Biostatistics

Andy Chang
Youngstown State University

Introduction

What is Biostatistics?
What is Statistics?

- Lies and nothing but lies!?
- Scientific method for understanding and solving problem!

Why Statistics?

- Variability
- Uncertainty

Statistics is the science of understanding data and of making decisions in the face of “variability” and “uncertainty”.

Number of Hospital Admissions for Viral Pneumonia, 1983-1998, and 1990 Census Population: Females Aged 0-17, in Sacramento, San Francisco, and Los Angeles, California

<table>
<thead>
<tr>
<th></th>
<th>Cases</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sacramento</td>
<td>951</td>
<td>134,492 (.71%)</td>
</tr>
<tr>
<td>San Francisco</td>
<td>837</td>
<td>126,162 (.66%)</td>
</tr>
<tr>
<td>Los Angeles</td>
<td>3743</td>
<td>140,913 (2.66%)</td>
</tr>
</tbody>
</table>

Course Information

- Syllabus & Web Sites
  - SPSS & Lab Location
  - Lecture Notes
  - Term Project
  - Videos (Recorded Lectures, AAO, SPSS)

- Where to find help?
  - WebCT Chatroom/Discussions/E-mail
  - Biostatistics Tutors
AIDS Epidemic

Findings:
- Out of 840 subjects randomly selected in the study, 32 tested HIV positive, (95% CI, [2.7%, 4.9%] or 3.8%±1.1%)
- The difference between the percentages of subjects tested HIV positive from regions A and B is statistically significant at 5% level.

Statistics in Broader Sense

Statistics is a field of study concerned with the
1) data collection, [Producing data]
2) organization, summarization, examination and providing an overview of the general features of data, [Exploring Data]
3) and the drawing of inferences about a body of data (population) based on the properties of a part of the data (sample) observed. [Statistical Inference]

Areas of Statistics

“1)” → Sampling
“2)” → Descriptive statistics
“3)” → Inferential statistics

- Estimation
- Hypothesis Testing

Definitions

Data: the facts and figures that are collected, and analyzed in a statistical study.
Population (n.): the collection of all individuals of interest in a statistical study. [Census (v.): obtain information from the whole population.]
Sample (n.): any subset of individuals from the population. [Sample (v.): obtain a sample from the population.]
In health and medical (or clinical) study, researchers investigate a sample of subjects to understand the effectiveness of a treatment or an intervention on target population.

*Therefore, statistical inference is very important in health and medical research.*

Public health is fundamentally concerned with preventing disease, disability, and premature death in human population or community.

*Therefore, statistical inference is very important in public health research.*

Goal 2 of the Healthy People 2010:

**Eliminate Health Disparities**

Eliminate health disparities among segments of the population, including differences that occur by gender, race or ethnicity, disability, geographic location, or sexual orientation.

Some Basic Terms In Statistics

**Individuals (subjects, experimental unit):** the entities on which data are collected.

**Variable:** a characteristic of interest for the individual which takes on different values in different individual.

**Variable Types**

- **Quantitative Variables** (numeric)
  - **Continuous:** a variable that has an uncountable number of possible values. (measurements)
  - **Discrete:** a variable that has a countable number of possible values. (counts)

- **Qualitative (Categorical) Variables**
  [hair color, gender, ...]

**Measurement Scales**

- **Nominal:** consists of labels, names or categories.
- **Ordinal:** data that the order or rank is meaningful.
- **Interval:** numerical data that arithmetic operations are meaningful.
- **Ratio:** data that the ratio of two data is meaningful.
**Univariate**
data set consists of observations on a single variable.

**Birth Weights**
of 1448 new born babies in a study.

**Mothers’ Marital Status**
of 1448 new born babies in a study.

**Bivariate**
data set is a special multivariate case that consists of two variables.

**Examine correlation between smoking and lung cancer**

**Average annual temperature and the mortality index**
for a type of breast cancer in women in certain region of Europe

---

19

Univariate

data set consists of observations on a single variable.

20

Birth Weights

of 1448 new born babies in a study.

21

Mothers' Marital Status

of 1448 new born babies in a study.

22

Bivariate
data set is a special multivariate case that consists of two variables.

23

Examine correlation between smoking and lung cancer

24

Average annual temperature and the mortality index for a type of breast cancer in women in certain region of Europe
Response/Explanatory Variables

- Response (Dependent, Outcome) Variable
  - Lung Cancer, Mortality Index
- Explanatory (Independent, Predictor) Variable
  - Smoking, Average Temperature

Multivariate
data set consists of observations on
two or more than variables.

Precipitation, minimum temperatures, and
hospitalizations of females for viral pneumonia:
Sacramento and Yolo counties, California,

Producing Data
(Sampling Methods)

Data in Public Health:
- Vital Statistics and the Census
- Public Health Surveillance
- Survey
- Registries
- Epidemic Investigations
- Research
- Program Evaluations

Data Collection Techniques:
- Personal interview,
- Telephone interview,
- Self-administrated questionnaire (person, mail, internet, ...),
- Direct observation,
- Clinical Trial, Design of Experiment, ...
Preparing Data

- Retain the raw data source.
- Study log: data received, investigation, statisticians, data source.
- Machine readable data base (save original files).
- Data cleaning (logic check, correcting, clarifying).

Sampling Techniques
(To obtain representative sample)

- Simple random sampling: A sample obtained from a population in such a manner that all samples of the same size have equal likelihood of being selected.
- Stratified random sampling: Classify the population into at least two strata, then draw a sample from each.
- Cluster sampling: Divide the population into sections, randomly select few of those sections, and then choose all members in them.
- Systematic sample: Select every i-th member in the population according to their ID.
- Convenient sample: Use results that are already available or using data that is convenient to obtain.

Bias

- Response bias
  - Influenced response (Bribe, Mislead,...)
  - Measurement error (Poor instrument, ...)
  - Lies (Sensitive questions, ...)
  - ...

- Non-response bias
  - Shy, Conservative, Conflict of interest
  - ...

Type of Studies

- Observational Study: conditions to which subjects are exposed are not controlled by the investigator. (No attempt is made to control or influence the variables of interest.)

- Experimental Study: conditions to which subjects are exposed are controlled by the investigator. (Manipulated treatments are used to see their effect on other variables.)

Results from observing behavior and outcomes from the use of medicine for 200 randomly selected patients. (Patients chose their medicine)

<table>
<thead>
<tr>
<th>Hypertension</th>
<th>Treatment</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Drug A</td>
<td>44</td>
<td>56</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Drug B</td>
<td>29</td>
<td>71</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>73</td>
<td>127</td>
<td>200</td>
</tr>
</tbody>
</table>

Drug A: 44/100 = 44%
Drug B: 29/100 = 29%

Hypertension

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Below 65</th>
<th>65+</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Drug A</td>
<td>5</td>
<td>18</td>
</tr>
<tr>
<td>Drug B</td>
<td>17</td>
<td>60</td>
</tr>
</tbody>
</table>

* Older patients prefer Drug A

OR <65: Drug A: 5/23 = 22%
Drug B: 17/77 = 22%

OR 65+: Drug A: 39/77 = 51%
Drug B: 12/23 = 52%
Confounding Effect

- Variables, whether part of a study or not, are said to be confounded when their effects on the outcome cannot be distinguished from each other.
- Age may affect the reaction to drug and may also affect drug choosing decision.

Randomized Comparative Experiments

- **Completely Randomized Design**: A design of experiment in which all the experimental units are allocated at random among all the treatments.
  - Control Group, Placebo Group
  - Treatment Group

Descriptive Statistics

- Data Presentation
  - Grouping tables
  - Graphical summary
- Numerical Summary
  - Center
  - Dispersion (Spread)

Exploratory Data Analysis

Data Presentation

**What type of statistical technique is appropriate for Data Presentation?**
- Categorical variable?
- Quantitative variable?

Data Sheet (Raw data)

<table>
<thead>
<tr>
<th>ID</th>
<th>Height(in)</th>
<th>Weight(lb)</th>
<th>BirthMonth</th>
<th>Exp</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>135</td>
<td>4</td>
<td>H</td>
<td>F</td>
</tr>
<tr>
<td>2</td>
<td>63</td>
<td>119</td>
<td>9</td>
<td>H</td>
<td>F</td>
</tr>
<tr>
<td>3</td>
<td>72</td>
<td>175</td>
<td>11</td>
<td>T</td>
<td>M</td>
</tr>
<tr>
<td>4</td>
<td>60</td>
<td>106</td>
<td>9</td>
<td>H</td>
<td>F</td>
</tr>
<tr>
<td>5</td>
<td>65</td>
<td>135</td>
<td>8</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>6</td>
<td>72</td>
<td>170</td>
<td>10</td>
<td>H</td>
<td>M</td>
</tr>
<tr>
<td>7</td>
<td>64</td>
<td>180</td>
<td>8</td>
<td>H</td>
<td>F</td>
</tr>
<tr>
<td>8</td>
<td>71</td>
<td>205</td>
<td>10</td>
<td>H</td>
<td>M</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A complete list

<table>
<thead>
<tr>
<th>ID</th>
<th>Height</th>
<th>Weight</th>
<th>BirthMonth</th>
<th>Exp.</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>135</td>
<td>4</td>
<td>H</td>
<td>F</td>
</tr>
<tr>
<td>2</td>
<td>63</td>
<td>119</td>
<td>9</td>
<td>H</td>
<td>F</td>
</tr>
<tr>
<td>3</td>
<td>72</td>
<td>175</td>
<td>11</td>
<td>T</td>
<td>M</td>
</tr>
<tr>
<td>4</td>
<td>60</td>
<td>106</td>
<td>9</td>
<td>H</td>
<td>F</td>
</tr>
<tr>
<td>5</td>
<td>65</td>
<td>135</td>
<td>8</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>6</td>
<td>72</td>
<td>170</td>
<td>10</td>
<td>H</td>
<td>M</td>
</tr>
<tr>
<td>7</td>
<td>64</td>
<td>180</td>
<td>8</td>
<td>H</td>
<td>F</td>
</tr>
<tr>
<td>8</td>
<td>71</td>
<td>205</td>
<td>10</td>
<td>H</td>
<td>M</td>
</tr>
<tr>
<td>9</td>
<td>75</td>
<td>195</td>
<td>6</td>
<td>T</td>
<td>M</td>
</tr>
<tr>
<td>10</td>
<td>71</td>
<td>185</td>
<td>8</td>
<td>H</td>
<td>M</td>
</tr>
<tr>
<td>11</td>
<td>71</td>
<td>182</td>
<td>6</td>
<td>T</td>
<td>M</td>
</tr>
<tr>
<td>12</td>
<td>65</td>
<td>108</td>
<td>8</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>13</td>
<td>73</td>
<td>150</td>
<td>4</td>
<td>H</td>
<td>M</td>
</tr>
<tr>
<td>14</td>
<td>67</td>
<td>128</td>
<td>6</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>15</td>
<td>74</td>
<td>175</td>
<td>6</td>
<td>H</td>
<td>M</td>
</tr>
<tr>
<td>16</td>
<td>66</td>
<td>160</td>
<td>9</td>
<td>H</td>
<td>F</td>
</tr>
<tr>
<td>17</td>
<td>65</td>
<td>143</td>
<td>9</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>18</td>
<td>72</td>
<td>190</td>
<td>11</td>
<td>T</td>
<td>M</td>
</tr>
<tr>
<td>19</td>
<td>64</td>
<td>180</td>
<td>2</td>
<td>H</td>
<td>M</td>
</tr>
<tr>
<td>20</td>
<td>61</td>
<td>195</td>
<td>5</td>
<td>T</td>
<td>M</td>
</tr>
<tr>
<td>21</td>
<td>72</td>
<td>220</td>
<td>7</td>
<td>H</td>
<td>M</td>
</tr>
<tr>
<td>22</td>
<td>69</td>
<td>295</td>
<td>7</td>
<td>H</td>
<td>M</td>
</tr>
</tbody>
</table>

Frequency Table and Charts (One Categorical Variable)

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>9</td>
<td>9/22 = .409 = 40.9%</td>
</tr>
<tr>
<td>Male</td>
<td>13</td>
<td>13/22 = .591 = 59.1%</td>
</tr>
<tr>
<td>Total</td>
<td>22</td>
<td>100%</td>
</tr>
</tbody>
</table>

Pareto chart

How do you describe yourself

* Bars arranged according to their frequencies.

Grouping and Displaying Quantitative Data

Frequency Distribution Table

<table>
<thead>
<tr>
<th>Class</th>
<th>Tally</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>90 - &lt;110</td>
<td>I</td>
<td>1</td>
</tr>
<tr>
<td>110 - &lt;130</td>
<td>II</td>
<td>2</td>
</tr>
<tr>
<td>130 - &lt;150</td>
<td>III</td>
<td>3</td>
</tr>
<tr>
<td>150 - &lt;170</td>
<td></td>
<td></td>
</tr>
<tr>
<td>170 - &lt;190</td>
<td></td>
<td></td>
</tr>
<tr>
<td>190 - &lt;210</td>
<td></td>
<td></td>
</tr>
<tr>
<td>210 - &lt;230</td>
<td></td>
<td></td>
</tr>
<tr>
<td>230 - &lt;250</td>
<td></td>
<td></td>
</tr>
<tr>
<td>250 - &lt;270</td>
<td></td>
<td></td>
</tr>
<tr>
<td>270 - &lt;290</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

270 to less than 290
Frequency Distribution Table

Data: 135, 119, 175, 106, 135, 170, 180, 205, 195, …

Class Tally Frequency
90 - <110 II 2
110 - <130 II 2
130 - <150 III 3
150 - <170 II 2
170 - <190 II 2
190 - <210 III 4
210 - <230 I 1
230 - <250 0 0
250 - <270 0 0
270 - <290 1 1
Total 22 22

Frequency Distribution Table

(From data sheet)

Class Frequency Relative Freq. Cumulative R.F.
90 - <110 2 2/22 = .091 2/22
110 - <130 2 2/22 = .091 4/22
130 - <150 3 3/22 = .136 7/22
150 - <170 2 2/22 = .091 9/22
170 - <190 7 7/22 = .318 16/22
190 - <210 4 4/22 = .182 20/22
210 - <230 1 1/22 = .045 21/22
230 - <250 0 0/22 = .000 21/22
250 - <270 0 0/22 = .000 21/22
270 - <290 1 1/22 = .045 22/22
Total 22 1.000

Frequency Distribution Table

(From data sheet with different boundaries)

Class Frequency Relative Freq. Cumulative R.F.
90< - 110 2 2/22 = .091 2/22
110< - 130 2 2/22 = .091 4/22
130< - 150 4 4/22 = .182 8/22
150< - 170 2 2/22 = .091 10/22
170< - 190 7 7/22 = .318 17/22
190< - 210 3 3/22 = .136 20/22
210< - 230 1 1/22 = .045 21/22
230< - 250 0 0/22 = .000 21/22
250< - 270 0 0/22 = .000 21/22
270< - 290 1 1/22 = .045 22/22
Total 22 1.000

Histogram (SPSS)

Std. Dev = 40.93
Mean = 169.1
N = 22.00

Cumulative R. F. Histogram

Cumulative R. F. Polygon
(Ogive)
What to observe in Histograms?

- **Outliers**: observations that stand out from the rest for some reason.
- **Center**: the “middle” of the data.
- **Spread**: the range; the extent of the data; how far the values are from each other.
- **Shape**: distribution pattern. [Skewness, symmetry, uniform, Normal, ...]

Histogram

Create Polygon

Density Curve

Density Curve
**Stemplots** (or Stem-and-leaf plots)

- leading digits are called stems
- final digits are called leaves

**Example:**
(number of hysterectomies performed by 15 male doctors)
27, 50, 33, 25, 86, 25, 85, 31, 37, 44, 20, 36, 59, 34, 28

```
2 | 0 5 5 7 8
3 | 1 3 4 6 7
→ 4 | 4
5 | 0 9
6 |
7 |
8 | 5 6
```

**Ordered Stemplot**

**Example:**
(number of hysterectomies performed by 15 male doctors)
27, 50, 33, 25, 86, 25, 85, 31, 37, 44, 20, 36, 59, 34, 28
by 10 female doctors, the numbers are:
5, 7, 10, 14, 18, 19, 25, 29, 31, 33

```
(Male) (Female)
2 | 0 5 5 7 8 0 | 5 7
3 | 1 3 4 6 7 1 | 0 4 8 9
4 | 4 2 | 5 9
5 | 0 9 3 | 1 3
6 |
7 |
8 | 5 6
```

**Back-to-back stem-plot**

(Female) (Male)
9840 | 1 1
95 | 2 0 5 5 7 8
31 | 3 1 3 4 6 7
| 4 4
| 5 0 9
| 6 |
| 7 |
| 8 | 5 6

**Box Plot**

```
Height
```

Intro - 11
Examine Bivariate Data

Odds of smoker to have cancer: $\frac{20}{30} = \frac{6}{9}$
Odds of nonsmoker to have cancer: $\frac{5}{45} = \frac{1}{9}$
Odds Ratio = $\frac{\frac{6}{9}}{\frac{1}{9}} = 6$

Contingency Table

<table>
<thead>
<tr>
<th></th>
<th>Cancer</th>
<th>No cancer</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>20</td>
<td>30</td>
<td>50</td>
</tr>
<tr>
<td>Non-Smoker</td>
<td>5</td>
<td>45</td>
<td>50</td>
</tr>
<tr>
<td>Column Total</td>
<td>25</td>
<td>75</td>
<td>100</td>
</tr>
</tbody>
</table>

Odds of smoker to have cancer: $\frac{20}{30} = \frac{6}{9}$
Odds of nonsmoker to have cancer: $\frac{5}{45} = \frac{1}{9}$
Odds Ratio = $\frac{\frac{6}{9}}{\frac{1}{9}} = 6$

Two Categorical Variables

Cluster bar chart

Two Quantitative variables

Data:

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Mortality Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>34</td>
<td>52</td>
</tr>
<tr>
<td>40</td>
<td>68</td>
</tr>
<tr>
<td>42</td>
<td>63</td>
</tr>
<tr>
<td>42</td>
<td>83</td>
</tr>
<tr>
<td>43</td>
<td>72</td>
</tr>
<tr>
<td>44</td>
<td>81</td>
</tr>
<tr>
<td>45</td>
<td>89</td>
</tr>
<tr>
<td>46</td>
<td>77</td>
</tr>
<tr>
<td>47</td>
<td>88</td>
</tr>
<tr>
<td>48</td>
<td>94</td>
</tr>
<tr>
<td>49</td>
<td>86</td>
</tr>
<tr>
<td>50</td>
<td>95</td>
</tr>
<tr>
<td>51</td>
<td>105</td>
</tr>
<tr>
<td>51</td>
<td>100</td>
</tr>
<tr>
<td>52</td>
<td>102</td>
</tr>
</tbody>
</table>

A Categorical & A Quantitative Variables

Side-by-side Boxplot
Time Plot

<table>
<thead>
<tr>
<th>Year</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>15.2</td>
</tr>
<tr>
<td>1996</td>
<td>15.1</td>
</tr>
<tr>
<td>1997</td>
<td>14.9</td>
</tr>
<tr>
<td>1998</td>
<td>16.2</td>
</tr>
<tr>
<td>1999</td>
<td>14.3</td>
</tr>
<tr>
<td>2000</td>
<td>13.2</td>
</tr>
<tr>
<td>2001</td>
<td>13.5</td>
</tr>
</tbody>
</table>

Time Plot

Youngstown Homicide Rate by Year

<table>
<thead>
<tr>
<th>Year</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>120</td>
</tr>
<tr>
<td>1981</td>
<td>110</td>
</tr>
<tr>
<td>1982</td>
<td>100</td>
</tr>
<tr>
<td>1983</td>
<td>90</td>
</tr>
<tr>
<td>1984</td>
<td>80</td>
</tr>
<tr>
<td>1985</td>
<td>70</td>
</tr>
<tr>
<td>1986</td>
<td>60</td>
</tr>
<tr>
<td>1987</td>
<td>50</td>
</tr>
<tr>
<td>1988</td>
<td>40</td>
</tr>
<tr>
<td>1989</td>
<td>30</td>
</tr>
<tr>
<td>1990</td>
<td>20</td>
</tr>
<tr>
<td>1991</td>
<td>10</td>
</tr>
<tr>
<td>1992</td>
<td>0</td>
</tr>
</tbody>
</table>

Different Scales

Gender

<table>
<thead>
<tr>
<th>Gender</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>270</td>
</tr>
<tr>
<td>Female</td>
<td>260</td>
</tr>
</tbody>
</table>

Different Scales

Gender

<table>
<thead>
<tr>
<th>Gender</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>300</td>
</tr>
<tr>
<td>Female</td>
<td>250</td>
</tr>
</tbody>
</table>

Incorrect and Misleading Chart