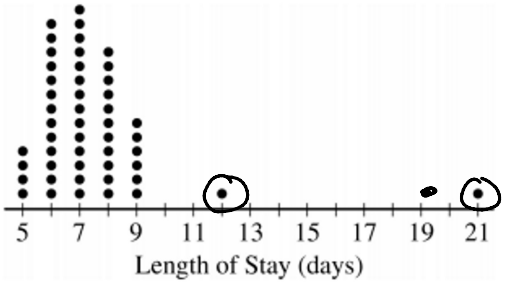
FRQ #1

Wednesday, May 19, 2021 10:38 AM

1. The length of stay in a hospital after receiving a particular treatment is of interest to the patient, the hospital, and insurance providers. Of particular interest are unusually short or long lengths of stay. A random sample of 50 patients who received the treatment was selected, and the length of stay, in number of days, was recorded for each patient. The results are summarized in the following table and are shown in the dotplot.

Length of stay (days)	5	6	7	8	9	12	21
Number of patients	4	13	14	11	6	1	1



50 patients

each quartile $\frac{50}{4} = 12.5$

(a) Determine the five-number summary of the distribution of length of stay.

min = 5

$Q_1 = 6$

Q_2 (median) = 7

$Q_3 = 8$

max = 21

(b) Consider two rules for identifying outliers, method A and method B. Let method A represent the $1.5 \times \text{IQR}$ rule, and let method B represent the 2 standard deviations rule.

(i) Using method A, determine any data points that are potential outliers in the distribution of length of stay. Justify your answer.

$IQR = Q_3 - Q_1 = 2$

$1.5\ IQR = 3$

$Q_1 - 1.5\ IQR = 6 - 3 = 3$

$Q_3 + 1.5\ IQR = 8 + 3 = 11$

because they are

$Q_3 + 1.5\,IQR = 8 + 3 = 11$

12, 21 days are outliers because they are above $Q_3 + 1.5\,IQR$ (upper fence)

(ii) The mean length of stay for the sample is 7.42 days with a standard deviation of 2.37 days. Using method B, determine any data points that are potential outliers in the distribution of length of stay. Justify your answer.

$7.42 \pm 2(2.37) =$ 2.68 to 12.16

21 days outlier because it is beyond 2 standard deviations from the mean

(c) Explain why method A might identify more data points as potential outliers than method B for a distribution that is strongly skewed to the right.

1) mean is more sensitive to outliers than Q_1 or Q_3 and will be shifted right due to the right skewness. This will shift the center of the interval.

2) standard deviation is more influenced by outliers than IQR. Thus, the ~~skew~~ extreme values will give a larger interval when based on standard deviation.

FRQ #2

Wednesday, May 19, 2021 10:39 AM

2. Researchers will conduct a year-long investigation of walking and cholesterol levels in adults. They will select a random sample of 100 adults from the target population to participate as subjects in the study.

   (a) One aspect of the study is to record the number of miles each subject walks per day. The researchers are deciding whether to have subjects wear an activity tracker to record the data or to have subjects keep a daily journal of the miles they walk each day. Describe what bias could be introduced by keeping the daily journal instead of wearing the activity tracker.

more ~~a~~ inclined to report when you did walk a lot versus ~~repor~~ less likely to report when you didn't walk

During the course of the study, the subjects will have their cholesterol levels measured each month by a doctor. The researchers will perform a significance test at the end of the study to determine whether the average cholesterol level for subjects who walk fewer miles each day is greater than for those who walk more miles each day.

(b) Selecting a random sample creates a reasonable representative sample of the target population. Explain the benefit of using a representative sample from the population.

any benefits our results can be extrapolated to the population of interest.

(c) Suppose the researchers conduct the test and find a statistically significant result. Would it be valid to claim that increased walking causes a decrease in average cholesterol levels for adults in the target population? Explain your reasoning.

no. This is an observational study. We can conclude a correlation, not causation, between walking and lower cholesterol levels.

FRQ #3

Wednesday, May 19, 2021 10:56 AM

3. To increase morale among employees, a company began a program in which one employee is randomly selected each week to receive a gift card. Each of the company's 200 employees is equally likely to be selected each week, and the same employee could be selected more than once. Each week's selection is independent from every other week.

(a) Consider the probability that a particular employee receives at least one gift card in a 52-week year. $P(X \geq 1)$

(i) Define the random variable of interest and state how the random variable is distributed.

X = # gift cards an employee receives per year

binomial distribution

$n = 52$

$p = \frac{1}{200} = .005$

(ii) Determine the probability that a particular employee receives at least one gift card in a 52-week year. Show your work.

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \left(\frac{199}{200}\right)^{52} = \boxed{0.2295}$$

(b) Calculate and interpret the expected value for the number of gift cards a particular employee will receive in a 52-week year. Show your work.

$$E(X) = np = 52(.005) = \boxed{0.26 \text{ gift cards}}$$

An employee would expect to win on average 0.26 gift cards per year when participating in this for a large number of years.

(c) Suppose that Agatha, an employee at the company, never receives a gift card for an entire 52-week year. Based on her experience, does Agatha have a strong argument that the selection process was not truly random? Explain your answer.

(c) Suppose that Agatha, an employee at the company, never receives a gift card for an entire 52-week year. Based on her experience, does Agatha have a strong argument that the selection process was not truly random? Explain your answer.

no. there is a .7705 chance that Agatha would not receive a gift card in a year, ~~so it is~~ which is very likely

FRQ #4

Wednesday, May 19, 2021 11:03 AM

4. The manager of a large company that sells pet supplies online wants to increase sales by encouraging repeat purchases. The manager believes that if past customers are offered $10 off their next purchase, more than 40 percent of them will place an order. To investigate the belief, 90 customers who placed an order in the past year are selected at random. Each of the selected customers is sent an e-mail with a coupon for $10 off the next purchase if the order is placed within 30 days. Of those who receive the coupon, 38 place an order.

(a) Is there convincing statistical evidence, at the significance level of $\alpha = 0.05$, that the manager's belief is correct? Complete the appropriate inference procedure to support your answer.

1 – proportion test

conditions:

independence = sample size ≤ 10% of population ✓

90 ≤ 10% of pop

reasonable to assume pop of customers ≥ 900

randomness ⇒ given ✓

normality ⇒ #success, #failures ≥ 10

38 ≥ 10 52 ≥ 10 ✓

⇒ H_0: $p = 0.4$ $n = 90$ $\hat{p} = 0.4222$ $\left(\frac{38}{90}\right)$

H_A: $p > 0.4$ $\bar{x} = 38$ $z = 0.43$

p-value = 0.333 > α

not enough evidence to conclude that $p > 0.4$

(b) Based on your conclusion from part (a), which of the two errors, Type I or Type II, could have been made? Interpret the consequence of the error in context.

H_A is actually true.

Type II error is possible. We failed to reject H_0 when we should have.

The true proportion is greater than 0.4 but we did not have enough evidence to conclude that.

The $10 off coupon was more effective than we concluded.

FRQ #5

Wednesday, May 19, 2021 11:03 AM

5. A research center conducted a national survey about teenage behavior. Teens were asked whether they had consumed a soft drink in the past week. The following table shows the counts for three independent random samples from major cities.

	Baltimore	Detroit	San Diego	Total
Yes	727	1,232	1,482	3,441
No	177	431	798	1,406
Total	904	1,663	2,280	4,847

(a) Suppose one teen is randomly selected from each city's sample. A researcher claims that the likelihood of selecting a teen from Baltimore who consumed a soft drink in the past week is less than the likelihood of selecting a teen from either one of the other cities who consumed a soft drink in the past week because Baltimore has the least number of teens who consumed a soft drink. Is the researcher's claim correct? Explain your answer.

prop for Baltimore $= \frac{727}{904} = 0.804$

prop for Detroit $= \frac{1232}{1663} = 0.74$

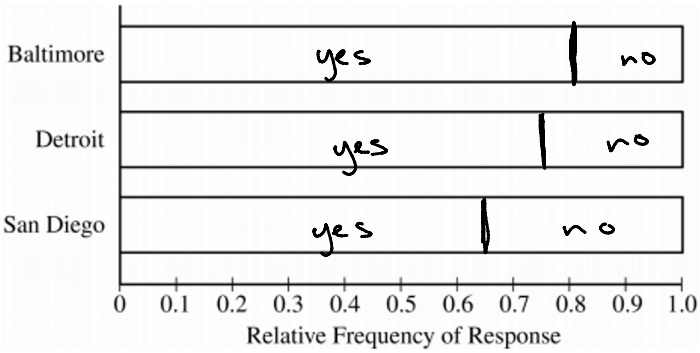
prop for SD $= \frac{1482}{2280} = 0.65$

the researcher is incorrect

(b) Consider the values in the table.

(i) Construct a segmented bar chart of relative frequencies based on the information in the table.

$\frac{727}{904} = 0.804$



$\frac{1232}{1663} = 0.74$

$\frac{1482}{2280} = 0.65$

(ii) Which city had the smallest proportion of teens who consumed a soft drink in the previous week? Determine the value of the proportion.

San Diego 0.65

(c) Consider the inference procedure that is appropriate for investigating whether there is a difference among the three cities in the proportion of all teens who consumed a soft drink in the past week.

(i) Identify the appropriate inference procedure.

Chi Squared Test for Homogeneity

(ii) Identify the hypotheses of the test.

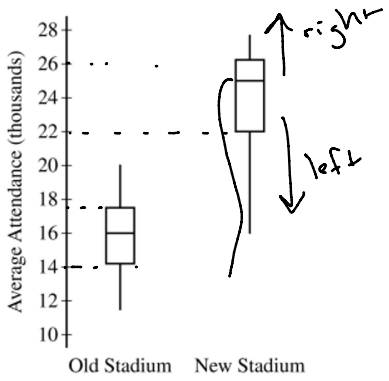
H_0: ~~all~~ proportion of teens who drink soda in the past week is the same among all cities

H_A: at least one proportion in the cities has a different value

FRQ #6

Wednesday, May 19, 2021 11:04 AM

6. Attendance at games for a certain baseball team is being investigated by the team owner. The following boxplots summarize the attendance, measured as average number of attendees per game, for 47 years of the team's existence. The boxplots include the 30 years of games played in the old stadium and the 17 years played in the new stadium.



(a) Compare the distributions of average attendance between the old and new stadiums.

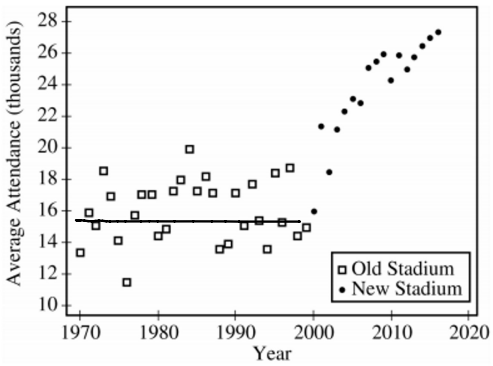
Center: new stadium has a higher median # attendeeds (25) compared to old stadium (16)

Spread: similar IQR (4) but new stadium does have a larger range (12 vs 8)

outliers: no apparent outliers

Shape: old stadium has no skew & new stadium is skewed left

The following scatterplot shows average attendance versus year.



(b) Compare the trends in average attendance over time between the old and new stadium.

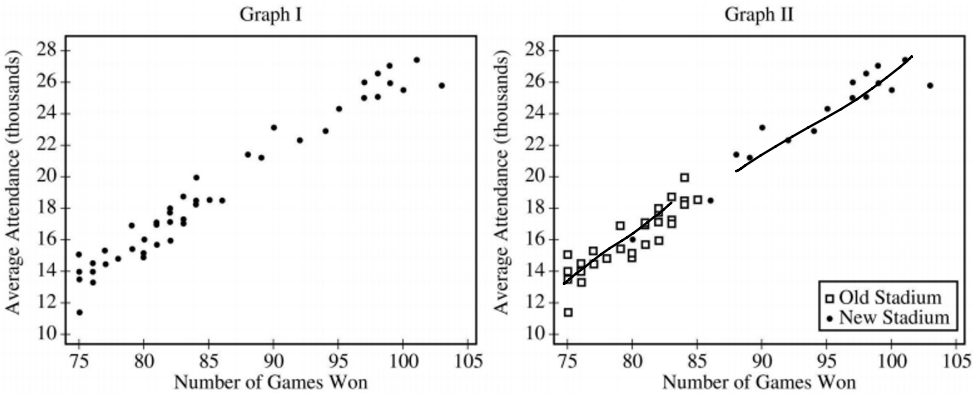
old stadium weak/no trend over time with (no change over time)

(b) Compare the trends in average attendance over time between the old and new stadium.

old stadium, weak/no trend over time with (no change over time)
a larger variability.

new stadium, strong positive linear trend over time (more attendees over time)
neither has any notable outliers

(c) Consider the following scatterplots.



(i) Graph I shows the average attendance versus number of games won for each year. Describe the relationship between the variables.

Strong, positive, linear association between # games won and average attendance with no noticeable outliers

(ii) Graph II shows the same information as Graph I, but also indicates the old and new stadiums. Does Graph II suggest that the rate at which attendance changes as number of games won increases is different in the new stadium compared to the old stadium? Explain your reasoning.

rate of attend

The rate at which attendance changes as number of games won

is about the same between the new and old stadium. The slope of the best fit line looks to be approximately the same.

(d) Consider the three variables: number of games won, year, and stadium. Based on the graphs, explain how one of those variables could be a confounding variable in the relationship between average attendance and the other variables.

When comparing the new and old stadium, attendance is higher. But that may due to the fact that more games are being won now, which is causing the higher attendance. And the newness of the stadium is just a coincidence.

Alternative:
Attendance is higher because they're winning more games, but that could be due to the new stadium which has better equipment for training. So it is the stadium that is causing the increased in attendance, not just the number of wins.