Midterm1_Review

Elaine Chiu

1/8/2022

A Road Map - Concepts

- Graphical Summaries of Data
- Dot plot
- stem leaf plot
- Histogram
- Box plot
- bar chart
- pie chart
- Numerical summaries of data
- sample average
- sample standard deviation
- range/min/max/IQR(interquartile range)
- Probabilities/Definition/properties/addition Rule/multiplication Rule/complement Rule / Independence/ Conditional Probability
- Normal Curve/ Z Score/ Standardization

How to prepare for the midterm

- Redo or read solutions for past discussion/exam/assignment problems.
- Make sure you understand the above terms.
- You should know how to draw each type of graphs.
- You should know what is the suitable type of graphs to use depending on the data type/ sample size.
- How to estimate the percentage in an interval given a histogram.
- Know how to calculate sample average and sample standard deviation given a data set, also, standard deviation has two formulas.
- know how to calculate range/min/max/IQR. Identify them on the box plot.
- You know how to calculate the probabilities/ conditional probabilities/ how to check Independence.
- Know how to use normal curve to infer the percentage below or above a given quantity.
- Know the meaning of z-score, and can convert between z score and raw quantity.

Exercises with Solutions

Problem 1 - Probability

- What is the probability that in a randomly chosen year your birthday is on a weekend(Saturday or Sunday)?

(a). 50%
(b). 28.6%
(c). 34.5%
(d). 14.3
(e). Not enough information to solve.

Solution:
We assume that the probability of your birthday landing on a specific day in a week is equally likely. Therefore, recall that the probability of an event is then $\frac{\text{number of events}}{\text{number of all outcomes}}$. The total probability is 100%, and out of seven days in a week, Saturday and Sunday are two out of seven days, thus the probability is $100\% \times \frac{2}{7} = 28.6\%$.

Problem 2 - Probability

- Imagine a bag containing 10 balls: 3 Green, 4 Blue, 2 Yellow, and 1 Red If we draw 2 balls from the bag, with replacement, what is the probability the first ball is either Red or Blue and the second ball is either Green or Yellow?

(a). 25%
(b). 16%
(c). 6%
(d). 75%
(e). 100%

Solution: The probability of the first ball being either Red or Blue is $\frac{4+1}{10} = \frac{1}{2}$, this is because we assume that the probability of getting any ball from the bag is equally likely, so the probability of first ball being either Red or Blue is the fraction of Blue and Red balls from all the balls, with the same reasoning, the probability of getting a ball being either Green or Yellow is $\frac{3+2}{10} = \frac{1}{2}$. By multiplication Rule, the chance of two events happening is the chance of the first event multiplied by the chance of the second event given the first event. $\frac{1}{2} \times \frac{1}{2} = 0.25$.

Problem 3 - Probability and Independence

- A sample of 250 high school students were asked, "If you had $\$1,000$ to contribute to one kind of charitable organization, which type of organization would you choose? Below is a two-way table of responses to this question and gender.

	Education	Environment	Health	International Aid	Other	Total
Female	19	33	50	28	10	140
Male	23	29	28	17	13	110
Total	42	62	78	45	23	250

Which of the following conclusions seems to be supported by the data?
(a). Most of the females who chose a health organization would have chosen an environmental organization as

their second choice, had they been asked.
(b). The proportion of males who said they would contribute to an environmental organization was higher than the proportion of females who said they would contribute to such an organization.

(c). A randomly chosen student is more likely to choose an environmental organization than any other.

(d). If a randomly chosen student is male, he is less likely to donate to education than international aid.

(e). The type of charity organizations one wants to donate to and the gender are independent.

Solution:
The correct answer is (b).
For (a), it is false since we don't know what people would have chosen as second choice.
For (b), P(environmental given male)= $\frac{29}{110} = 0.264$ P(environmental given female)=$\frac{33}{140} = 0.236$, so the statement is correct.
For (c), Environmental: $\frac{62}{250}$ is not larger than health $\frac{78}{250}$.
For (d), P(Education given male)=$\frac{23}{110}$ > P(international aid given male)=$\frac{17}{110}$.
For (e), recall that you have three ways to check if two events are independent or not. Events A and B are independent if:
- P(A given B) = P(A)
- P(B given A) = P(B)
- P(A and B) = P(A) $\times$ P(B)
In this case, since P(donate to Education) = $\frac{42}{250} \neq \frac{23}{110} =$P(Donate to Education given male), so two events are not independent.

Problem 4 - Numercal Summaries of Data

- The frying pans currently available at Hilldale Target are priced at: 12, 12, 16, 23, 26, and 37. They have an average price of 21. What priced frying pan could Target add that would result in the smallest overall standard deviation?
  (a). 21
  (b). 12
  (c). 24
  (d). 34
  (e). 37
  Solution:
  Since we know that standard deviation measures how far on average it is for a data point away from the sample average. We should add a value which is closest to the average to result in a smallest overall standard deviation. Therefore, the correct answer is (a) 21.

Problem 5 - Numerical Summaries of Data

- A vineyard is interested in the number of wasps on their grape vines. An employee counts and records the number of wasps on a sample of 50 clumps of vines. After calculating statistics on the data, the employee notices that he accidentally wrote down the maximum data value incorrectly. The highest number of wasps that he observed in a clump of vines was 42, not the recorded 51. He changed the incorrectly entered 51 to 42. The value of 42 is now the highest number of wasps observed in the sample of 50 clumps of vines. Which of the following statistics for his data does he need to recalculate?
  (I). Range
  (II). IQR
  (III). Standard Deviation
  (IV). Average
  (V). Median

(a). I, II, III only
(b). IV, V only
(c). II, V only
(d). I, IV, V only
(e). I, III, IV only
Solution:
The range is max - min, therefore, if max is changed, the range would also change; IQR is the third quantile subtract the first quantile, so would not be changed. The standard deviation and average are sensitive to any change of a single data point, therefore, the final answer is (e) I, III, IV only.

Problem 6 - Numerical Summaries of Data

- Which of these statements is true?
  (a). The standard deviation can never be larger than the average.
  (b). The range can never be larger than the average.
  (c). The mean can never be larger than the median.
  (d). The interquartile range can never be larger than the range.
  (e). The average can never be larger than the standard deviation
  Solution: (d) is correct. For (a), consider the counter-example $\{-10, -10, 10, 10\}$, the average is 0 but the standard deviation is obviously greater than 0, because 0 is the smallest possible standard deviation for any data set, and it only happens when all the data points are equal. for(e), consider the counter-example $\{10, 10, 10, 10\}$, the average is 10 while the standard deviation is 0.

Problem 7 - Numerical Summaries of Data

- A medical researcher collects health data on many women in each of several countries. One of the variables measured for each woman in the study is her weight in pounds. The following list gives the five-number summary (Min, Q1, Median, Q3, Max) for the weights of adult women in one of the countries.
  Country A: $92, 110, 120, 160, 240$
  About what percent of Country A women weigh between 110 and 240 pounds?

(a). 50% (b). 65% (c). 75% (d). 85% (e). 95%
Solution:

c. is correct, because 110 is Q1, and Max is 240.There are 25% between Q1 and Median, 25% between Median and $Q3$, 25% between $Q3 $and Max

Problem 8 - Normal Curve and Z-Score

- Fill in the blank. Suppose you have calculated a z-score of -4.3. The area under the standard normal curve to the left of the this z-score (-4.3) is ____ the area under the curve to the right of a z-score of 4.29?
  (a). Greater than.
  (b). Less than.
  (c). The same as.
  (d). Not enough information provided to answer.
  Solution:
  The correct answer is (b), less than, recall that standard normal curve is symmetric around 0.

Problem 9 - Normal Curve and Z-Score

- Assume the time for an emergency medical squad to arrive at the sports center at the edge of town is normally distributed with mean =17 minutes and standard deviation of 3 minutes. Which 6 minute arrival period duration has the highest probability?(Hint: Draw a picture)

(a). between 17 and 23 minutes
(b). between 11 and 17 minutes
(c). between 14 and 20 minutes
(d). between 10 and 16 minutes
(e). between 12 and 19 minutes
Solution: For a general normal curve, it is symmetric around its mean, and the density is highest around the mean. Therefore, we should choose an interval which is closest to the mean. (c) is correct.

Problem 10 - Normal Curve and Z-Score

- Which of the following statements are true about the standardized (z-scores) calculated on a normal population?
  (a). Z-scores tell us how many standard deviations above or below the mean a value falls.
  (b). Z-scores are always symmetric about the mean.
  (c). Z-scores can never take a value as extreme as -4.
  (d). A and B
  (e). A, B, and C
  Solution: (a) is the correct definition of z score.

Problem 11 - Normal Curve and Z-Score

- A fire department in a rural county reports that its response time to fires is approximately Normally distributed with a mean of 22 minutes and a standard deviation of 11.9 minutes. Approximately what proportion of their response times is over 30 minutes?

(a). 0.03
(b). 0.21
(c). 0.25
(d). 0.75
(e). 0.79
Solution:
This is a typical problem of normal curve. The first step is standardize 30. $\frac{30-22}{11.9} = 0.6723$, this means that 30 is 0.6723 standard deviation above the mean. Checking how big is the area above/to the right of z score 0.6723, it is 0.2514. If you do not have a table showing you the area to the right of 0.6723, you can check that the area below the curve to the left of 0.6723 is 0.7486, and you know the whole area is 100%, so the final answer is $100\% - 0.7486 = 0.2514$, the correct answer is (c) 0.25.

Problem 12 - Normal Curve and Z-Score

- The weights of adult women are approximately normally distributed about a mean of 142 lbs with a standard deviation of 14 lbs. If Renee is at the 96th percentile in weight for adult women, then her weight, in lbs, is closest to:
  (a). 167
  (b). 163
  (c). 170
  (d). 172
  (e). 174
  A: Look up 0.9600 in table gives us a z-score of 1.75, this means that Renee is 1.75 standard deviation above the average. Solve, 1.75=$\frac{(\text{observed value} - 142)}{14}$ so Observed value=$1.75 \times 14 + 142 = 166.5$, the correct answer is (a) 167.

Problem 13 - Graphical Summaries of Data

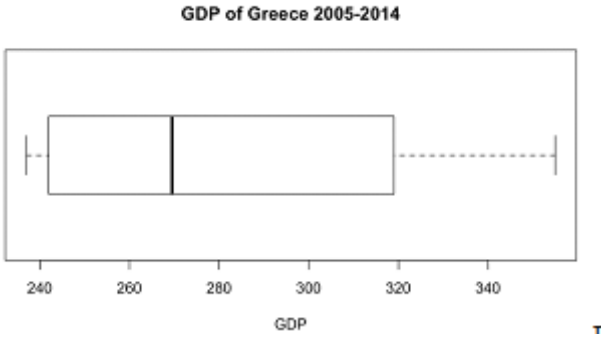
- Greece's economy has faltered in recent years. The gross domestic product (GDP) of Greece between 2005 and 2014 is listed below (rounded to nearest billion, in billions)

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
GDP	248	237	319	355	329	299	289	250	242	237

  Name the 5 numerical values necessary to construct a boxplot and give their corresponding values for the GDP data. After this, construct the box plot.

Solution:
Step 1: order the data
Order values: 237, 237, 242, 248, 250 , 289, 299, 319, 329, 355

Min: 237
Q1:242
Median: (250+289)/2=269.5 (because we have even number of data points)
Q3: 319
Max: 355
The box plot is:



Problem 14 - Graphical Summaries of Data

- States that experience four seasons tend to have more issues with potholes because of the changes in temperature and variety of precipitation. The Department of Transportation for Michigan is considering how much money they need to allocate to ensure the state's roads are safe for travel. They've taken a random sample of 500 roads from the state and then selected a random mile of that road to evaluate for repairs. On each mile stretch, they've recorded the length (in meters) that requires repairs. Below is the distribution table for the length in meters in need of repair:

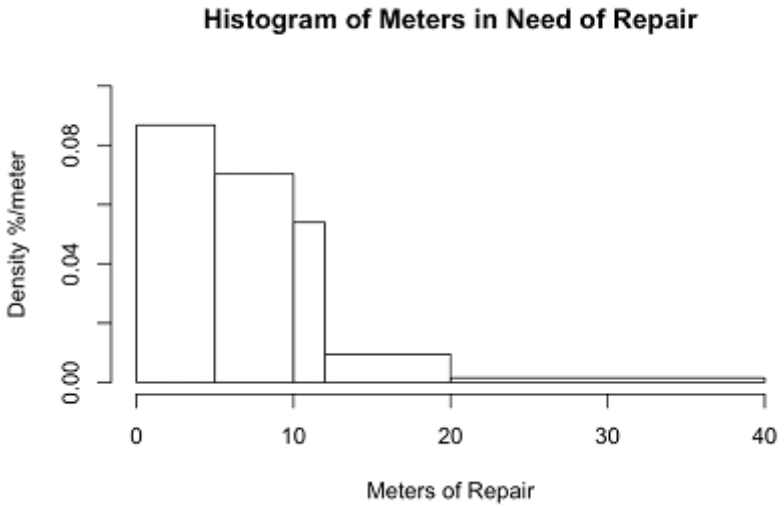
length	0—5	5—10	10—12	12—20	20—40
frequency	217	176	54	38	15

Draw the (density) histogram of length in need of repair. Furthermore, estimate the percentage of roads sampled with between 10 and 15 meters in need of repair.

Solution:
You need to first create the distribution table.

Bin	Count	%	Bin Length	Density %/meter
[0-5)	217	217/500=.434=43.4%	5	43.4/5=8.68
[5-10)	176	176/500=.352=35.2%	5	35.2/5=7.04
[10-12)	54	54/500=.108=10.8%	2	10.8/2=5.4
[12-20)	38	38/500=.076=7.6%	8	7.6/8=0.95
[20-40)	15	15/500=.03=3%	20	3/20=0.15

and the histogram itself looks like:



As for the percentage of roads sampled with between 10 and 15 meters in need of repair.
$\% = \%/\text{meter} \times \text{meter}$ so the interval $[10 - 12]$ gives 10.8%. from the interval of $[12 - 15]$ is 3 meters * $(0.95\%/\text{meter}) = 2.85\%$ for a total of $10.8 + 2.85 = 13.65\%$, the final answer is 13.65%.