**CHAPTER 2** and 3

# Displaying and Describing Categorical Data

## Five W's and one H

To provide context to data, we need the five W's and the one H:

- Who
  - tells use the individual about which (or whom) we have collected the data
  - **Respondents** - individuals who answer a survey
  - **Subjects/participants** - people on whom we experiment
  - **Experimental units** - animals, plants, and inanimate objects
- What (in what units)

    - **Variables** are characteristics that are recorded about each individual
        - should have a name to identify what data is being measured/studied
- When
- Where
- Why (if possible)
- How
    - how data is collected is important to make meaningful conclusions

# Summarizing and Displaying a Single Categorical Variable

## Three Rules of Data Analysis

1. **Make a picture** - things may be revealed that are not obvious in raw data.
2. **Make a picture** - well designed display will show the important features and patterns of the data
3. **Make a picture** - best way to tell others about your data is with a well chosen picture

## The Area Principle

> The area occupied by a part of the graph should correspond to the magnitude of the value that it represents.

## Frequency Tables

To display a categorical variable, we need to organize the number of cases associated with each category.

A **frequency table** records the totals for each category names

> Example
>
> the number of people in an airplane:
>
> | Class | Count |
> | --- | --- |
> | first | 325 |
> | second | 285 |
> | third | 706 |
> | crew | 885 |

A **relative frequency** displays the proportions or percentages, rather than the counts of the values in each category

> Example from above:
>
> | Class | % |
> |---|---|
> | first | 14.77 |
> | second | 12.95 |
> | third | 32.08 |
> | crew | 40.21 |

## Bar Charts

- stays true to the area principle
- displays the distribution of a categorical variable
    - showing counts for each category
- **Relative frequency bar chart** is used to draw attention to the relative proportions of the categories

## Pie Charts

- categories need to be **mutually exclusive**

# Exploring the Relationship Between Two Categorical Variables

A **contingency table** looks at two categorical variables together:

- how are the cases distributed along each variable, contingent on the  value of the other variable?

> Example questions:
>
> - is supporting green infrastructure investment contingent on a person's education level?
> - is scholarly achievement in ENVS 178 contingent on studying?

> Example Contingency Table

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

ie: the second cell in the crew column tells us that 673 crew members died when the Titanic sunk

- each cell of the table gives the count for a combination of the two values
- each frequency distribution is called a **marginal distribution** of its respective variable

## Conditional Distributions

- a conditional distribution shows the distribution of one variable for just the individuals who satisfy some condition on another variable

Example: Titanic sinking

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

ie: 28.6% of the total alive people were from the first class

## Segmented Bar Charts

- treats each bar as the "whole" and divides it proportionally into segments corresponding to the percentage in each group
- pretty much the same shit as a pie chart

```
<div style="page-break-after: always;"></div>
```

# Displaying and Summarizing Quantitative Data

# Building Histograms

- easy-to-understand summary of the distribution of a quantitative variable
    - but don't show the values themselves

## 1. Histogram Components

1. Decide on number of classes; consider lowest and highest value
2. Determine class width - difference between two lower class limits
3. Define lower class limits (CLlower) - the smallest value in a class
4. Define upper class limits (CLupper) - the largest value in a class
5. Define class midpoints - (CLlower + CLupper)/2
6. Determine the class boundaries - midpoint between two classes, as well as the upper and lower boundaries of range

## 2. Frequency Distribution Table

1. List classes
2. List the frequency values (#)
3. Calculate relative frequency (%)
4. Calculate cumulative frequency (#)

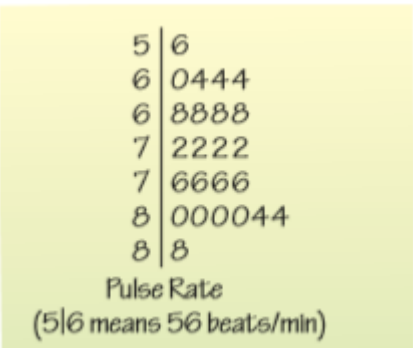
## 3. Plot Data

- Plot data with classes on x-axis frequency (or relative frequency) on y-axis

# Stem-and-Leaf Displays

- like a histogram but it shows individual values
- easier to make by hand

> Example: pulse rates of 24 women



## Constructing a Stem-and-Leaf Display

- cut each data value into leading digits ("stems") and trailing digits ("leaves")
- use stems to label the bins
- use only one digit for each leaf
    - either round or truncate the data values to one decimal place after the stem

> Example: distribution of scores for sample class
>
> n = 23
>
> 54, 64.5, 65.5, 68, 69, 70, 73, 73.5, 75, 77, 78, 80, 80, 80, 83, 83, 84, 84.5, 86, 86.5, 88.5, 91.5, 93

| leaf | stem |
|---|---|
| 5 | 4 |
| 6 | 4.5 5.5 8 9 |
| 7 | 0 3 3.5 5 7 8 |
| 8 | 0 0 0 3 3 4 4.5 6 6.5 8.5 |
| 9 | 1.5 3 |

# Shape, Center, and Spread

- when describing a distribution, make sure to always talk about shape, center, and spread

## Shape

- symmetry
    - how symmetrical is the data?
    - thinner ends of a distribution are called **tails**
- skew
    - if one tail stretches out farther than the other, the histogram is said to be **skewed** to the side of the longer tail
- peaks
- outliers
    - points that stand away from the body of the distribution

## Center

- the middle value that divides the histogram into two equal areas, is called the **median**

# Spread

## Range

> *range = max - min*

- the difference between the max and min values
- the range is a single number, NOT an interval of values
- this data does not really represent the data overall

## The Interquartile Range

- it may be better to describe the spread of a variable by ignoring the extremes and concentrating on the middle of that data
- One quarter of the data lies below the **lower quartile** and one quarter of the data lies above the **upper quartile** , HENCE, half the data lies between them
- **Percentile** - the value that leaves that percentage of the data below it
    - 25% of the data falls below the lower quartile (25th percentile)
    - 75% of the data falls below the upper quartile (75th percentile)
    - the median is the 50th percentile
- the difference between the quartiles tells us how much space the middle half of the data covers and is called the **interquartile range**

> *IQR = upper quartile - lower quartile*

# The Five - Number Summary

- the five-number summary of a distribution refers to its:
    - max
    - Q3
    - median
    - Q1
    - min

## Boxplots

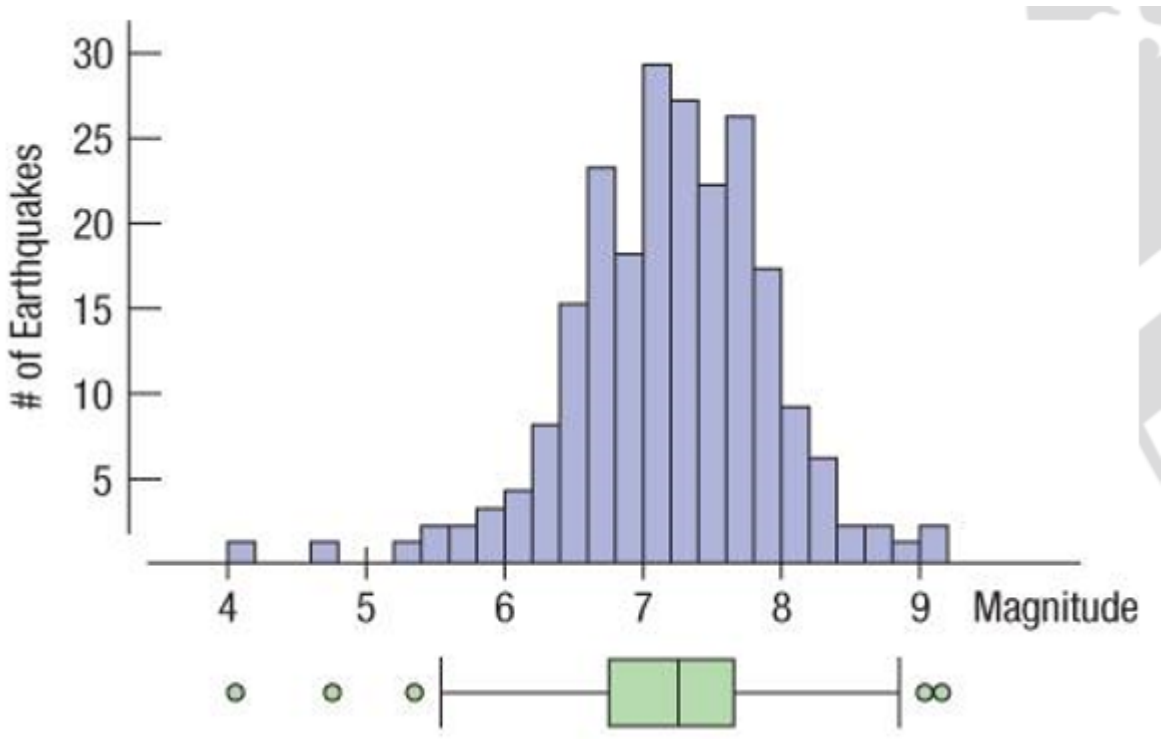
- graphical display of the five-number summary

#### Constructing Bloxplots

1. Draw a single vertical axis, spanning the range of the data. Draw short horizontal lines at the lower and upper quartiles and at the median. Then connect them with vertical lines to form a box.
2. Erect "fences" around the main part of the data

- upper fence is 1.5 IQRs above the upper quartile
- lower fence is 1.5 IQRs below the lower quartile
- fences are only to help constructing the boxplot, should NOT appear in the final display

3. Use fences to grow "whiskers"
   - draw lines from the ends of the box up and down the most extreme data values found within the fences
   - if a data value falls outside one of the fences, we do NOT connect it with a whisker
4. Add the outliers by displaying the data values beyond the fences with special symbols

> Example five-number summary to boxplot
>
> **Max** - 9.1
>
> **Q3** - 7.6
>
> **Median** - 7.2
>
> **Q1** - 6.7
>
> **Min** - 4.0
>
> Calculated IQR = 7.6 - 6.7 = 0.9
>
> Calculated upper fence = upper quartile + 1.5(IQR) = 7.6 + 1.5(0.9) = 8.95
>
> Calculated lower fence = lower quartile - 1.5(IQR) = 6.7 - 1.5(0.9) = 5.35
>
> 
> 

## Special Case: Symmetric Distributions

## The Center of Symmetric Distributions - The Mean

- when a distribution is symmetric, the mean and median are **approximately the same in value**, so either measure of the center may be used
- the the distribution is skewed or has outliers, its better to report the median

## The Spread of Symmetric Distributions: The Standard Deviation

- like the mean, the SD is appropriate only for symmetric data
- how far is each data value from the mean?
    - the difference is called a *deviation*

### Variance

- adding up squared deviations and finding their average:

$$s^2 = \frac{\sum(y - \bar{y})^2}{n - 1}$$

### Standard Deviation

- SD is the square root of variance

$$s = \sqrt{\frac{\sum(y - \bar{y})^2}{n - 1}}$$