Assignment 2

URBAN 520: Quantitative Methods in Urban Design and Planning

Due: November 17, 2026 11:59 pm Pacific Time

Total: 70 points, plus 1 points of extra credit

- You need to submit the following via Canvas: a PDF or HTML report (handwritten answers may be scanned into a PDF) with the answers in order, together with all files needed to reproduce your results, such as `.Rmd`, SPSS syntax (`.sps`), an Excel file (`.xlsx`), or a Tableau packaged workbook (`.twbx`).
- Each student is expected to submit their own assignment. Study groups are allowed, but you are expected to acknowledge the names of your collaborators, along with a short description of the types of collaboration, at the beginning of each submission.
- The only AI tool you may use for this assignment is UW's Purple AI. You should treat AI as a human collaborator instead of directly asking for answers or copying solutions. If you use AI in completing the assignment, you need to acknowledge how you used it, cite it appropriately, and submit the exported TXT conversation history.
- Whatever you submit, you must be able to explain how you did it and why. We may ask you to walk us through any part of your work at any time, and being unable to explain your own submission will be treated as if the work were not yours.
- You will have 6 penalty-free late days in total for assignments (max 3 late days per submission). Any submission delayed beyond the first 3 days will be penalized 10% per day for that specific assignment, but those days will not count toward your used late days.

Question 1 Basic Probability (6 points)

A. (2 points) If we toss a six-sided fair die twice:

1. How many possible outcomes exist?
2. What's the probability that the sum of the two numbers is 6?

B. (2 points) A survey shows that 33% of people take the light rail, 57% of people take the bus, and 63% of people take at least one of them.

1. What is the probability that a person takes both the light rail and the bus?
2. Given that a person takes the light rail, what is the probability that they also take the bus?

C. (2 points) A planning firm is applying for a zoning amendment to a new transit-oriented housing project in Seattle. Initially, planners estimate a 50% prior probability of approval. However, the city requests a Supplemental Environmental Impact Statement (SEIS). Historically, 75% of approved projects and 40% of denied projects were asked to submit this report.

1. Given a project is approved, what is the conditional probability that an SEIS was requested?
2. Given the SEIS is requested, calculate the posterior probability of project approval.

Question 2 Probability Distributions (13 points)

A. (4 points) A city is reviewing 12 housing development proposals submitted under the new zoning incentive program. The random variable X is defined as the number of proposals that allocate 20% or more of their units to low-income affordable housing. We know that for each proposal, the probability of meeting the requirement is 0.4, and the proposals are independent.

1. What are the possible values that this random variable X can take?
2. What type of probability distribution best models this random variable X?
3. What is the probability that exactly 5 proposals qualify?
4. What is the expectation $E(X)$ and the standard deviation $SD(X)$ of qualifying proposals among the 12 submissions?

B. (3 points) A transportation planner is working on an adaptive traffic light. The random variable X represents the number of minutes ($x = 1, 2,$ or 3) a vehicle has to wait during peak hours before the light turns green. The planner proposes the following probability mass function:

$$f(x) = \frac{x}{6}, \quad \text{for } x = 1, 2, \text{ or } 3$$

1. Is this a valid probability function and why?
2. What is the probability that a vehicle waits at least 2 minutes?
3. Calculate the expectation $E(X)$ and variance $\text{Var}(X)$.

C. (2 points) In Seattle, most buses arrive strictly every 15 minutes (hopefully). If a commuter walks to the bus stop at a completely random time, their waiting time x (in minutes) follows a continuous uniform distribution between 0 and 15 minutes.

1. Calculate $P(3 \leq x \leq 12)$ (waiting between 3 and 12 minutes).
2. Calculate expectation $E(x)$ (average waiting time) and variance $\text{Var}(x)$.

D. (2 points) A random variable X is normally distributed with mean $\mu = 50$ and standard deviation $\sigma = 5$. Using Figure 1 and answer:

1. What is the probability that X falls between 40 and 60?
2. What is the probability that X is larger than 55?

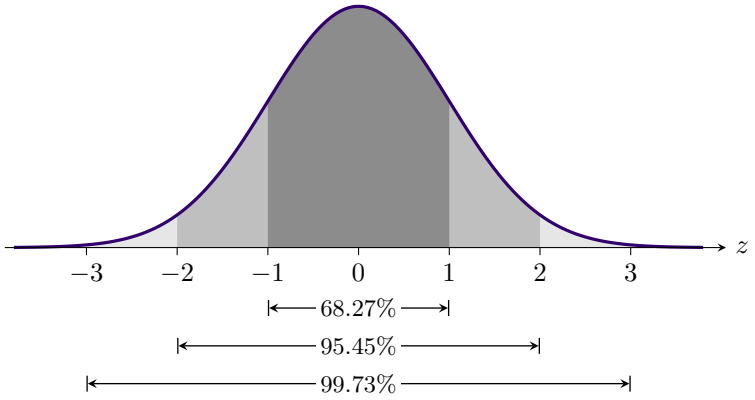


Figure 1: The standard normal distribution $N(0, 1)$.

E. (2 points) You are asked to do a study of shelters for abused and battered women to determine the necessary capacity in your city to provide housing for most of these women. After recording data for a whole year, you find that the mean number of women in shelters each night is 250, with a standard deviation of 75. Fortunately, the distribution of the number of women in the shelters each night is normal, so you can answer the following questions posed by the city council:

1. If the city's shelters have a capacity of 400, will that be enough places for abused women on 95% of all nights? What capacity would be enough on about 99.9% of nights?
2. The current capacity is 175 openings, as some shelters have been closed. What is the percentage of nights that the number of abused women seeking shelter will exceed current capacity?

Question 3 Sampling and Sampling Distributions (10 points)

A. (3 points) The following three common scenarios involve selecting a sample and understanding how a sample relates to a population.

1. A friend interviews every 10th shopper who passes by her as she stands outside one entrance of a major department store in a shopping mall. What type of sample is she selecting? How might you define the population from which she is selecting the sample?
2. A political polling firm samples 50 potential voters from a list of registered voters in each county in a state to interview for an upcoming election. What type of sample is this? Do you have enough information to tell?
3. Another political polling firm in the same state selects potential voters from the same list of registered voters with a very different method. First, they alphabetize the list of last names, then pick the first 20 names that begin with an A, the first 20 that begin with a B, and so on until Z (the sample size is thus 20×26, or 520). Is this a probability sample?

B. (4 points) In an urban building energy study, the annual electricity consumption intensity of commercial office buildings is analyzed. The energy data shows a population mean of 200 kWh/m^2/year, with a population standard deviation of 50 kWh/m^2/year. An energy auditor randomly selects a simple random sample of 100 office buildings to estimate the regional population mean using the sample mean $\bar{x}$.

1. What is the expected value of $\bar{x}$?
2. What is the standard error (standard deviation) of $\bar{x}$?
3. Describe the sampling distribution of $\bar{x}$.
4. What does this sampling distribution indicate about the relationship between the sample mean and the population mean?

C. (3 points) The mean commute time for city residents is $\mu = 26$ minutes with a population standard deviation of $\sigma = 9$ minutes. A simple random sample of $n = 50$ residents is selected. What are the specific time intervals around the population mean for:

1. Approximately 68% of the sample means (within 1 standard error).
2. Approximately 99.7% of the sample means (within 3 standard errors).
3. Commute times are not normal; why can you use Figure 1 in these questions? Could you do the same with a sample of $n = 10$ residents? What if commute times were normally distributed?

Question 4 Estimates and Confidence Intervals (12 points)

A. (2 points) A planner is conducting a study on neighborhood accessibility. A simple random sample of 6 residents was selected, and their walking times (in minutes) to the nearest community park were recorded as follows: 5, 8, 10, 7, 10, and 14. What is the point estimate of the population mean walking time and population standard deviation of walking times?

B. (3 points) As part of a study on gentrification, a researcher randomly samples 212 long-term low-income households who lived in a rapidly developing downtown neighborhood at the beginning of the year. By the end of the survey period, they find that 27 of these households have been displaced.

1. What parameter is being estimated? And what is the point estimate for the parameter?
2. Compute the statistic we use to measure the uncertainty of the point estimate.
3. The historical city-wide displacement rate in stable neighborhoods is 10%. Should the researcher be surprised by the observed displacement rate in this gentrifying neighborhood?

C. (2 points) The researcher in Q4B plans the same survey in another gentrifying neighborhood and wants to estimate its displacement rate within ±4 percentage points at the 95% confidence level.

1. With no prior information about the rate, how many households should the researcher sample? How many would be needed for ±2 percentage points? What happens to the required sample size when the margin of error is cut in half?
2. Using the displacement rate from Q4B as the planning value, how many households are needed for ±4 percentage points? Compare the numbers with no prior information scenario.

D. (2 points) In 2016, the Pew Research Center conducted a survey of 1,004 Canadians and 1,003 Americans to assess their opinion of climate change. The data show that 51% of Canadians and 45% of Americans believe climate change is a very serious problem.

1. Estimate the proportion of all Canadians and all Americans who believe climate change is a very serious problem at the 95% confidence interval.
2. Would you be confident in reporting that the majority (more than 50%) of Canadians and Americans believe climate change is a very serious problem? Why?

E. (3 points) The American Community Survey (ACS) publishes every estimate with a margin of error (MOE) at the 90% confidence level. Two neighboring census tracts in Ballard report these median gross rents (ACS 2020–2024 5-year estimates, Table B25064):

Tract	Median gross rent	MOE (90%)
47.02	$2,594	±$649
47.03	$2,112	±$219

1. For each tract, convert the MOE to a standard error, and compute a 95% confidence interval.
2. A neighborhood profile states that median rent in tract 47.02 is "about $480 higher" than in tract 47.03. Test whether the two medians differ at the ACS's 90% level.
3. Rewrite the profile's sentence so that it reports these numbers honestly.

Question 5 Hypothesis Testing (4 points)

A. (4 points) A food safety inspector is called upon to investigate a restaurant with a few customer reports of poor sanitation practices. The food safety inspector uses a hypothesis testing framework to evaluate whether regulations are not being met. If he decides the restaurant is in gross violation, its license to serve food will be revoked. (OpenIntro Statistics 5.29).

1. Write the hypotheses in words.
2. What are Type 1 and Type 2 errors in this context?
3. Which error is more problematic for the restaurant owner? What about for the diners?
4. As a diner, would you prefer that the food safety inspector requires strong evidence or very strong evidence of health concerns before revoking a restaurant's license?

Question 6 Land Use, Income, and Travel Mode (25 points)

Puget Sound VISION 2050 concentrates growth in Regional Growth Centers, in part so that more trips can be made without a car. Meanwhile, lower-income households may rely more on transit and walking. The file `psrc_trips_2023.csv` contains weekday trips (Monday–Thursday) made within the region by households in the 2023 PSRC Household Travel Survey, one row per trip. Each row includes the trip's `mode_class` and the household's `household_id`, `hhincome_broad`, and `home_rgcname`.

The survey is not a simple random sample. PSRC drew a stratified sample of addresses that deliberately over-samples some areas, such as Seattle and neighborhoods with high walking, biking, and transit use:

households in RGCs make up 20% of the households but only about 11% of the region's households. PSRC provides weights to correct for this. But, to keep the analysis simple, ignore the weights and treat the data as simple random samples.

Table 1: Variables in PSRC Trip Dataset

Variable	Description	Values
`trip_id`	Trip identifier	One value per row
`household_id`	Household identifier	3,474 households
`person_id`	Person identifier; its first eight digits are the `household_id`	5,872 people
`mode_class`	Main mode of the trip, as classified by PSRC	Drive SOV, Drive HOV2, Drive HOV3+, Ride Hail, Transit, Walk, Bike, Micromobility, School Bus, Other, Missing Response
`hhincome_broad`	Annual household income	Under $25,000; $25,000–$49,999; $50,000–$74,999; $75,000–$99,999; $100,000–$199,999; $200,000 or more; Prefer not to answer
`home_rgcname`	Regional Growth Center the home is in	"Not RGC", or one of 26 RGC names

Part 1. Preparing the Data

A. (2 points) Create `mode` with the three categories below and drop all other trips. How many trips remain, and why is it useful to collapse the modes before comparing households?

`mode`	`mode_class`
Car	Drive SOV, Drive HOV2, Drive HOV3+, Ride Hail
Transit	Transit
Walk/Bike	Walk, Bike, Micromobility
Dropped	School Bus, Other, Missing Response

B. (2 points) Create `income` with four groups: Under $50k, $50–100k, $100–200k, and $200k+ ("Prefer not to answer" is missing). Create `rgc`: *1* if `home_rgcname` is not "Not RGC", otherwise *0*.

C. (3 points) Aggregate the trips to one row per household, recording the **number of trips** and the **non-car share**, the **share of the household's trips made by transit or walk/bike**. How many households are there, and what is the range of trips per household? Should we use trip or household as the unit of analysis and why?

D. (3 points) Report n, mean, and SD of non-car share by `rgc` and by `income`, and a contingency table of `income` by `rgc`. About half of the households have a non-car share of 0, so its distribution is far from normal. If we wish to use a two-sample t-test to compare the mean non-car share of households in and outside RGCs, can we still use it? Why or why not?

Part 2. Testing

For each test, report the hypotheses (H_0 and H_1), test statistic, degrees of freedom, p-value, decision at $\alpha = 0.05$, and a one-sentence interpretation.

A. (5 points) Do households living in RGCs make a larger share of their trips without a car? Run a two-sample t-test (unequal variances) and give a 95% confidence interval for the difference.

B. (5 points) Does non-car share differ across the four income groups? Run a one-way ANOVA and check the equal-variance assumption using the group SDs. What does a significant F tell you, and what does it not tell you? Using the group means from Part 1D, describe the pattern.

C. (5 points) Is household income associated with living in an RGC? Run a chi-square test of independence on `income` by `rgc` and report the expected counts. Using this result, explain why the income difference you found in Part 2B may partly reflect where households live rather than income itself.

Question 7 Extra Credit (1 extra credit points)

A. (1 points, extra credit) How many hours did you spend on this assignment?