Optional Practice Problems: Lecture 07-10

URBAN 520: Quantitative Methods in Urban Design and Planning

- These problems are optional and not graded. There is nothing to submit.
- All problems are from *OpenIntro Statistics*, 4th edition (Diez, Çetinkaya-Rundel, and Barr, 2019), free at openintro.org. The number after each problem is its exercise number in the book. Answers to the odd-numbered exercises are in Appendix A of the book.
- ☆ marks the problems, which the approaches/ways of thinking used may not be fully covered in the assignments or example sections in the lecture notes.

Question 1 Basic Probability

A. ☆ **[Disjoint Events]** *True or false.* Determine if the statements below are true or false, and explain your reasoning. (OpenIntro 3.1)

(a) If a fair coin is tossed many times and the last eight tosses are all heads, then the chance that the next toss will be heads is somewhat less than 50%.
(b) Drawing a face card (jack, queen, or king) and drawing a red card from a full deck of playing cards are mutually exclusive events.
(c) Drawing a face card and drawing an ace from a full deck of playing cards are mutually exclusive events.

B. ☆ **[Addition Rule]** *Swing voters.* A Pew Research survey asked 2,373 randomly sampled registered voters their political affiliation (Republican, Democrat, or Independent) and whether or not they identify as swing voters. 35% of respondents identified as Independent, 23% identified as swing voters, and 11% identified as both. (OpenIntro 3.7)

(a) Are being Independent and being a swing voter disjoint, i.e. mutually exclusive?
(b) Draw a Venn diagram summarizing the variables and their associated probabilities.
(c) What percent of voters are Independent but not swing voters?
(d) What percent of voters are Independent or swing voters?
(e) What percent of voters are neither Independent nor swing voters?
(f) Is the event that someone is a swing voter independent of the event that someone is a political Independent?

C. **[Conditional Probability]** *Global warming.* A Pew Research poll asked 1,306 Americans "From what you've read and heard, is there solid evidence that the average temperature on earth has been getting warmer over the past few decades, or not?". The table below shows the distribution of responses by party and ideology, where the counts have been replaced with relative frequencies. (OpenIntro 3.15)

	Response			
Party and Ideology	Earth is warming	Not warming	Don't Know Refuse	Total
Conservative Republican	0.11	0.20	0.02	0.33
Mod/Lib Republican	0.06	0.06	0.01	0.13
Mod/Cons Democrat	0.25	0.07	0.02	0.34
Liberal Democrat	0.18	0.01	0.01	0.20
Total	0.60	0.34	0.06	1.00

(a) Are believing that the earth is warming and being a liberal Democrat mutually exclusive?
(b) What is the probability that a randomly chosen respondent believes the earth is warming or is a liberal Democrat?
(c) What is the probability that a randomly chosen respondent believes the earth is warming given that he is a liberal Democrat?
(d) What is the probability that a randomly chosen respondent believes the earth is warming given that he is a conservative Republican?
(e) Does it appear that whether or not a respondent believes the earth is warming is independent of their party and ideology? Explain your reasoning.
(f) What is the probability that a randomly chosen respondent is a moderate/liberal Republican given that he does not believe that the earth is warming?

D. [**Independence**] *Educational attainment of couples.* The table below shows the distribution of education level attained by US residents by gender based on data collected in the 2010 American Community Survey. (OpenIntro 3.11)

	Gender	
Highest education attained	Male	Female
Less than 9th grade	0.07	0.13
9th to 12th grade, no diploma	0.10	0.09
HS graduate (or equivalent)	0.30	0.20
Some college, no degree	0.22	0.24
Associate's degree	0.06	0.08
Bachelor's degree	0.16	0.17
Graduate or professional degree	0.09	0.09
Total	1.00	1.00

(a) What is the probability that a randomly chosen man has at least a Bachelor's degree?
(b) What is the probability that a randomly chosen woman has at least a Bachelor's degree?
(c) What is the probability that a man and a woman getting married both have at least a Bachelor's degree? Note any assumptions you must make to answer this question.
(d) If you made an assumption in part (c), do you think it was reasonable? If you didn't make an assumption, double check your earlier answer and then return to this part.

E. [**Bayes' Rule**] *HIV in Swaziland.* Swaziland has the highest HIV prevalence in the world: 25.9% of this country's population is infected with HIV. The ELISA test is one of the first and most accurate tests for HIV. For those who carry HIV, the ELISA test is 99.7% accurate. For those who do not carry HIV, the test is 92.6% accurate. If an individual from Swaziland has tested positive, what is the probability that he carries HIV? (OpenIntro 3.41)

F. [**Bayes' Rule**] *It's never lupus.* Lupus is a medical phenomenon where antibodies that are supposed to attack foreign cells to prevent infections instead see plasma proteins as foreign bodies, leading to a high risk of blood clotting. It is believed that 2% of the population suffer from this disease. The test is 98% accurate if a person actually has the disease. The test is 74% accurate if a person does not have the disease. There is a line from the Fox television show *House* that is often used after a patient tests positive for lupus: "It's never lupus." Do you think there is truth to this statement? Use appropriate probabilities to support your answer. (OpenIntro 3.21)

Question 2 Probability Distributions

A. ☆ **[Probability Model]** *Roulette winnings.* In the game of roulette, a wheel is spun and you place bets on where it will stop. One popular bet is that it will stop on a red slot; such a bet has an 18/38 chance of winning. If it stops on red, you double the money you bet. If not, you lose the money you bet. Suppose you play 3 times, each time with a $1 bet. Let Y represent the total amount won or lost. Write a probability model for Y. (OpenIntro 4.35)

B. **[Independent Trials]** *Married women.* The American Community Survey estimates that 47.1% of women ages 15 years and over are married. (OpenIntro 4.41)

(a) We randomly select three women between these ages. What is the probability that the third woman selected is the only one who is married?
(b) What is the probability that all three randomly selected women are married?
(c) On average, how many women would you expect to sample before selecting a married woman? What is the standard deviation?
(d) If the proportion of married women was actually 30%, how many women would you expect to sample before selecting a married woman? What is the standard deviation?
(e) Based on your answers to parts (c) and (d), how does decreasing the probability of an event affect the mean and standard deviation of the wait time until success?

C. **[Binomial]** *Underage drinking, Part I.* 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. (OpenIntro 4.17)

(a) Suppose a random sample of ten 18-20 year olds is taken. Is the use of the binomial distribution appropriate for calculating the probability that exactly six consumed alcoholic beverages? Explain.
(b) Calculate the probability that exactly 6 out of 10 randomly sampled 18- 20 year olds consumed an alcoholic drink.
(c) What is the probability that exactly four out of ten 18-20 year olds have *not* consumed an alcoholic beverage?
(d) What is the probability that at most 2 out of 5 randomly sampled 18-20 year olds have consumed alcoholic beverages?
(e) What is the probability that at least 1 out of 5 randomly sampled 18-20 year olds have consumed alcoholic beverages?

Underage drinking, Part II. We learned in Exercise 4.17 that about 70% of 18-20 year olds consumed alcoholic beverages in any given year. We now consider a random sample of fifty 18-20 year olds. (OpenIntro 4.19)

(a) How many people would you expect to have consumed alcoholic beverages? And with what standard deviation?
(b) Would you be surprised if there were 45 or more people who have consumed alcoholic beverages?
(c) What is the probability that 45 or more people in this sample have consumed alcoholic beverages? How does this probability relate to your answer to part (b)?

D. ☆ **[Z-Scores]** *GRE scores, Part I.* Sophia who took the Graduate Record Examination (GRE) scored 160 on the Verbal Reasoning section and 157 on the Quantitative Reasoning section. The mean score for Verbal Reasoning section for all test takers was 151 with a standard deviation of 7, and the mean score for the Quantitative Reasoning was 153 with a standard deviation of 7.67. Suppose that both distributions are nearly normal. (OpenIntro 4.3)

(a) Write down the short-hand for these two normal distributions.
(b) What is Sophia's Z-score on the Verbal Reasoning section? On the Quantitative Reasoning section? Draw a standard normal distribution curve and mark these two Z-scores.

(c) What do these Z-scores tell you?
(d) Relative to others, which section did she do better on?
(e) Find her percentile scores for the two exams.
(f) What percent of the test takers did better than her on the Verbal Reasoning section? On the Quantitative Reasoning section?
(g) Explain why simply comparing raw scores from the two sections could lead to an incorrect conclusion as to which section a student did better on.
(h) If the distributions of the scores on these exams are not nearly normal, would your answers to parts (b) - (f) change? Explain your reasoning.

GRE scores, Part II. In Exercise 4.3 we saw two distributions for GRE scores: $N(\mu = 151, \sigma = 7)$ for the verbal part of the exam and $N(\mu = 153, \sigma = 7.67)$ for the quantitative part. Use this information to compute each of the following: (OpenIntro 4.5)

(a) The score of a student who scored in the 80^{th} percentile on the Quantitative Reasoning section.
(b) The score of a student who scored worse than 70% of the test takers in the Verbal Reasoning section.

E. ☆ **[Normal Percentile]** *Auto insurance premiums.* Suppose a newspaper article states that the distribution of auto insurance premiums for residents of California is approximately normal with a mean of \$1,650. The article also states that 25% of California residents pay more than \$1,800. (OpenIntro 4.39)

(a) What is the Z-score that corresponds to the top 25% (or the 75^{th} percentile) of the standard normal distribution?
(b) What is the mean insurance cost? What is the cutoff for the 75th percentile?
(c) Identify the standard deviation of insurance premiums in California.

F. **[Normal, Binomial]** *Heights of 10 year olds, Part III.* Heights of 10 year olds, regardless of gender, closely follow a normal distribution with mean 55 inches and standard deviation 6 inches. (OpenIntro 4.47)

(a) What fraction of 10 year olds are taller than 76 inches?
(b) If there are 2,000 10 year olds entering Six Flags Magic Mountain in a single day, then compute the expected number of 10 year olds who are at least 76 inches tall. (You may assume the heights of the 10-year olds are independent.)
(c) Using the binomial distribution, compute the probability that 0 of the 2,000 10 year olds will be at least 76 inches tall.

G. ☆ **[Normal Approximation]** *University admissions.* Suppose a university announced that it admitted 2,500 students for the following year's freshman class. However, the university has dorm room spots for only 1,786 freshman students. If there is a 70% chance that an admitted student will decide to accept the offer and attend this university, what is the approximate probability that the university will not have enough dormitory room spots for the freshman class? (OpenIntro 4.37)

Question 3 Sampling and Sampling Distributions

A. **[Parameter, Statistic]** *Relaxing after work.* The General Social Survey asked the question, "After an average work day, about how many hours do you have to relax or pursue activities that you enjoy?" to a random sample of 1,155 Americans. The average relaxing time was found to be 1.65 hours. Determine which of the following is an observation, a variable, a sample statistic (value calculated based on the observed sample), or a population parameter. (OpenIntro 1.17)

(a) An American in the sample.

(b) Number of hours spent relaxing after an average work day.
(c) 1.65.
(d) Average number of hours all Americans spend relaxing after an average work day.

B. ☆ **[Sampling Methods]** *Evaluate sampling methods.* A university wants to determine what fraction of its undergraduate student body support a new $25 annual fee to improve the student union. For each proposed method below, indicate whether the method is reasonable or not. (OpenIntro 1.23)

(a) Survey a simple random sample of 500 students.
(b) Stratify students by their field of study, then sample 10% of students from each stratum.
(c) Cluster students by their ages (e.g. 18 years old in one cluster, 19 years old in one cluster, etc.), then randomly sample three clusters and survey all students in those clusters.

C. ☆ **[Sampling Bias]** *Sampling strategies.* A statistics student who is curious about the relationship between the amount of time students spend on social networking sites and their performance at school decides to conduct a survey. Various research strategies for collecting data are described below. In each, name the sampling method proposed and any bias you might expect. (OpenIntro 1.27)

(a) He randomly samples 40 students from the study's population, gives them the survey, asks them to fill it out and bring it back the next day.
(b) He gives out the survey only to his friends, making sure each one of them fills out the survey.
(c) He posts a link to an online survey on Facebook and asks his friends to fill out the survey.
(d) He randomly samples 5 classes and asks a random sample of students from those classes to fill out the survey.

D. **[Bias, Causation]** *Flawed reasoning.* Identify the flaw(s) in reasoning in the following scenarios. Explain what the individuals in the study should have done differently if they wanted to make such strong conclusions. (OpenIntro 1.39)

(a) Students at an elementary school are given a questionnaire that they are asked to return after their parents have completed it. One of the questions asked is, "Do you find that your work schedule makes it difficult for you to spend time with your kids after school?" Of the parents who replied, 85% said "no". Based on these results, the school officials conclude that a great majority of the parents have no difficulty spending time with their kids after school.
(b) A survey is conducted on a simple random sample of 1,000 women who recently gave birth, asking them about whether or not they smoked during pregnancy. A follow-up survey asking if the children have respiratory problems is conducted 3 years later. However, only 567 of these women are reached at the same address. The researcher reports that these 567 women are representative of all mothers.
(c) An orthopedist administers a questionnaire to 30 of his patients who do not have any joint problems and finds that 20 of them regularly go running. He concludes that running decreases the risk of joint problems.

E. **[Standard Error]** *Hen eggs.* The distribution of the number of eggs laid by a certain species of hen during their breeding period has a mean of 35 eggs with a standard deviation of 18.2. Suppose a group of researchers randomly samples 45 hens of this species, counts the number of eggs laid during their breeding period, and records the sample mean. They repeat this 1,000 times, and build a distribution of sample means. (OpenIntro 7.51)

(a) What is this distribution called?
(b) Would you expect the shape of this distribution to be symmetric, right skewed, or left skewed? Explain your reasoning.
(c) Calculate the variability of this distribution and state the appropriate term used to refer to this value.

(d) Suppose the researchers' budget is reduced and they are only able to collect random samples of 10 hens. The sample mean of the number of eggs is recorded, and we repeat this 1,000 times, and build a new distribution of sample means. How will the variability of this new distribution compare to the variability of the original distribution?

F. ☆ **[Variance of $\bar{X}$]** *Variance of a mean, Part I-III.*

(a) Suppose we have independent observations X_1 and X_2 from a distribution with mean μ and standard deviation σ. What is the variance of the mean of the two values: $\frac{X_1+X_2}{2}$? (OpenIntro 3.45-3.47)
(b) Suppose we have 3 independent observations X_1, X_2, X_3 from a distribution with mean μ and standard deviation σ. What is the variance of the mean of these 3 values: $\frac{X_1+X_2+X_3}{3}$?
(c) Suppose we have n independent observations X_1, X_2, ..., X_n from a distribution with mean μ and standard deviation σ. What is the variance of the mean of these n values: $\frac{X_1+X_2+\cdots+X_n}{n}$?

G. **[Sampling Distribution]** *Repeated water samples.* A nonprofit wants to understand the fraction of households that have elevated levels of lead in their drinking water. They expect at least 5% of homes will have elevated levels of lead, but not more than about 30%. They randomly sample 800 homes and work with the owners to retrieve water samples, and they compute the fraction of these homes with elevated lead levels. They repeat this 1,000 times and build a distribution of sample proportions. (OpenIntro 5.5)

(a) What is this distribution called?
(b) Would you expect the shape of this distribution to be symmetric, right skewed, or left skewed? Explain your reasoning.
(c) If the proportions are distributed around 8%, what is the variability of the distribution?
(d) What is the formal name of the value you computed in (c)?
(e) Suppose the researchers' budget is reduced, and they are only able to collect 250 observations per sample, but they can still collect 1,000 samples. They build a new distribution of sample proportions. How will the variability of this new distribution compare to the variability of the distribution when each sample contained 800 observations?

H. **[CLT, Proportions]** *CLT for proportions.* Define the term "sampling distribution" of the sample proportion, and describe how the shape, center, and spread of the sampling distribution change as the sample size increases when $p = 0.1$. (OpenIntro 5.34)

I. **[CLT Conditions]** *Orange tabbies.* Suppose that 90% of orange tabby cats are male. Determine if the following statements are true or false, and explain your reasoning. (OpenIntro 6.3)

(a) The distribution of sample proportions of random samples of size 30 is left skewed.
(b) Using a sample size that is 4 times as large will reduce the standard error of the sample proportion by one-half.
(c) The distribution of sample proportions of random samples of size 140 is approximately normal.
(d) The distribution of sample proportions of random samples of size 280 is approximately normal.

Question 4 Estimates and Confidence Intervals

A. **[Point Estimate]** *Quality control.* 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. (OpenIntro 5.3)

(a) What population is under consideration in the data set?
(b) What parameter is being estimated?
(c) What is the point estimate for the parameter?

(d) What is the name of the statistic we use to measure the uncertainty of the point estimate?
(e) Compute the value from part (d) for this context.
(f) The historical rate of defects is 10%. Should the engineer be surprised by the observed rate of defects during the current week?
(g) Suppose the true population value was found to be 10%. If we use this proportion to recompute the value in part (e) using $p = 0.1$ instead of $\hat{p}$, does the resulting value change much?

B. **[Margin of Error]** *Fireworks on July 4^{th}.* A local news outlet reported that 56% of 600 randomly sampled Kansas residents planned to set off fireworks on July 4^{th}. Determine the margin of error for the 56% point estimate using a 95% confidence level. (OpenIntro 6.7)

C. **[Proportion CI]** *Website registration.* A website is trying to increase registration for first-time visitors, exposing 1% of these visitors to a new site design. Of 752 randomly sampled visitors over a month who saw the new design, 64 registered. (OpenIntro 5.13)

(a) Check any conditions required for constructing a confidence interval.
(b) Compute the standard error.
(c) Construct and interpret a 90% confidence interval for the fraction of first-time visitors of the site who would register under the new design (assuming stable behaviors by new visitors over time).

D. **[Confidence Level]** *Life after college.* 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. (OpenIntro 6.45)

(a) Describe the population parameter of interest. What is the value of the point estimate of this parameter?
(b) Check if the conditions for constructing a confidence interval based on these data are met.
(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.
(d) What does "95% confidence" mean?
(e) Now calculate a 99% confidence interval for the same parameter and interpret it in the context of the data.
(f) Compare the widths of the 95% and 99% confidence intervals. Which one is wider? Explain.

E. ☆ **[CI Interpretation]** *Waiting at an ER, Part I.* A hospital administrator hoping to improve wait times decides to estimate the average emergency room waiting time at her hospital. She collects a simple random sample of 64 patients and determines the time (in minutes) between when they checked in to the ER until they were first seen by a doctor. A 95% confidence interval based on this sample is (128 minutes, 147 minutes), which is based on the normal model for the mean. Determine whether the following statements are true or false, and explain your reasoning. (OpenIntro 5.11)

(a) We are 95% confident that the average waiting time of these 64 emergency room patients is between 128 and 147 minutes.
(b) We are 95% confident that the average waiting time of all patients at this hospital's emergency room is between 128 and 147 minutes.
(c) 95% of random samples have a sample mean between 128 and 147 minutes.
(d) A 99% confidence interval would be narrower than the 95% confidence interval since we need to be more sure of our estimate.
(e) The margin of error is 9.5 and the sample mean is 137.5.
(f) In order to decrease the margin of error of a 95% confidence interval to half of what it is now,

we would need to double the sample size. (Hint: the margin of error for a mean scales in the same way with sample size as the margin of error for a proportion.)

F. **[CI Width]** *Relaxing after work.* The General Social Survey asked the question: "After an average work day, about how many hours do you have to relax or pursue activities that you enjoy?" to a random sample of 1,155 Americans. A 95% confidence interval for the mean number of hours spent relaxing or pursuing activities they enjoy was (1.38, 1.92). (OpenIntro 5.27)

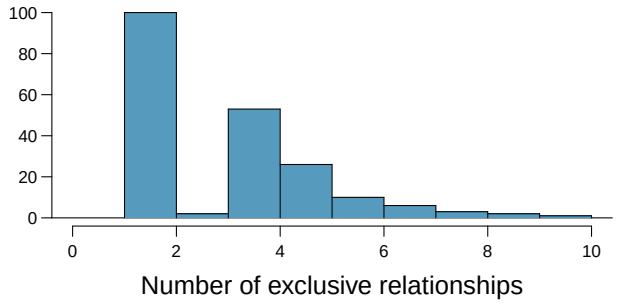
(a) Interpret this interval in context of the data.
(b) Suppose another set of researchers reported a confidence interval with a larger margin of error based on the same sample of 1,155 Americans. How does their confidence level compare to the confidence level of the interval stated above?
(c) Suppose next year a new survey asking the same question is conducted, and this time the sample size is 2,500. Assuming that the population characteristics, with respect to how much time people spend relaxing after work, have not changed much within a year. How will the margin of error of the 95% confidence interval constructed based on data from the new survey compare to the margin of error of the interval stated above?

G. ☆ **[Biased Sample]** *Study abroad.* A survey on 1,509 high school seniors who took the SAT and who completed an optional web survey shows that 55% of high school seniors are fairly certain that they will participate in a study abroad program in college. (OpenIntro 6.9)

(a) Is this sample a representative sample from the population of all high school seniors in the US? Explain your reasoning.
(b) Let's suppose the conditions for inference are met. Even if your answer to part (a) indicated that this approach would not be reliable, this analysis may still be interesting to carry out (though not report). Construct a 90% confidence interval for the proportion of high school seniors (of those who took the SAT) who are fairly certain they will participate in a study abroad program in college, and interpret this interval in context.
(c) What does "90% confidence" mean?
(d) Based on this interval, would it be appropriate to claim that the majority of high school seniors are fairly certain that they will participate in a study abroad program in college?

H. **[Claimed Value]** *Nutrition labels.* The nutrition label on a bag of potato chips says that a one ounce (28 gram) serving of potato chips has 130 calories and contains ten grams of fat, with three grams of saturated fat. A random sample of 35 bags yielded a confidence interval for the number of calories per bag of 128.2 to 139.8 calories. Is there evidence that the nutrition label does not provide an accurate measure of calories in the bags of potato chips? (OpenIntro 5.33)

I. **[t Interval]** *Exclusive relationships.* A survey conducted on a reasonably random sample of 203 undergraduates asked, among many other questions, about the number of exclusive relationships these students have been in. The histogram below shows the distribution of the data from this sample. The sample average is 3.2 with a standard deviation of 1.97. (OpenIntro 7.55)



Estimate the average number of exclusive relationships Duke students have been in using a 90% confidence interval and interpret this interval in context. Check any conditions required for inference, and note any assumptions you must make as you proceed with your calculations and conclusions.

J. **[Sample Size]** *College smokers.* We are interested in estimating the proportion of students at a university who smoke. Out of a random sample of 200 students from this university, 40 students smoke. (OpenIntro 6.43)

   (a) Calculate a 95% confidence interval for the proportion of students at this university who smoke, and interpret this interval in context. (Reminder: Check conditions.)
   (b) If we wanted the margin of error to be no larger than 2% at a 95% confidence level for the proportion of students who smoke, how big of a sample would we need?