2019-2

Introduction to Biostatistics - Lecture 1: Introduction and Descriptive Statistics

Jonggyu Baek
University of Massachusetts Medical School

Follow this and additional works at: https://escholarship.umassmed.edu/liberia_peer

Part of the Biostatistics Commons, Family Medicine Commons, Infectious Disease Commons, Medical Education Commons, and the Public Health Commons

Repository Citation
https://escholarship.umassmed.edu/liberia_peer/10

This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in PEER Liberia Project by an authorized administrator of eScholarship@UMMS. For more information, please contact Lisa.Palmer@umassmed.edu.
Introduction to Biostatistics

2/26/2019

Jonggyu Baek, PhD
Outline

• Purpose

• Introduction to biostatistics

• Descriptive Statistics
Purpose of the course

• Basic principles and applications of statistics to problems in clinical and public health settings.

• Will cover tools for statistical inference: t-test, chi-square tests, ANOVA, Linear regression.

• Interpretation and presentation of the results

• Introductory foundation for the implementation of those techniques using R or R studio software.
References

• Multiple authors, Openstax College

  Introductory Statistics

  Publisher: OpenStx, Pubdate: 2013

  https://open.umn.edu/opentextbooks/textbooks/introductory-statistics-2013

• Quick-R: https://www.statmethods.net/

• UCLA statistical computing: https://stats.idre.ucla.edu/
What is Statistics?

• **Statistics** is the science of learning from data, and of measuring, controlling, and communicating uncertainty; and it thereby provides the navigation essential for controlling the course of scientific and societal advances (Davidian, M. and Louis, T. A., 10.1126/science.1218685).

• **Statistics** is also an ART ...

  of conducting a study, analyzing the data, and derive useful conclusions from numerical outcomes about real life problems...
What is Biostatistics?

- **Biostatistics** is the application of statistics in medical research, e.g.:
  - Clinical trials
  - Epidemiology
  - Pharmacology
  - Medical decision making
  - Comparative Effectiveness Research
  - etc.
Statistical Analysis

Key steps for a complete and accurate statistical analysis:

• State a valid research question
• Collect information (DATA) for answering this question
• Validate/clean/organize the collected information
• Exploratory Data Analysis (EDA)
• Analyze this information
• Translate numerical results into answers
• Interpret results and derive conclusions
• Present the results and communicate with people
Terms in Biostatistics

• **Data**:  
  – all the information we collect to answer the research question

• **Variables**:  
  – Outcome, treatment, study population characteristics

• **Subjects**:  
  – units on which characteristics are measured

• **Observations**:  
  – data elements

• **Population**:  
  – all the subjects of interest

• **Sample**:  
  – a subset of the population for which data are collected

**Sample from Population**

<table>
<thead>
<tr>
<th>Descriptive Measure</th>
<th>Population</th>
<th>Sample</th>
<th>Summary of a characteristic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Size</strong></td>
<td>$N$</td>
<td>$n$</td>
<td>Total # of subjects</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>$\mu$</td>
<td>$\bar{x}$</td>
<td>Average</td>
</tr>
<tr>
<td><strong>Variance</strong></td>
<td>$\sigma^2$</td>
<td>$s^2$</td>
<td>Variance</td>
</tr>
</tbody>
</table>

Impossible/impractical to analyze the entire population →

→ thus we only analyze a sample
Statistical Inference

Collect and analyze data from samples

→ Calculate summary statistics

→ Make Inference about unknown population parameters (e.g., population average from sample mean)
The Framingham Heart Study
https://www.framinghamheartstudy.org/fhs-about/history/epidemiological-background/

• “a long term prospective study of the etiology of cardiovascular disease among a population of free living subjects in the community of Framingham, Massachusetts.”

• Identifying risk factors for cardiovascular disease (CVD)
• N=4,434 participants (subset of the original sample)
• Follow-up period: 1956 – 1968
• Longitudinal data: measurements approximately every 6 years
• 1 to 3 observations for each participant (total 11,627 obs)
The Framingham Heart Study

• Information:
  – ID
  – Age
  – Sex
  – Period (1\textsuperscript{st}, 2\textsuperscript{nd}, or 3\textsuperscript{rd} exam)
  – Systolic Blood Pressure (mmHg)
  – Diastolic Blood Pressure (mmHg)
  – Use of Anti-hypertensive medication at exam (yes/no)
  – Current smoking status (yes/no)
  – Average number of cigarettes smoked/day
  – Prevalent coronary Heart disease (yes/no)
  – ... etc
The Framingham Heart Study
Statistical Concepts: Example 1
The Framingham Heart Study

• Data :
  
• Variables :
  
• Subjects :
  
• Observations :
  
• Population :
  
• Sample :
Statistical Concepts: Example 1

The Framingham Heart Study

- **Data**: all the collected information for the purposes of this study
- **Variables**: 
- **Subjects**: 
- **Observations**: 
- **Population**: 
- **Sample**: 
Statistical Concepts: Example 1
The Framingham Heart Study

• Data:
  – all the collected information for the purposes of this study

• Variables:
  – “randid”, “period”, “sex”, “age”, “totchol”, “cursmoke”, .., etc

• Subjects:
  –

• Observations:

• Population:

• Sample:
Statistical Concepts: Example 1

The Framingham Heart Study

• Data:
  – all the collected information for the purposes of this study

• Variables:
  – “randid”, “period”, “sex”, “age”, “totchol”, “cursmoke”, .., etc

• Subjects:
  – participants (each one with unique ID number “randid”)

• Observations:

• Population:

• Sample:
Statistical Concepts: Example 1

The Framingham Heart Study

• **Data:**
  – all the collected information for the purposes of this study

• **Variables:**
  – “randid”, “period”, “sex”, “age”, “totchol”, “cursmoke”, .., etc

• **Subjects:**
  – participants (each one with unique ID number “randid”)

• **Observations:**
  – Each element of the dataset, e.g. for participant with “randid”=9428:
    • “period”=2, “totchol”=283, “age”=54, ... etc.

• **Population:**

• **Sample:**
Statistical Concepts: Example 1
The Framingham Heart Study

• **Data:**
  – all the collected information for the purposes of this study

• **Variables:**
  – “randid”, “period”, “sex”, “age”, “totchol”, “cursmoke”, .., etc

• **Subjects:**
  – participants (each one with unique ID number “randid”)

• **Observations:**
  – Each element of the dataset, e.g. for participant with “randid”=9428:
    • “period”=2, “totchol”=283, “age”=54, ... etc.

• **Population:**
  – ... “a population of free living subjects in the community of Framingham, Massachusetts.” ...

• **Sample:**
Statistical Concepts: Example 1
The Framingham Heart Study

• **Data**:  
  – all the collected information for the purposes of this study

• **Variables**:  
  – “randid”, “period”, “sex”, “age”, “totchol”, “cursmoke”, .., etc

• **Subjects**:  
  – participants (each one with unique ID number “randid”)

• **Observations**:  
  – Each element of the dataset, e.g. for participant with “randin”=9428:  
    • “period”=2, “totchol”=283, “age”=54, ... etc.

• **Population**:  
  – ... “a population of free living subjects in the community of Framingham, Massachusetts.” ...

• **Sample**:  
  – Subset of the population of size n=4,434
Classification of Variables

Discrete

Nominal
- Name only.
  Ex., *Gender*: Male, Female

Ordinal
- Order implied.
  Ex., *Self-reported health status*:
  - Excellent
  - Very good
  - Good
  - Fair
  - Poor

Continuous (interval, ratio)

Measurements in scale
- Ex., age, height, weight, cholesterol level, SBP, DBP, etc.

Reference: Introductory Statistics, Chapter 1.3
Classification of Variables: Example
The Framingham Heart Study

• Discrete Variables:
  – Nominal:
  – Ordinal:

• Continuous Variables:
**The Framingham Heart Study**

<table>
<thead>
<tr>
<th></th>
<th>sex</th>
<th>randid</th>
<th>totchol</th>
<th>age</th>
<th>sysbp</th>
<th>diabp</th>
<th>cursmoke</th>
<th>cigpday</th>
<th>bmi</th>
<th>diabetes</th>
<th>bpmeds</th>
<th>heartrte</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2448</td>
<td>195</td>
<td>39</td>
<td>106.0</td>
<td>70.0</td>
<td>0</td>
<td>0</td>
<td>26.97</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2448</td>
<td>209</td>
<td>52</td>
<td>121.0</td>
<td>66.0</td>
<td>0</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>6238</td>
<td>250</td>
<td>46</td>
<td>121.0</td>
<td>81.0</td>
<td>0</td>
<td>0</td>
<td>28.73</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>6238</td>
<td>260</td>
<td>52</td>
<td>105.0</td>
<td>69.5</td>
<td>0</td>
<td>0</td>
<td>29.43</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>6238</td>
<td>237</td>
<td>58</td>
<td>108.0</td>
<td>66.0</td>
<td>0</td>
<td>0</td>
<td>28.50</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>9428</td>
<td>245</td>
<td>48</td>
<td>127.5</td>
<td>80.0</td>
<td>1</td>
<td>20</td>
<td>25.34</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>9428</td>
<td>283</td>
<td>54</td>
<td>141.0</td>
<td>89.0</td>
<td>1</td>
<td>30</td>
<td>25.34</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>10552</td>
<td>225</td>
<td>61</td>
<td>150.0</td>
<td>95.0</td>
<td>1</td>
<td>30</td>
<td>28.58</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>10552</td>
<td>232</td>
<td>67</td>
<td>183.0</td>
<td>109.0</td>
<td>1</td>
<td>20</td>
<td>30.18</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>11252</td>
<td>285</td>
<td>46</td>
<td>130.0</td>
<td>84.0</td>
<td>1</td>
<td>23</td>
<td>23.10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>11252</td>
<td>343</td>
<td>51</td>
<td>109.0</td>
<td>77.0</td>
<td>1</td>
<td>30</td>
<td>23.48</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Classification of Variables: Example
The Framingham Heart Study

• Discrete Variables:
  – Nominal: “sex”, “cursmoke”, etc.
  – Ordinal: “period”

• Continuous Variables:
  – “sysbp”, “bmi”, etc
Descriptive statistics for Discrete variables

- **Frequency (f):** Number (#) of subjects in each category.
- **Relative frequency \((\frac{f}{n} \cdot 100):** Proportion (%) of subjects in each category.

Example: calculate number/proportion of subjects in each period

<table>
<thead>
<tr>
<th>Period</th>
<th>Frequency (f)</th>
<th>Relative Frequency (%)</th>
<th>Cumulative Relative Frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4434</td>
<td>4434 (\frac{4434}{11627} \cdot 100 = 38.1)</td>
<td>38.1</td>
</tr>
<tr>
<td>2</td>
<td>3930</td>
<td>33.8</td>
<td>71.9</td>
</tr>
<tr>
<td>3</td>
<td>3260</td>
<td>28.1</td>
<td>100</td>
</tr>
<tr>
<td>Total</td>
<td>11627</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>
Descriptive statistics for Discrete variables

- **Frequency (f)**: Number (#) of subjects in each category.

- **Relative frequency \( \left( \frac{f}{n} \cdot 100 \right) \)**: Proportion (%) of subjects in each category.

Example: calculate number/proportion of subjects in each period in R

```r
# frequency and relative frequency of period #
tab1 = table(dat1$period)
n = sum(tab1)
rel_tab1 = tab1/n*100  # alternative way: prop.table(tab1)*100
cum_tab1 = cumsum(rel_tab1)
cbind(tab1, rel_tab1, cum_tab1)

> cbind(tab1, rel_tab1, cum_tab1)
   tab1 rel_tab1  cum_tab1
1 4434 38.13537  38.13537
2 3930 33.80064  71.93601
3 3263 28.06399 100.00000
```
Graphical Methods for Discrete variables

- **Bar plots**: indicate frequency or relative frequency distribution

```r
barplot(tab1, xlab="Period", ylab = "Frequency")
barplot(rel_tab1, xlab="Period", ylab="Proportion")
```
Descriptive statistics for Discrete variables

- Frequency and relative frequency \( \left( \frac{f}{n} \cdot 100 \right) \) by groups

Example: calculate number/proportion of subjects in each period in R by sex (female if sex=2)

```r
## period by sex ##
tab2 = table(dat1$period, dat1$sex)
tab2
rel_tab2 = prop.table(tab2, margin=2)*100 ## the option margin = 2 for column sum to be 100%
rel_tab2
cbind(tab2, rel_tab2)

> ## period by sex ##
> tab2 = table(dat1$period, dat1$sex)
> tab2

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1944</td>
<td>2490</td>
</tr>
<tr>
<td>2</td>
<td>1691</td>
<td>2239</td>
</tr>
<tr>
<td>3</td>
<td>1387</td>
<td>1876</td>
</tr>
</tbody>
</table>

> rel_tab2 = prop.table(tab2, margin=2)*100 ## the option margin = 2 for column sum to be 100%
> rel_tab2

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>38.70968</td>
<td>37.69871</td>
</tr>
<tr>
<td>2</td>
<td>33.67184</td>
<td>33.89856</td>
</tr>
<tr>
<td>3</td>
<td>27.61848</td>
<td>28.40273</td>
</tr>
</tbody>
</table>

> cbind(tab2, rel_tab2)

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1944</td>
<td>2490</td>
<td>38.70968</td>
<td>37.69871</td>
</tr>
<tr>
<td>2</td>
<td>1691</td>
<td>2239</td>
<td>33.67184</td>
<td>33.89856</td>
</tr>
<tr>
<td>3</td>
<td>1387</td>
<td>1876</td>
<td>27.61848</td>
<td>28.40273</td>
</tr>
</tbody>
</table>
```
Descriptive statistics for Continuous variables

<table>
<thead>
<tr>
<th>Measures of location</th>
<th>Measures of dispersion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indicate where the collected values of a variable are “located” compared to the range of possible values it can take.</td>
<td>Indicate how dispersed the collected values of a variable are.</td>
</tr>
</tbody>
</table>
## Descriptive statistics for Continuous variables

<table>
<thead>
<tr>
<th>Measures of location</th>
<th>Measures of dispersion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indicate where the collected values of a variable are “located” compared to the range of possible values it can take.</td>
<td>Indicate how dispersed the collected values of a variable are.</td>
</tr>
<tr>
<td>• Mean</td>
<td>• Range</td>
</tr>
<tr>
<td>• Median</td>
<td>• Variance</td>
</tr>
<tr>
<td>• Quartiles</td>
<td>• Standard Deviation</td>
</tr>
<tr>
<td>• Mode</td>
<td>• Interquartile range (IQR)</td>
</tr>
<tr>
<td>• Min</td>
<td>• Mean Absolute Deviation (MAD)</td>
</tr>
<tr>
<td>• Max</td>
<td>• Coefficient of variation</td>
</tr>
</tbody>
</table>
# Measures of Location: Mean ($\bar{x}$)

## Definition

- Average value.
- A typical value for the variable of interest.

## Formula

\[
\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n}
\]

## Sample of $n=7$

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>$X_6$</th>
<th>$X_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>121</td>
<td>110</td>
<td>114</td>
<td>100</td>
<td>160</td>
<td>130</td>
<td>130</td>
</tr>
</tbody>
</table>
Measures of Location: Mean ($\bar{x}$)

<table>
<thead>
<tr>
<th>Definition</th>
<th>Formula</th>
</tr>
</thead>
</table>
| • Average value.  
• A typical value for the variable of interest. | $\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n}$ |

- Sample of $n=7$
- $X =$ Systolic Blood Pressure in mmHg:

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>$X_6$</th>
<th>$X_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>121</td>
<td>110</td>
<td>114</td>
<td>100</td>
<td>160</td>
<td>130</td>
<td>130</td>
</tr>
</tbody>
</table>

$$\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + X_3 + \cdots + X_7}{n} = \frac{121 + 110 + 114 + \cdots + 130}{7} = \frac{865}{7} = 123.57 \approx 123.6$$
Measures of Location: Median

<table>
<thead>
<tr>
<th>Definition</th>
<th>Formula</th>
</tr>
</thead>
</table>
| • The middle value of the variable of interest.  
  • 50% of the collected values are less and 50% are greater than the median. | • If n odd:  
  the $\frac{(n+1)^{th}}{2}$ observation  
  • If n even:  
  mean of the $\frac{n^{th}}{2}$ and the $\frac{n}{2} + 1)^{th}$ observations in the ordered sample |

<table>
<thead>
<tr>
<th>Unordered</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>$X_6$</th>
<th>$X_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>121</td>
<td>110</td>
<td>114</td>
<td>100</td>
<td>160</td>
<td>130</td>
<td>130</td>
</tr>
<tr>
<td>Ordered</td>
<td>$X_{(1)}$</td>
<td>$X_{(2)}$</td>
<td>$X_{(3)}$</td>
<td>$X_{(4)}$</td>
<td>$X_{(5)}$</td>
<td>$X_{(6)}$</td>
<td>$X_{(7)}$</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>110</td>
<td>114</td>
<td>121</td>
<td>130</td>
<td>130</td>
<td>160</td>
</tr>
</tbody>
</table>
# Measures of Location: Median

<table>
<thead>
<tr>
<th>Definition</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>The middle value of the variable of interest.</td>
<td></td>
</tr>
<tr>
<td>50% of the collected values are less and 50% are greater than the median.</td>
<td></td>
</tr>
<tr>
<td>• If n odd:</td>
<td></td>
</tr>
<tr>
<td>the (\frac{(n+1)t}{2}) observation</td>
<td></td>
</tr>
<tr>
<td>• If n even:</td>
<td></td>
</tr>
<tr>
<td>mean of the (\frac{n}{2}) and the (\frac{n}{2} + 1) (th)</td>
<td></td>
</tr>
<tr>
<td>observations in the ordered sample</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unordered</th>
<th>(X_1)</th>
<th>(X_2)</th>
<th>(X_3)</th>
<th>(X_4)</th>
<th>(X_5)</th>
<th>(X_6)</th>
<th>(X_7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(X_1)</td>
<td>121</td>
<td>110</td>
<td>114</td>
<td>100</td>
<td>160</td>
<td>130</td>
<td>130</td>
</tr>
<tr>
<td>Ordered</td>
<td>(X_{(1)})</td>
<td>(X_{(2)})</td>
<td>(X_{(3)})</td>
<td>(X_{(4)})</td>
<td>(X_{(5)})</td>
<td>(X_{(6)})</td>
<td>(X_{(7)})</td>
</tr>
<tr>
<td>(X_{(1)})</td>
<td>100</td>
<td>110</td>
<td>114</td>
<td>121</td>
<td>130</td>
<td>130</td>
<td>160</td>
</tr>
</tbody>
</table>

\(n=7 \rightarrow \text{odd \#} \rightarrow \text{median:} \frac{(7+1)}{2} = 4^{th} \text{ observation in the ordered sample} \rightarrow \text{median} = X_{(4)} = 121
# Measures of Location: Median

<table>
<thead>
<tr>
<th>Unordered</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>$X_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>121</td>
<td>110</td>
<td>114</td>
<td>100</td>
<td>160</td>
<td>130</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ordered</th>
<th>$X_{(1)}$</th>
<th>$X_{(2)}$</th>
<th>$X_{(3)}$</th>
<th>$X_{(4)}$</th>
<th>$X_{(5)}$</th>
<th>$X_{(6)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>110</td>
<td>114</td>
<td>121</td>
<td>130</td>
<td>130</td>
</tr>
</tbody>
</table>
Measures of Location: Median

\[ \text{n=6 } \rightarrow \text{ even } \# \quad \rightarrow \text{ median: mean of the } \left( \frac{6}{2} \right) = 3^{\text{th}} \text{ and the } \left( \frac{6}{2} + 1 \right) = 4^{\text{th}} \text{ observations in the ordered sample} \]

\[ \rightarrow \text{ median} = \frac{X_{(3)} + X_{(4)}}{2} = \frac{114 + 121}{2} = 117.5 \]
Measures of Location: Quartiles

<table>
<thead>
<tr>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>• First ($Q_1$): 25% of the collected values are less than $Q_1$.</td>
</tr>
<tr>
<td>• Second ($Q_2$): 50% of the collected values are less than $Q_2$ (median).</td>
</tr>
<tr>
<td>• Third ($Q_3$): 75% of the collected values are less than $Q_3$.</td>
</tr>
</tbody>
</table>
# Measures of Location: Percentiles

<table>
<thead>
<tr>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>• $q_p$: $p%$ of the collected values are less than $q_p$.</td>
</tr>
<tr>
<td>• E.g., $q_1$ is that value of the population (or sample) with 1% of the observed values being less and 99% being greater than it.</td>
</tr>
</tbody>
</table>
Measures of Location: Mode / Min / Max

<table>
<thead>
<tr>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Min</strong>: the minimum of the collected values ($X_{(1)}$).</td>
</tr>
<tr>
<td><strong>Max</strong>: the maximum of the collected values ($X_{(n)}$).</td>
</tr>
<tr>
<td><strong>Mode</strong>: the most frequent of the collected values.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unordered</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>$X_6$</th>
<th>$X_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>121</td>
<td>110</td>
<td>114</td>
<td>100</td>
<td>160</td>
<td>130</td>
<td>130</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ordered</th>
<th>$X_{(1)}$</th>
<th>$X_{(2)}$</th>
<th>$X_{(3)}$</th>
<th>$X_{(4)}$</th>
<th>$X_{(5)}$</th>
<th>$X_{(6)}$</th>
<th>$X_{(7)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>110</td>
<td>114</td>
<td>121</td>
<td>130</td>
<td>130</td>
<td>160</td>
</tr>
</tbody>
</table>
Measures of Location: Mode / Min / Max

### Definition

- **Min**: the minimum of the collected values ($X_{(1)}$).
- **Max**: the maximum of the collected values ($X_{(n)}$).
- **Mode**: the most frequent of the collected values.

<table>
<thead>
<tr>
<th>Unordered</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>$X_6$</th>
<th>$X_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>121</td>
<td>110</td>
<td>114</td>
<td>100</td>
<td>160</td>
<td>130</td>
<td>130</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ordered</th>
<th>$X_{(1)}$</th>
<th>$X_{(2)}$</th>
<th>$X_{(3)}$</th>
<th>$X_{(4)}$</th>
<th>$X_{(5)}$</th>
<th>$X_{(6)}$</th>
<th>$X_{(7)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>110</td>
<td>114</td>
<td>121</td>
<td>130</td>
<td>130</td>
<td>160</td>
</tr>
</tbody>
</table>

Min: 100
Max: 160
Mode: 130
# Measures of Dispersion: Variance ($s^2$)

<table>
<thead>
<tr>
<th>Definition</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Average squared deviation from the mean.</td>
<td>[ S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} ]</td>
</tr>
</tbody>
</table>

- $\overline{X} = 123.6$

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>$X_6$</th>
<th>$X_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>121</td>
<td>110</td>
<td>114</td>
<td>100</td>
<td>160</td>
<td>130</td>
<td>130</td>
</tr>
</tbody>
</table>
Measures of Dispersion: 
Variance ($s^2$)

<table>
<thead>
<tr>
<th>Definition</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Average squared deviation from the mean.</td>
<td>$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$</td>
</tr>
</tbody>
</table>

- $\overline{X} = 123.6$

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>$X_6$</th>
<th>$X_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>121</td>
<td>110</td>
<td>114</td>
<td>100</td>
<td>160</td>
<td>130</td>
<td>130</td>
</tr>
</tbody>
</table>

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} = \frac{(X_1 - \overline{X})^2 + \cdots + (X_7 - \overline{X})^2}{n-1} = \frac{(121 - 123.6)^2 + \cdots + (130 - 123.6)^2}{7-1} =$$

$$= \frac{2247.72}{6} = 374.62 \approx 374.6$$
Other Measures of Dispersion:

<table>
<thead>
<tr>
<th>Definition</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Standard deviation</td>
<td>( s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1}} )</td>
</tr>
<tr>
<td>• Mean Absolute Deviation (MAD)</td>
<td>( \text{MAD} = \frac{\sum_{i=1}^{n}</td>
</tr>
<tr>
<td>• Range</td>
<td>( \text{Max} - \text{Min} )</td>
</tr>
<tr>
<td>• Interquartile Range (IQR)</td>
<td>( Q_3 - Q_1 )</td>
</tr>
<tr>
<td>• Coefficient of variation</td>
<td>( \frac{s}{\bar{X}} )</td>
</tr>
</tbody>
</table>
Descriptive Statistics for Continuous Variables

Example: The Framingham Heart Study

SEX = 1 for male, 2 for female

\[ \text{Std.dev} = Var(X_i) \] to explain variation of sysbp

\[ \text{SE.mean} = \sqrt{Var(\bar{X})} \] to explain variation of MEAN sysbp
Graphical Methods for Continuous variables

- **Histogram**: indicate the distribution of the values of a continuous variable.

```r
## Histogram of sysbp by sex ##
dat_m = subset(dat1, sex==1) ## get a subset for male
dat_f = subset(dat1, sex==2) ## get a subset for female
par(mfrow = c(1,2)) ## to draw two plots side by side
hist(dat_m$sysbp, main="Histogram of sysbp for male")
hist(dat_f$sysbp, main="Histogram of sysbp for female")
```
Graphical Methods for Continuous variables

**Box - Plot**: indicate the distribution of the values of a continuous variable, pointing out the following quantities:
Outliers

• Observations above Q3 + 1.5IQR or below Q1 – 1.5IQR are called, “outliers”, in the box plot.

• Outliers are not caused by typo or errors.

• Outliers are simply part of data, which can not be ignored.

• Outliers explain how many extreme values are located at tails of a distribution.
Graphical Methods for Continuous variables

- **Box-Plot**: the distribution of the values of a continuous variable.

```r
## A box plot of sysbp by sex ##
par(mfrow = c(1,1))
boxplot(sysbp ~ sex, data=dat1, main="Box plot of sysbp by sex")
```