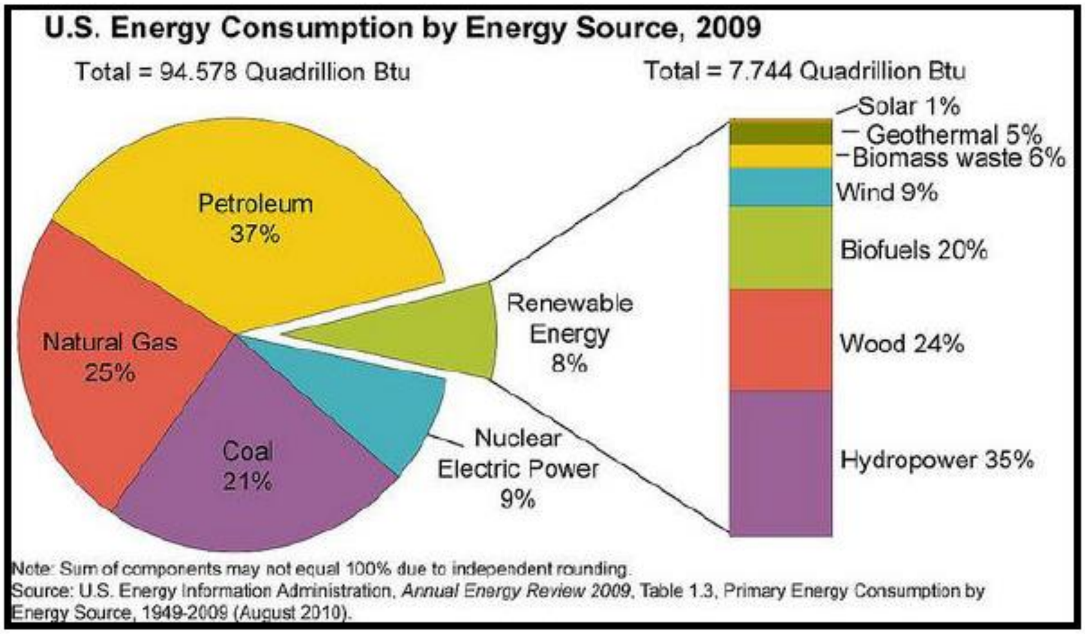
Week in Review – Stat 201
Session 3 – Tuesday, September 10th, 2024
Dr. Prasenjit Ghosh

TEXAS A&M UNIVERSITY
Math Learning Center

Problem 1

The pie chart below shows the US Energy Consumption by Energy Source for the year 2009[1].



Note: Sum of components may not equal 100% due to independent rounding.
Source: U.S. Energy Information Administration, *Annual Energy Review 2009*, Table 1.3, Primary Energy Consumption by Energy Source, 1949-2009 (August 2010).

1.1 The source with the highest consumption was

(a) petroleum.
(b) natural gas.
(c) coal.
(d) renewable energy.

As the pie chart shows energy consumption by energy source, identifying the source with the highest consumption is as simple as identifying which source is the largest chunk of the pie (or, in other words, the largest percentage): petroleum.

1.2 The combined percent of petroleum and natural gas was

(a) less than 25% of the total energy consumption.
(b) between 25% and 50% of the total energy consumption.
(c) between 50% and 75% of the total energy consumption.
(d) more than 75% of the total energy consumption.

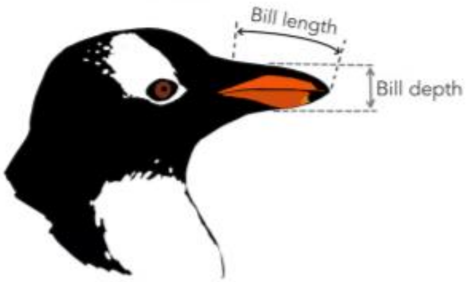
From the pie chart we can see that petroleum consumption was 37% of the total energy consumption and natural gas consumption was 25% of the total energy consumption.

[1] Duke, Çetinkaya-Rundel

Thus, the combined percentage of the two is 37% + 25% = 62% (which is between 50% and 75%).

Problem 2

Data were collected on 344 penguins living on three islands (Torgersen, Biscoe, and Dream) in the Palmer Archipelago, Antarctica. In addition to which island each penguin lives on, the data contains information on the species of the penguin (Adelie, Chinstrap, or Gentoo), its bill length, bill depth, and flipper length (measured in millimeters), its body mass (measured in grams), and the sex of the penguin (female or male). Bill length and depth are measured as shown in the image.15 (Gorman et al., 2014a) [1].



2.1 How many cases were included in the data?

344 cases (penguins) are included in the data. This is conveyed in the first sentence: “Data were collected on 344 penguins living on three islands…”

2.2 How many numerical variables are included in the data? Indicate what they are, and if they are continuous or discrete.

There are 4 numerical variables in the data: bill length, bill depth, and flipper length (measured in millimeters) and body mass (measured in grams). They are all continuous.

2.3 How many categorical variables are included in the data, and what are they? List the corresponding levels (categories) for each.

There are 3 categorical variables in the data: species (Adelie, Chinstrap, Gentoo), island (Torgersen, Biscoe, and Dream), and sex (female and male).

[1] Duke, Çetinkaya-Rundel

Problem 3

A survey of a random sample of 100 nurses working at a large hospital asked how many years they had been working in the profession. Their answers are summarized in the following (incomplete) table. Fill in the blanks in the table and round your answers to two decimal places:

# of years	Frequency	Cumulative frequency	Relative frequency	Cumulative relative frequency
<5	Using the relative frequency and the total number of nurses: .25 * 100 = **25**	This should be equal to the frequency because it is the first bin (i.e. there is no prior frequency to add to this one). .25 * 100 = **25**	.25	This should be equal to cumulative frequency divided by the total number of nurses: 25 / 100 = **.25**
5-10	30	This should be equal to the sum of the frequency of the first bin (25) and the frequency of this bin (30) 25 + 30 = **55**	This should be equal to frequency divided by the total number of nurses: **30 / 100 =** **.3**	Using similar logic to the above answer: 55 / 100 = **.55**
>10	Since we know there are 100 total nurses and each nurse should belong to one of these groups: 100 - 25 - 30 = **45**	This should be equal to the sum of the frequency of the first bin (25) and the frequency of the second bin (30) and the frequency of this bin (45): 25 + 30 + 45 = **100** *** Note: the cumulative frequency of the last bin/group should always be equal to the total	This should be equal to frequency divided by the total number of nurses: **45 / 100 =** **.45**	Using similar logic to the above answer: 100 / 100 = **1** *** Note: the cumulative frequency of the last bin/group should always be equal to 1

3.1 What proportion of nurses have five or more years of experience?

0.75. This is the sum of the relative frequencies of nurses that have 5-10 years of experience and the nurses that have 10+ years of experience: 0.3 + 0.45 = 0.75

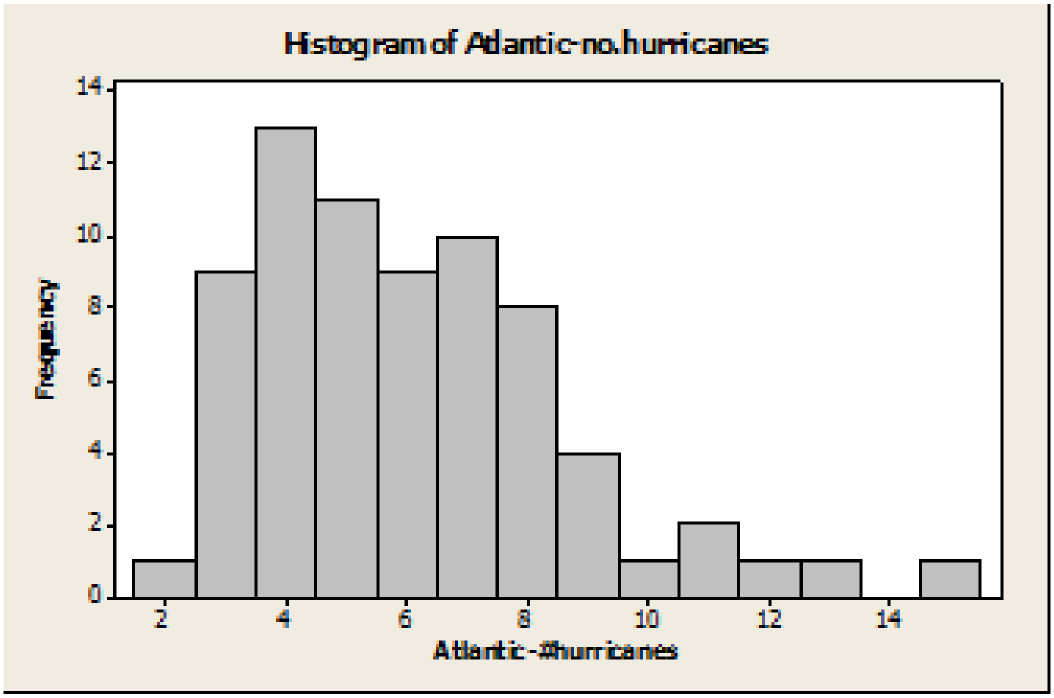
Another way to think about this is to simply subtract the relative frequency of the group that doesn't meet the criteria (5+ years of experience). Numerically speaking, 1 - 0.25 = 0.75

3.2 What proportion of nurses have ten or fewer years of experience?

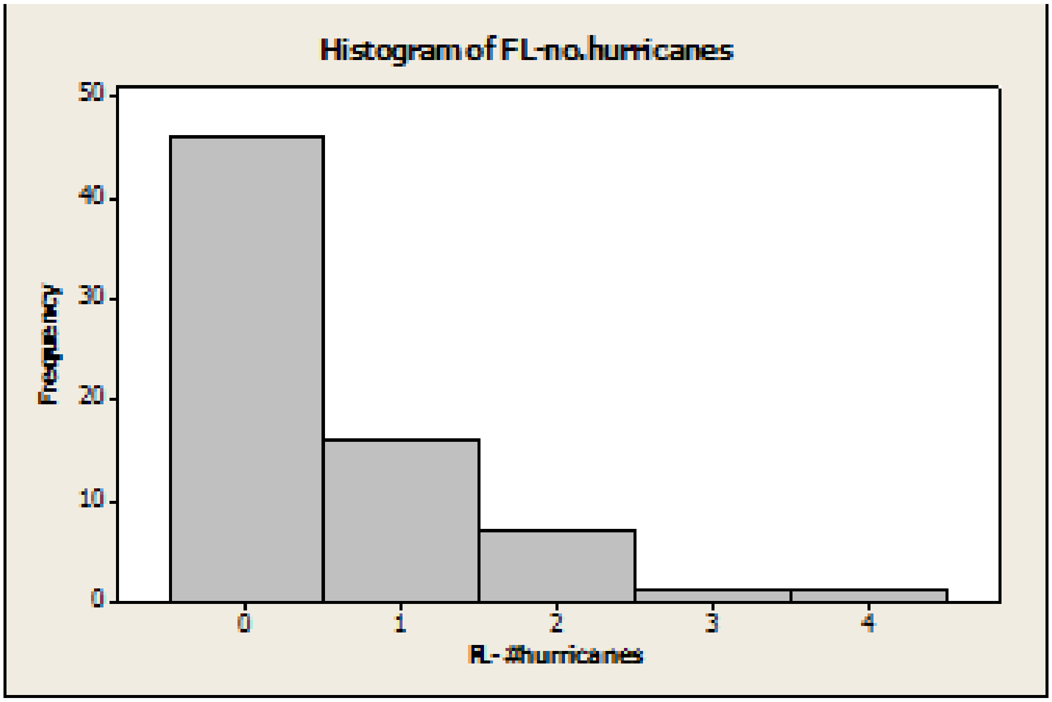
0.55. While there are also several ways to reach this conclusion, it's easiest to note that this is literally the definition of cumulative relative frequency.

Problem 4

The graphs below summarize hurricane activity for the 71 years in the period from 1940 to 2010. The top graph has the data on the number of hurricanes that form each year in the whole Atlantic Basin, and the bottom graph has information on how many of those hurricanes hit the state of Florida (Ripol).



1 Duke, Çetinkaya-Rundel



4.1 Which is the variable for the top graph? Which type of variable is it?

The annual number of hurricanes in the Atlantic. It is numerical and discrete (we cannot have half a hurricane).

4.2 Which is the parameter of interest for the top graph?

The mean number of hurricanes in the Atlantic per year.

4.3 On an average year we estimated to see about five (5) hurricanes in the Atlantic Basin and none (0) of them hitting Florida. Are the two numbers 5 and 0 hurricanes parameters or statistics?

a. 5: parameter; 0: statistic
b. 5: statistic; 0: parameter
c. Both are statistics
d. Both are parameters
e. It is impossible to tell without knowing which is μ and which is $\overline{X}$

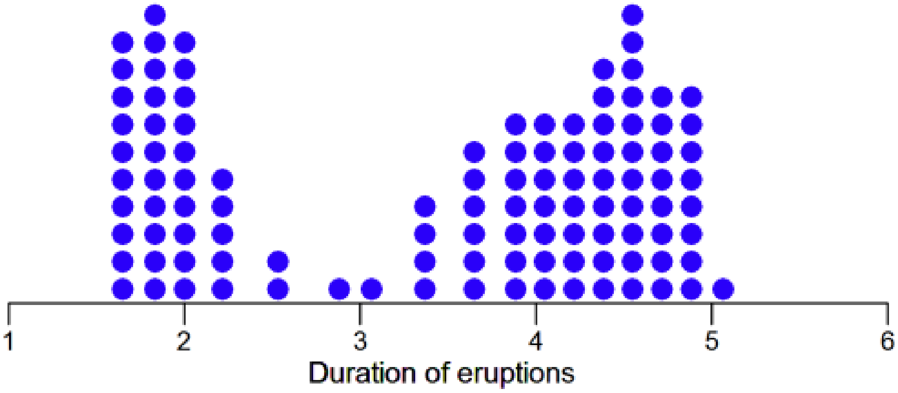
4.4 Which is the value of the median in Florida?

1 Duke, Çetinkaya-Rundel

Zero hurricanes. More than 50% of the data falls in the first bin/group on the second plot. Thus, the 50th percentile/median must be zero.

Problem 5

The following plot shows a dot plot of the duration of eruptions in minutes for the Old Faithful Geyser in Yellowstone National Park, Wyoming, USA (Dang).

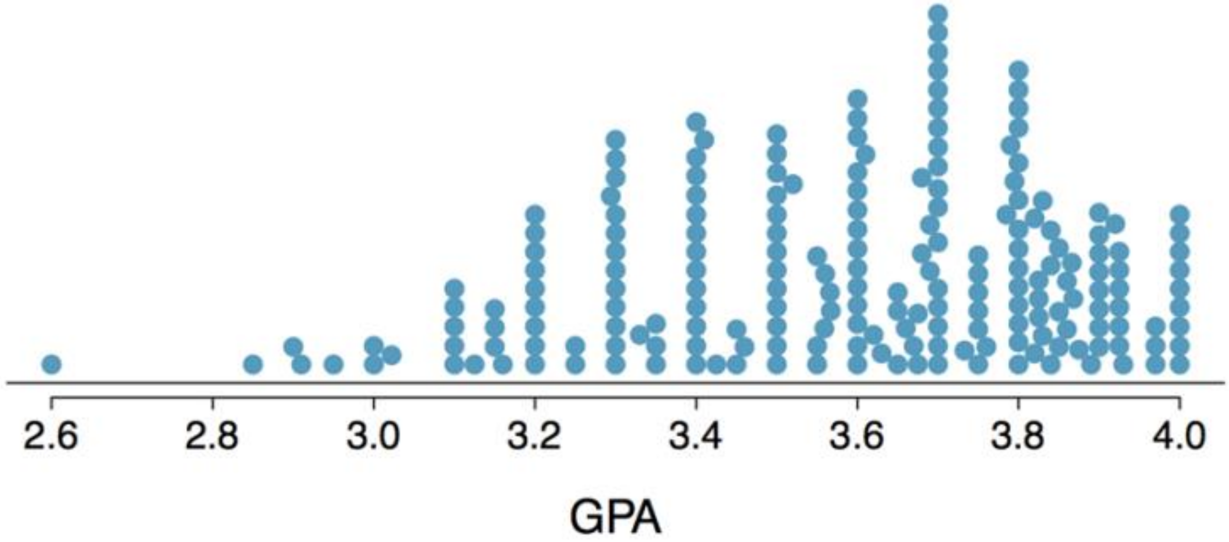


5.1 What is the minimum value approximately? 1.6 minutes
5.2 What is the maximum value approximately? 5.1 minutes
5.3 What is the range of the data? Approximately 5.1-1.6=3.5 minutes
5.4 Describe the shape of the distribution. Bimodal

1 Duke, Çetinkaya-Rundel

Problem 6

The following stacked dot plot represents the GPA scores of a group of students in an undergraduate course curriculum at a university.



6.1 Which of the following statements is true for the above stacked dot plot?

(a) Mode < Median < Mean
(b) Mode < Mean < Median
(c) Mean < Mode < Median
(d) Mean < Median < Mode
(e) Median < Mean < Mode
(f) Median < Mode < Mean

6.2 What measures of center and spread are best for the above stacked dot plot?

(a) Center: Mean; Spread: IQR
(b) Center: Median; Spread: IQR
(c) Center: Median; Spread: Range
(d) Center: Mean; Spread: Standard Deviation
(e) Center: Median; Spread: Standard Deviation

[1] Duke, Çetinkaya-Rundel

Problem 7

A May 2001 Gallup Poll found that many Americans believe in ghosts and other supernatural phenomena. The poll was based on telephone responses from 1,012 randomly selected adults. The table shows the percentages of people who expressed belief in various phenomena.

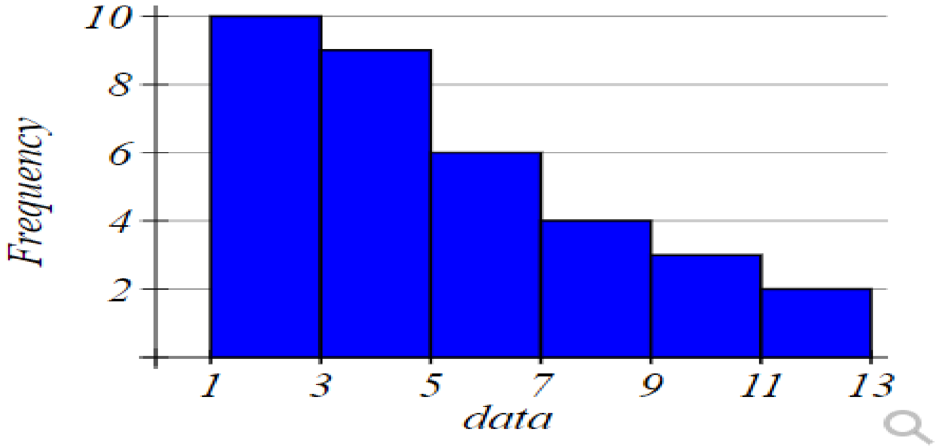
Phenomenon	Expressing Belief
Psychic Healing	54%
ESP	50%
Ghosts	38%
Astrology	28%
Channeling	15%

Is it reasonable to conclude that 66% of those polled expressed belief in either ghosts or astrology?

NO. Although the % of people expressing belief in ghosts + the % of people expressing belief in astrology = 66%, the groups MUST be overlapping with each other since the total % exceeds 100%.

Problem 8

Based on the histogram below:



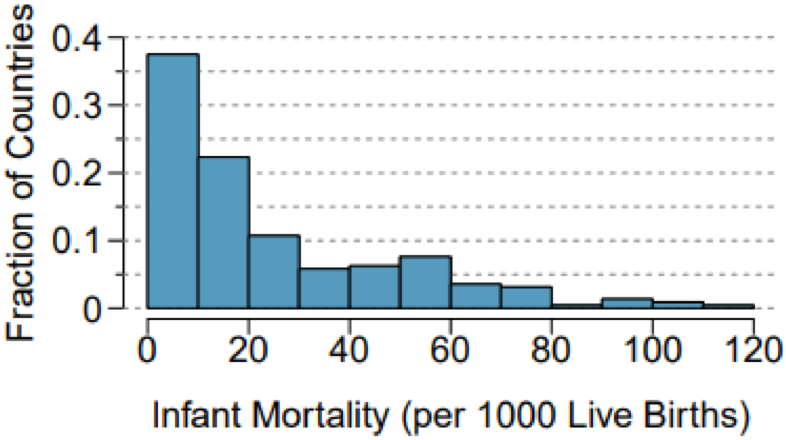
8.1 What is the class or bin width?
2 units. The width of each bin (evident on the x-axis).

8.2 What is the sample size approximately?
34 observations. The sum of all bin frequencies: **(10 + 9 + 6 + 4 + 3 + 2 = 34)**

[1] Duke, Çetinkaya-Rundel

Problem 9

The infant mortality rate is defined as the number of infant deaths per 1,000 live births. This rate is often used as an indicator of the level of the health sector in a country. The relative frequency histogram below shows the distribution of estimated infant death rates for 224 countries for which such data were available in 2014[1]. (Use this information to answer the following two problems).



9.1 Which is the variable of study? Which type of variable is it?
The number of infant deaths per 1,000 live births. It is a numerical and discrete variable.

9.2 What information does the histogram provide?
The histogram shows the distribution of estimated infant death rates for 224 countries for which such data were available in 2014

9.3 What proportion of countries has between 0 and 20 infant deaths per 1,000 live births?
Approximately 0.6 (0.375+0.225 = the sum of the first two bins' relative frequencies)

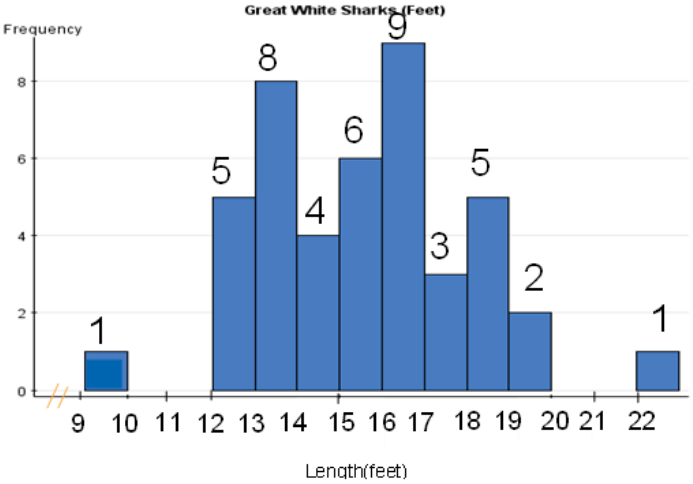
9.4 Would you expect the mean of this data set to be smaller or larger than the median? Explain your reasoning
The mean will be larger than the median because the distribution is skewed to the right.

9.5 Which numerical summary is a good measure of the center of the distribution? Explain why.
The median because the distribution is skewed to the right and the median is a robust/resistant measure of center.

[1] Duke, Çetinkaya-Rundel

Problem 10

Below is a histogram of the lengths (measured in feet) of 44 Great White Sharks.
(Use this information to answer the following two problems).



10.1 What proportion of sharks were between 15 and 20 feet? (n=44)

0.568 [(6+ 9 + 3 + 5 + 2)/44 = 0.568]

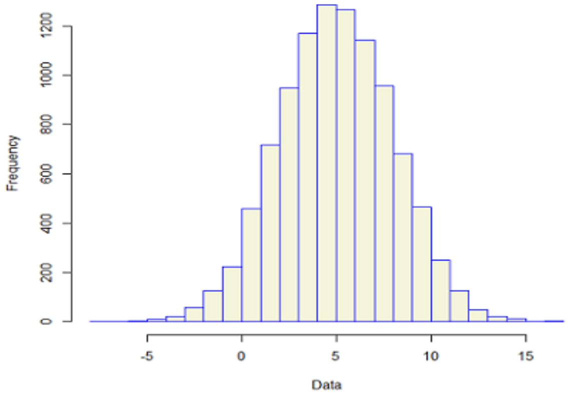
10.2 Which is the exact value of $\overline{x}$?

It cannot be estimated as we don’t have the values of the 44 observations

Problem 11

Which numerical summary is a good measure of center for the following distribution?

a) Median
b) Mean
c) Mode
d) Both median and mean
e) Median, mean, and mode



This is because the distribution lacks outliers and is symmetric. Essentially, with an approximately normal distribution, the mean ≈ median ≈ mode.

Problem 12

12.1 What is a reasonable action if an outlier is a legitimate data value and represents natural variability for the group and variable measured?

The value should not be discarded; in fact, it may be one of the more interesting values in the data set.

12.2 List statistics that give information only about the location of a dataset.

Mean, and median. Mode may be considered as a measure of location if the distribution is unimodal, otherwise use of a mode as a measure of location may be misleading.

12.3 List statistics that give information only about the spread of a dataset.

IQR, variance, standard deviation, and range. (Mean absolute deviation is another important measure of spread.)

Problem 13

In a survey, students are asked how many hours they study in a typical week. The five-number summary measures of the responses are:

2, 9, 14, 20, 60.

(Use this information to answer the following two problems).

13.1 Determine which of the following statements are true for the given dataset. More than one may apply.

a) The data set has no outliers.
b) The data set has exactly one outlier.
c) The data set has at least one outlier.
d) The data set has exactly two outliers: 2 and 60
e) About 50% of the students spend between 9 hours to 14 hours each week studying.

Here, IQR = Q_3-Q_1=20-9=11. Hence, Lower Fence = Q_1 - 1.5*IQR = 9 - 1.5*11 = -7.5, and Upper Fence = Q_3 + 1.5*IQR = 20 + 1.5*11 = 36.5. Since we are given 60 as the maximum, we know there is at least one outlier on the right (60 > Upper Fence).

13.2 Fill in the blank in the following sentence. About 75% of the students spent at least ____ hours studying in a typical week.

1 Duke, Çetinkaya-Rundel

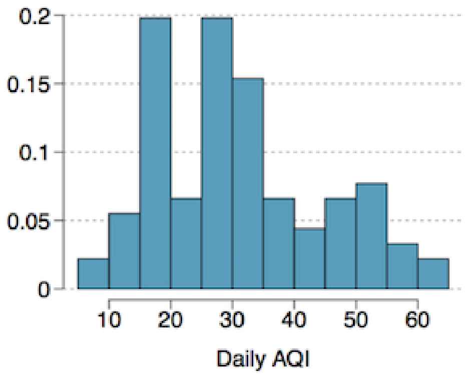
a) 9
b) 14
c) 20
d) 45
e) 60

This question requires that you identify the first quartile. The answer should be 9 hours, as you have approximately 75% studying 9 or more hours.

13.3 Which measure of center is not resistant to an outlier in the data? Mean

<u>Problem 14</u>

Daily air quality is measured by the air quality index (AQI) reported by the Environmental Protection Agency. This index reports the pollution level and what associated health effects might be a concern. The index is calculated for five major air pollutants regulated by the Clean Air Act and takes values from 0 to 300, where a higher value indicates lower air quality. AQI was reported for a sample of 91 days in 2011 in Durham, NC. The histogram below shows the distribution of the AQI values on these days.



14.1 Estimate the median AQI value of this sample.

Between 25 to 30 AQI

14.2 Would you expect the mean AQI value of this sample to be higher or lower than the median? Explain your reasoning.

Since the distribution is right skewed the mean is higher than the median.

Problem 15

Statistics exam scores for 33 students are as follows:

45, 59, 61, 63, 63, 63, 66, 67, 68, 70, 72, 72, 72, 72, 72, 73, 74, 75, 75, 76, 77, 77, 78, 78, 81, 83, 84, 84, 84, 87, 90, 93, 98.

Compute the three quartiles, and the IQR. Check if there is any outlier present in the data.

Steps:

1. The data is arranged in ascending order. So, no need to arrange them further.
2. Since $n = 33 \Rightarrow Q_2 =$ the 17th observation in the ordered arrangement $= x_{17} = 74$.
3. Divide the data set into two halves: $x_1, \ldots, x_{17}$, and $x_{17}, \ldots, x_{33}$.
4. $Q_1 =$ median of the first half consisting of observations $x_1, \ldots, x_{17} = x_9 = 68$.
5. $Q_3 =$ median of the second half consisting of observations $x_{17}, \ldots, x_{33} = x_{25} = 81$.
6. $IQR = Q_3 - Q_1 = 81 - 68 = 13$.
7. Lower Fence $= Q_1 - 1.5IQR = 68 - 1.5 \times 13 = 48.5$.
8. Upper Fence $= Q_3 + 1.5IQR = 81 + 1.5 \times 13 = 100.5$.
9. Since the score $45 <$ Lower Fence, it's an outlier.

Problem 16

AIDS data indicating the number of months a patient with AIDS lives after taking a new antibody drug are as follows:

3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47.

Compute the three quartiles, and the IQR. Check if there is any outlier present in the data.

[1] Duke, Çetinkaya-Rundel

Steps:

1. The data is arranged in ascending order. So, no need to arrange them further.
2. Since n = 40 ⇒ $Q_2 = (x_{20} + x_{21}) /2 = (24 + 24)/2 = 24$
3. Divide the data set into two halves: $x_1, ..., x_{20}$, and $x_{21}, ..., x_{40}$
4. Q_1 = median of $x_1, ..., x_{20} = (x_{10} + x_{11}) /2 = (15 + 15) /2 = 15$
5. Q_3 = median of $x_{21}, ..., x_{40} = (x_{30} + x_{31}) /2 = (32 + 33) /2 = 32.5$
6. $IQR = Q_3 - Q_1 = 32.5 - 15 = 17.5$
7. Lower Fence = $Q_1 - 1.5IQR = 15 - 1.5 \times 17.5 = -11.25$
8. Upper Fence = $Q_3 + 1.5IQR = 32.5 + 1.5 \times 17.5 = 58.75$
9. Since all the observations lie between the Lower Fence and the Upper Fence, there’s not any outliers.

Problem 17

On a timed math test, the first quartile for time it took to finish the exam was 35 minutes. Interpret the first quartile in this context.

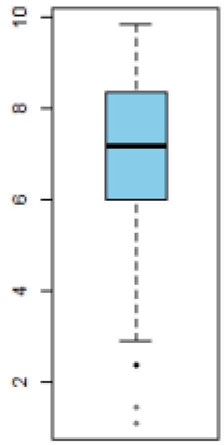
About 25 % of the total number of students finished the exam in 35 minutes or less, and around 75% of them took 35 minutes or more to submit the test.

A lower percentile is considered good in this context, as finishing more quickly on a timed exam is desirable. (If you take too long, you might not be able to finish.)

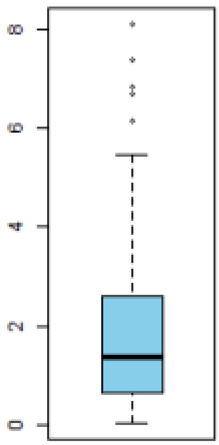
1 Duke, Çetinkaya-Rundel

Problem 18

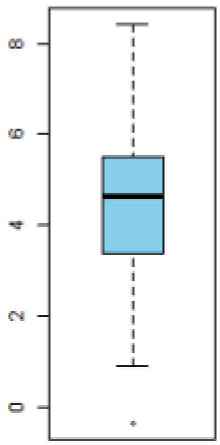
Identify the correct match of the box plots with the histograms given below:



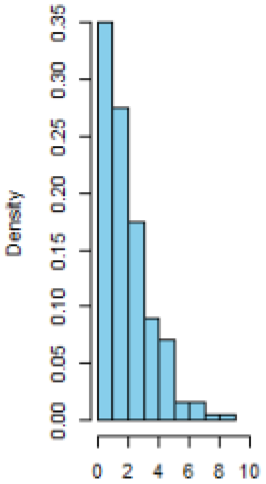
Boxplot A



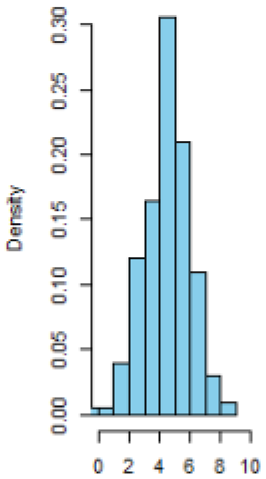
Boxplot B



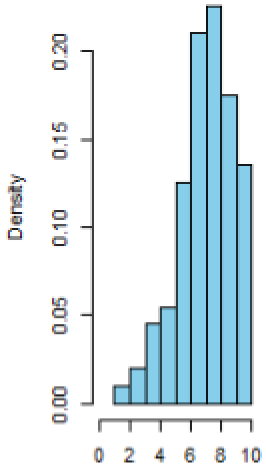
Boxplot C



Histogram 1



Histogram 2



Histogram 3

a) (Histogram 1, Boxplot C); (Histogram 2, Boxplot B); (Histogram 3, Boxplot A).
b) (Histogram 1, Boxplot B); (Histogram 2, Boxplot A); (Histogram 3, Boxplot C).
c) (Histogram 1, Boxplot A); (Histogram 2, Boxplot C); (Histogram 3, Boxplot B).
d) (Histogram 1, Boxplot B); (Histogram 2, Boxplot C); (Histogram 3, Boxplot A).
e) (Histogram 1, Boxplot C); (Histogram 2, Boxplot A); (Histogram 3, Boxplot B)

[1] Duke, Çetinkaya-Rundel

Problem 19

Which of the following data sets has the largest standard deviation?

a) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
b) 11.75, 12.75, 13.75, 14.75, 15.75, 16.75, 17.75, 18.75, 19.75, 20.75
c) 0, 0, 0, 0, 0, 10, 10, 10, 10, 10
d) 51, 52, 53, 54, 55, 56, 57, 58, 59, 60
e) 88.5, 89.5, 90.5, 91.5, 92.5, 92.5, 93.5, 94.5, 95.5, 96.5, 97.5

Problem 20

Which of the following statements about the sample standard deviation would be incorrect?

(a) A small value of the sample standard deviation indicates a small amount of variation among the data points.
(b) If the sample standard deviation is zero, all the observations must be equal to the sample mean, and conversely. (In case of no variability, a measure of spread must be zero.)
(c) A large value of the sample standard deviation indicates a large amount of variability among the data points, and there are some observations which are away from the sample mean.
(d) It is resistant to the presence of outliers.
(e) It works best for symmetric or nearly symmetric distributions.