1) a) When should you choose a bar graph over a histogram?

**Bar graphs are for multiple (categorical) variables, histograms are for a single quantitative variable.**

b) When should you choose a histogram over a stemplot?

**When there is a lot of data.**

d) When would you choose a time plot over a histogram?

**When time is a relevant factor, for instance in weather patterns.**

d) Which is usually better at describing a skewed distribution, the five number summary or the mean and standard deviation?

**The five number summary since the median and IQR are resistant to outliers.**

2) Sketch a skewed right and skewed left distribution (label each). Indicate where the mean and median would lie on each curve.

See textbook page 56, for an example.

3) a) Give the 5 point number summary for the following data: 8,10,2,14,16 and draw a boxplot.

First, we need to reorder the data:

2,8,10,14,16.

Now, Min = 2, Q₁ = \( \frac{2+8}{2} = 5 \), Median = 10, Q₃ = \( \frac{14+16}{2} = 15 \), and Max = 16.

![Boxplot](image)

b) Determine (mathematically) if there are any outliers for the data in 3a)

IQR = 15 – 5 = 10.

(1.5) 10 = 15.

Since Q₁ – 15 = –10 and Q₃ + 15 = 30, there are no outliers.

c) Find the standard deviation for the data in 3a)

\( \bar{x} = \frac{8+10+2+14+16}{5} = 10. \)

\( s^2 = \frac{2^2+0^2+8^2+4^2+6^2}{4} = 30 \Rightarrow s = \sqrt{30}. \)

4) Define a standard normal distribution (if you use normal notation explain what each number means)

\( N(0,1) \) where the mean is 0 and the standard deviation is 1.

5) The distribution of ACT scores for more than 1 million students in a recent high school graduating class was roughly Normal with mean \( \mu = 20 \) and standard deviation \( \sigma = 5 \). The SAT scores for 1.4 million students in the same graduating class were roughly Normal with mean \( \mu = 1000 \) and standard deviation \( \sigma = 200 \).

A) Tonya scores 1300 on the SAT. Jermaine scores 25 on the ACT. Assuming that both tests measure the same thing, who has the higher score? Report the z-scores for both students.

For Tonya we have \( \frac{1300-1000}{200} = \frac{3}{2} = 1.5 \) and for Jermaine we have \( \frac{25-20}{5} = 1 \).
So Tonya did better. 
B) What score on on the ACT is equivalent to Tonya’s score from part A)?

\[ \frac{3}{2} = \frac{x - 20}{5} \Rightarrow \frac{15}{2} = x - 20 \Rightarrow x = 27.5. \]

6) The length of human pregnancies from conception to birth varies according to a distribution that is approximately normal with mean 260 days and standard deviation 20 days.

a) What percent of pregnancies last less than 240 days (about 8 months)?

\[ \frac{240 - 260}{20} = -1 = z. \]

From table A) we can see that this means that 15.87% had pregnancies that last less than 240 days.

b) What percent of pregnancies last between 240 and 270 days (between 8 and 9 months)?

\[ \frac{270 - 260}{20} = \frac{1}{2} = z. \] The cumulative distribution for this value is 69.15%

From part a) we found that 240 has a \( z \) score of -1 with cumulative distribution 15.87%. Therefore,

\[ (69.15 - 15.87) \% = 53.28\% \]

c) How long do the longest 30.85% of pregnancies last?

From table A) we can see this corresponds to a \( z \)-score of .5.

So we solve \( \frac{x - 260}{20} = \frac{1}{2} \), Solution is: 270.

7) Fill in the blanks:

68% of the observations fall within one standard deviation of the mean.

95% of the observations fall within two standard deviations of the mean.

99.7% of the observations fall within three standard deviations of the mean.

8) 
A) What is a scatterplot mainly used for?

Determining the relationship (if any) between two quantitative variables measured on the same set of individuals.

B) Does there have to be an explanatory/response association between variables of a scatterplot?

No.

C) If there is an explanatory/response association between variables, what axis should they be plotted on in a scatterplot?

Explanatory on the x-axis, Response on the y-axis.

D) Draw an example of a scatterplot with a strong positive association.

See notes.

9) Suppose we’re have the following set of data for some variable \( x: 10, 12 \).

Suppose we’re have the following set of data for some variable \( y: 5, 11 \).

a) What is the correlation between \( x \) and \( y \)?

\[ \bar{x} = \frac{10 + 12}{2} = 11, \bar{y} = \frac{5 + 11}{2} = 8. \]

\[ s_x^2 = \frac{(10 - 11)^2 + (12 - 11)^2}{1} = s_x = \sqrt{2} \text{ and } s_y^2 = \frac{3^2 + 3^2}{2} = s_y = \sqrt{18}. \]

\[ r = \frac{1}{\sqrt{18} \sqrt{2}} \left( -1(-3) + 1(3) \right) = \frac{6}{\sqrt{36}} = 1. \]

b) Give the equation for the least squares regression line.

\[ b = \frac{1 \sqrt{13}}{\sqrt{2}} = \sqrt{3} = 3. \]
\[ \hat{y} = 3x + a \Rightarrow 8 = 33 + a \Rightarrow a = -25. \]
Therefore, \( \hat{y} = 3x - 25. \)

10) A least squares regression line is a line through the point \((\bar{x}, \hat{y})\) with slope \(\_\_\_\_\_. \)

\[ b = r \frac{s_y}{s_x}. \]

11) A study found a correlation of \( r = -0.61 \) between the gender of a worker and his or her income. Determine whether each of the following conclusions regarding this correlation coefficient is true or false.

a) women earn more than men on average \(F\)

b) women earn less than men on average \(F\)

c) an arithmetic mistake was made. Correlation must be positive \(F\)

d) The measurement makes no sense; \( r \) can only be measured between two quantitative variables. \(T\)

12) Suppose we’re given the correlation \( r \). How do we interpret \( r^2 \)?

**As the percentage of the variation that is explained by the least squares regression line.**

13) **True** or False? In order to draw a regression line there must be an explanatory/response relation between the variables.

14) **Extrapolation** is the use of a regression line for prediction far outside of the range of values of the explanatory variable used to obtain the line.