Summaries of Categorical Variables

- A frequency distribution is a list of the observed categories and a count of the number of observations in each.
- A frequency distribution may be displayed with a table or a bar chart.
- For ordinal categorical random variables, it is conventional to order the categories in the display (table or bar chart) in the meaningful order.
- For non-ordinal variables, two conventional choices are alphabetical and by size of the counts.
- The vertical axis of a bar chart may show frequency or relative frequency.
- It is conventional to leave space between bars of a bar chart of a categorical variable.

Summary of Blood Type Data

Here is a frequency table.

```r
> summary(BloodType)
A   AB  B  I dont know  O
 27  10   9     34     16
```

Here is a bar chart.

![Bar chart of Blood Type Data]

Summary of Majors

```
[,1]
Animal Science      5
Bacteriology        3
Biochemistry         4
Biological Aspects of Conservation  2
Biology             24
Biomedical Engineering 4
Botany               1
Business-accounting  1
Chemical Engineering  1
Dairy Science        1
Genetics            30
Kinesiology          2
Life Science Communication 1
Medical Microbiology and Immunology 2
Molecular Biology    1
Nutritional Science  1
Pharmacology/Toxicology 1
Psychology          2
Undecided           2
Wildlife Ecology     2
Wildlife Ecology - Natural Resources 1
Zoology             5
```

Summary of Second Majors

```
[,1]
Bacteriology        2
Biochemistry         2
Biological Aspects of Conservation  3
Biology             1
Chemistry           1
Classics in Humanities 1
French              2
Genetics            2
German              1
I might add or switch to Biology Education 1
Kinesiology         1
Latin American Studies 1
Life Science Communication 1
Molecular Biology   1
Music               68
Philosophy          1
Psychology          1
Spanish             4
Women Studies Certificate 1
Zoology             2
```

Exploratory Data Analysis

- Exploratory Data Analysis involves both graphical displays of data and numerical summaries of data.
- A data set is often represented as a matrix.
- There is a row for each unit.
- There is a column for each variable.
- A unit is an object that can be measured, such as a person, or a thing.
- A variable is a characteristic of a unit that can be assigned a number or a category.
- For the survey data, each respondent is a unit.
- Variables include sex, major, year in school, miles from home, height, and blood type.

Variables

- Variables are either quantitative or categorical.
- In a categorical variable, measurements are categories.
- Examples include blood type, sex.
- The variable year in school is an example of an ordinal categorical variable, because the levels are ordered.
- Quantitative variables record a number for each unit.
- Examples include height, which is continuous and number of sisters, which is discrete.
- Often, continuous variables are rounded to a discrete set of values (such as heights to the nearest inch or half inch).
- We can also make a categorical variable from a continuous variable by dividing the range of the variable into classes (So, for example, height could be categorized as short, average, or tall).
- Identifying the types of variables can be important because some methods of statistical analysis are appropriate only for a specific type of variable.

Samples

- A sample is a collection of units on which we have measured one or more variables.
- The number of observations in a sample is n.
- Common notation for the sample size is n.
- The textbook adopts the convention of using uppercase letters for variables and lower case letters for observed values.
**Summaries of Quantitative Variables**

- Quantitative variables from very small samples can be displayed with a dotplot.
- Histograms are a more general tool for displaying the distribution of quantitative variables.
- A histogram is a bar graph of counts of observations in each class, but no space is drawn between classes.
- If classes are of different widths, the bars should be drawn so that areas are proportional to frequencies.
- Selection of classes is arbitrary. Different choices can lead to different pictures.
- Too few classes is an over-summary of the data.
- Too many classes can cloud important features of the data with noise.

**Summary of Miles from Home for Students within 250 miles**

**Summary of Height**

**A Dotplot of Hours of Sleep**

**Summary of Miles from MSC**

**Stem-and-Leaf Diagrams**

- Stem-and-Leaf diagrams are useful for showing the shape of the distribution of small data sets without losing any (or much) information.
- Begin by rounding all data to the same precision.
- The last digit is the leaf.
- Anything before the last digit is the stem.
- In a stem-and-leaf diagram, each observation is represented by a single digit to the right of a line.
- Stems are shown only once.
- Show stems to fill gaps!
- Combining or splitting stems can lead to a better picture of the distribution.

**Summary of Miles from Home for Students within 250 miles**

**Summary of Height**

**A Dotplot of Hours of Sleep**

**Summary of Miles from MSC**

**Stem-and-Leaf Diagram of Brothers and Sisters**

The decimal point is at the !
0 | 00000000000000000000000000000000000000000000000000000000000000000000
1 | 00000000000000000000000000000000000000000000000000000000000000000000
2 | 00000000000000000000000000000000000000000000000000000000000000000000
3 | 00000000000000000000000000000000000000000000000000000000000000000000
4 | 0
The decimal point is at the !
0 | 00000000000000000000000000000000000000000000000000000000000000000000
2 | 00000000000000000000000000000000000000000000000000000000000000000000
4 | 0
Quantiles

- Quantiles are a generic name for positions in the distribution of a quantitative variable.
- For example, percentiles divide a distribution into 100 equal sized parts.
- Quartiles, which divide a distribution into four equal sized parts, are a common statistical tool.
- The first quartile, \( Q_1 \), is the location that separates the smallest quarter of the data from the rest. This is also known as the 25th percentile.
- The third quartile, \( Q_3 \), is the location that separates the top quarter of the data and is also known as the 75th percentile.
- The median is the second quartile.
- Different authors and statistical software packages have different definitions of quantiles.
- The definition I prefer is that a value \( x \) is a \( p \)-quantile of a sample if the proportion of observations less than or equal to \( x \) is at least \( p \) and if the proportion of observations greater than or equal to \( x \) is at least \( 1 - p \).

5 Number Summary

- The minimum, first quartile, median, third quartile, and maximum are called the five-number summary of a quantitative variable.
- The interquartile range (IQR) is the difference between the third and first quartiles.
  \[ \text{IQR} = Q_3 - Q_1 \]
- The range is the difference between the maximum and the minimum.
- Graphical displays of five-number summaries are called boxplots.

Example:

```r
> fivenum(MilesClass)
(1) 0.20 1.00 1.00 1.25 15.00
```

Skewness

- Histograms show several qualitative features of a quantitative variable, such as the number of modes and skewness.
- A distribution is approximately symmetric if the left and right halves are approximately mirror images of each other.
- A distribution is skewed to the right if the right half of the data (the larger values) are more spread out than the left half of the data.
- A distribution is skewed to the left if the left half of the data (the smaller values) are more spread out than the right half of the data.
- It is fairly common for biological data to be skewed to the right. Often times there is a barrier below which there can be no values, but no upper limit.

Measures of Center

- There are two common measures of center for quantitative variables, the mean and the median.
- For sample data \( y_1, y_2, \ldots, y_n \), the sample mean is
  \[ \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \]
- The sample mean is the balancing point of the variable.
- The sample median is the middle value — at least half the values are larger and at least half are smaller.
- If a sample size is odd, only one number will be the median.
- If a sample size is even, any number between the middle two numbers could be the median. By convention, we take the average of the middle two numbers.
- Imagine a histogram made of a uniform solid material. The mean is about the point at which the solid would balance. The median is about at a point that would divide the area of the histogram exactly in half.

Boxplot of Hours of Sleep

```
```

Comparing the mean and the median

- The mean and median of a symmetric distribution are the same.
- The median is more resistant to outliers than the mean. For example, the mean and median of the numbers 1, 2, 3 are 2, but for the data set 1, 2, 30, the median is still 2, but the mean is 11, far away from each observation.
- The median can be a better measure of a ‘typical value’ than the mean especially for strongly skewed variables.
- If a variable is skewed to the right, the mean will typically be larger than the median.
- The opposite is true if the variable is skewed to the left.

Example:

```r
> mean(MilesHome)
(1) 616.7901
> median(MilesHome)
(1) 132.5
```
The Empirical Rule

For many variables (especially those that are nearly symmetric and bell-shaped), the following empirical rule is often a very good approximation.

- About 68% of the observations are within 1 SD of the mean.
- About 95% of the observations are within 2 SDs of the mean.
- Nearly all observations are within 3 SDs of the mean.

Example:

```r
> sleep = Sleep[!is.na(Sleep)]
> m = mean(sleep)
> s = sd(sleep)
> c(m, s)
> sum(abs(sleep - m) < s)/length(sleep)
[1] 0.6875
> sum(abs(sleep - m) < 2*s)/length(sleep)
[1] 0.9375
> sum(abs(sleep - m) < 3*s)/length(sleep)
[1] 1
```

Samples and Populations

- The previous techniques are useful for describing a data set, or a sample of data.
- It is often of interest to generalize findings from a sample to a larger group that statisticians call a population.
- This generalization is called statistical inference.
- Statistical inference is often concerned with using statistics, characteristics that can be calculated from sample data, to estimate parameters, characteristics of populations.
- Examples:
  - \( p \) = population proportion, \( \hat{p} \) = sample proportion
  - \( \mu \) = population mean, \( \bar{y} \) = sample mean
  - \( \sigma \) = population standard deviation, \( s \) = sample standard deviation

Boxplot of Miles from Class

```r
> boxplot(MilesClass)
```

Side-by-side boxplot of height versus sex

```r
> boxplot(Height ~ Sex)
```

Measures of Dispersion

- The standard deviation or SD is the most common statistical measure of dispersion.
- A deviation from the mean is the signed distance of an observation from the mean.
  \[
  \text{deviation} = \text{value of observation} - \bar{y}
  \]
  Observations greater than the mean have positive deviations while those less than the mean have negative deviations.
- The standard deviation is a special type of average deviation
  \[
  s = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}
  \]
  This is almost the square root of the mean squared deviation from the mean.

Measures of Dispersion

- Statisticians use \( n - 1 \) instead of \( n \) in the denominator for a technical mathematical reason of historical, if not practical, importance.
- The standard deviation can often be interpreted as the size of a typical deviation from the mean.