STAT 340

Mid-term Exam 1 Review

Spring 2024

EXAM INSTRUCTIONS

(1) The exam will be on Wednesday, Feb. 21, 11:30am-12:20pm in class.

(2) The exam will cover chapters 1, 2, 3, 4, and 6 (up to slide 25).

(3) You need to bring a calculator for the exam. You are not allowed to use your cellphone's calculator.

(4) You may have one 8 ½" by 11" (front and back) sheet of paper with formulas, definitions, or whatever you think it is important.

(5) You must show work for possible partial credit.

(6) All work on this exam must be completely on your own. Cheating will be penalized with a score of zero.

1. Classify the following measurements observed in the study as nominal, ordinal, continuous or discrete.

   a. Risk of experiencing complications in pregnancy (low, medium, high)

      Ordinal

   b. Temperature (in °F)

      Continuous

   c. Blood type (A, B, AB, O)

      Nominal

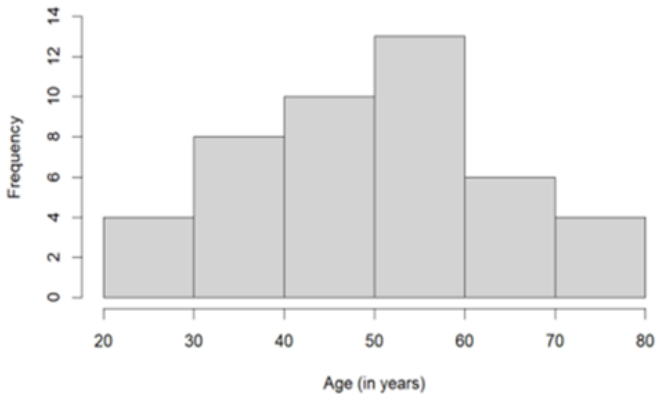
   d. Number of doctor visits in a month (0, 1, 2, …)

      Discrete

   e. Blood sodium levels (in milliequivalents per liter mEq/L)

      Continuous

2. Here is a histogram of the age of **45 people** participated in an online survey, grouped into bins of length 10. The median age must be



a. between 30 to 40
b. between 40 to 50
c. between 50 to 60
d. between 60 to 70

The median is the middle value of the data points. Since this histogram is roughly symmetric, mean=median=mode. Therefore, 50-60 has the median.

However, when you have a skewed histogram, you can simply use the following method.

Odd: $\left(\frac{n+1}{2}\right)$th data point as the median

Even: The average of $\left(\frac{n}{2}\right)$th data point and $\left(\frac{n}{2}+1\right)$th data point

Here $n = 45$, an odd number

Median is $\frac{45+1}{2} = 23^{rd}$ data point.

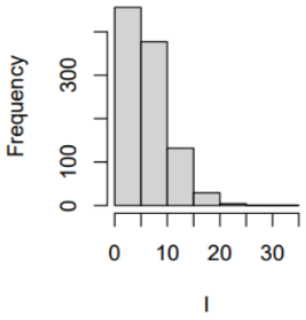
Now find the frequency of each interval/bar and get the cumulative frequency.

Interval	Frequency	Cumulative frequency	
20-30	4	4	This bar has the first 4 datapoints
30-40	8	12	This bar contains 5^{th} – 12^{th} data points
40-50	10	22	This bar has 13^{th} – 22^{nd} data points
50-60	13	35	This bar has 23^{rd} – 35^{th} data points
60-70	6	41	This bar has 36^{th} – 41^{st} data points
70-80	4	45	This bar has 42^{nd} – 45^{th} data points

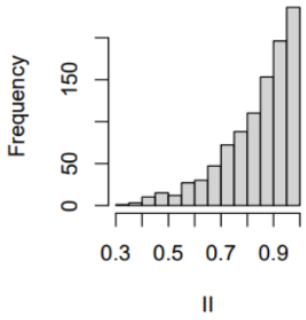
This interval has the median (23^{rd} data point)

Therefore, the median age must be between 50 to 60.

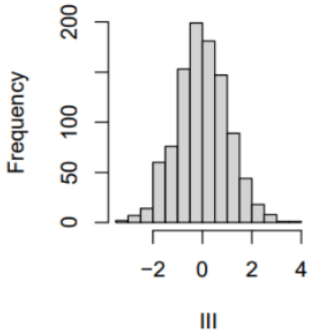
3. For each histogram, classify it as symmetric, negatively skewed, or positively skewed.



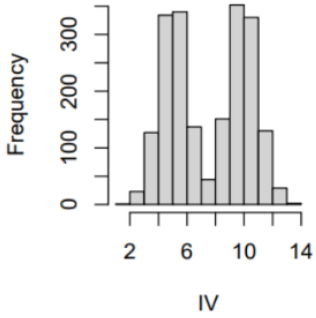
Positively skewed



Negatively skewed



Symmetric, this has only one mode (unimodal)



This is also symmetric, but this has two modes (bimodal)

4. Here are the lengths of the 20 Torrey pine needles arranged in increasing order:

12.6	21.2	21.6	21.7	23.1	23.7	24.2	24.2	25.5	26.6
26.8	28.9	29.0	29.7	29.7	30.2	32.5	33.7	33.7	39.2

a. What is the median for this dataset?

The average of $\left(\frac{n}{2}\right)$th and $\left(\frac{n}{2}+1\right)$th data points, i.e. average of 10th and 11th data points.

$$Q_2 = \frac{26.6 + 26.8}{2} = 26.7$$

b. Find the first quartile, third quartile and the interquartile range.

For the first quartile, consider the first 10 data points (again even number of data points) as a different data set and find the median (middle value)

12.6 21.2 21.6 21.7 23.1 23.7 24.2 24.2 25.5 26.6

$$Q_1 = \frac{23.1 + 23.7}{2} = 23.4$$

For the third quartile, consider the last 10 data points (again even number of data points) as a different data set and find the median (middle value)

26.8 28.9 29.0 29.7 29.7 30.2 32.5 33.7 33.7 39.2

$$Q_3 = \frac{29.7 + 30.2}{2} = 29.95$$

IQR = $Q_3 - Q_1 = 29.95 - 23.4 = 6.55$

c. Does the 1.5 × IQR rule identify any suspected outliers?

$$Q_1 - 1.5 \times IQR = 23.4 - (1.5 \times 6.55) = 13.575$$

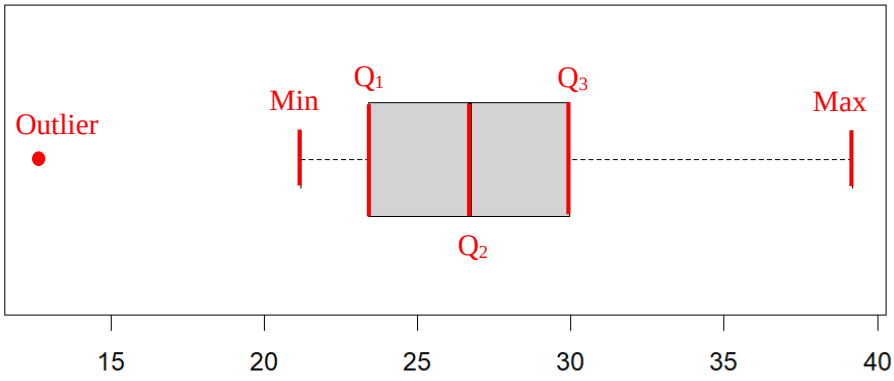
$$Q_3 + 1.5 \times IQR = 29.7 + (1.5 \times 6.55) = 39.775$$

Any value less than 13.575 or greater than 39.775 is an outlier.
12.6 is less than 13.575, therefore it's an outlier.

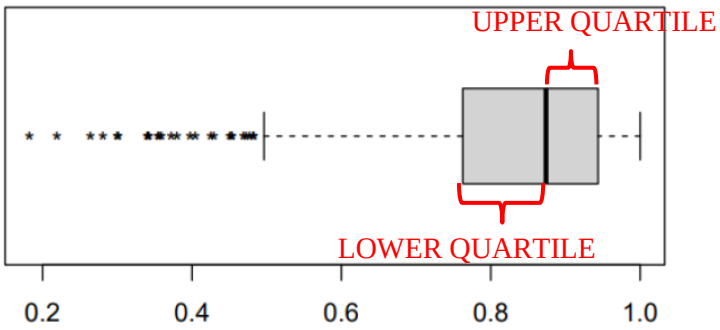
d. Draw a horizontal box plot for this data set. You can use the following scale.

Five-number summary: Min = 21.2, Q_1 = 23.4, Q_2 =26.7, Q_3 = 29.95, Max = 39.2

Make sure to mark the outlier(s) as well. In this case, 12.6 .



5. Based on the box plot below, which of the following do you expect to be true?



This is negatively skewed (upper quartile is smaller than lower quartile)

a. Mean > Median
b. Mean = Median
c. Mean < Median
d. Can't know without calculating mean and median

- Positively skewed: Mean > Median > Mode (Upper quartile > lower quartile)
- Negatively skewed: Mean < Median < Mode (Upper quartile < lower quartile)
- Symmetric: Mean = Median = Mode (Upper quartile = lower quartile)

6. The following dataset is a sample containing n = 7 observations:

2, 1, 3, 4, 3, 5, 3

a. Calculate the sample mean and the sample standard deviation for this dataset.

Sample mean $\bar{x} = \frac{2+1+3+4+3+5+3}{7} = 3$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2	2-3= -1	1
1	1-3 = -2	4
3	3-3 = 0	0
4	4-3 = 1	1
3	3-3 = 0	0
5	5-3 = 2	4
3	3-3 = 0	0
Total		10

Sample standard deviation = $s = \sqrt{\frac{10}{7-1}} = 1.291$

b. Provide an interpretation for the value of standard deviation that you obtained in (a).

The average distance between the mean 3 and the data points is 1.291. When compared with the data points 1.291 is not a large value, therefore we can say the data points are somewhat closer to the mean.

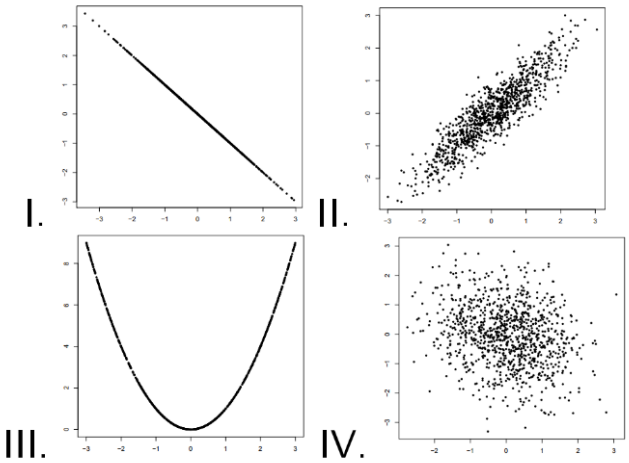
7. What are all the values that a correlation r can possibly take?

   a. $r \geq 0$
   b. $0 \leq r \leq 1$
   c. $-1 \leq r \leq 1$

8. If the points on a scatterplot lie very close to a straight line. The correlation between x and y is close to

   a. -1
   b. 1
   c. either -1 or 1, depending on the direction.

9. For each scatter plot, match the value of correlation coefficient.



   a. $r = -0.2$
   b. $r = -1$
   c. $r = 0$
   d. $r = 0.9$

I	r = -1 (perfect negative correlation)
II	r = 0.9 (strong positive correlation)
III	r = 0 (this is a curve, i.e. no linear relationship)
IV	r = - 0.2 (weak negative correlation)

r is used to measure the direction (positive or negative) and strength of **LINEAR** relationship between two quantitative variables. If you see any non-linear relationship (i.e. curves, cluster etc.) , then r = 0.

10. The toto toucan possesses the largest beak relative to body size of all birds. This exaggerated feature has received various interpretations, such as being a refined adaptation for feeding. However, the large surface area may also be an important mechanism for radiating heat (and hence cooling the bird) as outdoor temperature increases.

    To investigate the relationship between outdoor temperature (x) (in ℃) and percentage heat loss from beak (y), we compute the following quantities.

    - $\bar{x}$ = mean of the values of x = 22.5
    - $\bar{y}$ = mean of the values of y = 47.375
    - s_x = standard deviation of the values of x = 4.761
    - s_y = standard deviation of the values of y = 10.751
    - r = correlation between x and y = 0.914

    a. What is the equation of the least-squares regression line for predicting beak heat loss, as a percentage of total body heat loss from all sources, from temperature? (Give your answers in three decimal places.)

    $$Slope:\ \ b = r\frac{s_y}{s_x} = 0.914 \times \frac{10.751}{4.761} = 2.064$$

    $$Intercept:\ a = \bar{y} - b\bar{x} = 47.375 - (2.064 \times 22.5) = 0.935$$

    $$Least-squares\ regression\ line:\ \hat{y} = 0.935 + 2.064\,x$$

    b. Explain in specific language what the slope and intercept of this line say about the relationship between beak heat loss and temperature.

    Slope: For every 1℃ increase in the outdoor temperature, the toucan will lose about 2.064% more heat through its beak.

    Intercept: When the outdoor temperature is 0℃, the toucan will lose about 0.935% heat through its beak.

    c. Use the equation that you found in (a) to predict beak heat loss, as a percentage of total heat loss from all sources, at a temperature of 25℃.

    $$\hat{y} = 0.935 + (2.064 \times 25) = 52.535$$

    d. What percentage of the variation in beak heat loss is explained by the straight-line relationship with temperature?

    The question asks the value of coefficient of determination. $r^2 = 0.914^2 = 0.835$

    83.5% of the variation in beak heat loss is explained by the straight-line relationship with temperature.

11. List two types of good and two types of bad sampling designs.

    Good types of sampling designs: Simple random sampling (SRS), Stratified sampling, Multistage sampling

    Bad types of sampling designs: Convenience sample, Voluntary response sample, Made-up sample

12. A researcher is interested in understanding Kansas State University undergraduate student's perceptions of the nutritional quality of campus food in an academic year. It isn't practical to contact all students. He wants to collect data from a sample of students. Classify the following sampling approaches as convenience, voluntary, SRS, stratified, or multistage sampling.

    a. He will prepare a small questionnaire and send it to all the undergraduate KSU students via email and collect data from the responses.

       Voluntary sampling
       The participants make the choice to fill in or not fill in the questionnaire.

    b. The researcher has prepared a list with all the **undergraduate students in KSU** and will randomly select 1000 of them to collect data.

       Simple random sampling (SRS)
       We select the students randomly from the desired population.

    c. He will collect data from undergraduate students who participated in a similar study in the previous academic year.

       Convenience sampling
       The researcher chooses the sample according to his/her convenience. He/she uses an existing sample.

    d. He will group the students into academic year (freshman, sophomore, junior, senior). Then he will select students from each group with the proportion of overall students selected from a given academic year being equal to the proportion of students in that academic year at the university. The total number of selected students should be 1000.

       Stratified sampling
       Includes students from all the academic year (strata) using the population proportion.

    e. He will group the undergraduate students into college (10 - agriculture, engineering, arts and science etc.) in the first stage and academic year (4) in the second stage. He will select 20 students from each college-academic year combination.

       Multistage sampling
       Use SRS to select 10 colleges and then use all 4 academic years. So, 10 x 4 = 40 combinations.
       Chooses only 20 students from each combination.
       40 x 20 = 800 students in the sample.

13. A survey is carried out at Kansas State University to estimate the proportion of all undergraduate students living at home during the current term. Of all the undergraduate students enrolled at the campus, a random sample of 100 was surveyed.

    a. What is the population of interest?

    Undergraduate students in KSU

    b. What is the sample?

    100 undergraduate students in KSU

14. Educational policy researchers randomly selected 400 teachers at random from the National Science Teachers Association database of members and asked them whether or not they believed that evolution should be taught in public schools. They received responses from 252 teachers.

    a. What is the population of interest?

    Teachers from National Science Teachers Association (members of NSTA)

    b. What is the sample?

    400 teachers from NSTA

15. Toxoplasmosis is an infection with a parasite called Toxoplasma gondii. In the USA, roughly 9% of all people who are 12-49 years old have antibodies, suggesting infection. A sample of 100 people were selected for a study to test the effectiveness of a new drug for treating Toxoplasmosis.

    a. What is the population of interest?

    People affected by Toxoplasmosis

    b. What is the sample?

    100 people affected by Toxoplasmosis