About the Tutorials

• Every Monday (10:30am – 12noon)

• Purpose: discussing statistics issue related to research methods.
  – Approximately one chapter per tutorial.
  – Teaching + Assignments + in-class problem discussion.

• In-class quizzes (TBD).

• Website: http://web.cs.dal.ca/~anwar/ds/ds.html.

• Book (Readings Statistics and Research, Schuyler W. Huck, 5th / 6th Edition), mandatory.

• Book Website: http://www.readingstats.com/fifth/index.htm

• Book Quizzes: Should be done prior to coming to class for the applicable chapter(s).
Typical Format of a Journal Article

- **Abstract**
- **Introduction**
  - Summary of background
  - Research problem
  - Research question
  - Research hypothesis
  - Remainder ...
- **Related Work** (aka Research Background/Research Rationale, Background, ... etc.).
- **Method**
  - Who participated? What instruments? What was done by participants?
- **Materials** (aka: equipment, apparatus, instruments, or scales)
  - Things other participants used in study.
- **Procedure** (How the study was conducted).
- **Results** (text, tables, figures)
- **Discussion** (What the results mean with respect to question/hypothesis/statements of purpose)
- **Conclusion**
- **References**
Statistics and Research Design

• Statistics: Theory and method of analyzing quantitative data from samples of observations ... to help make decisions about hypothesized relations.
  – Tools used in research design

• Research Design: Plan and structure of the investigation so as to answer the research questions (or hypotheses)
Statistics and Research Design, cont’d

• Analogy:
  
  – Research design is the blueprint of the study.
  
  – In quantitative designs, statistical design and procedures are the craft and tools used to conduct quantitative studies.
  
  – The logic of hypothesis testing is the decision-making process that links statistical design to research design.
Statistics

• There are two types of statistics

  – **Descriptive Statistics**: involve tabulating, depicting, and describing data
    • Summarizes a set of data
  – **Inferential Statistics**: predicts or estimates characteristics of a population from a knowledge of the characteristics of only a sample of the population
    • Help us draw conclusions about a population
In a web search task, the sites users viewed and the search queries they submitted were collected.

<table>
<thead>
<tr>
<th>Participants</th>
<th>Sites viewed</th>
<th>Queries Submitted</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>56</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>39</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>34</td>
<td>13</td>
</tr>
<tr>
<td>7</td>
<td>33</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>32</td>
<td>14</td>
</tr>
<tr>
<td>9</td>
<td>28</td>
<td>18</td>
</tr>
<tr>
<td>4</td>
<td>23</td>
<td>20</td>
</tr>
<tr>
<td>1</td>
<td>22</td>
<td>19</td>
</tr>
<tr>
<td>8</td>
<td>17</td>
<td>28</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>34</td>
</tr>
</tbody>
</table>

Spearman Correlation $r = -0.944$, $p < 0.0001$
In statistics the term **population** includes all members of a defined group that we are studying or collecting information on for data driven decisions.

A *part* of the population is called a **sample**.

A **sample** is a proportion of the population, a slice of it, a part of it and all its characteristics.

A sample is a scientifically drawn group that actually *possesses the same characteristics as* the population – **if it is drawn randomly**. (This may be hard for you to believe, but it is true!)

**Randomly** drawn samples must have *two characteristics*:

- Every person has an equal opportunity to be selected for the sample; and,
- selection of one person is independent of the selection of another person.

Chapter 2

Describing, Exploring and Comparing Data
Outline

2-2
- 2-1 Introduction
- 2-2 Frequency Distributions
- 2-3 Visualizing Data
Objectives

2-3

• Organize data using frequency distributions.
• Represent data in frequency distributions graphically using histograms, frequency polygons and ogives.
2-2 Frequency Distributions

2-5

• When data are collected in original form, they are called **raw data**.

• When the raw data is organized into a **frequency distribution**, the frequency will be the number of values in a specific class of the distribution.
2-2 Frequency Distributions

2-6

A *frequency distribution* is the organizing of raw data in table form, using classes and frequencies.

- The following slides show an example of a frequency distribution.
• **Categorical frequency distributions** - can be used for data that can be placed in specific categories, such as nominal- or ordinal-level data.

• **Examples** - political affiliation, religious affiliation, blood type etc.
### 2-2 Blood Type Frequency Distribution - Example

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>B</td>
<td>7</td>
<td>28</td>
</tr>
<tr>
<td>O</td>
<td>9</td>
<td>36</td>
</tr>
<tr>
<td>AB</td>
<td>4</td>
<td>16</td>
</tr>
</tbody>
</table>
2-2 Ungrouped Frequency Distributions

2-9

- **Ungrouped frequency distributions** - can be used for data that can be enumerated and when the range of values in the data set is not large.

- **Examples** - number of miles your instructors have to travel from home to campus, number of girls in a 4-child family etc.
### Number of Miles Traveled - Example

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>24</td>
</tr>
<tr>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>15</td>
<td>10</td>
</tr>
</tbody>
</table>
2-2 Grouped Frequency Distributions

- **Grouped frequency distributions** - can be used when the range of values in the data set is very large. The data must be grouped into classes that are more than one unit in width.

- **Examples** - the life of boat batteries in hours.
### Example

#### Lifetimes of Boat Batteries

<table>
<thead>
<tr>
<th>Class limits</th>
<th>Class Boundaries</th>
<th>Frequency</th>
<th>Cumulative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>24 - 30</td>
<td>23.5 - 37.5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>38 - 51</td>
<td>37.5 - 51.5</td>
<td>14</td>
<td>18</td>
</tr>
<tr>
<td>52 - 65</td>
<td>51.5 - 65.5</td>
<td>7</td>
<td>25</td>
</tr>
</tbody>
</table>
Terms Associated with a Grouped Frequency Distribution

- Class limits represent the smallest and largest data values that can be included in a class.
- In the lifetimes of boat batteries example, the values 24 and 30 of the first class are the class limits.
- The lower class limit is 24 and the upper class limit is 30.
2-2 Terms Associated with a Grouped Frequency Distribution

- The **class boundaries** are used to separate the classes so that there are no gaps in the frequency distribution.
The class width for a class in a frequency distribution is found by subtracting the lower (or upper) class limit of one class minus the lower (or upper) class limit of the previous class.
Guidelines for Constructing a Frequency Distribution

• There should be between 5 and 20 classes.
• The class width should be an odd number.
• The classes must be mutually exclusive.
2-2 Guidelines for Constructing a Frequency Distribution

- The classes must be continuous.
- The classes must be exhaustive.
- The class must be equal in width.
2-2 Procedure for Constructing a Grouped Frequency Distribution

- Find the highest and lowest value.
- Find the range.
- Select the number of classes desired.
- Find the width by dividing the range by the number of classes and rounding up.
2-2 Procedure for Constructing a Grouped Frequency Distribution

- Select a starting point (usually the lowest value); add the width to get the lower limits.
- Find the upper class limits.
- Find the boundaries.
- Tally the data, find the frequencies and find the cumulative frequency.
In a survey of 20 patients who smoked, the following data were obtained. Each value represents the number of cigarettes the patient smoked per day. Construct a frequency distribution using six classes. (The data is given on the next slide.)
2-2 Grouped Frequency Distribution - Example

<table>
<thead>
<tr>
<th>10</th>
<th>8</th>
<th>6</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>13</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td>11</td>
<td>9</td>
<td>18</td>
<td>14</td>
</tr>
<tr>
<td>13</td>
<td>12</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>16</td>
<td>11</td>
</tr>
</tbody>
</table>
2-2 Grouped Frequency Distribution - Example

- **Step 1:** Find the highest and lowest values: \( H = 22 \) and \( L = 5 \).

- **Step 2:** Find the range:
  \[
  R = H - L = 22 - 5 = 17.
  \]

- **Step 3:** Select the number of classes desired. In this case it is equal to 6.
2-2 Grouped Frequency Distribution - Example

• **Step 4:** Find the class width by dividing the range by the number of classes. \( \text{Width} = \frac{17}{6} = 2.83 \). This value is rounded up to 3.
Step 5: Select a starting point for the lowest class limit. For convenience, this value is chosen to be 5, the smallest data value. The lower class limits will be 5, 8, 11, 14, 17 and 20.
2-2  Grouped Frequency Distribution - Example

- **Step 6:** The upper class limits will be 7, 10, 13, 16, 19 and 22. For example, the upper limit for the first class is computed as 8 - 1, etc.
• **Step 7:** Find the class boundaries by subtracting 0.5 from each lower class limit and adding 0.5 to the upper class limit.
• **Step 8:** Tally the data, write the numerical values for the tallies in the frequency column and find the cumulative frequencies.

• The grouped frequency distribution is shown on the next slide.
Who Smoked how many?

Note: The dash “-” represents “to”.

<table>
<thead>
<tr>
<th>Class Limits</th>
<th>Class Boundaries</th>
<th>Frequency</th>
<th>Cumulative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>05 to 07</td>
<td>4.5 - 7.5</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>08 to 10</td>
<td>7.5 - 10.5</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>11 to 13</td>
<td>10.5 - 13.5</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>14 to 16</td>
<td>13.5 - 16.5</td>
<td>5</td>
<td>16</td>
</tr>
<tr>
<td>17 to 19</td>
<td>16.5 - 19.5</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>20 to 22</td>
<td>19.5 - 22.5</td>
<td>1</td>
<td>20</td>
</tr>
</tbody>
</table>
2-2 Stem-and-Leaf Displays

• **Stem-and-Leaf frequency distributions**
  • Grouped Frequencies carry loss of information (hidden into the groups).
  • Stem-and-leaf does group and *keep* the original data.
• It is usually ordered.
Stem-and-leaf Example

• Subjects in a psychological study were timed while completing a certain task. Complete a stem-and-leaf plot for the following list of times:
  • 7.6, 8.1, 9.2, 6.8, 5.9, 6.2, 6.1, 5.8, 7.3, 8.1, 8.8, 7.4, 7.7, 8.2
  • First, I'll reorder this list:
  • 5.8, 5.9, 6.1, 6.2, 6.8, 7.3, 7.4, 7.6, 7.7, 8.1, 8.1, 8.2, 8.8, 9.2

http://www.purplemath.com/modules/stemleaf2.htm
2-3 Visualizing Data

The three most commonly used graphs in research are:

- The histogram.
- The frequency polygon.
- The cumulative frequency graph, or ogive (pronounced o-jive).
The histogram is a graph that displays the data by using vertical bars of various heights to represent the frequencies.
Example of a Histogram

Number of Cigarettes Smoked per Day

Frequency
A **frequency polygon** is a graph that displays the data by using lines that connect points plotted for frequencies at the midpoint of classes. The frequencies represent the heights of the midpoints.
Example of a Frequency Polygon

Frequency Polygon

Number of Cigarettes Smoked per Day

Frequency
A **cumulative frequency graph** or **ogive** is a graph that represents the cumulative frequencies for the classes in a frequency distribution.
Example of an Ogive

<table>
<thead>
<tr>
<th>Class Limits</th>
<th>Class Boundaries</th>
<th>Frequency</th>
<th>Cumulative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>05 to 07</td>
<td>4.5 - 7.5</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>08 to 10</td>
<td>7.5 - 10.5</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>11 to 13</td>
<td>10.5 - 13.5</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>14 to 16</td>
<td>13.5 - 16.5</td>
<td>5</td>
<td>16</td>
</tr>
<tr>
<td>17 to 19</td>
<td>16.5 - 19.5</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>20 to 22</td>
<td>19.5 - 22.5</td>
<td>1</td>
<td>20</td>
</tr>
</tbody>
</table>
• **Time series graph** - A time series graph represents data that occur over a specific period of time.
2-3 Other Types of Graphs - Time Series Graph

PORT AUTHORITY TRANSIT RIDERSHIP

Ridership (in millions)

Year


Yea

Ridersh

(ina mili

o

9

89 87 85 83 81 79 77 75


Ye

Ridersh

(millions)
• **Pie graph** - A pie graph is a circle that is divided into sections or wedges according to the percentage of frequencies in each category of the distribution.
Pie Chart of the Number of Crimes Investigated by Law Enforcement Officers in U.S. National Parks During 1995

- Assaults (164, 68.3%)
- Rape (34, 14.2%)
- Robbery (29, 12.1%)
- Homicide (13, 5.4%)
2-4 Other Types of Graphs - Box and Whisker Plots (Boxplots)

- An exploratory graphic, created by John W. Tukey, used to show the distribution of a dataset (at a glance).

Box and Whisker from: Michelson–Morley experiment
http://en.wikipedia.org/wiki/Box_plot
Reading a Box-and-Whisker Plot

- **Outliers**: Any datum that is more than \((3/2)\) times the upper quartile is ‘usually’ considered an outlier. The same applies for any datum \((3/2)\) times less than the lower quartile.
- **Maximum**: Any data greater than the upper quartile and NOT into outliers.
- **Upper quartile**: is the 25% data greater than the median.
- **Median**: Middle datum (50% of data is greater than the median and 50% is less than the median).
- **Lower quartile**: 25% of the data that is less than the median.
- **Minimum**: Any data less than the lower quartile and NOT into outliers.

Please refer to [http://www.5min.com/Video/How-to-Interpret-Box-and-Whisker-Plots-275636648](http://www.5min.com/Video/How-to-Interpret-Box-and-Whisker-Plots-275636648) for a nice video.
Measures of Central Tendency

• The **mean** is the average value of all the data in the set.

• The **median** is the middle value in a data set that has been arranged in numerical order so that exactly half the data is above the median and half is below it.

• The **mode** is the value that occurs most frequently in the set.

• In a **normal distribution**, mean, median and mode are identical in value.

Credits: (Stats. Canada) → [http://www.statcan.gc.ca/edu/power-pouvoir/ch11/5214867-eng.htm](http://www.statcan.gc.ca/edu/power-pouvoir/ch11/5214867-eng.htm)