Homework 1 Problems

(**1**) Define the following terms: * **(a)** Population - the collection of all possible persons, events, or objects of interest * **(b)** Sample - the subset of the population that we actually observe (the observed observations) * **(c)** Parameter - A numerical characteristic of a population (alternatively: A function of the entire population) * **(d)** Statistic - A numerical characteristic of a sample (alternatively: A function of the sample)

(**2**) Explain the difference between a **qualitative** (categorical) and a **quantitative** variable {A qualitative variable is a non-numeric characteristic such as a name or label while a quantitative variable is numerical characteristic such as a measurement or count}

(**3**) Explain the difference between a **discrete** and a **continuous** variable and give an example of each

(**4**) Explain the difference between a **nominal** and an **ordinal** variable and give an example of each

(**5**) **At What age did women marry?** A historian wants to estimate the average age at marriage of women in New England in the early 19th century. Within her state archives, she finds marriage records for the years $1800 - 1820$, which she treats as a sample of all marriage records from the early 19th century. The average age of the women in the records is 24.1 years of age. Using the appropriate statistical method, she estimates that the average age of brides in the early 19^{th}- century New England was between 23.5 and 24.7 years of age.

- **(a)** Which part of this example gives a descriptive summary of the data?
- **(b)** Which part of this example draws an inference about a population?
- **(c)** What population is the historian is studying?
- **(d)** The average age the historian computed from the historical records was 24.1 years of age. Is 24.1 years of age a statistic or a parameter? Why?

(**6**) Consider the following data obtained from flipping a coin 15 times. A value of H denotes the coin was heads whereas a value of T denotes the coin flip was tails.

coin.flip.result.

H

H

H

T

H

T

T

T

H

H

T

T

T

H

T

Compute the frequency table for the variable “coin flip result” and answer the following questions:

Result

Frequency

Relative Frequency

H

T

- **(a)** What type of variable is “Coin Flip Result”?
- **(b)** What proportion of the coin flips were heads?
- **(c)** Why do we not compute the cumulative relative frequency for this variable?
- **(d)** What the best graphical display for this variable and why?

(**7**) A survey about color preferences reported the age distribution of the people who responded. Below are the results

1

Age Group (Years)

1-18

19-24

25-25

36-60

51-69

70 and over

2

Counts

10

97

70

36

14

5

Use this table to answer parts a - d

- **(a)** Compute the relative frequency for each age group
- **(b)** Make a bar graph where the heights of the bars are relative frequencies
- **(c)** Describe the distribution
- **(d)** Explain why your bar graph is not a histogram

(**8**) Email spam is the curse of the internet. The table below gives a compilation of the most common types of spam

Type of Spam

Percentage

Adult

14.5

Financial

16.2

Health

7.3

Leisure

7.8

Products

21.0

Scams

14.2

Use the table to answer the questions a and b

- **(a)** Report the modal spam category
- **(b)** Construct a **Pareto chart** using the table above

(**9**)

(**10**) Consider the following data from a survey conducted on college students in the state of Florida. As part of their research, surveyors recorded the high school performance (measured in grade point average - GPA) of 60 college students from across the state. For your convenience the data have been sorted from least to greatest:

$$2.0, 2.1, 2.3, 2.8, 3.0, 3.0, 3.0, 3.0, 3.0, 3.0,$$
$$3.0, 3.0, 3.0, 3.1, 3.2, 3.3, 3.3, 3.4, 3.4, 3.4,$$
$$3.4, 3.5, 3.5, 3.5, 3.5, 3.6, 3.6, 3.7, 3.7, 3.8,$$
$$3.8, 3.8, 3.8, 4.0, 4.0$$

The following frequency table gives the distribution of the variable high school GPA (with values rounded to nearest 0.5). Fill in the table and answer the following questions:

GPA

Frequency

Relative frequency

Cumulative RF

2.0

3

0.05

2.5

2

3.0

0.4

3.5

25

0.42

4.0

11

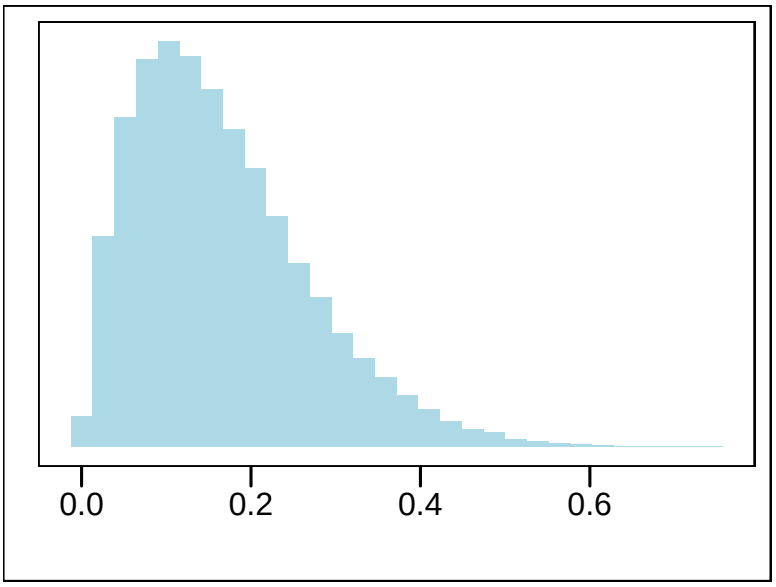
0.18

- **(a)** What type of variable is GPA?
- **(b)** What proportion of college students had a high school GPA > 3.0?
- **(c)** What proportion of college students had a high school GPA < 3.0?
- **(d)** What is the **mean** GPA in this sample?
- **(e)** What is the **median** GPA in this sample?
- **(f)** What is the **mode** of GPA in this sample?
- **(g)** Construct a dot plot for the variable GPA (hint use the frequency table)
- **(h)** Construct a histogram for the variable GPA (hint use the frequency table)
- **(i)** Construct a stem and leaf plot for the variable GPA

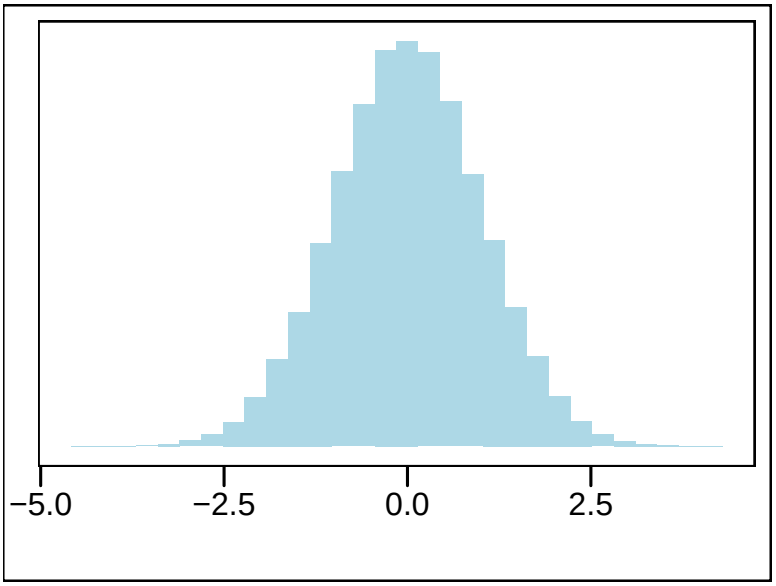
(**11**) Which statistic is more resistant to outliers, the mean or median? Why?

(**12**) Describe the shape of the following distributions and for each distribution identify if the mean will be larger, smaller or the same as the median.

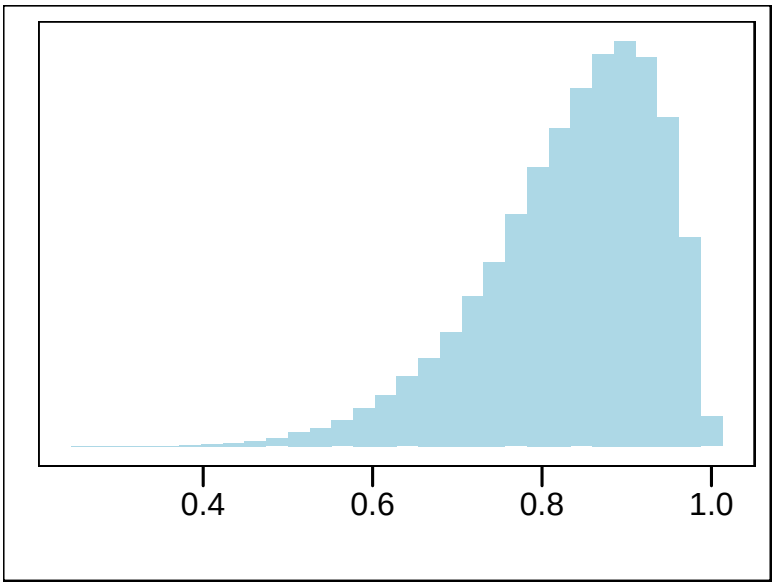
(a)



(b)

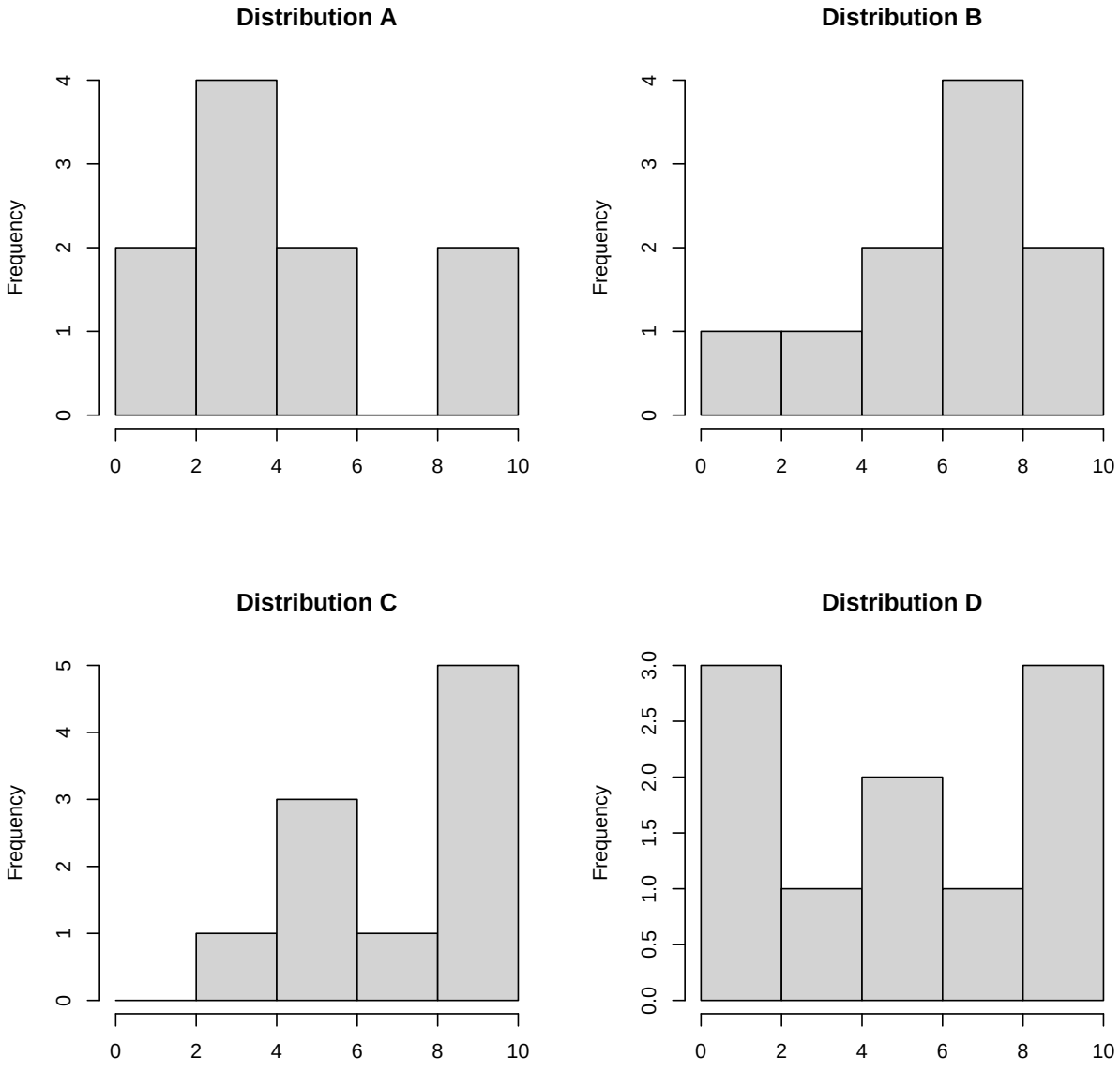


(c)



(**13**) Consider the following four sets of observations of a quantitative variable x. For your convenience the observations have been sorted in increasing order. Match datasets $1-4$ with the correct histogram (labeled $A-D$)

$$\text{Dataset } 1 = \{0.1, 1.1, 2.6, 2.7, 3.4, 3.4, 4.1, 4.4, 8.8, 9.6\}$$
$$\text{Dataset } 2 = \{0.1, 0.3, 1.2, 2.4, 4.4, 4.5, 8.0, 8.9, 9.3, 9.3\}$$
$$\text{Dataset } 3 = \{1.1, 3.8, 5.3, 6.0, 6.2, 6.9, 7.9, 7.9, 8.1, 8.7\}$$
$$\text{Dataset } 4 = \{3.4, 4.5, 5.4, 5.6, 7.0, 8.5, 8.9, 9.2, 9.7, 9.7\}$$



(**14**) Consider the following set of 10 observations of a variable X sorted from least to greatest:

$$3.3, 3.8, 4.0, 4.8, 4.8, 5.1, 5.2, 5.6, 5.7, 6.9$$

Use the data to answer parts a-b

- **(a)** Given the data, compute the 40^{th} percentile of X
- **(b)** Compute Interquartile Range (IQR) range of X

(**15**) Consider the following $n = 20$ observations of the sugar and sodium content of several popular cereal brands and answer questions a - g:

Brand

Sodium (mg)

Sugar (g)

Type

Frosted Mini Wheats

0

11

A

Raisin Bran

340

18

A

All Bran

70

5

A

Apple Jacks

140

14

C

Cap'n Crunch

200

12

C

Cheerios

180

1

C

Cinnamon Toast Crunch

210

10

C

Crackling Oat Bran

150

16

A

Fiber One

100

0

A

Frosted Flakes

130

12

C

Froot Loops

140

14

C

Honey Bunches of Oats

180

7

A

Honey Nut Cheerios

190

9

C

Life

160

6

C

Rice Krispies

290

3

C

Honey Smacks

50

15

A

Special K

220

4

A

Wheaties

180

4

A

Corn Flakes

200

3

A

Honeycomb

210

11

C

- **(a)** Construct a histogram of the data using the following bins: $\leq 2(g)$, $2-4(g)$, $4-6(g)$, $6-8(g)$,$8-10(g)$, $10-12(g)$, $12-14(g)$,$14-16(g)$,$> 16(g)$ (hint start by creating a frequency table)
- **(b)** Compute the mean, median, and mode of the variable "Sugar (g)"
- **(c)** Compute the variance and standard deviation of the variable "Sugar (g)"
- **(d)** Compute the quartiles (Q1, Q2, Q3) and interquartile range of the variable "Sugar (g)"
- **(e)** Create a box and whisker plot for the variable "Sugar (g)" and mark the Q1, Q2, and Q3 quartiles, the median and any potential outliers
- **(f)** Plot the Cumulative Distribution of the variable "Sugar (g)"
- **(g)** What proportion of cereals have a sugar content less than or equal to 10 grams?