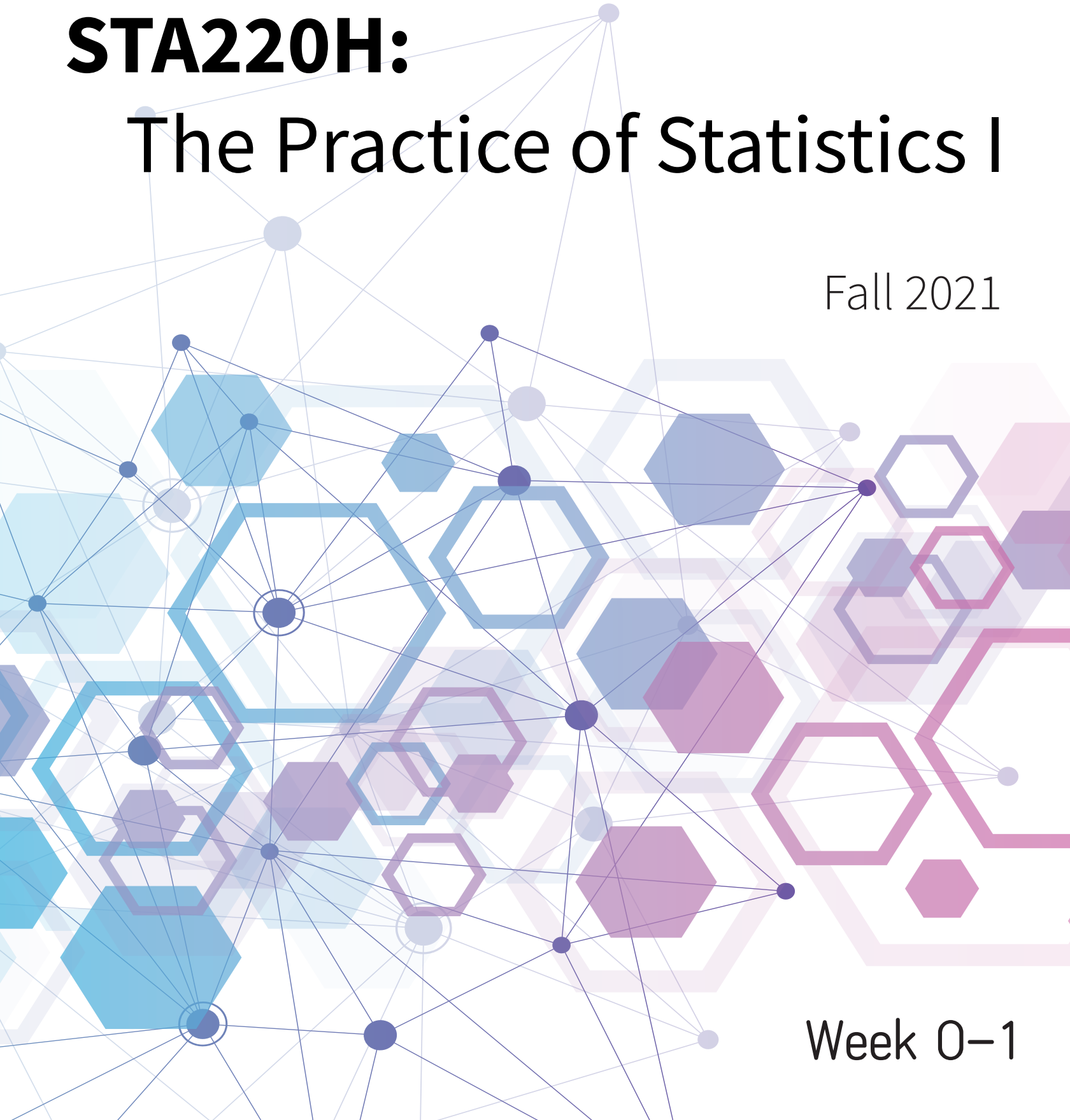
Talent
Education

STA220H:

The Practice of Statistics I

Fall 2021



Week 0–1

TABLE OF CONTENTS

TYPES OF STATISTICS

1. Descriptive statistics
   - Use numerical and graphical methods to look for patterns in a data set, to summarize the information revealed in a data set and to present that information in a convenient form.
   - Examples: Average, spread, range, mode, scatter plot, etc.
2. Inferential statistics
   - Use sample data to make estimates, decisions, predictions or other generalizations about a larger set of data.
   - Examples: Hypothesis test, confidence interval, etc.

KEY ELEMENTS OF STATISTICS

VARIABLE

Definition: Variable is a characteristic or property of an individual unit in the population.

TYPES OF VARIABLES

1. Qualitive (Categorical): The value of the data is a **category**.

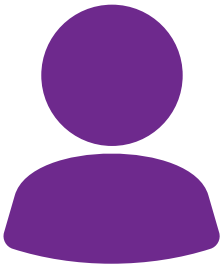
	Properties	Examples
Nominal	• Categories • **No ordering**	• Yes, No, Don't know • White, Black, Other
Ordinal	• Categories • **Order is important**	• Grade: A, B, C, D • Agree, Neutral, Disagree

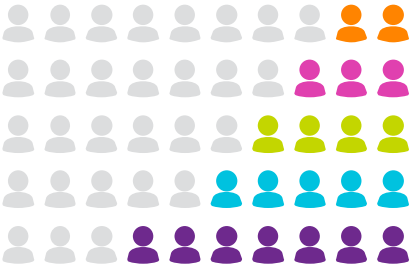
2. Quantitative (Numerical): The value of the data is a **number**.

	Properties	Examples
Discrete	• Quantitative • **Countable**	• # of book on bookshelf {0, 1, 2, ...}
Continuous	• Quantitative • **Not countable**	• Height • Waiting time

Note: Quantitative data can be converted into qualitive variable.

POPULATION & SAMPLE





1. Population: A set of units (usually people, objects, transactions, or events) that we are interested in studying.
2. Sample: A sample is a subset of the population.
3. Individual case: The subject which we collect data on.

DESCRIPTIVE STATISTICS

DESCRIBE CATEGORICAL DATA

FREQUENCY TABLE

Class	Count
First	325
Second	285
Third	706
Crew	885

Class frequency:

- Class: is one of the categories into which qualitative data can be classified.
- Class frequency: Number of observations in each class.

Class	%
First	14.77
Second	12.95
Third	32.08
Crew	40.21

Class relative frequency:

- Class frequency divided by the total number of observations in the data set.

$$\text{Class relative frequency} = \frac{\text{Class frequency}}{n}$$

Class percentage:

- Class relative frequency multiplied by 100%.

CONTINGENCY TABLE

The contingency table (Two-way table) is used to look **two different categorical variables**.

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	203	118	178	212	711
	Dead	122	167	528	673	1490
	Total	325	285	706	885	2201

The contingency table can show:

- Counts.
- Frequency and relative frequency.

1. Marginal distribution: Distribution of either variable alone.

		Sex		
		Male	Female	Total
Response	Game	279	200	**479**
	Commercial	81	156	**237**
	Won't Watch	132	160	**292**
	Total	**492**	**516**	**1008**

Marginal distribution of response:

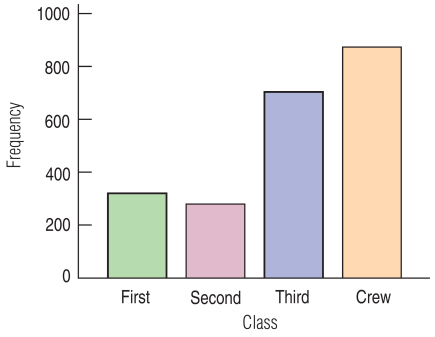
$$\frac{479}{1008} = 47.5\% \quad \frac{237}{1008} = 23.5\% \quad \frac{292}{1008} = 29.0\%$$

2. Conditional distribution: The relative frequency of each category of variable, given a specific value of the other variable in the contingency table.

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	**203** 28.6%	**118** 16.6%	**178** 25.0%	**212** 29.8%	**711** **100%**
	Dead	**122** 8.2%	**167** 11.2%	**528** 35.4%	**673** 45.2%	**1490** **100%**

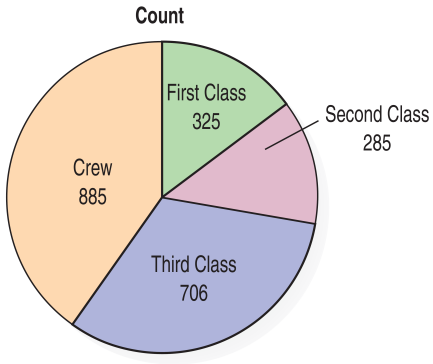
CHARTS

1. Bar charts



A bar chart displays the distribution of a categorical variable, showing the counts for each category next to each other for easy comparison.

2. Pie charts



Pie charts show the whole group of cases as a circle and slice the circle into pieces whose sizes are proportional to the fraction of the whole in each category.

DESCRIBE NUMERICAL DATA

CENTRAL TENDENCY

1. Mean
   a. **Mean is the average of a particular data set.**
   b. Notation:
      i. Sample mean is denoted as $\bar{x}$.
      ii. Population mean is denoted as μ.

c. Formula:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

d. Mean is sensitive to extreme values.

2. Median
   a. **Median is the middle value of a particular data set.**
   b. Median is resistant to extreme values.
   c. Note: If there is an even number of observations, the median is the average of the two middle values.

3. Mode
   a. **Mode is the data which has the highest frequency.**

LOCATION

1. Percentile
   a. A percentile (or a centile) is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall.
   b. Formula: p^{th} percentile at i^{th} position.

$$i = \frac{p}{100}(n+1)$$

2. Quartile
   a. Quartiles in statistics are values that divide your data into quarters.
   b. Common quartiles:
      i. **Lower quantile (QL) or 1st quantile: 25th percentile.**
      ii. Middle quantile(M) or median: 50th percentile.
      iii. **Upper quantile (QU) or 3rd quantile: 75th percentile.**
   c. Interquartile range (IQR): Difference between the 3rd quantile and 1st quantile.

Five-number summary:

1. Minimum
2. 1st quartile
3. Median
4. 3rd quartile
5. Maximum

Note: No Mean in the Five-number summary.

VARIATION

1. Range
   a. Range is the difference between the maximum and minimum value.

b. Easy to calculate but no robust and informative.

2. Variance
   a. Variance is the expectation of the squared deviation of a random variable from its mean.
   b. Notation:
      i. Sample variance is denoted by $\mathbf{s^2}$.
      ii. Population variance is denoted by $\boldsymbol{\sigma^2}$.
   c. Formula:

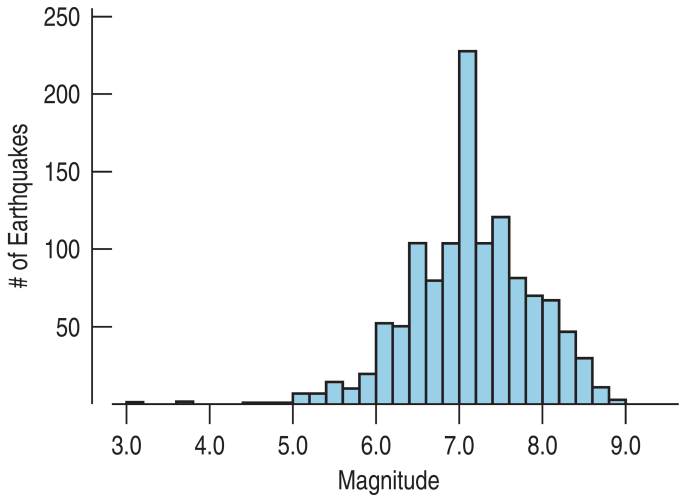
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

3. Standard deviation
   a. Standard deviation is simply the square root of the variance
   b. Notation:
      i. Sample standard deviation is denoted by $\mathbf{s}$.
      ii. Population standard deviation is denoted by $\boldsymbol{\sigma}$.
   c. Formula:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

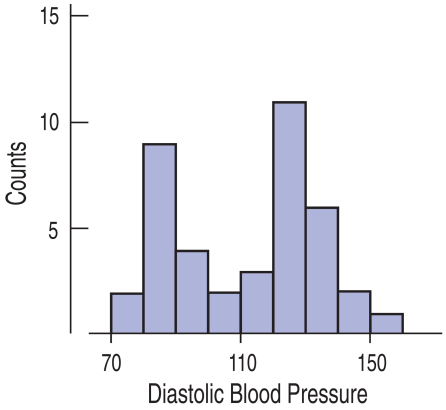
DIAGRAMS

1. Histogram
   a. A histogram is an accurate representation of the distribution of numerical data.
   b. Shows the frequency (or relative frequency) for each bin.
   c. It can demonstrate the distribution of samples.



d. Shapes of distribution:

i. Number of peaks



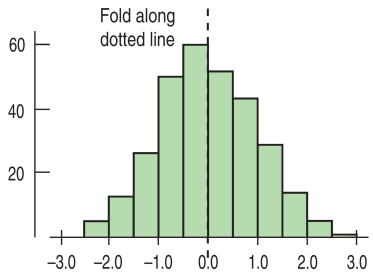
Number of peaks:

1. 1 Peak: Unimodal.
2. 2 Peaks: Bimodal.
3. More than 3 peaks: Multimodal.

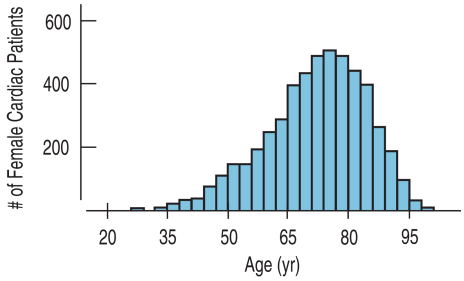
Note: If all bars are approximate same height, it is called **uniform**.

ii. Symmetry

1) Symmetric distribution

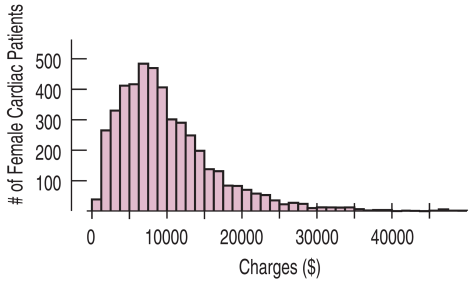


2) Skewed distribution



Skewed-Left distribution:

1. Negative skewness.
2. Skewness is determined by the tail, long left tail.
3. **Mean < Median < Mode**



Skewed-Right distribution:

1. Positive skewness.
2. Skewness is determined by the tail, long right tail.
3. **Mean > Median > Mode**

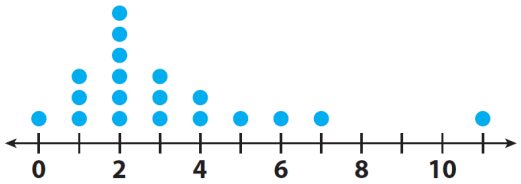
2. Stem-and-Leaf display

a. Stem-and-leaf display is like a histogram, but it shows the individual values.

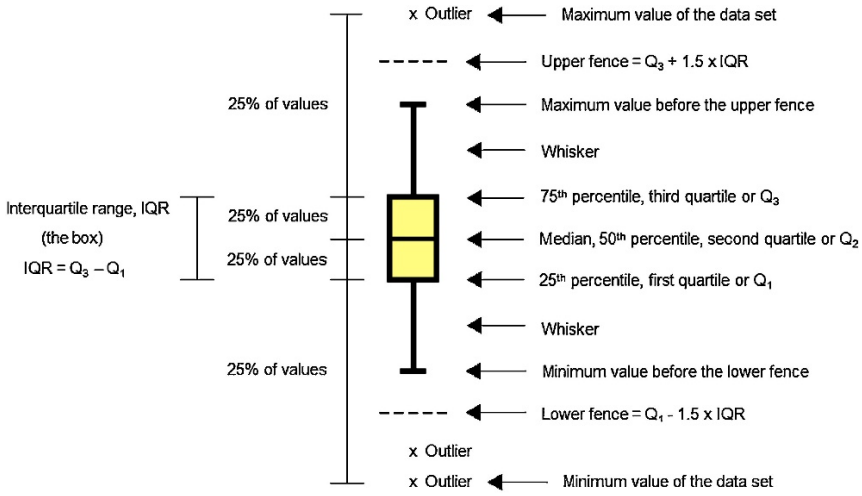
b. Stem-and Leaf plot shows raw data in two parts: **Stem** and **Leaf**.
c. Data are arranged in order from smallest to largest.

Stem	Leaf
2	3, 4, 6
3	2, 4, 6, 7, 8
4	2, 3, 3, 4, 6, 7, 9
5	0, 3, 6, 7
6	0, 3
7	1
8	2, 7, 9

3. Dot plot
   a. Replace the data on the stem-and-leaf display with dots.



4. Modified boxplot & standard boxplot
   a. Box plot uses five-number summary to represent the data set.
   b. Types of boxplot
      i. Standard boxplot: The standard boxplot includes ALL data points including outliers.
      ii. Modified boxplot: The modified boxplot shows outliers in individual data points.
   c. Sample's variability is interpreted by the box:
      i. Wider box: Higher variation.
      ii. Smaller box: Lower variation.
   d. Length of the whisker can infer skewness of the distribution:
      i. If left (lower) whisker is longer, the distribution is left skewed.
      ii. If right (upper) whisker is longer, the distribution is right skewed.



Outlier: An observation point that is distant from other observations.

PRACTICE

1. Draw stem-and-leaf plot for following data:
   41, 52, 6, 19, 22, 10, 40, 55, 60, 35

2. Find the mean, mode, and median of the following data values:
   a. 13, 10, 11, 7, 8, 12, 9, 11, 12, 10, 10, 5, 70

   b. 10, 11, 12, 10, 10, 11, 12, 13, 8, 9, 7

3. Find the range, variance, standard deviation of the following data values:
   68, 69, 93, 76, 87, 79, 88, 90, 74

4. Find the five-number summary of the following data values:
   78, 93, 68, 84, 90, 74

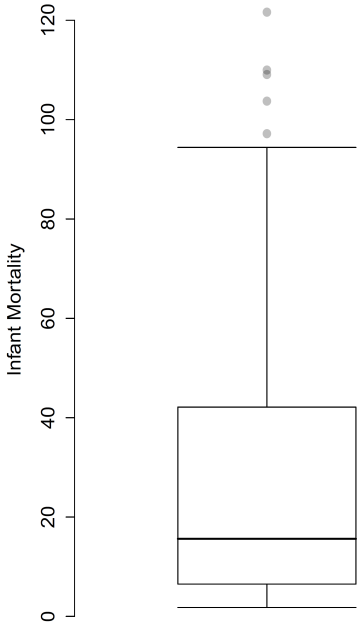
5. Draw a boxplot and modified boxplot for following data values:
   79, 68, 88, 69, 90, 74, 87, 93, 76

6. The boxplot below shows the distribution of estimated infant mortality rates for 224 countries in 2014.

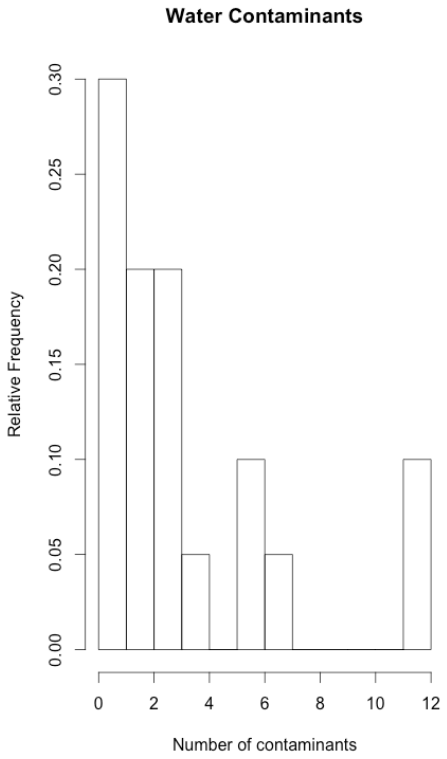
   a. Estimate the median infant mortality rate and the variability in infant mortality rate.

   b. Do you expect the mean of this data set to be smaller, larger, or equal to the median? Explain.

   c. Are there any countries that have infant mortality rates that seem extreme relative to the other countries? Explain.



7. A histogram of the number of contaminants identified in twenty inspections of a municipalities' water supply is shown below.
    a. Estimate the median number of contaminants for this sample?



    b. What is the shape of the distribution?

    c. Would you expect the mean to be greater than or less than the median? Explain.

    d. Estimate Q_1, Q_3, and IQR for the distribution.

    e. If you were to draw a modified boxplot then would this boxplot have any observations labeled outside the whiskers as outliers? Explain.