Lecture 2
Graphical and Numerical Summaries
### Types of Data

- A subset of the data from a study of a series of male patients from Greenlane Hospital in Auckland after a heart attack.

- Goal of the study: How long will the patient live after the heart attack?

<table>
<thead>
<tr>
<th>ID</th>
<th>EJEC</th>
<th>SYS-VOL</th>
<th>DIA-VOL</th>
<th>OCCLU</th>
<th>STEN</th>
<th>TIME</th>
<th>OUTCOME</th>
<th>AGE</th>
<th>SMOKE</th>
<th>BETA</th>
<th>CHOL</th>
<th>SURG</th>
</tr>
</thead>
<tbody>
<tr>
<td>390</td>
<td>72</td>
<td>36</td>
<td>131</td>
<td>0</td>
<td>0</td>
<td>143</td>
<td>0</td>
<td>49</td>
<td>2</td>
<td>2</td>
<td>59</td>
<td>0</td>
</tr>
<tr>
<td>279</td>
<td>52</td>
<td>74</td>
<td>155</td>
<td>37</td>
<td>63</td>
<td>143</td>
<td>0</td>
<td>54</td>
<td>2</td>
<td>2</td>
<td>68</td>
<td>1</td>
</tr>
<tr>
<td>391</td>
<td>62</td>
<td>52</td>
<td>137</td>
<td>33</td>
<td>47</td>
<td>16</td>
<td>2</td>
<td>56</td>
<td>2</td>
<td>2</td>
<td>52</td>
<td>0</td>
</tr>
<tr>
<td>201</td>
<td>50</td>
<td>165</td>
<td>329</td>
<td>33</td>
<td>30</td>
<td>143</td>
<td>0</td>
<td>42</td>
<td>2</td>
<td>2</td>
<td>39</td>
<td>0</td>
</tr>
<tr>
<td>202</td>
<td>50</td>
<td>47</td>
<td>95</td>
<td>0</td>
<td>100</td>
<td>143</td>
<td>0</td>
<td>46</td>
<td>2</td>
<td>2</td>
<td>74</td>
<td>1</td>
</tr>
<tr>
<td>69</td>
<td>27</td>
<td>124</td>
<td>170</td>
<td>77</td>
<td>23</td>
<td>143</td>
<td>0</td>
<td>57</td>
<td>2</td>
<td>2</td>
<td>NA</td>
<td>2</td>
</tr>
<tr>
<td>310</td>
<td>60</td>
<td>86</td>
<td>215</td>
<td>7</td>
<td>50</td>
<td>40</td>
<td>0</td>
<td>51</td>
<td>2</td>
<td>2</td>
<td>58</td>
<td>0</td>
</tr>
<tr>
<td>392</td>
<td>72</td>
<td>37</td>
<td>132</td>
<td>40</td>
<td>10</td>
<td>14</td>
<td>5</td>
<td>53</td>
<td>2</td>
<td>2</td>
<td>75</td>
<td>0</td>
</tr>
<tr>
<td>311</td>
<td>60</td>
<td>65</td>
<td>163</td>
<td>0</td>
<td>40</td>
<td>142</td>
<td>0</td>
<td>45</td>
<td>2</td>
<td>2</td>
<td>72</td>
<td>0</td>
</tr>
<tr>
<td>393</td>
<td>63</td>
<td>52</td>
<td>140</td>
<td>0</td>
<td>10</td>
<td>142</td>
<td>0</td>
<td>46</td>
<td>2</td>
<td>2</td>
<td>90</td>
<td>0</td>
</tr>
<tr>
<td>70</td>
<td>29</td>
<td>117</td>
<td>164</td>
<td>50</td>
<td>0</td>
<td>142</td>
<td>0</td>
<td>48</td>
<td>2</td>
<td>2</td>
<td>72</td>
<td>0</td>
</tr>
<tr>
<td>203</td>
<td>48</td>
<td>69</td>
<td>133</td>
<td>0</td>
<td>27</td>
<td>142</td>
<td>0</td>
<td>54</td>
<td>2</td>
<td>2</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td>394</td>
<td>59</td>
<td>54</td>
<td>133</td>
<td>30</td>
<td>13</td>
<td>142</td>
<td>0</td>
<td>39</td>
<td>2</td>
<td>1</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td>204</td>
<td>50</td>
<td>67</td>
<td>135</td>
<td>37</td>
<td>63</td>
<td>141</td>
<td>0</td>
<td>49</td>
<td>2</td>
<td>2</td>
<td>86</td>
<td>2</td>
</tr>
<tr>
<td>280</td>
<td>53</td>
<td>65</td>
<td>138</td>
<td>0</td>
<td>33</td>
<td>140</td>
<td>0</td>
<td>58</td>
<td>2</td>
<td>1</td>
<td>49</td>
<td>0</td>
</tr>
<tr>
<td>55</td>
<td>17</td>
<td>184</td>
<td>221</td>
<td>57</td>
<td>13</td>
<td>5</td>
<td>1</td>
<td>50</td>
<td>2</td>
<td>2</td>
<td>70</td>
<td>2</td>
</tr>
<tr>
<td>79</td>
<td>37</td>
<td>68</td>
<td>140</td>
<td>37</td>
<td>47</td>
<td>118</td>
<td>5</td>
<td>58</td>
<td>2</td>
<td>2</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td>205</td>
<td>45</td>
<td>106</td>
<td>193</td>
<td>33</td>
<td>43</td>
<td>140</td>
<td>0</td>
<td>47</td>
<td>1</td>
<td>1</td>
<td>38</td>
<td>1</td>
</tr>
<tr>
<td>206</td>
<td>43</td>
<td>85</td>
<td>150</td>
<td>0</td>
<td>13</td>
<td>23</td>
<td>5</td>
<td>51</td>
<td>2</td>
<td>2</td>
<td>61</td>
<td>0</td>
</tr>
<tr>
<td>312</td>
<td>60</td>
<td>59</td>
<td>149</td>
<td>7</td>
<td>37</td>
<td>139</td>
<td>0</td>
<td>43</td>
<td>2</td>
<td>1</td>
<td>56</td>
<td>0</td>
</tr>
<tr>
<td>80</td>
<td>39</td>
<td>41</td>
<td>168</td>
<td>47</td>
<td>43</td>
<td>100</td>
<td>1</td>
<td>55</td>
<td>2</td>
<td>2</td>
<td>62</td>
<td>1</td>
</tr>
<tr>
<td>281</td>
<td>57</td>
<td>53</td>
<td>124</td>
<td>0</td>
<td>57</td>
<td>140</td>
<td>0</td>
<td>58</td>
<td>2</td>
<td>1</td>
<td>93</td>
<td>0</td>
</tr>
<tr>
<td>207</td>
<td>44</td>
<td>68</td>
<td>121</td>
<td>27</td>
<td>60</td>
<td>139</td>
<td>0</td>
<td>55</td>
<td>2</td>
<td>2</td>
<td>63</td>
<td>1</td>
</tr>
<tr>
<td>282</td>
<td>51</td>
<td>53</td>
<td>109</td>
<td>0</td>
<td>77</td>
<td>139</td>
<td>0</td>
<td>41</td>
<td>2</td>
<td>2</td>
<td>45</td>
<td>4</td>
</tr>
<tr>
<td>396</td>
<td>63</td>
<td>58</td>
<td>157</td>
<td>0</td>
<td>13</td>
<td>139</td>
<td>0</td>
<td>51</td>
<td>2</td>
<td>2</td>
<td>60</td>
<td>0</td>
</tr>
<tr>
<td>208</td>
<td>49</td>
<td>81</td>
<td>157</td>
<td>13</td>
<td>13</td>
<td>139</td>
<td>0</td>
<td>49</td>
<td>2</td>
<td>2</td>
<td>60</td>
<td>0</td>
</tr>
<tr>
<td>209</td>
<td>48</td>
<td>58</td>
<td>112</td>
<td>0</td>
<td>0</td>
<td>72</td>
<td>1</td>
<td>56</td>
<td>2</td>
<td>2</td>
<td>57</td>
<td>0</td>
</tr>
<tr>
<td>283</td>
<td>58</td>
<td>71</td>
<td>167</td>
<td>27</td>
<td>0</td>
<td>138</td>
<td>0</td>
<td>45</td>
<td>2</td>
<td>1</td>
<td>46</td>
<td>0</td>
</tr>
<tr>
<td>210</td>
<td>42</td>
<td>92</td>
<td>159</td>
<td>0</td>
<td>0</td>
<td>139</td>
<td>0</td>
<td>57</td>
<td>2</td>
<td>2</td>
<td>58</td>
<td>0</td>
</tr>
<tr>
<td>397</td>
<td>68</td>
<td>50</td>
<td>156</td>
<td>0</td>
<td>100</td>
<td>138</td>
<td>0</td>
<td>51</td>
<td>2</td>
<td>1</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td>211</td>
<td>43</td>
<td>146</td>
<td>259</td>
<td>47</td>
<td>33</td>
<td>3</td>
<td>1</td>
<td>56</td>
<td>2</td>
<td>2</td>
<td>70</td>
<td>0</td>
</tr>
<tr>
<td>398</td>
<td>67</td>
<td>43</td>
<td>130</td>
<td>0</td>
<td>70</td>
<td>138</td>
<td>0</td>
<td>49</td>
<td>2</td>
<td>2</td>
<td>NA</td>
<td>3</td>
</tr>
<tr>
<td>284</td>
<td>52</td>
<td>70</td>
<td>146</td>
<td>0</td>
<td>23</td>
<td>137</td>
<td>0</td>
<td>47</td>
<td>1</td>
<td>2</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td>286</td>
<td>54</td>
<td>62</td>
<td>133</td>
<td>33</td>
<td>23</td>
<td>137</td>
<td>0</td>
<td>38</td>
<td>2</td>
<td>2</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td>71</td>
<td>37</td>
<td>93</td>
<td>148</td>
<td>47</td>
<td>0</td>
<td>137</td>
<td>0</td>
<td>59</td>
<td>2</td>
<td>2</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td>286</td>
<td>51</td>
<td>65</td>
<td>133</td>
<td>43</td>
<td>7</td>
<td>136</td>
<td>0</td>
<td>54</td>
<td>2</td>
<td>2</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td>212</td>
<td>42</td>
<td>95</td>
<td>163</td>
<td>40</td>
<td>10</td>
<td>109</td>
<td>3</td>
<td>57</td>
<td>2</td>
<td>2</td>
<td>NA</td>
<td>4</td>
</tr>
<tr>
<td>400</td>
<td>66</td>
<td>49</td>
<td>144</td>
<td>10</td>
<td>50</td>
<td>65</td>
<td>1</td>
<td>52</td>
<td>2</td>
<td>2</td>
<td>55</td>
<td>0</td>
</tr>
<tr>
<td>287</td>
<td>54</td>
<td>66</td>
<td>145</td>
<td>7</td>
<td>40</td>
<td>136</td>
<td>0</td>
<td>47</td>
<td>2</td>
<td>2</td>
<td>62</td>
<td>0</td>
</tr>
<tr>
<td>81</td>
<td>39</td>
<td>144</td>
<td>237</td>
<td>13</td>
<td>87</td>
<td>156</td>
<td>3</td>
<td>39</td>
<td>2</td>
<td>2</td>
<td>56</td>
<td>3</td>
</tr>
<tr>
<td>813</td>
<td>63</td>
<td>52</td>
<td>141</td>
<td>0</td>
<td>47</td>
<td>43</td>
<td>3</td>
<td>48</td>
<td>2</td>
<td>2</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td>68</td>
<td>30</td>
<td>219</td>
<td>314</td>
<td>33</td>
<td>45</td>
<td>76</td>
<td>1</td>
<td>53</td>
<td>1</td>
<td>2</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td>288</td>
<td>59</td>
<td>39</td>
<td>94</td>
<td>0</td>
<td>0</td>
<td>135</td>
<td>0</td>
<td>47</td>
<td>1</td>
<td>2</td>
<td>63</td>
<td>0</td>
</tr>
<tr>
<td>407</td>
<td>67</td>
<td>39</td>
<td>117</td>
<td>0</td>
<td>73</td>
<td>53</td>
<td>1</td>
<td>57</td>
<td>2</td>
<td>2</td>
<td>62</td>
<td>2</td>
</tr>
</tbody>
</table>

*NA = Not Available (missing data code).
**Types of variables**

**Quantitative**
- (Age, time, weight, etc.)
- May take any value from some interval (weight)

**Discrete**
- May take values from some grid (age in years)

**Categorical**
- (Color, surgery outcome, smoking, etc.)
- No order (surgery outcome)

**Ordinal**
- Order (Letter grade)
Type of variable

is determined by data you have and problem you consider

Aging of a person is a continuous process
Age is a quantitative, continuous variable

Age in years (18, 25, 63,…) is quantitative, discrete

Age as (Kid, Young, Middle-age, Senior) is qualitative, ordinal
Simple Plots - Graphical Summaries of Data
Simple plots: dot plot

- Original data: {8  3  6  4.5  4  4.5}
- Sorted data: {3  4  4.5  4.5  6  8}
- Simple graph: dot plot

Simple plots: dot plot

- Interesting features of the data emphasized by the dot plot

Dot plot showing special features.

Simple plots: histogram

- Histogram is the most widely used statistical graph
Simple plots: histogram

- Divide observational interval into subintervals, (also called bins, class intervals)
- Calculate number of observation within each bin
- Draw a rectangle w/height = number of observations = frequency
- Relative frequency is the number of observations within a bin divided by the total number of observations
### Frequency vs. Relative Frequency

#### Frequency

- **Number of observations**

```
1 2 3 4 3 1 2 4
```

- **Measurement**

```
.05 .1 .15 .2 .15 .05 .1 .2
```

- **$n = 20$ (sample size)**
- **$k = 8$ (# of bins)**

#### Relative Frequency

```
\sum_{i=1}^{k} \frac{n_i}{n} = \frac{n}{n} = 1
```

- **$n = 20$ (sample size)**
- **$k = 8$ (# of bins)**
Figure 2.3.10  Features to look for in histograms and stem-and-leaf plots.

Example. According to the National Center for Health Statistics, the 6 leading causes of death in 1995 are: heart disease, cancer, stroke, pulmonary diseases, accidents, and others.

<table>
<thead>
<tr>
<th>Cause of death</th>
<th>Count (k)</th>
<th>percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>heart diseases</td>
<td>738</td>
<td>31.92</td>
</tr>
<tr>
<td>cancer</td>
<td>538</td>
<td>23.27</td>
</tr>
<tr>
<td>stroke</td>
<td>158</td>
<td>6.83</td>
</tr>
<tr>
<td>pulmonary diseases</td>
<td>103</td>
<td>4.46</td>
</tr>
<tr>
<td>accidents</td>
<td>93</td>
<td>4.02</td>
</tr>
<tr>
<td>others</td>
<td>682</td>
<td>29.5</td>
</tr>
<tr>
<td>All causes</td>
<td>2,312</td>
<td>100</td>
</tr>
</tbody>
</table>
Numerical Summaries: Central value and spread

\[ x_0 \] is the central value, characteristic value.

spread, most of the observed values
Sample 1

Different locations!

Sample 2
Sample 1

Sample 2

Different spread!
Different location and different spread!

Data from “Smart Student” Questionnaire
Measuring Center of the Data Set

• Example. You want to buy a 3-4 bdrm house. Need info on real estate sales in Reno, say on 200 recently sold 3-4bdrm houses.

• Data: $325,300, $287,650, $589,900, $230,900, …, $455,800.

• Q: What is the “average” selling price for a 3-4 bdrm house? What does AVERAGE mean?

• Most common? Most frequent? Mode

• Dividing selling prices in half, i.e. half are lower and higher that the “average”? Median

• Arithmetic average of all selling prices. Mean
Section 1.2: Summary Statistics

Let $X_1, \ldots, X_n$ be a sample.

- **Sample Mean:**
  \[
  \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i
  \]

- **Sample Variance:**
  \[
  s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{1}{n-1} \left( \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \right)
  \]

- **Sample standard deviation is the square root of the sample variance.**
More on Summary Statistics

- If \( X_1, \ldots, X_n \) is a sample, and \( Y_i = a + bX_i \), where \( a \) and \( b \) are constants, then
  \[ \bar{Y} = a + b\bar{X} \]

- If \( X_1, \ldots, X_n \) is a sample, and \( Y_i = a + bX_i \), where \( a \) and \( b \) are constants, then
  \[ s_y^2 = b^2 s_x^2, \text{ and } s_y = |b| s_x. \]
Measures of Relative Standing: quartiles and percentiles

- **Q₁ (First Quartile)** separates the bottom (smallest) 25% of sorted values from the top (largest) 75%.

- **Q₂ (Second Quartile)** same as the median; separates the bottom (smallest) 50% of sorted values from the top (largest) 50%.

- **Q₃ (Third Quartile)** separates the bottom (smallest) 75% of sorted values from the top (largest) 25%.

**Interquartile Range (or IQR):** $Q₃ - Q₁$
The median is another measure of center, like the mean. It divides the data into two halves.

Computing median: Order the $n$ data points from smallest to largest.

- If $n$ is odd, the sample median is the number in position $\frac{n+1}{2}$.

- If $n$ is even, the sample median is the average of the numbers in positions $\frac{n}{2}$ and $\frac{n}{2} + 1$. 
Percentiles

• The \textit{\textit{p}th percentile} of a sample, divides the sample so that as nearly as possible \textit{\textit{p}}\% of the sample values are less than the \textit{\textit{p}th percentile}, and (100 – \textit{\textit{p}}\%) are greater.

• Please \textit{read} about the computation of percentiles in the course text.
Sample Statistics and Population Parameters

- A numerical summary of a sample is called a statistic.
- A numerical summary of a population is called a parameter.
- Statistics are often used to estimate parameters.
The 5-number summary of a data set consists of (1) the minimum value; (2) $Q_1$; (3) the median ($Q_2$); (4) $Q_3$; and (5) the maximum value.

A boxplot (or box-and-whisker-diagram) is a graph of a data set that visualizes the 5-point summary.

Boxplot example, data=ages of actresses who got best Actress Award.
Simple plots: boxplot

- **3-rd quartile, Q3**
- **1-st quartile, Q1**
- **Median**
- **Maximal data point within 1.5 IQR of Q3**
- **Minimal data point within 1.5 IQR of Q1**
- **Outliers**
Simple plots: comparative boxplot
Simple plots: comparative boxplot

Comparative boxplot for coyote lengths

Length in cm

Females

Males
Symmetry and Skewness of a Distribution

Symmetric

distribution of data is symmetric if the left half of its histogram is roughly a mirror image of its right half.

Skewed

distribution of data is skewed if it is not symmetric and if it extends more to one side than the other.
Skewness

Mode = Mean = Median

MEAN = MEDIAN

(b) Symmetric

Mean

Mode

Median

(a) Skewed to the Left
(Negatively)

MEAN < MEDIAN

Mode

Mean

Median

(c) Skewed to the Right
(Positively)

MEAN > MEDIAN
Boxplots, histograms and symmetry

(a) Normal (bell-shaped) distribution
1000 heights (in.) of women

(b) Uniform distribution
1000 rolls of a die

(c) Skewed distribution
Incomes (thousands of dollars) of 1000 statistics professors
Where do outliers come from?

- Errors of measurements or recording – in those cases, people tend to disregard them.

- Natural order of things – they point to very important phenomena like floods, heat waves, hurricanes, etc. Should not be discarded but studied.

Example. Waiting times (in minutes) for a bus, 100 observations.
Scatterplot

- Data for which items consists of a pair of values is called **bivariate**.
- The graphical summary for bivariate data is a **scatterplot**.
- Display of a scatterplot:
John Wilder Tukey
(June 16, 1915 - July 26, 2000)