Statistics

Statistics deals with the collection and analysis of data to solve real-world problems.

History

Ancient Babylonians & Egyptians

Marine insurance rates determined by past statistical data (fourteenth century)

People

Blasé Pascal (1623-1662) | Pierre de Fermat (1601-1665)

People

Jacob Bernoulli (1654-1705)

Law of large numbers

\[
\frac{1}{n}(X_1 + X_2 + \cdots + X_n) \rightarrow E(X)
\]

People

Pierre Simon Laplace (1749-1827)

Karl Friedrich Gauss (1777-1855)
People

Sir Francis Galton (1822 – 1911)

- Height of Son
- Height of Father
- Height of the sons of fathers regressed towards the mean height of the population
- Regression Analysis

Founders of modern statistics

Karl Pearson (1857-1936)

- Biology
- Evolution
- Genetics

Ronald Fisher (1892-1962)

- MLE
- ANOVA
- Correlation
- Chi-square test

Egon Pearson (1895-1980)

- Testing Hypothesis
- Confidence Intervals

Jerzy Neyman (1894-1981)

Era of information

- Demographics
- Transaction data
- Customer behavior
- Human genome
- Satellite photographs

Huge datasets

- Data mining
- Data reduction

Multidisciplinary statistics methodology

Elements of Statistics

Statistics

- Data Collection
  - Experimental Design
  - Survey Sampling
  - Observational Study

- Descriptive Statistics
  - And Statistical Graphics

- Statistical Inference
  - Estimation
  - Testing Hypothesis

Descriptive Statistics

Data: a set of numbers representing characteristics of observations

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>90.00</td>
<td>90.00</td>
<td>90.00</td>
<td>90.00</td>
<td>90.00</td>
<td>90.00</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>80.00</td>
<td>80.00</td>
<td>80.00</td>
<td>80.00</td>
<td>80.00</td>
<td>80.00</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>85.00</td>
<td>85.00</td>
<td>85.00</td>
<td>85.00</td>
<td>85.00</td>
<td>85.00</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>90.00</td>
<td>90.00</td>
<td>90.00</td>
<td>90.00</td>
<td>90.00</td>
<td>90.00</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>95.00</td>
<td>95.00</td>
<td>95.00</td>
<td>95.00</td>
<td>95.00</td>
<td>95.00</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>105.00</td>
<td>105.00</td>
<td>105.00</td>
<td>105.00</td>
<td>105.00</td>
<td>105.00</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>110.00</td>
<td>110.00</td>
<td>110.00</td>
<td>110.00</td>
<td>110.00</td>
<td>110.00</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>115.00</td>
<td>115.00</td>
<td>115.00</td>
<td>115.00</td>
<td>115.00</td>
<td>115.00</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>120.00</td>
<td>120.00</td>
<td>120.00</td>
<td>120.00</td>
<td>120.00</td>
<td>120.00</td>
</tr>
</tbody>
</table>

- Year
- Tutorial Quiz
- Midterm Final
- Overall Grade

- A+
- A
- B
- C

- Data mining
- Data reduction

Multidisciplinary statistics methodology
Descriptive Statistics

Scales of Measurement

Nominal Scale: Male/Female, Chinese/British/American/...

Ordinal Scale: disagree/agree/strongly agree, low/medium/high

Interval Scale: °C, °F, altitude of a place

Ratio Scale: length, weight

Statistical Graphics

Frequency Distribution

Table: Grade of Students

<table>
<thead>
<tr>
<th>Grade</th>
<th>Count</th>
<th>CumCnt</th>
<th>Percent</th>
<th>CumPct</th>
</tr>
</thead>
<tbody>
<tr>
<td>A+</td>
<td>1</td>
<td>7</td>
<td>1.39</td>
<td>1.39</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>26</td>
<td>5.17</td>
<td>6.56</td>
</tr>
<tr>
<td>A-</td>
<td>3</td>
<td>57</td>
<td>11.33</td>
<td>17.89</td>
</tr>
<tr>
<td>B+</td>
<td>4</td>
<td>92</td>
<td>18.29</td>
<td>36.18</td>
</tr>
<tr>
<td>B</td>
<td>5</td>
<td>89</td>
<td>17.69</td>
<td>53.88</td>
</tr>
<tr>
<td>B-</td>
<td>6</td>
<td>76</td>
<td>15.11</td>
<td>68.99</td>
</tr>
<tr>
<td>C+</td>
<td>7</td>
<td>58</td>
<td>11.53</td>
<td>80.52</td>
</tr>
<tr>
<td>C</td>
<td>8</td>
<td>36</td>
<td>7.16</td>
<td>87.67</td>
</tr>
<tr>
<td>C-</td>
<td>9</td>
<td>23</td>
<td>4.57</td>
<td>92.25</td>
</tr>
<tr>
<td>F</td>
<td>10</td>
<td>34</td>
<td>6.76</td>
<td>99.01</td>
</tr>
<tr>
<td>N=</td>
<td>503</td>
<td>43</td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Summary Statistics for Discrete Variables

<table>
<thead>
<tr>
<th>Vehilces</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cars</td>
<td>45</td>
<td>50</td>
</tr>
<tr>
<td>Lorries</td>
<td>22</td>
<td>29</td>
</tr>
<tr>
<td>Motorcycles</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>Buses</td>
<td>86</td>
<td>100</td>
</tr>
<tr>
<td>Cars</td>
<td>1076</td>
<td>100</td>
</tr>
<tr>
<td>Lorries</td>
<td>222</td>
<td>100</td>
</tr>
<tr>
<td>Motorcycles</td>
<td>5945</td>
<td>100</td>
</tr>
</tbody>
</table>

Bar Chart

Histogram

Pie Chart
Percentile

"Joining Mensa Society requires scoring in the 98th percentile in one of the standard intelligence test."

Percentile from raw data

\[
(n+1)p = r.f
\]

integer 

fractional

\[
(100p)\text{th percentile} = X_{(r)} + f(X_{(s)} - X_{(r)})
\]

Dataset of \( n \) observations

approximately \( np \) observations less than or equal to \((100p)\)th percentile

<table>
<thead>
<tr>
<th>Percentile from histogram</th>
</tr>
</thead>
<tbody>
<tr>
<td>area on the left of ((100p))th percentile = ( np )</td>
</tr>
</tbody>
</table>

\[
|X_{(r)} = L_r + \frac{f}{d}(np - F_{r-1})|
\]

5th percentile

\[
30th \text{ percentile} = 30 \%
\]

75th percentile

\[
75th \text{ percentile} = \frac{2(151 - 14)}{3} = 64.67
\]

Boxplot

Five number summary

\[
\text{Min} \quad Q_1 \quad \text{Median} \quad Q_3 \quad \text{Max}
\]

Boxplots

MTB > Boxplot 'Overall'*'Year';
SUBC> Transpose;
SUBC> Box;
SUBC> Type 1;
SUBC> Color 2;
SUBC> Symbol;
SUBC> Outlier;
SUBC> Title "Overall scores of Math244 students";
SUBC> Title "by year";
Detecting Outliers

- Inner lower fence
- Inner upper fence
- Outer lower fence
- Outer upper fence

1.5 IQR

Stem-and-Leaf Display

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
<th>N</th>
<th>Leaf unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>8</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

N = 10
Leaf unit = 1

Stem-and-Leaf Display

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
<th>N</th>
<th>Leaf unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

N = 15
Leaf unit = 1

Lying with Statistical Graphics

Market penetration of four brands of cigarette: A, B, C, D

<table>
<thead>
<tr>
<th>Monthly sales of the four brands (in $m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td>D</td>
</tr>
</tbody>
</table>

Bad Graphic Designs

- The bars representing white cars are actually red in color.
- The bars representing white cars are actually red in color.
Bad Graphic Designs

Measure of Central Location

Percentage of college enrolment with age 25 and over

Measures of Central Location

mean / average / arithmetic mean

\[ \bar{X} = \frac{1}{n} (X_1 + X_2 + \cdots + X_n) \]  
(raw data)

\[ \bar{X} = \frac{\sum f_i X_i}{\sum f_i} \]  
(ungrouped frequency table)

\[ \bar{X} = \frac{\sum f_i m_i}{\sum f_i} \]  
(grouped frequency table)

\( m = \) midpoints of \( i^{th} \) class

Measures of Variation

\[ \text{range} = \text{max} - \text{min} \]

\[ \text{IQR} = Q_3 - Q_1 \]

\[ \text{MAD} = \frac{1}{n} \sum |x_i - \bar{X}| \]

\[ \text{SD} = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{X})^2} \]

Skewness

skewed to right (+ve skewed)

\[ \gamma_1 > 0 \]

\[ \gamma_2 > 0 \]

skewed to left (-ve skewed)

\[ \gamma_1 < 0 \]

\[ \gamma_2 < 0 \]

Standard Deviation

\[ s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{X})^2} \]

Why do we use \((n-1)\) instead of \(n\)?

\[ n=1 \Rightarrow SD = 0 \text{  } s = \text{ undefined \downarrow more reasonable} \]

\[ \bar{X} = \frac{1}{n-1} \sum x_i \]

only \((n-1)\) pieces of information in \( \sum (x_i - \bar{X}) \)