**Chapter 1 – Exploring Data**

**Introduction (pp. 2-7)**

Hyena Lab

**Statistics** is the science of data. It is the practice or science of collecting and analyzing numerical data in large quantities, esp. for the purpose of inferring proportions in a whole from those in a representative sample.

* Collect information (data - sample)
* Analyze the information (compute statistics, make plots, etc)
* Make conclusions (infer characteristics of a population based upon a sample)

Statistics is “**customer driven**” – always a question to be answered.

Any set of data contains info on **individuals**. The characteristics of individuals are referred to as **variables**.

* **Individuals** are the objects described by a set of data. People, animals, things.
* **Variables** are characteristics of an individual. Can take on different values for different individuals.

**Example** –

Whenever you receive data, ask:

* **Who** are the individuals described by the data? How many are there?
* **What** are the variables? What units are involved?
* We will eventually extend the questioning to **Why, when, where,** and **how** were the data produced?

**Types of Variables**

* **Categorical** –
* **Quantitative** –

**Example** – Table on page 3.

* **Who?** 10 Canadian students who took the survey.
* **What** variables?
  + **Province**
  + **Gender**
  + **Dominant hand**
  + **Height**
  + **Wrist circum**
  + **Preferred communication**
  + **Travel time to school**
* **Highlighted row?**

When examining data sets we are going to be concerned about the **distribution** of the variables in the data set.

**Distribution** – tells us what values the variable takes on and how often it does so.

In Statistics we are going to be interested in drawing conclusions that go beyond the data at hand. This is called **inference** – the 3rd step in Statistics.

**Homework:**

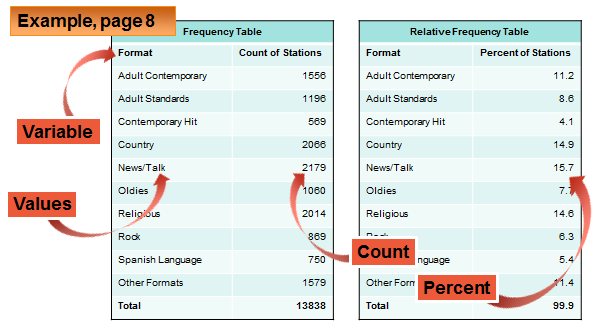
**Section 1.1 – Analyzing Categorical Data (pp. 7-24)**

**Review**

* Definition of Individual and Variable
* Types of Variables
* Statistics: Collect data, analyze it, make inferences

**Distributions/Frequency Tables/Relative Frequency Tables**

The **distribution** of the values of a **categorical** variable lists the count or percent of the individuals that fall into each category.

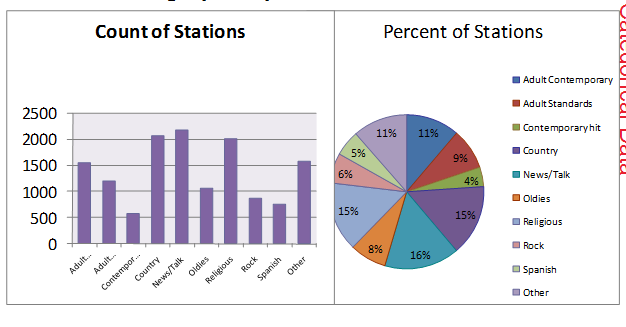


* Discuss individual data points.
* Discuss how to build relative frequency table from frequency table.
* Discuss rounding errors

**Bar Graphs and Pie Charts**

A picture is worth a thousand words…….

(Page 9)



* Discuss tables
* Pie charts must contain all of the categories that make up the whole
* Bar charts are easier to make and are also more flexible than pie charts – a bar chart can display any set of quantities that are measured in the same units (do not have to add to 100%)

**Graphs: Good and Bad**

* Bars should be the same width
* Bars should not be pictographs
* Y-axis should start at 0 and not be compressed.

Examples on page 11.

**Teams** – Do problem 16 on pp 21-22

16. The audience for movies – Here are data on the percent of people in several age groups who attended a movie in the past 12 months:

Age Group Movie Attendance

18-24 83%  
 25-34 73%  
 35-44 68%  
 45-54 60%  
 55-64 47%  
 65-74 32%  
 75 and up 20%

(a) Display these data in a bar graph. Describe what you see.

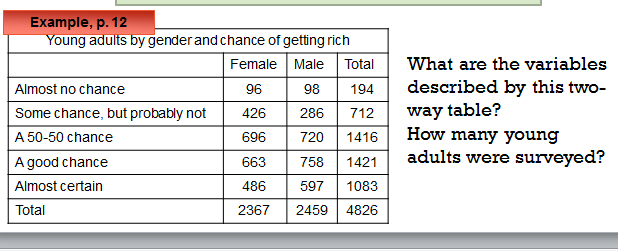
(b) Would it be correct to make a pie chart of these data? Why or why not?

(c) A movie studio wants to know what percent of the total audience for movies is 18-24 year olds. Explain why these data do not answer this question.

**Two-Way Tables**

A two-way table is a table that describes two categorical variables. They have a *row variable* and a *column variable*.

Example on page 12

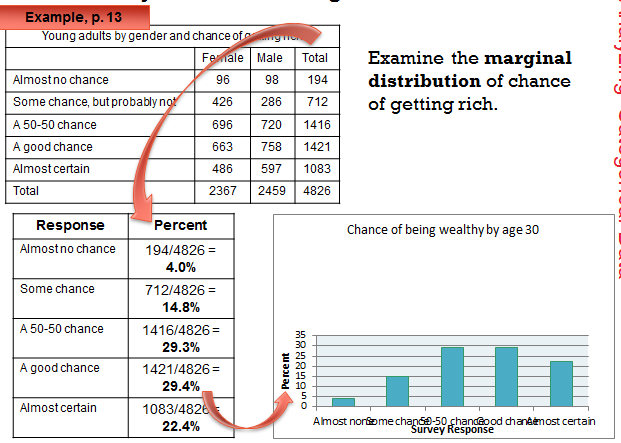


**Marginal Distributions**

In order to grasp how the variables compare we will compute a *marginal distribution*. The **marginal distribution** of one of the categorical variables in a two-way table of counts is the distribution of values of that variable *among all individuals described by the table*. It will be in the form of percents.

Percents are better than counts to make comparisons especially when comparing groups of different sizes.

**Example**: Steps: (1) Use the data in the table to calculate the marginal distribution; (2) make a graph of the marginal distribution.



**Teams** – Check your understanding p. 14

**Relationships between categorical variables: Conditional Distributions**

Marginal dist’s do not tell us anything about the relationship between two variables. To do this we must calculate some well-chosen percents.

Look at females alone in the table. Now we are only looking at 2367 individuals.

|  |  |
| --- | --- |
| p. 14 | This gives us the **conditional distribution for females**.  A **conditional distribution** of a variable describes the values of that variable among individuals who have a specific value of another variable. There is a separate conditional dist for each value of the other variable. |
| **Example**. Conditional dist for men:p. 15 |  |

**Organizing a Statistical Problem** – 4 Step Process

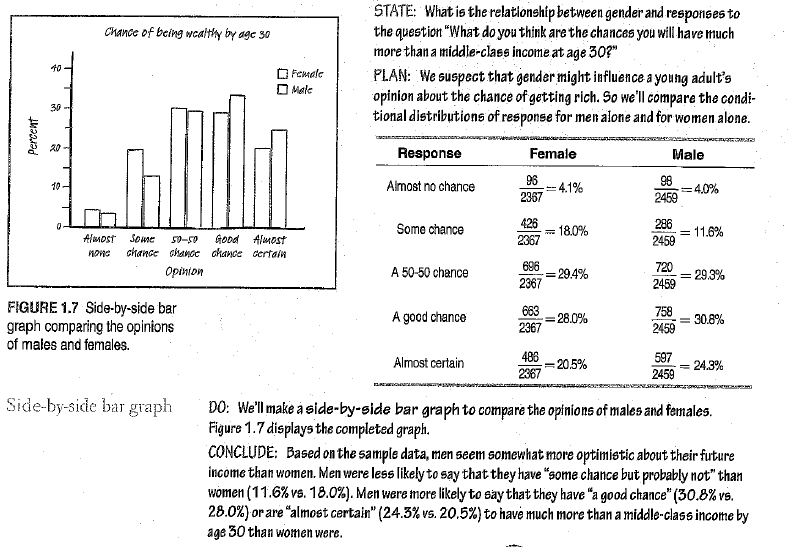
1. **State**: What is the question that you are trying to answer?

2. **Plan**: How will you go about answering the question? What statistical techniques does the problem call for? Have you met the conditions and assumptions necessary to use those techniques?

3. **Do**: Make graphs and carry out the calculations

4. **Conclude**: Give your **practical** conclusion in the **context** of the real-world problem.

**Example p. 17 – Can we conclude that young men and young women differ in their opinions about the likelihood of future wealth? Give appropriate evidence to support your answer.**



**Association** – We say there is **association** between two variables if specific values of one variable tend to occur in common with specific values of the other.

Caution: Just because an association exists does not mean one variable *causes* another variable to act in a certain way. Also, there may be other variables *lurking in the background*.

**Homework: 11-25 odd, 27-34**

**Section 1.2 – Displaying Quantitative Data with Graphs (pp. 25-48)**

**Review Categorical vs. Quantitative variables**

**Dotplots** (Small data sets)

Example: Number of turnovers for 2009 Oakland Raiders during 16 regular-season NFL games:

3, 0, 3, 3, 3, 2, 4, 1, 2, 3, 1, 0, 1, 2, 3, 2

(1) Draw horizontal axis and label variable name; (2) Scale the axis; (3) Make a dot above location for each individual.

**How to Examine the Distribution of a Quantitative Variable** (SOCS – AP EXAM!!)

In any graph, look for the **overall pattern** and for striking **departures** from the pattern.

**S**hape: Concentrate on main features. Look for **clusters** and obvious **gaps**. Look for **potential** **outliers**. Look for **rough** **symmetry** or **clear** **skewness**. Look for number of **modes** or **peaks**. (Discuss **unimodal, bimodal, multimodal** – examples on pp 30-31)

* A distribution is roughly **symmetric** if the right and left sides of the graph are approximately the same
* A distribution is **skewed to the right** if the right side of the graph is much longer than the left side. It is **skewed to the left** if the left side of the graph is much longer than the right side. (The direction of the long tail gives the direction of skewness.)

**C**enter: Find a value that divides the observations in half. We will use the **mean** and **median** to do this.

**S**pread: The spread of the distribution tells us how much **variability** there is in the data. Describe smallest and largest values. Also use the range (Max - Min). (Other measures later.)

**O**utliers: What values differ markedly from the bulk of the data? (Rules later)

**Teamwork**:   
1. Describe the distribution of the Raiders’ Turnovers.

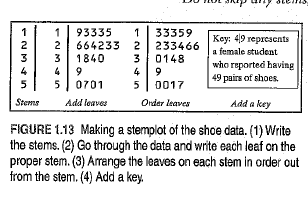
2. p. 29

|  |  |
| --- | --- |
|  |  |

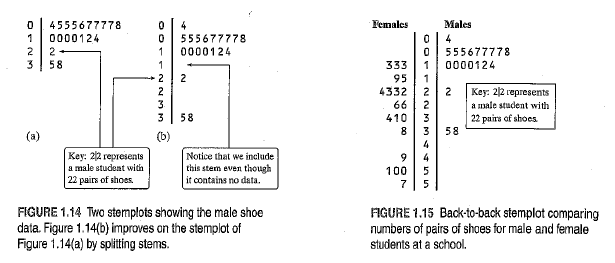
**Comparing Distributions (Very important – example on p. 30)**

|  |  |
| --- | --- |
|  | Include explicit comparison words, eg, the center of \_\_\_\_ is greater than the center of \_\_\_\_.  A very common mistake on the AP Exam is describing the characteristics of the distributions separately w/o making these explicit comparisons. |

**Stemplots** – Another simple method for displaying fairly small data sets. (p. 31)



Discuss **Splitting Stems** (p. 32) Discuss **back-to-back stemplots** (p. 32)



(Common mistake on AP Exam: forget the KEY and the Labels.)

**Description back-to-back above (CYU #1 p. 32)**: In general, it appears that females have more pairs of shoes than males. The median value for the males was 9 pairs while the female median was 26 pairs. The females also have a larger range of 57-13=44 in comparison to the range of 38-4=34 for the males. Finally, both the males and females have distributions that are skewed right, although the distribution of the males is more heavily skewed, as evidenced by the three likely outliers at 22, 35 and 38. The females do not have any likely outliers.

|  |  |
| --- | --- |
|  | 2. B  3. B  4. B  CYU page 35 |

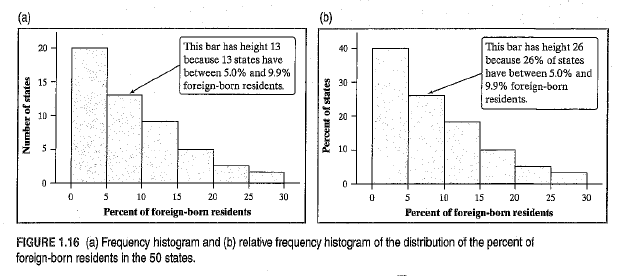
**Histograms** – Better for large data sets. Groups data into classes of equal width. Sometimes distribution is clearer if nearby values are grouped together.

(1) Divide the range of the data into classes of equal width.

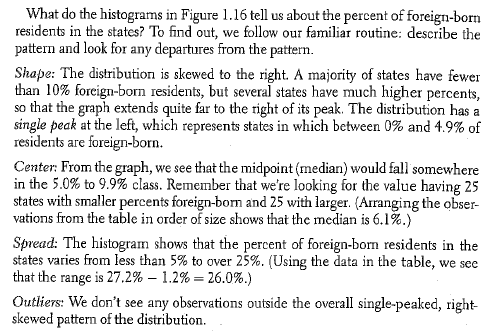
* Often times, it is best to make a dotplot first to decide how wide to make the classes.

(2) Find the count (frequency) or percent (relative frequency) of individuals in each class.

(3) Label axes and scale and draw histogram



Discuss *frequency histogram* versus *relative frequency histogram*. Relative frequency histogram are typically more useful because they make it easier to compare two distributions especially when the number of individuals is very different.

P. 35  


**Discuss Calculator graphing** – p 17, NTA

**Using Histograms Wisely**

* Do not confuse histograms and bar graphs
  + Histograms are for quantitative variables
  + Bar graphs are for categorical variables
  + Histograms have no space between bars; bar graphs have a blank space between bars
* Do not use counts or percents as data.
* Use percents instead of counts when comparing distributions with different numbers of observations
* Just because a graph looks nice, it is not necessarily a meaningful display of data.

Homework: 37-45 odd; 52, 54, 59, 69-74

**Section 1.3 – Describing Quantitative Data with Numbers (pp. 48-73)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sample** | **Males (B)** | **Sample** | **Males (B)** |
| 1 | ***2*** | 11 | ***5*** |
| 2 | ***2*** | 12 | ***2*** |
| 3 | ***1*** | 13 | ***3*** |
| 4 | ***2*** | 14 | ***1*** |
| 5 | ***4*** | 15 | ***8*** |
| 6 | ***4*** | 16 | ***3*** |
| 7 | ***2*** | 17 | ***1*** |
| 8 | ***2*** | 18 | ***2*** |
| 9 | ***1*** | 8 | ***2*** |
| 10 | ***2*** | 20 | ***5*** |

**Data Set**: 20 simple random samples of size 10 to determine proportion of male hyenas in Croatan NF pack.

**Measuring Center: The Mean**

Symbols/Formula:

Procedure:

Example:

Meaning:

Resistance:

**Measuring Center: The Median**

Procedure:

Example:

**2 2 1 1 4 4 2 2 1 2 5 2 3 1 8 3 1 2 2 5**

Meaning:

Resistance:

**Comparing the Mean and Median**

* The mean and median of a *roughly symmetric* distribution will be close together.
* If the distribution is *exactly symmetric*, the mean and median will be exactly the same.
* In a *skewed* distribution, the mean is usually farther out in the long tail than the median.

Example:

**Measuring Spread: The Interquartile Range (IQR)**

The *first quartile Q1:*

The *second quartile:*

The *third quartile:*

Procedure: (1) Arrange data in order; (2) Q1 is the median of the values left of the median; (3) Q3 is the median of the values right of the median; (4) IQR = Q3 – Q1.

Example: **1 1 1 1 1 2 2 2 2 2 2 2 2 3 3 4 4 5 5 8**

Resistance:

**Identifying Outliers** – An observation that falls more than 1.5 x IQR above Q3 or below Q1 is considered an *outlier*.

Example:

**Five-Number Summary & Boxplots**

The *5-Number Summary* consists of:

Example:

These numbers roughly divide the distribution into *quarters.*

A **boxplot** graphically depicts the 5-number summary.

Procedure: (1) Draw and label a horizontal axis; (2) Draw a box from Q1 to Q3; (3) Mark the median  
 in the box with a vertical line segment; (4) Draw line segments (whiskers) from box to  
 minimum values (consider outliers).

Example:

**Calculator Procedures** - NTA p. 16

**Team Work:** Complete Check Your Understanding on p. 59.

**Measuring Spread: The (vaunted) Standard Deviation**

Symbols/Formula:

The *standard deviation* sx measures the *average* distance of the   
observations from their mean.

Procedure: The *standard deviation* is calculated by finding the  
 average of the squared distances and then taking  
 the square root. The average *squared difference*   
 is called the ***variance***.

Example: These are the foot lengths (in cm) for a random sample of seven 14-year-olds from the United Kingdom: 25 22 20 25 24 24 28

The mean foot length is 24 cm.

|  |  |  |
| --- | --- | --- |
| **x** | **xi -** |  |
| 25 |  |  |
| 22 |  |  |
| 20 |  |  |
| 25 |  |  |
| 24 |  |  |
| 24 |  |  |
| 28 |  |  |

**Properties of the Standard Deviation**

**Check your understanding, p. 63.**

**\*\*\*\*\*Choosing Measures of Center and Spread\*\*\*\*\***

* Skewed Distributions:
* Distributions with strong outliers:
* Reasonably symmetric distributions:

**\*\*\*\*\*Resistance\*\*\*\*\***

* Median:
* IQR:
* IQR:

**Analyzing Data Sets**

From this point on, whenever you are analyzing data sets, in the “Do” step you should:

* Plot the distribution
* Create a numerical summary which includes:
  + Mean
  + Standard deviation
  + 5-Number Summary (min, Q1, median, Q3, max)

**Homework:** 79-91 odd, 97, 103, 105, 107-110