degree at this university, and interpret it in the context of the data.

    Solution:

    $$87\% \pm 1.96\sqrt{\frac{0.87(1-0.87)}{400}} = 87\% \pm 3.3\%.$$

    So we can be 95% sure that the percent of graduates who found jobs within one year is between 83.7% and 90.3%.

    (d) What does "95% confidence" mean?

    Solution: It means we can be 95% sure that the confidence interval contains the parameter of interest.

    (e) Now calculate a 99% confidence interval for the same parameter and interpret it in the context of the data.

Solution:

$$87\% \pm 2.576\sqrt{\frac{0.87(1-0.87)}{400}} = 87\% \pm 4.3\%.$$

So we can be 99% sure that the percent of graduates who found jobs within one year is between 82.7% and 91.3%.

(f) Which has a larger margin of error, the 95% or the 99% confidence interval?

Solution: The 99% confidence interval has a larger margin of error.