AIU FOR 6520
Statistical Research Design & Methods in Forensic Psychology

Instructor: Ivo Dinov, Asst. Prof. of Statistics, Neurology, Psychology
AIU, UCLA, Winter 2003
http://www.stat.ucla.edu/~dinov/courses_students.html

Review of Research & Design I – Fall'02
- Intro to stats, vocabulary & intro to SPSS
- Displaying data
- Central tendency and variability
- Normal z-scores, standardized distribution
- Probability, Samples & Sampling error
- Type I and Type II errors; Power of a test
- Intro to hypothesis testing
- One sample tests & Two independent samples tests
- Two sample tests - dependent samples & Estimation
- Correlation and regression techniques
- Non-parametric statistical tests

Coverage of Research & Design II – Spring’03
- Applications of Central Limit Theorem, Law of Large Numbers.
- Design of studies and experiments.
- Fisher's F-Test & Analysis Of Variance (ANOVA, 1- or 2-way).
- Principle Component Analysis (PCA).
- $\chi^2$ (Chi-Square) Goodness-of-fit test.
- Multiple linear regression
- General Linear Model
- Bootstrapping and Resampling

Newtonial science vs. chaotic science
Science we encounter at schools deals with crisp certainties (e.g., prediction of planetary orbits, the periodic table as a descriptor of all elements, equations describing area, volume, velocity, position, etc.)
As soon as uncertainty comes in the picture it shakes the foundation of the deterministic science, because only probabilistic statements can be made in describing a phenomenon (e.g., roulette wheels, chaotic dynamic weather predictions, Geiger counter, earthquakes, etc.)
What is then science all about – describing absolutely certain events and laws alone, or describing more general phenomena in terms of their behavior and chance of occurring? Or may be both!
**Intro to stats, vocabulary & intro to SPSS**
- Central tendency and variability
- Normal z-scores, standardized distribution
- Probability, Samples & Sampling error
- Type I and Type II errors; Power of a test
- Intro to hypothesis testing
- One sample tests & Two independent samples tests
- Two sample tests - dependent samples & Estimation
- Correlation and regression techniques
- Non-parametric statistical tests

---

**Errors in Samples …**
- **Selection bias**: Sampled population is not a representative subgroup of the population really investigated.
- **Non-response bias**: If a particular subgroup of the population studied does not respond, the resulting responses may be skewed.
- **Question effects**: Survey questions may be slanted or loaded to influence the result of the sampling.
- **Target population**: Entire group of individuals, objects, units we study.
- **Study population**: A subset of the target population containing all “units” which could possibly be used in the study.
- **Sampling protocol**: Procedure used to select the sample.
- **Sample**: The subset of “units” about which we actually collect info.

---

**More definitions …**
- How could you implement the lottery method to randomly sample 10 students from a class of 2507? List all names, assign numbers 1-2507 to all students. Use a random-number generator to choose (10×) a number in range [0,2507]. Process students drawn.
- **Random or chance error**: The difference between the sample value and the true population value (e.g., 49% vs. 69%, in the above body overweight example).
- **Non-sampling errors**: (e.g., non-response bias) in the census may be considerable larger than in a comparable survey, since surveys are much smaller operations and easier to control.
- **Sampling errors**: Arising from a decision to use a sample rather than entire population.
- **Unbiased procedure/protocol**: (e.g., using the proportion of left-handers from a random sample to estimate the corresponding proportion in the population).
- **Cluster sampling**: A cluster of individuals/units are used as a sampling unit, rather than individuals.

---

**Variation in sample percentages**

**Poll**: Do you consider yourself overweight?

<table>
<thead>
<tr>
<th>Samples of 20 people</th>
<th>Target: True population percentage = 50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

We are getting closer to the population mean, as

<table>
<thead>
<tr>
<th>Sample percentage</th>
<th>Sample percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>70</td>
</tr>
<tr>
<td>80</td>
<td>90</td>
</tr>
</tbody>
</table>

**Figure 1.1.1** Comparing percentages from 10 different surveys each of 20 people with those from 10 surveys each of 500 people (all surveys from same population).

---

**More terminology …**
- **Census** — attempt to sample the entire population.
- **Parameter** — numerical characteristic of the population, e.g., income, age, etc. Often we want to estimate population parameters.
- **Statistic** — a numerical characteristic of the sample. *(Sample) statistic is used to estimate a corresponding population parameter.*
- **Why do we sample at random?** We draw “units” from the study population at random to avoid bias. Every subject in the study sample is equally likely to be selected. Also random-sampling allows us to calculate the likely size of the error in our sample estimates.
Questions ...

- Give an example where non-representative information from a survey may be useful. Non-representative info from surveys may be used to estimate parameters of the actual sub-population which is represented by the sample. E.g., Only about 2% of dissatisfied customers complain (most just avoid using the services), these are the most-vocal reps. So, we can not make valid conclusions about the stereotype of the dissatisfied customer, but we can use this info to track down changes in levels of complaints over years.
- Why is it important to take a pilot survey?
- Give an example of an unsatisfactory question in a questionnaire. (In a telephone study: What time is it? Do we mean Eastern/Central/Mountain/Pacific?)

Questions ...

- Random allocation – randomly assigning treatments to units, leads to representative sample only if we have large # experimental units.
- Completely randomized design - the simplest experimental design, allows comparisons that are unbiased (not necessarily fair). Randomly allocate treatments to all experimental units, so that every treatment is applied to the same number of units. E.g., if we have 12 units and 3 treatments, and we study treatment efficacy, we randomly assign each of the 3 treatments to 4 units exactly.
- Blocking - grouping units into blocks of similar units for making treatment-effect comparisons only within individual groups. E.g., Study of human life expectancy perhaps income is clearly a factor, we can have high- and low-income blocks and compare, say, gender differences within these blocks separately.

Questions ...

- Why should we try to “blind” the investigator in an experiment?
- Why should we try to “blind” human experimental subjects?
- The basic rule of experimenter: "Block what you can and randomize what you cannot."

Questions ...

- What is the difference between a designed experiment and an observational study? (no control of the design in observational studies)
- Can you conclude causation from an observational study? Why or why not? (not in general!)
- How do we try to investigate causation questions using observational studies? In a smoking-lung-cancer study: try to divide all subjects, in the obs. study, into groups with equal, or very similar levels of all other factors (age, stress, income, etc.) - i.e. control for all outside factors. If rate of lung-cancer is still still higher in smokers we get a stronger evidence of causality.
- What is the idea of controlling for a variable, and why is it used? Effects of this variable in the treatment/control groups are similar.
- Epidemiology – science of using statistical methods to find causes or risk factors for diseases.

The Subject of Statistics

Statistics is concerned with the process of finding out about the world and how it operates -
- in the face of variation and uncertainty
- by collecting and then making sense (interpreting) of data.
Displaying data

- Intro to stats, vocabulary & intro to SPSS
- **Displaying data**
  - Central tendency and variability
  - Normal z-scores, standardized distribution
  - Probability, Samples & Sampling error
  - Type I and Type II errors; Power of a test
  - Intro to hypothesis testing
  - One sample tests & Two independent samples tests
  - Two sample tests - dependent samples & Estimation
  - Correlation and regression techniques
  - Non-parametric statistical tests

Types of variable

- **Quantitative** variables are *measurements* and counts
  - Variables with **few repeated values** are treated as **continuous**.
  - Variables with **many repeated values** are treated as **discrete**.
- **Qualitative** variables (a.k.a. factors or class-variables) describe **group membership**

Distinguishing between types of variable

<table>
<thead>
<tr>
<th>Types of Variables</th>
<th>Quantitative (measurements and counts)</th>
<th>Qualitative (define groups)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous</td>
<td>Discrete (few repeated values)</td>
<td>Categorical (no idea of order)</td>
</tr>
<tr>
<td></td>
<td>Discrete (many repeated values)</td>
<td>Ordinal (fall in natural order)</td>
</tr>
</tbody>
</table>

Questions …

- What is the difference between quantitative and qualitative variables?
- What is the difference between a discrete variable and a continuous variable?
- Name two ways in which observations on qualitative variables can be stored on a computer. (strings/indexes)
- When would you treat a discrete random variable as though it were a continuous random variable?
  - Can you give an example? ($34.45, bill)

Different graphs of the same set of numbers – percentages of the world’s gold production in 1991

(a) Bar graph
(b) Pie chart
(c) Segmented bar

Questions …

- For what two purposes are tables of numbers presented? (convey information about trends in the data, detailed analysis)
- When should you round numbers, and when should you preserve full accuracy?
- How should you arrange the numbers you are most interested in comparing? (Arrange numbers you want to compare in columns, not rows. Provide written/verbal summaries/footnotes. Show row/column averages.)
- Should a table be left to tell its own story?
The dot plot

```
3 4 5 6 7 8
```

Dot plot.

Atypical obs.

**Figure 2.3.1** Dot plot. From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

**Figure 2.3.2** Dot plot showing special features.

**Figure 2.3.3** Grading of a university course.

**Figure 2.3.4** Dot plot with and without a scale break.

**Figure 2.3.5** Forecast of percent growth in GDP for 1990 for some South-East Asian and Pacific countries.

**Example of a stem-and-leaf plot**

<table>
<thead>
<tr>
<th>Stem</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaf</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Values 52, 54 and their frequencies

Stem-plot of the 45 obs’s of the Ejection variable in the Heart Attack data table.

**Traffic death-rates data**

<table>
<thead>
<tr>
<th>Countries</th>
<th>Traffic Death-Rates (per 100,000 Population)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>17.4</td>
</tr>
<tr>
<td>Austria</td>
<td>19.8</td>
</tr>
<tr>
<td>Belgium</td>
<td>12.5</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>15.8</td>
</tr>
<tr>
<td>Czechoslovakia</td>
<td>13.0</td>
</tr>
<tr>
<td>Denmark</td>
<td>20.0</td>
</tr>
<tr>
<td>Finland</td>
<td>11.6</td>
</tr>
<tr>
<td>France</td>
<td>12.0</td>
</tr>
<tr>
<td>East Germany</td>
<td>13.1</td>
</tr>
<tr>
<td>West Germany</td>
<td>21.1</td>
</tr>
<tr>
<td>Greece</td>
<td>15.3</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>9.4</td>
</tr>
<tr>
<td>Hungary</td>
<td>17.1</td>
</tr>
<tr>
<td>Ireland</td>
<td>10.3</td>
</tr>
<tr>
<td>Japan</td>
<td>10.4</td>
</tr>
<tr>
<td>Kuwait</td>
<td>26.8</td>
</tr>
<tr>
<td>Netherlands</td>
<td>11.3</td>
</tr>
<tr>
<td>New Zealand</td>
<td>20.1</td>
</tr>
<tr>
<td>Norway</td>
<td>10.5</td>
</tr>
<tr>
<td>Poland</td>
<td>12.6</td>
</tr>
<tr>
<td>Portugal</td>
<td>25.6</td>
</tr>
<tr>
<td>Singapore</td>
<td>9.8</td>
</tr>
<tr>
<td>South Korea</td>
<td>12.6</td>
</tr>
<tr>
<td>South Africa</td>
<td>12.9</td>
</tr>
<tr>
<td>United States</td>
<td>10.5</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>10.1</td>
</tr>
</tbody>
</table>

Data for 1983, 1984 or 1985 depending on the country (prior to reunification of Germany)

Source: Hutchinson [1987, page 3].
FOR6520, AIU, Ivo Dinov

Slide 31

Round-off
Units: 17 | 4 = 17.4 deaths per 100,000

FIGURE 2.3.7
Two stem-and-leaf plots for the traffic deaths data

(a)  Histogram of the female coyote-lengths data.

(b)  Stem-and-leaf plot rotated

Table 2.3.3 Frequency Table for Female Coyote Lengths

<table>
<thead>
<tr>
<th>Class Interval</th>
<th>Tally</th>
<th>Frequency</th>
<th>Stem-and-leaf plot</th>
</tr>
</thead>
<tbody>
<tr>
<td>70-75</td>
<td></td>
<td>2</td>
<td>71 4</td>
</tr>
<tr>
<td>75-80</td>
<td>0</td>
<td>4</td>
<td>80 1 4 4</td>
</tr>
<tr>
<td>80-85</td>
<td>6</td>
<td>6</td>
<td>85 5 6 7 7 7 7</td>
</tr>
<tr>
<td>85-90</td>
<td>12</td>
<td>6</td>
<td>90 0 1 2 3 4 4 4</td>
</tr>
<tr>
<td>90-95</td>
<td>13</td>
<td>5</td>
<td>95 6 7 8</td>
</tr>
<tr>
<td>95-100</td>
<td>5</td>
<td>1</td>
<td>100 8</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>40</td>
<td></td>
</tr>
</tbody>
</table>

Body length (cm)

Females

<table>
<thead>
<tr>
<th>Length (cm)</th>
<th>N</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>70-75</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>75-80</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>80-85</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>85-90</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>90-95</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>95-100</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>100-105</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td></td>
</tr>
</tbody>
</table>

Males

<table>
<thead>
<tr>
<th>Length (cm)</th>
<th>N</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>70-75</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>75-80</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>80-85</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>85-90</td>
<td>8</td>
<td>5 5 5 6 7 7 7 7 8 8 9</td>
</tr>
<tr>
<td>90-95</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>95-100</td>
<td>8</td>
<td>6 7 8 8 8 8 8 8 9 9 9 9</td>
</tr>
<tr>
<td>100-105</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td></td>
</tr>
</tbody>
</table>

Coyotes captured in Nova Scotia, Canada. Data courtesy of Dr Vera Eastwood.

Table 2.3.2 Coyote Lengths Data (cm)

<table>
<thead>
<tr>
<th>Length (cm)</th>
<th>Females</th>
<th>Males</th>
</tr>
</thead>
<tbody>
<tr>
<td>93.0</td>
<td>97.0</td>
<td></td>
</tr>
<tr>
<td>97.0</td>
<td>92.0</td>
<td>94.1</td>
</tr>
<tr>
<td>92.0</td>
<td>101.6</td>
<td></td>
</tr>
<tr>
<td>93.0</td>
<td>93.0</td>
<td>90.8</td>
</tr>
<tr>
<td>84.5</td>
<td>91.5</td>
<td>101.0</td>
</tr>
<tr>
<td>90.5</td>
<td>86.4</td>
<td>94.5</td>
</tr>
<tr>
<td>91.4</td>
<td>94.5</td>
<td>88.5</td>
</tr>
<tr>
<td>83.5</td>
<td>88.5</td>
<td>86.5</td>
</tr>
<tr>
<td>88.0</td>
<td>91.7</td>
<td>90.0</td>
</tr>
<tr>
<td>91.7</td>
<td>91.0</td>
<td>91.0</td>
</tr>
</tbody>
</table>

Questions …

What advantages does a stem-and-leaf plot have over a histogram? (S&L Plots return info on individual values, quick to produce by hand, provide data sorting mechanisms. But, histograms are more attractive and more understandable).

The shape of a histogram can be quite drastically altered by choosing different class-interval boundaries. What type of plot does not have this problem? (density trace) What other factor affects the shape of a histogram? (bin-size)

What was another reason given for plotting data on a variable, apart from interest in how the data on that variable behaves? (shows features, cluster/gaps, outliers; as well as trends.)

Interpreting Stem-plots and Histograms
Interpreting Stem-plots and Histograms

(j) Spike in pattern
(k) Outliers
(l) Truncation plus outlier

Features to look for in histograms and stem-and-leaf plots.

Fascinations with histograms – Histogram of heights using the actual people

Subjects are university genetics students, females in white and males in dark tops.

Skewness & Kurtosis

What do we mean by symmetry and positive and negative skewness? Kurtosis?

Skewness = \[ \frac{\sum (Y_i - \bar{Y})^3}{(N-1)SD^3} \]
Kurtosis = \[ \frac{\sum (Y_i - \bar{Y})^4}{(N-1)SD^4} - 3 \]

Skewness is linearly invariant Sk(aX+b)=Sk(X)
Skewness is a measure of unsymmetry
Kurtosis is (also linearly invariant) a measure of flatness
Both are used to quantify departures from StdNormal
Skewness(StdNorm)=0; Kurtosis(StdNorm)=3

Descriptive statistics from computer programs like STATA

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>TrMean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>45</td>
<td>50.133</td>
<td>51.000</td>
<td>50.366</td>
<td>6.092</td>
<td>0.908</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Q1</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>36.000</td>
<td>59.000</td>
<td>46.500</td>
<td>56.000</td>
</tr>
</tbody>
</table>

Box plot compared to dot plot

Data

Construction of a box plot
TABLE 2.5.1  Word Lengths for the First 100 Words on a Randomly Chosen Page

<table>
<thead>
<tr>
<th>Value u</th>
<th>Frequency f</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
</tr>
<tr>
<td>4</td>
<td>22</td>
</tr>
<tr>
<td>5</td>
<td>13</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
</tr>
</tbody>
</table>

For data visualization:

- Frequency Table
- Histogram of random numbers generated
- Bar graph for species data.

Describing data with pictures and two numbers

- Random Number generation: frequency histogram
- Descriptive statistics
  - Central tendency (Mode, Median, Mean)
  - Variability (Variance, Standard deviation)

Central tendency and variability

- Intro to stats, vocabulary & intro to SPSS
- Displaying data
- Central tendency and variability
  - Normal z-scores, standardized distribution
  - Probability, Samples & Sampling error
  - Type I and Type II errors; Power of a test
  - Intro to hypothesis testing
  - One sample tests & Two independent samples tests
  - Two sample tests - dependent samples & Estimation
  - Correlation and regression techniques
  - Non-parametric statistical tests

Random number generation [5:316]: ascending order

<table>
<thead>
<tr>
<th>Sx ID number</th>
<th>Sx ID number</th>
<th>Sx ID number</th>
<th>Sx ID number</th>
<th>Sx ID number</th>
<th>Sx ID number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>11</td>
<td>42</td>
<td>27</td>
<td>75</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>19</td>
<td>42</td>
<td>28</td>
<td>77</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>14</td>
<td>59</td>
<td>24</td>
<td>81</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>16</td>
<td>58</td>
<td>28</td>
<td>83</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
<td>15</td>
<td>57</td>
<td>25</td>
<td>83</td>
</tr>
<tr>
<td>6</td>
<td>33</td>
<td>16</td>
<td>58</td>
<td>26</td>
<td>83</td>
</tr>
<tr>
<td>7</td>
<td>33</td>
<td>17</td>
<td>58</td>
<td>27</td>
<td>83</td>
</tr>
<tr>
<td>8</td>
<td>38</td>
<td>17</td>
<td>68</td>
<td>27</td>
<td>97</td>
</tr>
<tr>
<td>9</td>
<td>42</td>
<td>19</td>
<td>60</td>
<td>29</td>
<td>101</td>
</tr>
<tr>
<td>10</td>
<td>42</td>
<td>20</td>
<td>68</td>
<td>30</td>
<td>102</td>
</tr>
</tbody>
</table>

Histogram of random numbers generated

Notes:
1. Histogram allows you to see distribution – estimate middle, spread.
2. Sacrifice some information (within intervals) to gain this perspective
3. How can we further reduce this information to a single number for middle and another for spread?
Central tendency: the middle in a single number

- **Mode**: The most frequent score in the distribution.
- **Median**: The centermost score if there are an odd number of scores or the average of the two centermost scores if there are an even number of scores.
- **Mean**: The sum of the scores divided by the number of scores.

---

**Beware of inappropriate averaging**

*Suggested by a 1977 cartoon in The New Yorker magazine by Dana Fradon.*

---

**Five number summary**

The five-number summary = (Min, Q1, Med, Q3, Max)

---

**Inter-quartile Range**

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

---

**Calculate the measures of central tendency**

---

**Number of licks: sorted from few to many**

Mode: 42
Median = 91
Arithmetic mean

\[ \frac{\sum x}{N} = \text{Sample mean} = \bar{X} (\text{pronounced "Xbar"}) \]

\[ \frac{\sum x}{N} = \text{Population mean} = \mu (\text{pronounced "mew"}) \]

<table>
<thead>
<tr>
<th>Ss ID number</th>
<th>X</th>
<th>X1</th>
<th>X1+2</th>
<th>X1+100</th>
<th>X2</th>
<th>X2</th>
<th>X2-100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ss ID number</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ss ID number</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ss ID number</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ss ID number</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ss ID number</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Median = 91; Mean = 111.8

Neat things about the mean

- If you add/subtract a constant to/from each score, you change the mean by adding/subtracting the constant to/from it.

- If you multiply/divide each score by a constant you change the mean by multiplying/dividing it by the constant.
Neat things about the mean

- If you add/subtract a constant to/from each score, you change the mean by adding/subtracting the constant to/from it.
- If you multiply/divide each score by a constant you change the mean by multiplying/dividing it by the constant.
- Summed deviations from the mean = 0, or $\sum (x_i - \bar{x}) = 0$

<table>
<thead>
<tr>
<th>$X_i$</th>
<th>$x_i - \bar{x}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

Sum => 6 0
Mean => 2 0

Neat things about the mean

- Sum of squared deviations from the mean (SS) is minimized.
- $\sum (x_i - \bar{x})^2$ = minimum
- $\sum x^2 - (\sum x)^2/N = $ minimum

<table>
<thead>
<tr>
<th>$X_i$</th>
<th>$x_i - \bar{x}$</th>
<th>$(x_i - \bar{x})^2$</th>
<th>$(x_i - 0)^2$</th>
<th>$(x_i - 3)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td>0</td>
</tr>
</tbody>
</table>

Sum => 6 0 2 14 5
Mean => 2 0

Height of 26 women in inches (real data)

- Number Gender Height (X)

<table>
<thead>
<tr>
<th>Number</th>
<th>Gender</th>
<th>Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>61</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>61</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>61</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>62</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>62</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>62</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>62</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>63</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>63</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>64</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>64</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>64</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>64</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>64</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>64</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>65</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>65</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>65</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>66</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>66</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>66</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>66</td>
</tr>
<tr>
<td>23</td>
<td>1</td>
<td>66</td>
</tr>
<tr>
<td>24</td>
<td>1</td>
<td>67</td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>69</td>
</tr>
<tr>
<td>26</td>
<td>1</td>
<td>70</td>
</tr>
</tbody>
</table>

Sum == 1672
Mode = 64
Median = 64
Mean = 64.3

Height of 26 women in inches (real data)

- Frequency histogram for women in 214

- Mode = 64
- Median = 64
- Mean = 64.3

Height of 26 women in inches (real data)

- Frequency histogram for women in 214

- Mode = 64
- Median = 64
- Mean = 64.3

Height of 26 women in inches (real data)

- Frequency histogram for women in 214

- Mode = 64
- Median = 64
- Mean = 64.3

Height of 26 women in inches (real data)

- Frequency histogram for women in 214

- Mode = 64
- Median = 64
- Mean = 64.3
Questions about measures of central tendency?

- Why is the mean our preferred measure of central tendency?
  - Adjusted for the number of scores.
  - Takes into account the numerical “weight” of each score.
    As scores of greater magnitude are added, the mean increases.
    As scores of lesser magnitude are added, the mean decreases.
  - Sum of squared deviations from the mean (SS) is minimized.
    SS is the square of the sum of each score’s difference from the mean.
    \[ \sum (x-\bar{x})^2 \]
    \[ \sum x^2 - (\sum x)^2/N = \text{minimum} \]

Variability

- Not only interested in a distribution’s middle.
- Also interested in its spread (or variability).
- Define distributions by:
  - Central tendency
  - Variability
- How can we describe variability with a single number?

Variability

- How do these distributions differ?
  - One is more spread out (greater variability) than another.
  - Spread out around what?
Describe variability around the mean with one number.

- Want to adjust for the number of scores.
- Take into account the numerical "weight" of each score.
  - As scores are farther from the mean, the index of variability should increase
  - As scores are closer to the mean, the index of variability should decrease
- Suppose we measured each score’s distance from its mean, and then used the average distance as our measure?
  - Using the average distance will adjust for the number of scores.
  - Measuring the distance from the mean should tell us how spread out each score is relative to the mean.

Try measure of variability with some simple numbers

- $X$ (a population) = \{2, 4, 6\}
- $\mu = 4$
- What is the average distance from the mean?
  - How far is $X_1$ away from the mean? $(2 - 4 = 2)$
  - How far is $X_2$ away from the mean? $(4 - 4 = 0)$
  - How far is $X_3$ away from the mean? $(6 - 4 = 2)$
- What is the sum of the distance from the mean? $\sum(x - \mu) = 6$
- How can we use the distance from the mean as a measure of variability?

Is 2.7 the average distance each score is from its mean?

- $X = \{2, 4, 6\}$
- In absolute terms:
  - $X_1$ is 2 away from the mean
  - $X_2$ is 0 away from the mean
  - $X_3$ is 2 away from the mean
- Shouldn’t average distance be about 4/3 or 1.33?
- Why is the variance ($\sigma^2$) as a measure of the average distance of each score from its mean so much bigger than our intuition (that is, why is $\sigma^2 = 2.7$ when the average distance from the mean is obviously closer to 1.3?"

BECAUSE WE SQUARED ALL THE DEVIATIONS!

How do we “unsquare” the variance?

- Unsquare the variance ($\sigma^2$) by taking the square root of it:
  \[ \sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}} \]
- Why? To get back to the original scale of $X$.
  \[ \sqrt{2.7} = 1.63 \]
  much closer to our intuitively derived 1.3
Generally

- **Population variance** (sigma squared, or \( \sigma^2 \)) is the average of the squared deviations from the mean:
  \[
  \frac{\sum(x-\mu)^2}{N}
  \]
  Note: also written \( \frac{SS}{N} \) or \( \frac{\sum x^2 - (\sum x)^2}{N} \)

- **Population standard deviation** (sigma, or \( \sigma \)) is the square root of the average of the squared deviations from the mean:
  \[
  \frac{\sqrt{\sum(x-\mu)^2}}{N}
  \]
  Note: also written \( \frac{SS}{N} \) or \( \frac{\sqrt{\sum x^2 - (\sum x)^2}}{N} \)

Questions?

- We know what happens to the mean when we add or subtract a constant to/from all the scores, but what happens to the variance and standard deviation?
- We know what happens to the mean when we multiply or divide all the scores by a constant, but what happens to the variance and standard deviation?

Remember these?

- More variability
  - Frequency histogram for women in 214
  - \( \bar{X} = 64.3 \)

- Less Variability
  - Another frequency histogram
  - \( \bar{X} = 64.3 \)

Questions?

- We know what happens to the mean when we add/subtract a constant to/from all the scores, but what happens to \( s^2 \) and \( s \)?
- We know what happens to the mean when we multiply or divide all the scores by a constant, but what happens to \( s^2 \) and \( s \)?
Remember these?

<table>
<thead>
<tr>
<th>More variability</th>
<th>Less Variability</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Frequency histogram for men in 214" /></td>
<td><img src="image" alt="Frequency histogram for women in 214" /></td>
</tr>
<tr>
<td>$\bar{X} = 64.3, S = 2.3$</td>
<td>$\bar{X} = 64.3, S = 0.9$</td>
</tr>
</tbody>
</table>

What have we learned?

- Mean is preferred measure of central tendency:
  \[ \frac{\sum x}{N} = \text{Sample mean} = \bar{X}, \text{more commonly } \frac{\sum x}{N} \]
- Standard deviation is preferred measure of variability:
  \[ \frac{\sum (x - \bar{X})^2}{N-1} = \text{Sample standard deviation (s), can also be written } SS/N-1 \]

What have we learned?

- Mean is preferred measure of central tendency:
  \[ \frac{\sum x}{N} = \text{Sample mean} = \bar{X}, \text{more commonly } \frac{\sum x}{N} \]
- Standard deviation is preferred measure of variability:
  \[ \frac{\sum (x - \bar{X})^2}{N-1} = \text{Sample standard deviation (s), can also be written } SS/N-1 \]

What have we learned?

- Mean is preferred measure of central tendency:
  \[ \frac{\sum x}{N} = \text{Sample mean} = \bar{X}, \text{more commonly } \frac{\sum x}{N} \]
- Standard deviation is preferred measure of variability:
  \[ \frac{\sum (x - \bar{X})^2}{N-1} = \text{Sample standard deviation (s), can also be written } SS/N-1 \]

What have we learned?

- Mean is preferred measure of central tendency:
  \[ \frac{\sum x}{N} = \text{Sample mean} = \bar{X}, \text{more commonly } \frac{\sum x}{N} \]
- Standard deviation is preferred measure of variability:
  \[ \frac{\sum (x - \bar{X})^2}{N-1} = \text{Sample standard deviation (s), can also be written } SS/N-1 \]

What have we learned?

- Mean is preferred measure of central tendency:
  \[ \frac{\sum x}{N} = \text{Sample mean} = \bar{X}, \text{more commonly } \frac{\sum x}{N} \]
- Standard deviation is preferred measure of variability:
  \[ \frac{\sum (x - \bar{X})^2}{N-1} = \text{Sample standard deviation (s), can also be written } SS/N-1 \]
Basic method for obtaining probabilities

- Sketch a Normal curve, marking the mean and other values of interest.
- Shade the area under the curve that gives the desired probability.
- Devise a way of getting the desired area from lower-tail areas.
- Obtain component lower-tail probabilities from a computer program.

(a) Computing $\Pr(160 < X < 180)$

We want $\Pr(160 < X < 180) = \Pr(X \leq 180) - \Pr(X \leq 160)$.

Obtaining an upper-tail probability

We want $\Pr(X > 25)$. Programs supply $\Pr(X \leq 25) = 0.2874$.

Since total area under curve = 1, $\Pr(X > 25) = 1 - \Pr(X \leq 25)$.

Generally, $\Pr(X > x) = 1 - \Pr(X \leq x)$.

The inverse problem – percentiles/quantiles

80% of people have height below the 80th percentile. This is EQ to saying there’s 80% chance that a random observation from the distribution will fall below the 80th percentile.

The inverse problem is what is the height for the 80th percentile/quantile? So far we studied given the height value what’s the corresponding percentile?

Review

- What is meant by the 60th percentile of heights?
- What is the difference between a percentile and a quantile? (percentile used in expressing results in %, whereas quantiles used to express results in term of probabilities)
- The lower quartile, median and upper quartile of a distribution correspond to special percentiles. What are they? Express in terms of quantiles. (25%, 50%, 75%)
- Quantiles are sometimes called inverse cumulative probabilities. Why?
The standard normal curve is described by the equation:

\[ y = \frac{e^{-\frac{x^2}{2}}}{{\sqrt{2\pi}}} \]

Where remember, the natural number \( e \approx 2.7182 \ldots \)

We say: \( X \sim \text{Normal}(\mu, \sigma) \), or simply \( X \sim \text{N}(\mu, \sigma) \)

The general normal curve is defined by:
- Where \( \mu \) is the average (of the symmetric) normal curve, and \( \sigma \) is the standard deviation (spread of the distribution).
- Why worry about a standard and general normal curves?
- How to convert between the two curves?

The \( z \)-score of \( x \) is the number of standard deviations \( x \) is from the mean. (Body-Mass-Index, BMI)

<table>
<thead>
<tr>
<th>TABLE 6.3.1 Examples of ( z )-Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x )</td>
</tr>
<tr>
<td>Male BMI values (kg/m(^2))</td>
</tr>
<tr>
<td>25</td>
</tr>
<tr>
<td>35</td>
</tr>
<tr>
<td>Female heights (cm)</td>
</tr>
<tr>
<td>155</td>
</tr>
<tr>
<td>180</td>
</tr>
</tbody>
</table>

Male BMI values: \( \mu = 27.3 \), \( \sigma = 4.1 \) \nFemale heights: \( \mu = 162.7 \), \( \sigma = 6.2 \)

Which ones of these are unusually large/small/away from the mean?
Working in standard units

Obtain $z$ from program
(Program returns 1.6449)

Review

Obtain $z$ from program
(Program returns 1.6449)

Summary

Continuous Variables and Density Curves

- There are no gaps between the values a continuous random variable can take.
- Random observations arise in two main ways: (i) by sampling populations; and (ii) by observing processes.

The density curve

- The probability distribution of a continuous variable is represented by a density curve.
  - Probabilities are represented by areas under the curve.
  - The probability that a random observation falls between $a$ and $b$ is equal to the area under the density curve between $a$ and $b$.
  - The total area under the curve equals 1.
  - The population (or distribution) mean $\mu_X = E(X)$, is where the density curve balances.
  - When we calculate probabilities for a continuous random variable, it does not matter whether interval endpoints are included or excluded.

For any random variable $X$

- $E(aX+b) = a \ E(X) + b$ and $SD(aX+b) = |a \ SD(X)|$
The Normal distribution

\[ X \sim \text{Normal}(\mu, \sigma) \]

Features of the Normal density curve:
- The curve is a symmetric bell-shape centered at \( \mu \).
- The standard deviation \( \sigma \) governs the spread.
  - 68.3% of the probability lies within 1 standard deviation of the mean
  - 95.4% within 2 standard deviations
  - 99.7% within 3 standard deviations

Probabilities

- Computer programs provide lower-tail (or cumulative) probabilities of the form \( \Pr(X \leq x) \)
  - We give the program the \( x \)-value; it gives us the probability.
- Computer programs also provide inverse lower-tail probabilities (or quantiles)
  - We give the program the probability; it gives us the \( x \)-value.
- When calculating probabilities, we shade the desired area under the curve and then devise a way of obtaining it via lower-tail probabilities.

Standard Units

The \( z \)-score of a value \( a \) is ....
- the number of standard deviations \( a \) is away from the mean
- positive if \( a \) is above the mean and negative if \( a \) is below the mean.

The standard Normal distribution has \( \mu = 0 \) and \( \sigma = 1 \).
- We usually use \( Z \) to represent a random variable with a standard Normal distribution.

Ranges, extremes and \( z \)-scores

Central ranges:
- \( \Pr(-z \leq Z \leq z) \) is the same as the probability that a random observation from an arbitrary Normal distribution falls within \( z \) SDs either side of the mean.

Extremes:
- \( \Pr(Z \geq z) \) is the same as the probability that a random observation from an arbitrary Normal distribution falls more than \( z \) standard deviations above the mean.
- \( \Pr(Z \leq -z) \) is the same as the probability that a random observation from an arbitrary Normal distribution falls more than \( z \) standard deviations below the mean.

Combining Random Quantities

Variation and independence:
- No two animals, organisms, natural or man-made objects are ever identical.
- There is always variation. The only question is whether it is large enough to have a practical impact on what you are trying to achieve.
- Variation in component parts leads to even greater variation in the whole.

Independence

We model variables as being independent ....
- if we think they relate to physically independent processes
- and if we have no data that suggests they are related.

Both sums and differences of independent random variables are more variable than any of the component random variables.
The mean height is 64 in and the standard deviation is 2 in. What percentage of the incoming recruits will be trained to operate armored combat vehicles (tanks)?

For independent random variables, (cf. Pythagorean theorem),
\[ SD(X + Y) = SD(X) + SD(Y) \]
\[ E(X + Y) = E(X) + E(Y) \]
[Aside: Sums and differences of independent Normally distributed random variables are also Normally distributed.]

For independent random variables, (cf. Pythagorean theorem),
\[ SD(X + Y) = SD(X) + SD(Y) \]
\[ E(X + Y) = E(X) + E(Y) \]
[Aside: Sums and differences of independent Normally distributed random variables are also Normally distributed.]

Areas under Standard Normal Curve – Normal Approximation, Scottish Army Recruits

- The mean height is 64 in and the standard deviation is 2 in.
  - Only recruits shorter than 65.5 in will be trained for tank operation.
  - About what percentage of the recruits will have no restrictions on duties?

- Recruits within \( \frac{1}{2} \) standard deviations of the mean will have no restrictions on training/duties.

- \( X \sim N(64, 4) \)
  - Percentage is 77.34%

- \( X \sim N(64, 4) \)
  - Percentage is 38.30%

Percentiles for Standard Normal Curve

- When the histogram of the observed process follows the normal curve Normal Tables (of any type, as described before) may be used to estimate percentiles. The \( N \)-th percentile of a distribution is \( P \) if \( N \% \) of the population observations are less than or equal to \( P \).

- Example, suppose the Math-part SAT scores of newly admitted freshmen at UCLA averaged 535 (out of [200:800]) and the SD was 100. Estimate the 95 percentile for the score distribution.

- Solution:
  \[ Z = \frac{X - \mu}{\sigma} = \frac{700 - 535}{100} = 1.65 \]
  \[ Z = 1.65 \]

Summary

1. The Standard Normal curve is symmetric w.r.t. the origin (0,0) and the total area under the curve is 100% (1 unit).
2. Std units indicate how many SD’s is a value below (-)/above (+) the mean.
3. Many histograms have roughly the shape of the normal curve (bell-shape).
4. If a list of numbers follows the normal curve the percentage of entries falling within each interval is estimated by: 1. Converting the interval to StdUnits and, 2. Computing the corresponding area under the normal curve (Normal approximation).
5. A histogram which follows the normal curve may be reconstructed just from \((\mu, \sigma^2)\), mean and variance=std dev^2.
6. Any histogram can be summarized using percentiles.
Example – work out in your notebooks

1. Compute the chance a random observation from a distribution (symmetric, bell-shaped, unimodal) with \( m = 75 \) and \( SD = 12 \) falls within the range \([53 : 71]\).

2. \( (53-75)/12 = -11/6 = -1.83 \) Std units
3. \( (71-75)/12 = -0.33(3) \) Std units
4. Area\([53:71]\) =
5. \( (\text{SN_area}\[-1.83:1.83\] – \text{SN_area}\[-0.33:0.33\])/2 \)
6. \( = (93\% - 25\%)/2 = 34\% \)
7. Compute the 90th percentile for the same data:
8. \( b+a=b=100\% \quad a=80\% \rightarrow A=0.8 \)
9. \( a+b=90\% \quad b=10\% \quad Z=1.3 \) SU
10. \( 90\% P = \sigma \sqrt[2]{1.3} + \mu = 12 \times 1.3 + 75 = 90.6 \)

Check Work

Should it be <50\% or >50\%?

General Normal Curve

The general normal curve is defined by:
- Where \( \mu \) is the average (of the symmetric) normal curve, and \( \sigma \) is the standard deviation (spread of the distribution).

- Why worry about a standard and general normal curve?
- How to convert between the two curves?

Areas under Standard Normal Curve

- Many histograms are similar in shape to the standard normal curve. For example, persons height. The height of all incoming female army recruits is measured for custom training and assignment purposes (e.g., very tall people are inappropriate for constricted space positions, and very short people may be disadvantages in certain other situations). The mean height is computed to be 64 in and the standard deviation is 2 in. Only recruits shorter than 65.5 in will be trained for tank operation and recruits within \( \frac{1}{2} \) standard deviations of the mean will have no restrictions on duties.

- What percentage of the incoming recruits will be trained to operate armored combat vehicles (tanks)?
- About what percentage of the recruits will have no restrictions on training/duties?

Standard Normal Curve – Table differences

- There are different tables and computer packages for representing the area under the standard normal curve. But the results are always interchangeable.
Let's Make a Deal Paradox – aka, Monty Hall 3-door problem

This paradox is related to a popular television show in the 1970's. In the show, a contestant was given a choice of three doors/cards of which one contained a prize (diamond). The other two doors contained gag gifts like a chicken or a donkey (clubs).

Let's Make a Deal Paradox. After the contestant chose an initial door, the host of the show then revealed an empty door among the two unchosen doors, and asks the contestant if he or she would like to switch to the other unchosen door. The question is should the contestant switch. Do the odds of winning increase by switching to the remaining door?

1. Pick One card
2. Show one Club Card
3. Change 1st pick?

Let's Make a Deal Paradox. The intuition of most people tells them that each of the doors, the chosen door and the unchosen door, are equally likely to contain the prize so that there is a 50-50 chance of winning with either selection? This, however, is not the case.

The probability of winning by using the switching technique is 2/3, while the odds of winning by not switching is 1/3. The easiest way to explain this is as follows:

Let's Make a Deal Paradox. The probability of picking the wrong door in the initial stage of the game is 2/3.

If the contestant picks the wrong door initially, the host must reveal the remaining empty door in the second stage of the game. Thus, if the contestant switches after picking the wrong door initially, the contestant will win the prize.

The probability of winning by switching then reduces to the probability of picking the wrong door in the initial stage which is clearly 2/3.

Demo: Applets.dir/StatGames.exe
Uncertainty ➔ Pick a door
Properties of probability distributions

- A sequence of numbers \( \{p_1, p_2, p_3, \ldots, p_n\} \) is a probability distribution for a sample space \( S = \{s_1, s_2, s_3, \ldots, s_n\} \), if \( \text{pr}(s_k) = p_k \), for each \( 1 \leq k \leq n \). The two essential properties of a probability distribution are:
  - \( p_i \geq 0 \)
  - \( \sum p_i = 1 \)

- How do we get the probability of an event from the probabilities of outcomes that make up that event?

- If all outcomes are distinct and equally likely, how do we calculate \( \text{pr}(A) \)? If \( A = \{a_1, a_2, a_3, \ldots, a_9\} \) and \( \text{pr}(a_1) = \text{pr}(a_2) = \ldots = \text{pr}(a_9) = p \), then
  \[ \text{pr}(A) = 9 \times \text{pr}(a_1) = 9p. \]

Example of probability distributions

- Tossing a coin twice. Sample space \( S = \{HH, HT, TH, TT\} \), for a fair coin each outcome is equally likely, so the probabilities of the 4 possible outcomes should be identical, \( p \). Since, \( \text{pr}(HH) = \text{pr}(HT) = \text{pr}(TH) = \text{pr}(TT) = p \) and \( p \geq 0 \), we have
  \[ p = 0.25. \]

Conditional Probability

The conditional probability of \( A \) occurring given that \( B \) occurs is given by

\[ \text{pr}(A|B) = \frac{\text{pr}(A \text{ and } B)}{\text{pr}(B)} \]

Suppose we select one out of the 400 patients in the study and we want to find the probability that the cancer is on the extremities given that it is of type nodular: \( \text{P} = \frac{73}{125} = \text{P}(\text{C. on Extremities | Nodular}) \)

<table>
<thead>
<tr>
<th>Site</th>
<th>Head and Neck</th>
<th>Trunk</th>
<th>Extremities</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hutchinson's melanomuc freckle</td>
<td>22</td>
<td>2</td>
<td>10</td>
<td>34</td>
</tr>
<tr>
<td>Superficial</td>
<td>16</td>
<td>54</td>
<td>115</td>
<td>185</td>
</tr>
<tr>
<td>Nodular</td>
<td>19</td>
<td>33</td>
<td>73</td>
<td>125</td>
</tr>
<tr>
<td>Indeterminant</td>
<td>11</td>
<td>17</td>
<td>28</td>
<td>56</td>
</tr>
<tr>
<td>Column Totals</td>
<td>68</td>
<td>106</td>
<td>226</td>
<td>400</td>
</tr>
</tbody>
</table>

Contingency table based on Melanoma histological type and its location

Melanoma – type of skin cancer – an example of laws of conditional probabilities

Multiplication rule- what’s the percentage of Israelis that are poor and Arabic?

\[ \text{pr}(A \text{ and } B) = \text{pr}(A|B) \text{pr}(B) = \text{pr}(B|A) \text{pr}(A) \]

0.0728

0.14

1.0

<table>
<thead>
<tr>
<th>All people in Israel</th>
</tr>
</thead>
<tbody>
<tr>
<td>14% of these are Arabic</td>
</tr>
<tr>
<td>52% of this 14% are poor</td>
</tr>
<tr>
<td>7.28% of Israelis are both poor and Arabic</td>
</tr>
</tbody>
</table>

Let’s Make a Deal Paradox.

- After the contestant chose an initial door, the host of the show then revealed an empty door among the two unchosen doors, and asks the contestant if he or she would like to switch to the other unchosen door. The question is should the contestant switch. Do the odds of winning increase by switching to the remaining door?

\[ \text{P(Win (swap strat.) | 1st is Club) = 1} \]

\[ \text{P(Win (swap strat.) & 1st is Club) =} \]

\[ = \text{P(Win (swap strat.) | 1st is Club) x P(1st is Club)} \]

\[ = 1 \times \frac{2}{3} = \frac{2}{3}. \]
Review

1. Proportions (partial description of a real population) and probabilities (giving the chance of something happening in a random experiment) may be identical - under the experiment choose-a-unit-at-random.

2. Properties of probabilities.

\[ pr(A) = 1 - pr(\overline{A}) \]

\[ pr(A \text{ and } B) = pr(A | B)pr(B) = pr(B | A)pr(A) \]

A tree diagram for computing conditional probabilities

Suppose we draw 2 balls at random one at a time without replacement from an urn containing 4 black and 3 white balls, otherwise identical. What is the probability that the second ball is black? Sample Spc?

\[ P(\{2-nd ball is black\}) = \]

\[ P(\{2-nd is black\} \& \{1-st is black\}) + P(\{2-nd is black\} \& \{1-st is white\}) = \]

\[ 4/7 \times 3/6 + 4/6 \times 3/7 = 4/7. \]

Conditional probabilities and 2-way tables

• Many problems involving conditional probabilities can be solved by constructing two-way tables

• This includes reversing the order of conditioning

\[ P(A \& B) = P(A | B) \times P(B) = P(B | A) \times P(A) \]

Proportional usage of oral contraceptives and their rates of failure

We need to complete the two-way contingency table of proportions

\[ pr(\text{Failed and Oral}) = \frac{pr(\text{Failed} \text{ and Oral})}{pr(\text{Oral})} \]

\[ pr(\text{Failed and IUD}) = \frac{pr(\text{Failed} \text{ and IUD})}{pr(\text{IUD})} \]

\[ pr(\text{Successful}) = \frac{pr(\text{Successful})}{pr(\text{Total})} \]

\[ pr(\text{Steril}) = \frac{pr(\text{Steril})}{pr(\text{Total})} \]

\[ pr(\text{Barrier}) = \frac{pr(\text{Barrier})}{pr(\text{Total})} \]

\[ pr(\text{Sperms}) = \frac{pr(\text{Sperms})}{pr(\text{Total})} \]

\[ pr(\text{Total}) = \frac{pr(\text{Total})}{pr(\text{Total})} \]

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Steril</th>
<th>Oral</th>
<th>Barrier</th>
<th>IUD</th>
<th>Sperms</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failed</td>
<td>0.38</td>
<td>0.32</td>
<td>0.24</td>
<td>0.03</td>
<td>0.03</td>
<td>1.00</td>
</tr>
<tr>
<td>Total</td>
<td>0.38</td>
<td>0.32</td>
<td>0.24</td>
<td>0.03</td>
<td>0.03</td>
<td>1.00</td>
</tr>
</tbody>
</table>

\[ pr(\text{Steril}) = \frac{0.38}{1.00} \]

\[ pr(\text{Barrier}) = \frac{0.24}{1.00} \]

\[ pr(\text{IUD}) = \frac{0.03}{1.00} \]

\[ pr(\text{Sperms}) = \frac{0.03}{1.00} \]

\[ pr(\text{Total}) = \frac{1.00}{1.00} \]
**Oral contraceptives cont.**

\[ \Pr(\text{Failed | Oral}) = \frac{\Pr(\text{Failed} \land \text{Oral})}{\Pr(\text{Oral})} = \frac{0.38}{0.32} = 0.57 \]

\[ \Pr(\text{IUD}) = \frac{\Pr(\text{Failed} \land \text{IUD})}{\Pr(\text{IUD})} = \frac{0.03}{0.03} = 1.00 \]

**Remarks …**

- **Type I & Type II errors – Power of a test**
  - Intro to stats, vocabulary & intro to SPSS
  - Displaying data
  - Central tendency and variability
  - Normal z-scores, standardized distribution
  - Probability, Samples & Sampling error
  - Type I and Type II errors; Power of a test
  - Intro to hypothesis testing
  - One sample tests & Two independent samples tests
  - Two sample tests - dependent samples & Estimation
  - Correlation and regression techniques
  - Non-parametric statistical tests

**HIV cont.**

\[ \Pr(\text{HIV} \land \text{Positive}) = \Pr(\text{Positive} | \text{HIV}) \times \Pr(\text{HIV}) \]

\[ \Pr(\text{Not HIV} \land \text{Negative}) = \Pr(\text{Negative} | \text{Not HIV}) \times \Pr(\text{Not HIV}) \]

**TABLE 4.6.6 Proportions by Disease Status and Test Result**

<table>
<thead>
<tr>
<th>Disease status</th>
<th>HIV</th>
<th>Not HIV</th>
<th>Test Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV</td>
<td>.98</td>
<td>N/A</td>
<td>Positive</td>
</tr>
<tr>
<td>Not HIV</td>
<td>N/A</td>
<td>.93</td>
<td>Negative</td>
</tr>
</tbody>
</table>

**TABLE 4.6.5 Number of Individuals Having a Given Mean Absorbance Ratio (MAR) in the ELISA for HIV Antibodies**

<table>
<thead>
<tr>
<th>MAR</th>
<th>Healthy Donor</th>
<th>HIV patients</th>
<th>Test cut-off</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;2</td>
<td>202</td>
<td>0</td>
<td>275</td>
</tr>
<tr>
<td>2 - 2.99</td>
<td>73</td>
<td>2</td>
<td>0 False-positives (FNE)</td>
</tr>
<tr>
<td>3 - 3.99</td>
<td>15</td>
<td>7</td>
<td>False-positives (FNE)</td>
</tr>
<tr>
<td>4 - 4.99</td>
<td>3</td>
<td>7</td>
<td>False-positives (FNE)</td>
</tr>
<tr>
<td>5 - 5.99</td>
<td>2</td>
<td>15</td>
<td>False-positives (FNE)</td>
</tr>
<tr>
<td>6 - 6.59</td>
<td>2</td>
<td>36</td>
<td>False-positives (FNE)</td>
</tr>
<tr>
<td>7+</td>
<td>0</td>
<td>21</td>
<td>False-positives (FNE)</td>
</tr>
</tbody>
</table>

**HIV – reconstructing the contingency table**

\[ \Pr(\text{HIV} \land \text{Positive}) = \Pr(\text{Positive} | \text{HIV}) \times \Pr(\text{HIV}) \]

\[ \Pr(\text{Not HIV} \land \text{Negative}) = \Pr(\text{Negative} | \text{Not HIV}) \times \Pr(\text{Not HIV}) \]

**TABLE 4.6.4 Table Constructed from the Data in Example 4.6.**

<table>
<thead>
<tr>
<th>Method</th>
<th>Steril.</th>
<th>Oral</th>
<th>Barrier</th>
<th>IUD</th>
<th>Sperm.</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failed</td>
<td>0</td>
<td>.01</td>
<td>.02</td>
<td>.03</td>
<td>?</td>
<td>.06</td>
</tr>
<tr>
<td>Didn’t</td>
<td>3.8</td>
<td>.01</td>
<td>.02</td>
<td>.03</td>
<td>?</td>
<td>.06</td>
</tr>
<tr>
<td>Total</td>
<td>3.8</td>
<td>.02</td>
<td>.04</td>
<td>.03</td>
<td>?</td>
<td>1.00</td>
</tr>
</tbody>
</table>
### Proportions of HIV infections by country

TABLE 4.6.7 Proportions Infected with HIV

<table>
<thead>
<tr>
<th>Country</th>
<th>No. AIDS Cases</th>
<th>Population (millions)</th>
<th>pr(HIV)</th>
<th>pr(HIV Positive)</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>218,301</td>
<td>252.7</td>
<td>252.7</td>
<td>0.00864</td>
</tr>
<tr>
<td>Canada</td>
<td>6,116</td>
<td>26.7</td>
<td>26.7</td>
<td>0.00229</td>
</tr>
<tr>
<td>Australia</td>
<td>3,238</td>
<td>16.8</td>
<td>16.8</td>
<td>0.00193</td>
</tr>
<tr>
<td>New Zealand</td>
<td>323</td>
<td>3.4</td>
<td>3.4</td>
<td>0.00095</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>5,451</td>
<td>57.3</td>
<td>57.3</td>
<td>0.00095</td>
</tr>
<tr>
<td>Ireland</td>
<td>142</td>
<td>3.6</td>
<td>3.6</td>
<td>0.00039</td>
</tr>
</tbody>
</table>

### Hypothesis testing

- Intro to stats, vocabulary & intro to SPSS
- Displaying data
- Central tendency and variability
- Normal z-scores, standardized distribution
- Probability, Samples & Sampling distribution
- Type I and Type II errors; Power of a test

### ESP (extra sensory perception) or just guessing?

**Deck of equal number of Zener/Rhine cards**

- **True value for just guessing (0.200)**
- **Pratt & Woodruﬀ's proportion (0.2082)**

Can sampling variations alone account for Pratt & Woodruﬀ's success rate = 20.82% correct vs. 20% expected.

### Was Cavendish’s experiment biased?

A number of famous early experiments of measuring physical constants have later been shown to be biased.

**Mean density of the earth**

- **True value = 5.517**
- **Cavendish’s data:** (from previous Example 7.2.2)
- **n = 23, sample mean = 5.483, sample SD = 0.1904**

Simulate taking 400 sets of 23 measurements from $N(5.517, 0.1904)$. Plotted are the results of the sample means.

- **21.5% of the means were smaller than this**

**Was Cavendish’s experiment biased?**

- **Cavendish values unusually**
- **Cavendish mean (5.483)**
- **True mean (5.517)**
- **SD=0.1904**
- **SD=0.1904**

**Figure 9.1.2** Sample means from 400 sets of observations from an unbiased experiment.
Cavendish: measuring distances in std errors

Figure 9.1.3 Sample *t* values from 400 unbiased experiments (each *t* value is distance between sample mean and 5.517 in std errors).

20.5% of samples had *t* values smaller than this.

Cavendish *t* value = 0.844

Cavendish data lies within the central 60% of the distribution.

Measuring the distance between the true-value and the estimate in terms of the SE

- Intuitive criterion: Estimate is credible if it’s not far away from its hypothesized true-value!
- But how far is far-away?
- Compute the distance in standard-terms:
  \[ \frac{T - \text{Estimator}}{SE} = \frac{\text{True Parameter Value}}{SE} \]
- Reason is that the distribution of *T* is known in some cases (*Student’s t*, or N(0,1)). The estimator (obs-value) is typical/atypical if it is close to the center/tail of the distribution.

Comments

- Why can't we (rule-in) prove that a hypothesized value of a parameter is exactly true? (Because when constructing estimates based on data, there’s always sampling and may be non-sampling errors, which are normal, and will effect the resulting estimate. Even if we do 60,000 ESP tests, as we saw earlier, repeatedly we are likely to get estimates like 0.2 and 0.200001, and 0.199999, etc. – none of which may be exactly the theoretically correct, 0.2.)
- Why use the rule-out principle? (Since, we can't use the rule-in method, we try to find compelling evidence against the observed/data-constructed estimate – to reject it.)
- Why is the null hypothesis & significance testing typically used? (H0: skeptical reaction to a research hypothesis; ST is used to check if differences or effects seen in the data can be explained simply in terms of sampling variation!)

Hypotheses

Guiding principles

We cannot rule in a hypothesized value for a parameter, we can only determine whether there is evidence to rule out a hypothesized value.

The null hypothesis tested is typically a skeptical reaction to a research hypothesis.

Comments

- How can researchers try to demonstrate that effects or differences seen in their data are real? (Reject the hypothesis that there are no effects)
- How does the alternative hypothesis typically relate to a belief, hunch, or research hypothesis that initiates a study? (H1: Ha = H0 specifies the type of departure from the null-hypothesis, H0, skeletal reaction, which we are expecting (research hypothesis itself).
- In the Cavendish’s mean Earth density data, null hypothesis was H0: \( \mu = 5.517 \). We suspected bias, but not bias in any specific direction, hence H1: \( \mu \neq 5.517 \).
In the ESP Pratt & Woodruff data, the null hypothesis was $H_0: \mu = 0.2$ (pure-guessing). We suspected bias, toward success rate being higher than that, hence the research hypothesis $H_a: \mu > 0.2$.

Other commonly encountered situations are:
- $H_0: \mu_1 - \mu_2 = 0$
- $H_a: \mu_1 - \mu_2 > 0$
- $H_0: \mu_{\text{rest}} - \mu_{\text{activation}} = 0$
- $H_a: \mu_{\text{rest}} - \mu_{\text{activation}} \neq 0$

The t-test

**Alternative hypothesis**

<table>
<thead>
<tr>
<th>Evidence against $H_0: \theta &gt; \theta_0$ provided by</th>
<th>$P$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_1: \theta &gt; \theta_0$</td>
<td>$P = \Pr(T \geq t_0)$</td>
</tr>
<tr>
<td>$H_1: \theta &lt; \theta_0$</td>
<td>$P = \Pr(T \leq t_0)$</td>
</tr>
<tr>
<td>$H_1: \theta \neq \theta_0$</td>
<td>$P = 2 \Pr(T \geq</td>
</tr>
</tbody>
</table>

where $T \sim \text{Student}(df)$

**Interpretation of the p-value**

<table>
<thead>
<tr>
<th>Approximate size of $P$-Value</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$&gt; 0.12$</td>
<td>No evidence against $H_0$</td>
</tr>
<tr>
<td>$0.10 - 0.05$</td>
<td>Weak evidence against $H_0$</td>
</tr>
<tr>
<td>$0.01$</td>
<td>Strong evidence against $H_0$</td>
</tr>
<tr>
<td>$0.001$</td>
<td>Very Strong evidence against $H_0$</td>
</tr>
</tbody>
</table>

The t-test

Using $\hat{\theta}$ to test $H_0: \theta = \theta_0$ versus some alternative $H_1$.

**STEP 1** Calculate the test statistic $t_0 = \frac{\hat{\theta} - \theta_0}{\text{se}(\hat{\theta})}$

where $T \sim \text{Student}(df)$

- This tells us how many standard errors the estimate is above the hypothesized value ($t_0$ positive) or below the hypothesized value ($t_0$ negative).

**STEP 2** Calculate the $P$-value using the following table.

**STEP 3** Interpret the $P$-value in the context of the data.
**Is a second child gender influenced by the gender of the first child, in families with >1 kid?**

**TABLE 9.3.4 First and Second Births by Sex**

<table>
<thead>
<tr>
<th>Second Child</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>3,202</td>
<td>2,776</td>
<td>5,978</td>
</tr>
<tr>
<td>Female</td>
<td>2,620</td>
<td>2,792</td>
<td>5,412</td>
</tr>
<tr>
<td>Total</td>
<td>5,822</td>
<td>5,568</td>
<td>11,390</td>
</tr>
</tbody>
</table>

- Research hypothesis needs to be formulated first before collecting/looking/interpreting the data that will be used to address it. Mothers whose 1st child is a girl are more likely to have a girl, as a second child, compared to mothers with boys as 1st children.
- Data: 20 yrs of birth records of 1 Hospital in Auckland, NZ.

---

**Analysis of the birth-gender data – data summary**

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of births</th>
<th>Number of girls</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Previous child was girl)</td>
<td>5412</td>
<td>2792</td>
</tr>
<tr>
<td>2 (Previous child was boy)</td>
<td>5978</td>
<td>2776</td>
</tr>
</tbody>
</table>

- Let $p_g$ = true proportion of girls in mothers with girl as first child, $p_b$ = true proportion of girls in mothers with boy as first child. Parameter of interest is $p_g - p_b$.
- H$_0$: $p_g - p_b = 0$ (skeptical reaction). H$_c$: $p_g - p_b > 0$ (research hypothesis)

---

**Hypothesis testing as decision making**

<table>
<thead>
<tr>
<th>Actual situation</th>
<th>Decision made</th>
<th>Type I error</th>
<th>Type II error</th>
</tr>
</thead>
<tbody>
<tr>
<td>H$_0$ is true</td>
<td>Accept H$_0$ as true</td>
<td>OK</td>
<td>Type II error</td>
</tr>
<tr>
<td>H$_0$ is false</td>
<td>Reject H$_0$ as false</td>
<td>Type I error</td>
<td>OK</td>
</tr>
</tbody>
</table>

- Sample sizes: $n_1 = 5412$, $n_2 = 5978$. Sample proportions (estimates) $\hat{p}_1 = 2792/5412 = 0.5159$, $\hat{p}_2 = 2776/5978 = 0.4644$.
- H$_0$: $p_g - p_b = 0$ (skeptical reaction). H$_c$: $p_g - p_b > 0$ (research hypothesis)

---

**Analysis of the birth-gender data**

- We have strong evidence to reject the $H_0$, and hence conclude mothers with first child a girl a more likely to have a girl as a second child.
- How much more likely? A 95% CI:
  
  $CI(p_g - p_b) = [0.033; 0.070]$. And computed by:
  
  $\hat{p}_1 - \hat{p}_2 \pm 1.96 \times SE\left(\hat{p}_1 - \hat{p}_2\right) =
  
  \left(\hat{p}_1 - \hat{p}_2\right) \pm 1.96 \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} =
  
  0.0515 \pm 1.96 \times 0.0093677 \approx [3\%, 7\%]$

---

**Hypotheses**

- The **null hypothesis**, denoted by $H_0$, is the (skeptical reaction) hypothesis tested by the statistical test.
- **Principle guiding the formulation of null hypotheses**: We cannot rule a hypothesized value in; we can only determine whether there is enough evidence to rule it out. Why is that?
- **Research (alternative) hypotheses** lay out the conjectures that the research is designed to investigate and, if the researchers hunches prove correct, establish as being true.
Example: Is there racial profiling or are there confounding explanatory effects??

- The book by Best (Dammed Lies and Statistics: Untangling Numbers from the Media, Politicians and Activists, Joel Best) shows how we can test for racial bias in police arrests. Suppose we find that among 100 white and 100 black youths, 10 and 17, respectively, have experienced arrest. This may look plainly discriminatory. But suppose we then find that of the 80 middle-class white youths 4 have been arrested, and of the 50 middle-class black youths 2 arrested, whereas the corresponding numbers of lower-class white and black youths arrested are, respectively, 6 of 20 and 15 of 50. These arrest rates correspond to 5 per 100 for white and 4 per 100 for black middle-class youths, and 30 per 100 for both white and black lower-class youths. Now, better analyzed, the data suggest effects of social class, not race as such.

One sample tests & Two independent samples tests

- Intro to stats, vocabulary & intro to SPSS
- Displaying data
- Central tendency and variability
- Normal z-scores, standardized distribution
- Probability, Samples & Sampling error
- Type I and Type II errors; Power of a test
- Intro to hypothesis testing
- One sample tests & Two independent samples tests
- Two sample tests - dependent samples & Estimation
- Correlation and regression techniques
- Non-parametric statistical tests

Analysis of two independent samples

Urinary androsterone levels – data, dot-plots and 95% CI. Relations between hormonal levels and homosexuality, Margolese, 1970. Hormonal levels are lower for homosexuals. Samples are independent, as unrelated. Results, P-value of t-test 0.004 with a CI (\(\mu_{het} - \mu_{homos}\)) = [0.4:1.7]. Normal hypothesis satisfied? Skewed?

<table>
<thead>
<tr>
<th>Urinary Androsterone Levels (mg/24 hr)</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
<th>95% CI</th>
<th>Normal hypothesis satisfied? Skewed?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homosexual: 25.5, 16.8, 3.9, 3.4, 2.3, 1.6, 2.5, 3.4, 1.6, 4.3, 2.8, 20.9, 1.8, 2.2, 3.1, 1.3</td>
<td>8</td>
<td>3.4</td>
<td>0.6</td>
<td>0.1</td>
<td>(3.0, 3.8)</td>
<td>Yes</td>
</tr>
<tr>
<td>Heterosexual: 3.9, 4.0, 3.8</td>
<td>4</td>
<td>3.6</td>
<td>0.3</td>
<td>0.1</td>
<td>(3.2, 4.0)</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Table 10.2.1 Urinary Androsterone Levels (mg/24 hr)**

**Heterosexual:** 3.9, 4.0, 3.8, 3.9, 2.9, 3.2, 4.6, 4.3, 3.1, 2.7, 2.3

**Homosexual:** 2.5, 4.0, 3.8, 3.9, 2.9, 3.2, 4.6, 4.3, 3.1, 2.7, 2.3

Important points

1. The distinction between a randomized experiment and an observational study is made at the time of result interpretation. The very same statistical analysis is carried for the two situations.
2. We’ve already stressed the importance of plotting data prior to stat-analysis. Plots have many important roles – prevent dangerous misconceptions from arising (data overlaps, clusters, outliers, skewness, trends in the data, etc.)

Comparing two means for independent samples

Suppose we have 2 samples/means/distributions as follows: \(\{x_1, N(\mu_1, \sigma_1)\}\) and \(\{x_2, N(\mu_2, \sigma_2)\}\). We’ve seen before that to make inference about \(\mu_1 - \mu_2\) we can use a T-test for \(H_0: \mu_1 - \mu_2 = 0\) with \(t = \frac{\bar{x}_1 - \bar{x}_2}{SE(\bar{x}_1 - \bar{x}_2)}\) and \(CI(\mu_1 - \mu_2) = \bar{x}_1 - \bar{x}_2 \pm t \times SE(\bar{x}_1 - \bar{x}_2)\)

And \(CI(\mu_1 - \mu_2) = \bar{x}_1 - \bar{x}_2 \pm t \times SE(\bar{x}_1 - \bar{x}_2)\)

If the 2 samples are independent we use the SE formula \(SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\) with \(df = \min(n_1 - 1, n_2 - 1)\)

This gives a conservative approach for hand calculation of an approximation to the what is known as the Welch procedure, which has a complicated exact formula.
Means for independent samples – equal or unequal variances?

Pooled T-test is used for samples with assumed equal variances. Under data Normal assumptions and equal variances of \( (\mu_1 - \mu_2 = 0) \), where

\[
SE = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad t = \frac{(X_1 - X_2)^2}{s_p^2} \quad \text{is exactly Student's t distributed with } df = \frac{(n_1 - 1) + (n_2 - 1)}{2}
\]

Here \( s_p \) is called the pooled estimate of the variance, since it pools info from the 2 samples to form a combined estimate of the single variance \( \sigma^2 = \sigma_1^2 = \sigma_2^2 = \sigma^2 \). The book recommends routine use of the Welch unequal variance method.

Comparing two means for independent samples

1. How sensitive is the two-sample t-test to non-Normality in the data? (The 2-sample T-tests and CI’s are even more robust than the 1-sample tests, against non-Normality, particularly when the shapes of the 2 distributions are similar and \( n_1 = n_2 = n \), even for small \( n \), remember \( df = n_1 + n_2 - 2 \).

2. Are there nonparametric alternatives to the two-sample t-test? (Wilcoxon rank-sum-test, Mann-Whitney test, equivalent tests, same \( P \)-values.)

3. What difference is there between the quantities tested and estimated by the two-sample t-procedures and the nonparametric equivalent? (Non-parametric tests are based on ordering, not size, of the data and hence use median, not mean, for the average. The equality of 2 means is tested and CI(\( \mu_1 - \mu_2 \)).

Two sample tests - dependent samples & Estimation

- Intro to stats, vocabulary & intro to SPSS
- Displaying data
- Central tendency and variability
- Normal z-scores, standardized distribution
- Probability, Samples & Sampling error
- Type I and Type II errors; Power of a test
- Intro to hypothesis testing
- One sample tests & Two independent samples tests
- Two sample tests - dependent samples & Estimation
- Correlation and regression techniques
- Non-parametric statistical tests

Moon illusion Data

<table>
<thead>
<tr>
<th>Subject</th>
<th>Eyes Elevated</th>
<th>Eyes Level</th>
<th>Difference (Elevated - Level)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.03</td>
<td>2.03</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>1.65</td>
<td>1.73</td>
<td>-0.08</td>
</tr>
<tr>
<td>3</td>
<td>1.00</td>
<td>1.06</td>
<td>-0.06</td>
</tr>
<tr>
<td>4</td>
<td>1.25</td>
<td>1.40</td>
<td>-0.15</td>
</tr>
<tr>
<td>5</td>
<td>1.05</td>
<td>0.95</td>
<td>0.10</td>
</tr>
<tr>
<td>6</td>
<td>1.02</td>
<td>1.13</td>
<td>-0.11</td>
</tr>
<tr>
<td>7</td>
<td>1.67</td>
<td>1.41</td>
<td>0.26</td>
</tr>
<tr>
<td>8</td>
<td>1.86</td>
<td>1.73</td>
<td>0.13</td>
</tr>
<tr>
<td>9</td>
<td>1.56</td>
<td>1.63</td>
<td>-0.07</td>
</tr>
<tr>
<td>10</td>
<td>1.73</td>
<td>1.56</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Source: Kaufman and Rock [1962].

Paired Comparisons

- Sometimes we have two data sets, which are not independent, but rather observations matched in pairs.
- Back to the Kaufman & Rock study of the Moon size illusion. Does the moon size appear different with eyes level and with eyes raised? Does eye position make a difference? Eyes elevated refers to raising the eye from horizontal to zenith position. 10 Subjects are tested under eye-level (control) condition, by physically moving the subject’s body from level to zenith position with fixed eye direction – horizontal. Ratios of the Moon size in level and zenith positions, for the two paradigms are given below.

Plotting Eyes elevated rations vs. eyes level rations
For paired data, analyze the differences. $H_0: \mu_{\text{diff}} = 0$

- Dot plot of differences for the moon illusion data (with a 95% CI for the mean difference).

<table>
<thead>
<tr>
<th>Differences (Elev. - Level)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.2</td>
</tr>
</tbody>
</table>

Test of $\mu = 0.0000$ vs $\mu > 0.0000$

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
<th>t-stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difference</td>
<td>10</td>
<td>0.0190</td>
<td>0.1371</td>
<td>0.0434</td>
<td>0.44</td>
<td>0.34</td>
</tr>
</tbody>
</table>

95% CI (-0.0791, 0.1171)

Can't reject $H_0$, no evidence eye position causes illusion.

2-sample $t$-tests and intervals for differences between means $\mu_1 - \mu_2$

Assume
- statistically independent random samples from the two populations of interest
- both samples come from Normal distributions
- Welch method also assumes $\sigma_1 = \sigma_2$
- Pooled method (unpooled) does not

Two-sample $t$-methods are
- remarkably robust against non-Normality
- can be sensitive to the presence of outliers in small to moderate-sized samples
- one-sided tests are reasonably sensitive to skewness.

The Wilcoxon or Mann-Whitney test is a nonparametric alternative to the two-sample $t$-test.

We know how to analyze 1 & 2 sample data. How about if we have than 2 samples – One-way ANOVA, $F$-test

One-way ANOVA refers to the situation of having one factor (or categorical variable) which defines group membership – e.g., comparing 4 reading methods, effects of different reading methods on reading comprehension, data: 50 – 13/14 y/o students tested.

Hypotheses for the one-way analysis-of-variance $F$-test

Null hypothesis: All of the underlying true means are identical.
Alternative: Differences exist between some of the true means.

ANOVA – One-Way

- Intro to stats, vocabulary & intro to SPSS
- Displaying data
- Central tendency and variability
- Normal z-scores, standardized distribution
- Probability, Samples & Sampling error
- Type I and Type II errors; Power of a test
- Intro to hypothesis testing
- One sample tests & Two independent samples tests
- Two sample tests - dependent samples & Estimation
- ANOVA
- Correlation and regression techniques
- Non-parametric statistical tests

Comparing 4 reading methods

Comparing 4 reading methods, effects of different reading methods on reading comprehension, data: 50 – 13/14 y/o students tested.
- Mapping: using diagrams to relate main points in text;
- Scanning: reading the intro and skimming for an overview before reading details;
- Mapping and Scanning;
- Neither.

Table below shows increases in test scores, of 4 groups of students taking similar exams twice, w/ & w/o using a reading technique.

Research question: Are the results better for students using mapping, scanning or both?
TABLE 10.3.1 Increase in Reading Age

<table>
<thead>
<tr>
<th></th>
<th>Both</th>
<th>0.1</th>
<th>3.2</th>
<th>4.3</th>
<th>-0.5</th>
<th>1.9</th>
<th>3.6</th>
<th>0.4</th>
<th>2.3</th>
<th>-1.4</th>
<th>-0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Map</td>
<td>1.0</td>
<td>-0.5</td>
<td>1.0</td>
<td>0.6</td>
<td>0.6</td>
<td>1.0</td>
<td>1.0</td>
<td>-1.4</td>
<td>2.2</td>
<td>3.6</td>
</tr>
<tr>
<td></td>
<td>Scan</td>
<td>1.0</td>
<td>3.3</td>
<td>1.4</td>
<td>-0.9</td>
<td>1.0</td>
<td>0.0</td>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Neither</td>
<td>-0.3</td>
<td>-1.3</td>
<td>1.6</td>
<td>-0.4</td>
<td>-0.7</td>
<td>0.6</td>
<td>-1.8</td>
<td>-2.0</td>
<td>-0.7</td>
<td></td>
</tr>
</tbody>
</table>

Figure 10.3.1 Increases in reading ages with individual 95% CIs.


The F-test indicates that there’s real evidence true differences exist it does not give indication of where the differences are or how large they are.

One-way Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grp</td>
<td>3</td>
<td>27.06</td>
<td>9.02</td>
<td>4.45</td>
<td>0.008</td>
</tr>
<tr>
<td>Error</td>
<td>46</td>
<td>93.35</td>
<td>2.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>49</td>
<td>120.41</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Individual 95% CIs For Mean Based on Pooled StDev

<table>
<thead>
<tr>
<th>Level</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>MapScan</td>
<td>22</td>
<td>1.459</td>
<td>1.544</td>
</tr>
<tr>
<td>MapOnly</td>
<td>12</td>
<td>1.233</td>
<td>1.441</td>
</tr>
<tr>
<td>ScanOnly</td>
<td>7</td>
<td>0.914</td>
<td>1.302</td>
</tr>
<tr>
<td>Neither</td>
<td>9</td>
<td>-0.556</td>
<td>1.135</td>
</tr>
</tbody>
</table>

Pooled StDev = 1.425

Interpreting the P-value from the F-test

(The null hypothesis is that all underlying true means are identical.)

A large P-value indicates that the differences seen between the sample means could be explained simply in terms of sampling variation.

A small P-value indicates evidence that real differences exist between at least some of the true means, but gives no indication of where the differences are or how big they are.

To find out how big any differences are we need confidence intervals.

Where did the F-statistics came from?

Let’s look at this example comparing groups. How do we obtain intuitive evidence against H0? Far separated sample means + differences of sample means are large compared to their internal (within) variability! Which of the following examples indicate group diff’s are “large”?

Form of a typical ANOVA table

TABLE 10.3.2 Typical Analysis-of-Variance Table for One-Way ANOVA

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>df</th>
<th>Mean sum of squares, F-statistic, P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>[ \sum (x_i - \overline{x})^2 ]</td>
<td>k-1</td>
<td>[ \frac{\sum (x_i - \overline{x})^2}{(k-1)s_B^2} ] = [ f_0 ]</td>
</tr>
<tr>
<td>Within</td>
<td>[ \sum (x_{ij} - \overline{x}_i)^2 ]</td>
<td>n - k</td>
<td>[ \frac{\sum (x_{ij} - \overline{x}_i)^2}{(n - k)s_W^2} ] = [ s^2 ]</td>
</tr>
<tr>
<td>Total</td>
<td>[ \sum (x_{ij} - \overline{x})^2 ]</td>
<td>n - 1</td>
<td>[ \frac{\sum (x_{ij} - \overline{x})^2}{(n - 1)s^2} ] = [ F ]</td>
</tr>
</tbody>
</table>

Mean sum of squares = (sum of squares)/df

● The F-test statistic, \( f_0 \), applies when we have independent samples each from \( k \) Normal populations, \( N(\mu_i, \sigma) \), note same variance is assumed.
More about the F-test

- $s_B^2$ is a measure of variability of sample means, how far apart they are.
- $s_W^2$ reflects the avg. internal Variability within the samples.
- The F-test statistic, $f_0$, tests $H_0$ by comparing the variability of the sample means (numerator) with the variability within the samples (denominator).
- Evidence against $H_0$ is provided by values of $f_0$ which would be unusually large if $H_0$ was true.

What are $x_1, x_\ldots, x_{\cdot j}$, etc.?

Need Online reference

Apple juice sales (units per week)

$H_0$: $\mu_1 = \mu_2 = \mu_3$

$H_A$: at least 2 means differ

$x_{i,j}$, $1 \leq i \leq n_j$; $1 \leq j \leq 3$

Sum of Squares for treatments (cities)

$$SS_T = \sum_{j=1}^{k} n_j (\overline{x}_j - \overline{x})^2$$

$$SS_T = 20(577.55 - 613.07)^2 + 20(653.00 - 613.07)^2 + 20(608.65 - 613.07)^2$$

$$= 57512.23$$

Sum of squares for the Error

Sum of Squares for Error: $SSE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{i,j} - \overline{x}_j)^2$

$SSE = 19(10,723.84) + 19(7,238.01) + 19(8,609.47)$

$= 506,967.88$

F-test

Test Statistic:

$$F = \frac{MST}{MSE} = \frac{SST/(k-1)}{SSE/(n-k)}$$

$$= \frac{57512.23/(3-1)}{506,967.88/(60 - 3)}$$

$$= 3.23$$

Rejection Region: $F_{(3,60)} > F_{.05,257} = 3.15$

Conclusion: Reject $H_0$
What are $x_i$, $x_{i-1}$, etc.?

One-Way Design ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of Freedom</th>
<th>Sum of Squares</th>
<th>Mean Squares</th>
<th>$F$ Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>$k-1$</td>
<td>$SST$</td>
<td>$MST$</td>
<td>$MST$/MSE</td>
</tr>
<tr>
<td>Error</td>
<td>$n-k$</td>
<td>$SSE$</td>
<td>$MSE$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$n-1$</td>
<td>$SS(Total)$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: $MST=SST/(k-1)$

$MSE=SSE/(n-k)$

$F = \sum\frac{n_j(\bar{X}_j - \mu)^2}{k-1}$

$W_i = \sum\frac{(X_i - \mu_j)^2}{n_i}$

$F$-test assumptions

1. Samples are independent, physically independent subjects, units, or objects are being studied.
2. Sample Normal distributions, especially sensitive for small $n_i$, number of observations, $N(\mu_j, \sigma_j)$.
3. Standard deviations should be equal within all samples, $\sigma_1 = \sigma_2 = \sigma_3 = ... = \sigma_k = \sigma$. ($1/2 \leq \sigma_i/\sigma_j \leq 2$)

How to check/validate these assumptions for your data?
- Independence is clear since different groups of students are used.
- Dot-plots of group data show no evidence of non-normality.
- Sample SD's are very similar, hence we assume population SD's are similar.

Chapter 12: Lines in 2D (Regression and Correlation)

- Vertical Lines
- Horizontal Lines
- Oblique lines
- Increasing/Decreasing
- Slope of a line
- Intercept
- $Y = \alpha X + \beta$, in general.

Math Equation for the Line?

Chapter 12: Lines in 2D (Regression and Correlation)

- Draw the following lines:
  - $Y = 2X + 1$
  - $Y = -3X - 5$
  - Line through $(X_1, Y_1)$ and $(X_2, Y_2)$.
  - $(Y - Y_1)(Y_2 - Y_1) = (X - X_1)(X_2 - X_1)$.

Math Equation for the Line?

F-test assumptions

- One-Way Design ANOVA Table

Skip - Pertussis data cont.

Vaccine used

<table>
<thead>
<tr>
<th>Analysis of variance procedure</th>
<th>$F$ statistic</th>
<th>$P$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>$DF$</td>
<td>Sum of Squares</td>
</tr>
<tr>
<td>Model</td>
<td>2</td>
<td>15.61477991</td>
</tr>
<tr>
<td>Error</td>
<td>87</td>
<td>111.47215764</td>
</tr>
<tr>
<td>Corrected Total</td>
<td>89</td>
<td>127.28919556</td>
</tr>
</tbody>
</table>

With the outlier included, the $P$-value increases to 0.023.

Correlation and regression techniques

- Intro to stats, vocabulary & intro to SPSS
- Displaying data
- Central tendency and variability
- Normal z-scores, standardized distribution
- Probability, Samples & sampling error
- Type I and Type II errors; Power of a test
- Intro to hypothesis testing
- One sample tests & Two independent samples tests
- Two sample tests - dependent samples & Estimation
- Correlation and regression techniques
- Non-parametric statistical tests

Chapter 12: Lines in 2D (Regression and Correlation)

<table>
<thead>
<tr>
<th>Analysis of variance procedure</th>
<th>$F$ statistic</th>
<th>$P$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>$DF$</td>
<td>Sum of Squares</td>
</tr>
<tr>
<td>Model</td>
<td>2</td>
<td>15.61477991</td>
</tr>
<tr>
<td>Error</td>
<td>87</td>
<td>111.47215764</td>
</tr>
<tr>
<td>Corrected Total</td>
<td>89</td>
<td>127.28919556</td>
</tr>
</tbody>
</table>
Approaches for modeling data relationships

**Regression and Correlation**

- There are random and nonrandom variables
- Correlation applies if both variables ($X, Y$) are random (e.g., We saw a previous example, systolic vs. diastolic blood pressure SISVOL/DIAVOL) and are treated symmetrically.
- Regression applies in the case when you want to single out one of the variables (response variable, $Y$) and use the other variable as predictor (explanatory variable, $X$), which explains the behavior of the response variable, $Y$.

**Regression relationship** = trend + residual scatter

A causal relationship is a way of studying relationships between variables (random/nonrandom) for predicting or explaining behavior of 1 variable (response) in terms of others (explanatory variables or predictors).

**Correlation Coefficient**

- Correlation coefficient ($-1 \leq R \leq 1$): a measure of linear association, or clustering around a line of multivariate data.
- Relationship between two variables ($X, Y$) can be summarized by: $(\mu_X, \sigma_X), (\mu_Y, \sigma_Y)$ and the correlation coefficient, $R$. $R = 1$, perfect positive correlation (straight line relationship), $R = 0$, no correlation (random cloud scatter), $R = -1$, perfect negative correlation.

Computing $R(X, Y)$: (standardize, multiply, average)

\[
R(X, Y) = \frac{1}{N-1} \sum_{k=1}^{N} \left( \frac{x_k - \mu_X}{\sigma_X} \right) \left( \frac{y_k - \mu_Y}{\sigma_Y} \right)
\]

Example:

\[
\begin{array}{c|ccc|c|c|c}
\text{Gender} & \text{Height} & \text{Weight} & x^2 & y^2 & x^2y & x^2 - \bar{x}^2 - \bar{y}^2 \\
\hline
1 & 137 & 64 & 8476 & 4096 & 88064 & 2133.24 \\
2 & 150 & 76 & 22500 & 5776 & 113360 & 2012.00 \\
3 & 165 & 87 & 27225 & 7569 & 158095 & 168.86 \\
4 & 165 & 95 & 27562 & 9025 & 158095 & 168.86 \\
5 & 150 & 65 & 22500 & 4225 & 83250 & 2012.00 \\
\hline
\text{Total} & 850 & 342 & 236 & 254334 & 1052.0 \\
\end{array}
\]

\[
\bar{x} = \frac{236}{5} = 47.2, \quad \bar{y} = \frac{254334}{5} = 50866.8
\]

\[
R(X, Y) = \frac{1}{N-1} \sum_{k=1}^{N} \left( \frac{x_k - \mu_X}{\sigma_X} \right) \left( \frac{y_k - \mu_Y}{\sigma_Y} \right)
\]

\[
\begin{align*}
\mu_X &= \frac{966}{6} = 161 \text{ cm}, & \mu_Y &= \frac{332}{6} = 55 \text{ kg}, \\
\sigma_X &= \sqrt{\frac{216}{5}} = 6.573, & \sigma_Y &= \sqrt{\frac{215.3}{5}} = 6.563
\end{align*}
\]

\[
Corr(X, Y) = R(X, Y) = 0.904
\]
Correlation is pseudo-invariant w.r.t. linear transformations of $X$ or $Y$

$$R(X, Y) = \frac{1}{N-1} \sum_{k=1}^{N} \left( \frac{x_k - \mu_x}{\sigma_x} \right) \left( \frac{y_k - \mu_y}{\sigma_y} \right)$$

$$R(aX + b, cY + d), \quad \text{since}$$

$$\left( \frac{ax_k + b - \mu_x + b}{\sigma_x} \right) = \left( \frac{ax_k + b - (a\mu_x + b)}{|a| \times \sigma_x} \right)$$

$$\left( \frac{a(x_k - \mu) + b - b}{|a| \times \sigma_x} \right) = \text{sign}(a) \left( \frac{x_k - \mu_x}{\sigma_x} \right)$$

1. $R$ measures the extent of linear association between two continuous variables.
2. Association does not imply causation - both variables may be affected by a third variable – age was a confounding variable.

Correlation is Associative

$$R(X, Y) = \frac{1}{N} \sum_{k=1}^{N} \left( \frac{x_k - \mu}{\sigma} \right) \left( \frac{y_k - \mu}{\sigma} \right) = R(Y, X)$$

Correlation measures linear association, NOT an association in general!!! So, Corr($X,Y$) could be misleading for $X$ & $Y$ related in a non-linear fashion.

Trend and Scatter - Computer timing data

- The major components of a regression relationship are trend and scatter around the trend.
- To investigate a trend – fit a math function to data, or smooth the data.
- Computer timing data: a mainframe computer has $X$ users, each running jobs taking $Y$ min time. The main CPU swaps between all tasks. $Y^*$ is the total time to finish all tasks. Both $Y$ and $Y^*$ increase with increase of tasks/users, but how?

<table>
<thead>
<tr>
<th>$X$</th>
<th>Number of terminals</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>45</th>
<th>40</th>
<th>10</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y^*$</td>
<td>Total Time (mins)</td>
<td>6.6</td>
<td>14.9</td>
<td>18.4</td>
<td>12.4</td>
<td>7.9</td>
<td>0.9</td>
<td>5.5</td>
</tr>
<tr>
<td>$Y$</td>
<td>Time Per Task (secs)</td>
<td>9.9</td>
<td>17.8</td>
<td>18.4</td>
<td>16.5</td>
<td>11.9</td>
<td>5.3</td>
<td>11</td>
</tr>
<tr>
<td>$X$</td>
<td>Number of terminals</td>
<td>50</td>
<td>30</td>
<td>65</td>
<td>40</td>
<td>65</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>$Y^*$</td>
<td>Total Time (mins)</td>
<td>12.6</td>
<td>6.7</td>
<td>23.6</td>
<td>9.2</td>
<td>20.2</td>
<td>21.4</td>
<td></td>
</tr>
<tr>
<td>$Y$</td>
<td>Time Per Task (secs)</td>
<td>15.1</td>
<td>13.3</td>
<td>21.8</td>
<td>13.8</td>
<td>18.6</td>
<td>19.8</td>
<td></td>
</tr>
</tbody>
</table>

Equation for the straight line – linear/affine function

$$\beta_0=\text{Intercept (the } y\text{-value at } x=0)$$

$$\beta_1=\text{Slope of the line (rise/run), change of } y \text{ for every unit of increase for } x.$$

We want to find reasonable models (descriptions) for these data!
The quadratic curve

**Quadratic Curve**

\[ Y = \beta_0 + \beta_1 x + \beta_2 x^2 \]

\( \beta_2 \) positive \hspace{1cm} \( \beta_2 \) negative

---

Other Non-linear model curves (trigonometric, piece-wise polynomial)

- Data from the Keck telescope in Hawaii (red points) show the variation over time of the radial velocity of the star Gliese 876. The white curve is the best fit to the data points, implying that there are two unseen planets perturbing the motion of the star and each other.

---

The exponential curve, \( y = a e^{bx} \)

- Used in population growth/decay models.

---

Choosing the “best-fitting” line

- Least-squares line: Choose line with smallest sum of squared prediction errors.
- Min: \( \sum (y_i - \hat{y}_i)^2 \)
- Parameters denoted: \( \hat{\beta}_0 \) (intercept), \( \hat{\beta}_1 \) (slope).

---

Fitting a line through the data

- Example: Data points and fitted line.
- Figure 12.3.1: Fitting a line by least squares.
The idea of a residual or prediction error

<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_i$</td>
<td>$\hat{y}_i$</td>
</tr>
</tbody>
</table>

Data point $(x_i, y_i)$

Residual $u_i = y_i - \hat{y}_i$

Trend

Least squares criterion

Least squares criterion: Choose the values of the parameters to minimize the sum of squared prediction errors (or sum of squared residuals),

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Least-squares line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Least-squares line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

The least squares line

Computer timings data – linear fit

**Figure 12.3.2** Two lines on the computer-timings data.

**TABLE 12.3.1 Prediction Errors**

<table>
<thead>
<tr>
<th></th>
<th>$y = 0.25x$</th>
<th>$y = 0.15x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>$\hat{y}$</td>
<td>$\hat{y}$</td>
</tr>
<tr>
<td>40</td>
<td>9.00</td>
<td>7.00</td>
</tr>
<tr>
<td>50</td>
<td>17.80</td>
<td>15.50</td>
</tr>
<tr>
<td>60</td>
<td>18.40</td>
<td>15.50</td>
</tr>
<tr>
<td>45</td>
<td>16.50</td>
<td>14.25</td>
</tr>
<tr>
<td>40</td>
<td>11.90</td>
<td>13.00</td>
</tr>
<tr>
<td>50</td>
<td>15.00</td>
<td>15.50</td>
</tr>
<tr>
<td>60</td>
<td>13.00</td>
<td>15.50</td>
</tr>
<tr>
<td>50</td>
<td>18.60</td>
<td>19.25</td>
</tr>
<tr>
<td>65</td>
<td>13.80</td>
<td>13.00</td>
</tr>
<tr>
<td>65</td>
<td>18.60</td>
<td>19.25</td>
</tr>
<tr>
<td>65</td>
<td>19.80</td>
<td>19.25</td>
</tr>
</tbody>
</table>

Sum of squared errors 37.46 90.36

**Computer timings data**

The sum of squared errors is 37.46 for $y = 0.25x$ and 90.36 for $y = 0.15x$. 

Computer timings data – linear fit
Adding the least squares line

The regression equation is
\[ y = 3.05 + 0.260x \]

\( \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \)

Figure 12.3.3 Computer-timings data with least-squares line.


1. The least-squares line \( \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \) passes through the points \((x = 0, \hat{y} = ?)\) and \((x = x, \hat{y} = ?)\). Supply the missing values.

\[
\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} ; \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}
\]

Hands – on worksheet!

1. \( X = \{-1, 2, 3, 4\}, \ Y = \{0, -1, 1, 2\} \)

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>( x - \bar{x} )</th>
<th>( y - \bar{y} )</th>
<th>( (x - \bar{x})(y - \bar{y}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>0</td>
<td>-0.5</td>
<td>-0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>2</td>
<td>-1</td>
<td>0</td>
<td>-1.5</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1.5</td>
<td>1.5</td>
<td>2.25</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( \hat{\beta}_0 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} ; \quad \hat{\beta}_0 = \frac{\sum (x_i - \bar{x})^2}{\sum (x_i - \bar{x})(y_i - \bar{y})} \)

1. \( X = \{-1, 2, 3, 4\}, \ Y = \{0, -1, 1, 2\}, \ x = 2, \ y = 0.5 \)

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>( x - \bar{x} )</th>
<th>( y - \bar{y} )</th>
<th>( (x - \bar{x})(y - \bar{y}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>0</td>
<td>-0.5</td>
<td>-0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>2</td>
<td>-1</td>
<td>0</td>
<td>-1.5</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1.5</td>
<td>1.5</td>
<td>2.25</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Course Material Review

1. Part I
2. Data collection, surveys.
3. Experimental vs. observational studies
4. Numerical Summaries (5-number-summary)
5. Binomial distribution (prob’s, mean, variance)
6. Probabilities & proportions, independence of events and conditional probabilities
7. Normal Distribution and normal approximation

Course Material Review – cont.

1. Part II
2. Central Limit Theorem – sampling distribution of \( \bar{X} \)
3. Confidence intervals and parameter estimation
4. Hypothesis testing
5. Paired vs. Independent samples
6. Analysis Of Variance (1-way-ANOVA, one categorical var.)
7. Correlation and regression
8. Best-linear-fit, least squares method
Review

1. What are the quantities that specify a particular line?
2. Explain the idea of a prediction error in the context of fitting a line to a scatter plot. To what visual feature on the plot does a prediction error correspond? (scatter-size)
3. What property is satisfied by the line that fits the data best in the least-squares sense?
4. The least-squares line passes through \((x = 0, y = \?)\) and \((x = \?, y = \?)\). Supply the missing values.

Motivating the simple linear model

Figure 12.4.1 Lathe tool lifetimes.

Data generated from \(Y = 6 + 2x + \text{error(U)}\)

Data generated from \(Y = 6 + 2x + \text{error(U)}\)

Data generated from \(Y = 6 + 2x + \text{error(U)}\)

Histograms of least-squares estimates from 1,000 data sets

Estimates of intercept, \(\hat{\beta}_0\)

Estimates of slope, \(\hat{\beta}_1\)

Figure 12.4.3 Data generated from the model \(Y = 6 + 2x + \text{error(U)}\) where \(U ~ \text{Normal}(\mu = 0, \sigma = 3)\).
For the simple linear model, least-squares estimates are unbiased \[ E(\beta) = \beta \] and Normally distributed.

Noisier data produce more-variable least-squares estimates.

1. Before considering using the simple linear model, what sort of pattern would you be looking for in the scatter plot? (linear trend with constant scatter spread across the range of X)

2. What assumptions are made by the simple linear model, SLM? (X is linearly related to the mean value of the Y obs’s at each X, \( \mu_Y = \beta_0 + \beta_1 X \); where \( \beta_0 \) & \( \beta_1 \) are the true values of the intercept and slope of the SLM; The LS estimates \( \hat{\beta}_0 \) & \( \hat{\beta}_1 \) estimate the true values of \( \beta_0 \) & \( \beta_1 \); and the random errors \( U = Y - \mu_Y \sim N(\mu, \sigma) \).

3. If the simple linear model holds, what do you know about the sampling distributions of the least-squares estimates? (Unbiased and Normally distributed)

4. In the simple linear model, what behavior is governed by \( \sigma \)? (the spread of scatter of the data around trend)

5. Our estimate of \( \sigma \) can be thought of as a sample standard deviation for the set of prediction errors from the least-squares line.

\[ \text{RMS Error for regression} \]
\[ \text{Error} = \text{Actual value} - \text{Predicted value} \]
\[ \text{The RMS error for the regression line} \ Y = \beta_0 + \beta_1 X \]
\[ \sqrt{\frac{(y_1 - \hat{y})^2 + (y_2 - \hat{y})^2 + \cdots + (y_n - \hat{y})^2}{n-2}} \]
where \( \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \)

\[ \text{Compute the RMS Error for this regression line} \]
\[ \text{Error} = \text{Actual value} - \text{Predicted value} \]
\[ \text{The RMS error for the regression line} \ Y = \beta_0 + \beta_1 X \]
\[ \sqrt{\frac{(y_1 - \hat{y})^2 + (y_2 - \hat{y})^2 + \cdots + (y_n - \hat{y})^2}{n-2}} \]
where \( \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \)
Compute the RMS Error for this regression line

1. First compute the LS linear fit (estimate $\hat{\beta}_0$ + $\hat{\beta}_1$)

$$X = x_{11}, x_{21}, \ldots, x_{n1}; Y = y_{11}, y_{21}, \ldots, y_{n1}; (X_{11}, \ldots, X_{n1})^2 = (x_{11} - \bar{x})^2; (Y_{11}, \ldots, Y_{n1})^2 = (y_{11} - \bar{y})^2;$$

2. Then compute the individual errors

Then compute the cumulative RMS measure.

Note on the Correlation coefficient formula,

Recall the correlation coefficient...

Another form for the correlation coefficient is:

$$R(X, Y) = \text{Corr}(X, Y) = \frac{\frac{N}{i=1} \left( x_i - \bar{x} \right) \left( y_i - \bar{y} \right)}{\sqrt{\left( \frac{N}{i=1} \left( x_i - \bar{x} \right)^2 \right) \left( \frac{N}{i=1} \left( y_i - \bar{y} \right)^2 \right)}}$$

Misuse of the correlation coefficient

Some patterns with $r = 0$

Linear Regression

1. Regression relationship = trend + residual scatter

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \text{Err}$$

2. Trend = best linear fit line (LS)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}; \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

3. Scatter = residual (prediction) error Err=Obs-Pred

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = (y_i - \hat{y}_i)^2 + (y_i - \hat{y}_i)^2 + \ldots + (y_i - \hat{y}_i)^2$$

Another Notation for the Slope of the LS line

1. Note that there is a slight difference in the formula for the slope of the Least-Squares best-linear fit line:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}; \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \text{Corr}(X, Y) \times \frac{SD(Y)}{SD(X)}; \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
Non-parametric statistical tests

- Intro to stats, vocabulary & intro to SPSS
- Displaying data
- Central tendency and variability
- Normal z-scores, standardized distribution
- Probability, Samples & Sampling error
- Type I and Type II errors; Power of a test
- Intro to hypothesis testing
- One sample tests & Two independent samples tests
- Two sample tests - dependent samples & Estimation
- Correlation and regression techniques
- Non-parametric statistical tests

---

**Helmet sizes for NZ Air Force – complete table**

<table>
<thead>
<tr>
<th>Recruit</th>
<th>Cardboard (mm)</th>
<th>Metal (mm)</th>
<th>Difference (Card-metal)</th>
<th>Sign of difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>146</td>
<td>145</td>
<td>1</td>
<td>+</td>
</tr>
<tr>
<td>2</td>
<td>151</td>
<td>153</td>
<td>-2</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>163</td>
<td>161</td>
<td>2</td>
<td>+</td>
</tr>
<tr>
<td>4</td>
<td>152</td>
<td>151</td>
<td>1</td>
<td>+</td>
</tr>
<tr>
<td>5</td>
<td>151</td>
<td>149</td>
<td>2</td>
<td>+</td>
</tr>
<tr>
<td>6</td>
<td>151</td>
<td>150</td>
<td>1</td>
<td>+</td>
</tr>
<tr>
<td>7</td>
<td>149</td>
<td>150</td>
<td>-1</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>166</td>
<td>163</td>
<td>3</td>
<td>+</td>
</tr>
<tr>
<td>9</td>
<td>149</td>
<td>147</td>
<td>2</td>
<td>+</td>
</tr>
<tr>
<td>10</td>
<td>155</td>
<td>154</td>
<td>1</td>
<td>+</td>
</tr>
<tr>
<td>11</td>
<td>155</td>
<td>155</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>156</td>
<td>156</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>162</td>
<td>161</td>
<td>1</td>
<td>+</td>
</tr>
<tr>
<td>14</td>
<td>150</td>
<td>152</td>
<td>-2</td>
<td>-</td>
</tr>
<tr>
<td>15</td>
<td>156</td>
<td>154</td>
<td>2</td>
<td>+</td>
</tr>
<tr>
<td>16</td>
<td>158</td>
<td>156</td>
<td>2</td>
<td>+</td>
</tr>
<tr>
<td>17</td>
<td>149</td>
<td>147</td>
<td>2</td>
<td>+</td>
</tr>
<tr>
<td>18</td>
<td>163</td>
<td>160</td>
<td>3</td>
<td>+</td>
</tr>
</tbody>
</table>

---

**Head sizes: Does type of caliper make a difference?**

Hypothesized value

<table>
<thead>
<tr>
<th>Differences (Cardboard - Metal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>-2</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>-1</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>6</td>
</tr>
</tbody>
</table>

**Figure 10.1.8**

Dot plot of differences in size (with 95% CI).


**Paired T-Test and Confidence Interval**

Paired T for cardboard - metal

<table>
<thead>
<tr>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardboard</td>
<td>18</td>
<td>154.56</td>
<td>5.82</td>
</tr>
<tr>
<td>Metal</td>
<td>18</td>
<td>152.94</td>
<td>5.54</td>
</tr>
<tr>
<td>Difference</td>
<td>18</td>
<td>1.611</td>
<td>2.146</td>
</tr>
</tbody>
</table>

95% CI for mean difference: (0.544, 2.678)

T-Test of mean difference=0 (vs not=0): T-Value=3.19

**Figure 10.1.9**

Minitab paired-t output for the size data.


**Review**

1. What is a paired-comparison experiment? (Observed data are matched in pairs).
2. In a paired-comparison experiment, why is it wrong to treat the two sets of measurements as independent data sets? (Data are usually taken from the same unit under different treatments, so observations should be related).
3. How do you analyze the data from a paired-comparison experiment? (Analyze the difference).
4. What situations is appropriate to use the paired-comparison method to analyze the data? (Pre- and post-treatment study using FDG PET imaging).

---

**The population median – μ̂**

- Pos. skewed
- Symmetric
- Neg. skewed

<table>
<thead>
<tr>
<th>Pos. skewed</th>
<th>Symmetric</th>
<th>Neg. skewed</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>50%</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>Median</td>
<td>Mean</td>
<td>Median</td>
</tr>
<tr>
<td>μ̂</td>
<td>μ̂</td>
<td>μ̂</td>
</tr>
<tr>
<td>Median</td>
<td>Mean</td>
<td>Median</td>
</tr>
</tbody>
</table>
**Definition of the population median**

1. The **population median** is defined as the number in the middle of the distribution of the RV, i.e., 50% of the data lies below and 50% above the median.
2. Under what circumstances is the population median the same as the population mean? (symmetry of the distribution.)
3. Why do we use the **population median** rather than the population mean in the **sign test**? (imperfect distribution, mean may not be representative, or may be outlier heavily influenced.)
4. Why is the model for the **sign test** like tossing a fair coin? (In the sign-test we test H0: \( \mu \sim 0 \), under H0 a random observation is as likely to be \( < \mu \) as to be \( > \mu \). So observation has + or - sign with the same probability, hence the coin-toss model, distribution-free, non-parametric approach. Testing H0 is just like testing biased/unbiased coin).

**Helmet paired head measurements**

From the cardboard vs. metal caliper tests, we see 14 + and 3 – signs, implying larger overall measurements using the cardboard calipers. It’s like tossing a coin 17 times and getting 14 heads. How likely is that?

If \( Y \sim \text{Binomial}(17, 0.5) \), number of successes (heads) in 17 fair coin tosses, then \( P(Y \geq 14) = 0.00636 \), hence if we test \( p = 0.5, \) vs. \( p! = 0.5 \), two-tailed test, the chance is \( 2P(Y \geq 14) = 0.0127 \).

**Comments**

5. What independence assumption must hold before the **sign test** is applicable? How important is it that this assumption is true? (requires that obs’s are independent (one-sample test) and different pairs are independent (paired data), very sensitive.)
6. What advantages and disadvantages does the **sign test** have in comparison with the **t-test**? (Main advantage - test is distribution-free and insensitive to outliers. Disadvantage - when hypothesis for T-test or a parametric test are met the \( t \) are shorter and the parametric tests are more likely to detect departure from normality.)

**Review**

7. Why is the **sign test** called a distribution-free test? Does this mean that distributions are not used in performing the test? (no assumptions on the data underlying distribution, but distributions are actually used, e.g., Binomial).
8. In applying the **sign test** to paired data, how do you handle situations where both observations are tied (indistinguishable)? (ignore them)

**Why Use Nonparametric Statistics?**

- Parametric tests are based upon assumptions that may include the following:
  - The data have the same variance, regardless of the treatments or conditions in the experiment.
  - The data are normally distributed for each of the treatments or conditions in the experiment.
  - What happens when we are not sure that these assumptions have been satisfied?

**How Do Nonparametric Tests Compare with the Usual z, t, and F Tests?**

- Studies have shown that when the usual assumptions are satisfied, nonparametric tests are about 95% efficient when compared to their parametric equivalents.
- When normality and common variance are not satisfied, the nonparametric procedures can be much more efficient than their parametric equivalents.
The Wilcoxon Rank Sum Test

- Suppose we wish to test the hypothesis that two distributions have the same center.
- We select two independent random samples from each population. Designate each of the observations from population 1 as an “A” and each of the observations from population 2 as a “B”.
- If \( H_0 \) is true, and the two samples have been drawn from the same population, when we rank the values in both samples from small to large, the A’s and B’s should be randomly mixed in the rankings.

What happens when \( H_0 \) is true?

- Suppose we had 5 measurements from population 1 and 6 measurements from population 2.
- If they were drawn from the same population, the rankings might be like this. ABABBABABBA
- In this case if we summed the ranks of the A measurements and the ranks of the B measurements, the sums would be similar.

What happens if \( H_0 \) is not true?

- If the observations come from two different populations, perhaps with population 1 lying to the left of population 2, the ranking of the observations might take the following ordering.

  AAABABABB

  In this case the sum of the ranks of the B observations would be larger than that for the A observations.

How to Implement Wilcoxon’s Rank Test

- Rank the combined sample from smallest to largest.
- Let \( T_1 \) represent the sum of the ranks of the first sample (A’s).
- Then, \( T^* \), defined below, is the sum of the ranks that the A’s would have had if the observations were ranked from large to small.

  \[
  T^* = n_1 (n_1 + 1) - T_1
  \]

The Wilcoxon Rank Sum Test

- The test statistic is the smaller of \( T_1 \) and \( T^* \).
- Reject \( H_0 \) if the test statistic is less than the critical value found in Table 7(a).
- Table 7(a) is indexed by letting population 1 be the one associated with the smaller sample size \( n_1 \), and population 2 as the one associated with \( n_2 \), the larger sample size.

Example

The wing stroke frequencies of two species of bees were recorded for a sample of \( n_1 = 4 \) from species 1 and \( n_2 = 6 \) from species 2. Can you conclude that the distributions of wing strokes differ for these two species? Use \( \alpha = .05 \).

<table>
<thead>
<tr>
<th>Species 1</th>
<th>Species 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>238 (10)</td>
<td>180 (3.5)</td>
</tr>
<tr>
<td>225 (9)</td>
<td>169 (1)</td>
</tr>
<tr>
<td>190 (8)</td>
<td>180 (3.5)</td>
</tr>
<tr>
<td>180 (7)</td>
<td>185 (6)</td>
</tr>
<tr>
<td>178 (2)</td>
<td>182 (5)</td>
</tr>
</tbody>
</table>

H0: the two species are the same
Ha: the two species are in some way different

1. The sample with the smaller sample size is called sample 1.
2. We rank the 10 observations from smallest to largest, shown in parentheses in the table.
The Bee Problem

Can you conclude that the distributions of wing strokes differ for these two species? \( \alpha = .05 \).

\[
\frac{9 + 10}{2} = 9.5 = T_1
\]

\[
5 + 1 = 6 = T_2
\]

1. The test statistic is \( T = 10 \).
2. The critical value of \( T \) from Table 7(b) for a two-tailed test with \( \alpha/2 = .025 \) is \( T = 12 \); \( H_0 \) is rejected if \( T \leq 12 \).

\[
\frac{1}{2} \left[ n_1 (n_1 + n_2 + 1) \right] = \frac{1}{2} \left[ 4 (4 + 6 + 1) \right] = 12
\]

\[
\frac{1}{2} \left[ n_2 (n_1 + n_2 + 1) \right] = \frac{1}{2} \left[ 6 (4 + 6 + 1) \right] = 18
\]

Minitab Output

Recall \( T_1 = 34, T_1^* = 10 \).

\[
\text{Mann-Whitney Test and CI: Species1, Species2}
\]

Species1 N = 4 Median = 207.50
Species2 N = 6 Median = 180.00
Point estimate for ETA1-ETA2 is 30.50
95.7 Percent CI for ETA1-ETA2 is (5.99, 56.01)
W = 34.0
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.0139 (adjusted for ties)

Minitab calls the procedure the Mann-Whitney U Test, equivalent to the Wilcoxon Rank Sum Test. The test statistic is \( W = T_1 = 34 \) and has \( p \)-value \( = .0142 \). Do not reject \( H_0 \) for \( \alpha = .05 \).

Large Sample Approximation: Wilcoxon Rank Sum Test

When \( n_1 \) and \( n_2 \) are large (greater than 10 is large enough), a normal approximation can be used to approximate the critical values in Table 7.

1. Calculate \( T_1 \) and \( T_2 \). Let \( T = \min(T_1, T_2) \).
2. The statistic \( z = \frac{T - \mu_T}{\sigma_T} \) has an approximate \( z \) distribution with

\[
\mu_T = \frac{n_1 (n_1 + n_2 + 1)}{2} \quad \text{and} \quad \sigma_T^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}
\]

Some Notes

When should you use the Wilcoxon Rank Sum test instead of the two-sample \( t \) test for independent samples?
- When the responses can only be ranked and not quantified (e.g., ordinal qualitative data)
- When the \( F \) test or the Rule of Thumb shows a problem with equality of variances
- When a normality plot shows a violation of the normality assumption

The Sign Test

- The sign test is a fairly simple procedure that can be used to compare two populations when the samples consist of paired observations.
- It can be used
  - when the assumptions required for the paired-difference test of Chapter 10 are not valid or
  - when the responses can only be ranked as “one better than the other”, but cannot be quantified.

For each pair, measure whether the first response—say, A—exceeds the second response—say, B.
- The test statistic is \( x \), the number of times that A exceeds B in the \( n \) pairs of observations.
- Only pairs without ties are included in the test.
- Critical values for the rejection region or exact \( p \)-values can be found using the cumulative binomial tables in Appendix I.
The Sign Test

H₀: the two populations are identical versus
H₁: one or two-tailed alternative
is equivalent to
H₀: p = P(A exceeds B) = .5 versus
H₁: p (≠, <, or >) .5

Test statistic: x = number of plus signs
Rejection region, p-values from Bin(n=size, p).

Example

Two gourmet chefs each tasted and rated eight different meals from 1 to 10. Does it appear that one of the chefs tends to give higher ratings than the other? Use α = .01.

<table>
<thead>
<tr>
<th>Meal</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chef A</td>
<td>6</td>
<td>4</td>
<td>7</td>
<td>8</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Chef B</td>
<td>8</td>
<td>5</td>
<td>4</td>
<td>7</td>
<td>3</td>
<td>7</td>
<td>9</td>
<td>8</td>
</tr>
</tbody>
</table>

Signs: + - - + - - - 0 0

H₀: the rating distributions are the same (p = .5)
H₁: the ratings are different (p ≠ .5)

Large Sample Approximation: The Sign Test

When n ≥ 25, a normal approximation can be used to approximate the critical values of Binomial distribution.

1. Calculate x = number of plus signs.
2. The statistic $z = \frac{x - .5n}{\sqrt{n}}$ has an approximate $z$ distribution.

Example

You record the number of accidents per day at a large manufacturing plant for both the day and evening shifts for n = 100 days. You find that the number of accidents per day for the evening shift $x₁$ exceeded the number of accidents for the day shift $x₂$ on 63 of the 100 days.

For a two-tailed test, we reject H₀ if $|z| > 1.96$ (5% level).

H₀ is rejected. There is evidence of a difference between the day and night shifts.

H₀: the distributions (# of accidents) are the same (p = .5)
H₁: the distributions are different (p ≠ .5)

Test statistic: $z = \frac{x₁ - .5n}{\sqrt{n}}$ = 2.69

Example

Chef A
Chef B
Meal

Which test should you use?

- We compare statistical tests using
  - **Definition**: Power = $1 - \beta$
    - $\beta$ = P(reject H₀ when H₀ is true)
  - **The power** of the test is the probability of rejecting the null hypothesis when it is false and some specified alternative is true.
  - **The power** is the probability that the test will do what it was designed to do—that is, detect a departure from the null hypothesis when a departure exists.

Y ~ Bin(n, p) ➔ E(Y) = np
Var(Y) = np(1-p)
Eliminate zero differences.

If either \( T^+ \) or \( T^- \) is unusually large, this provides evidence against the null hypothesis.

Tied observations are assigned average of the ranks they would have gotten if not tied.

For each pair, calculate the difference \( d = x_1 - x_2 \). Eliminate zero differences.

Rank the absolute values of the differences from 1 to \( n \). Tied observations are assigned average of the ranks they would have gotten if not tied.

- \( T^+ \) = rank sum for positive differences
- \( T^- \) = rank sum for negative differences

If the two populations are the same, \( T^+ \) and \( T^- \) should be nearly equal. If either \( T^+ \) or \( T^- \) is unusually large, this provides evidence against the null hypothesis.

The Wilcoxon Signed-Rank Test

\( T^- = \text{rank sum for negative differences} \)

\( T^+ = \text{rank sum for positive differences} \)

The Wilcoxon Signed-Rank Test is a more powerful nonparametric procedure that can be used to compare two populations when the samples consist of paired observations.

It uses the ranks of the differences, \( d = x_1 - x_2 \) that we used in the paired-difference test.

The Wilcoxon Signed-Rank Test

\( H_0: \) the two populations are identical versus \( H_a: \) one or two-tailed alternative

Test statistic: \( T = \min (T^+, T^-) \)

Critical values for a one or two-tailed rejection region can be found using Wilcoxon Signed-Rank Test Table.

Example

To compare the densities of cakes using mixes A and B, six pairs of pans (A and B) were baked side-by-side in six different oven locations. Is there evidence of a difference in density for the two cake mixes?

<table>
<thead>
<tr>
<th>Location</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cake Mix A</td>
<td>1.15</td>
<td>1.10</td>
<td>1.09</td>
<td>1.13</td>
<td>1.11</td>
<td>1.14</td>
</tr>
<tr>
<td>Cake Mix B</td>
<td>1.29</td>
<td>1.20</td>
<td>1.12</td>
<td>1.15</td>
<td>1.13</td>
<td>1.13</td>
</tr>
<tr>
<td>( d = x_1 - x_2 )</td>
<td>-0.04</td>
<td>-0.10</td>
<td>-0.11</td>
<td>-0.11</td>
<td>-0.14</td>
<td>-0.19</td>
</tr>
</tbody>
</table>

\( T^- = \min (5+4+3+1+6) = 19. \)

Do not reject \( H_0 \). There is insufficient evidence to indicate that there is a difference in densities for the two cake mixes.

Cake Densities

<table>
<thead>
<tr>
<th>Location</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cake Mix A</td>
<td>1.15</td>
<td>1.10</td>
<td>0.08</td>
</tr>
<tr>
<td>Cake Mix B</td>
<td>1.29</td>
<td>1.20</td>
<td>1.12</td>
</tr>
<tr>
<td>( d = x_1 - x_2 )</td>
<td>0.06</td>
<td>-0.18</td>
<td>-0.04</td>
</tr>
</tbody>
</table>

\( T^- = \min (5+4+3+1+6) = 19. \)

The test statistic is \( T = \min (T^+, T^-) = 2. \)

Rejection region: Use Table 8. For a two-tailed test with \( \alpha = .05 \), reject \( H_0 \) if \( T \leq 1 \).
When \( n \geq 25 \), a normal approximation can be used to approximate the critical values in Table 8.

1. Calculate \( T_i^+ \) and \( T_i^- \). Let \( T = \min(T_i^+, T_i^-) \).
2. The statistic \( z = \frac{T_i - \mu_i}{\sigma_i} \) has an approximate \( z \) distribution with 
   \[ \mu_i = \frac{n(n + 1)}{4} \] and \( \sigma_i^2 = \frac{n(n + 1)(2n + 1)}{24} \).

### The Kruskal-Wallis – \( H \) Test

- **The Kruskal-Wallis \( H \) Test** is a nonparametric procedure that can be used to compare more than two populations in a completely randomized design.
- **Non-parametric equivalent to ANOVA F-test!**
- All \( n = n_1 + n_2 + \ldots + n_k \) measurements are jointly ranked.
- We use the sums of the ranks of the \( k \) samples to compare the distributions.

- Rank the total measurements in all \( k \) samples from 1 to \( n \). Tied observations are assigned average of the ranks they would have gotten if not tied.
- Calculate
  - \( T_i = \) rank sum for the \( i^{th} \) sample \( i = 1, 2, \ldots, k \)
  - \( n = n_1 + n_2 + \ldots + n_k \)
- And the test statistic \( H \) is (analog to: \( F = \frac{MSST}{MSSE} \))
  \[ H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n + 1) \]

### Example

Four groups of students were randomly assigned to be taught with four different techniques, and their achievement test scores were recorded. Are the distributions of test scores the same, or do they differ in location?

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank</td>
<td>65</td>
<td>75</td>
<td>59</td>
<td>94</td>
</tr>
<tr>
<td>87</td>
<td>69</td>
<td>78</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>73</td>
<td>81</td>
<td>67</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>79</td>
<td>81</td>
<td>62</td>
<td>68</td>
<td></td>
</tr>
</tbody>
</table>

### Teaching Methods

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank</td>
<td>65</td>
<td>75</td>
<td>59</td>
<td>94</td>
</tr>
<tr>
<td>87</td>
<td>69</td>
<td>78</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>73</td>
<td>81</td>
<td>67</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>79</td>
<td>81</td>
<td>62</td>
<td>68</td>
<td></td>
</tr>
</tbody>
</table>

- **H0**: the distributions of scores are the same
- **H1**: the distributions differ in location

**Test statistic:** \( H = \frac{12}{n(n+1)} \sum \frac{T_i^2}{n_i} - 3(n + 1) \)

\[ H = \frac{12}{16(17)} \left( 31^2 + 35^2 + 15^2 + 55^2 \right) - 3(17) = 8.96 \]
Teaching Methods

H₀: the distributions of scores are the same
Hₐ: the distributions differ in location

Test statistic: $H = \frac{12}{n(n+1)} \sum \frac{T_i^2}{n_i} - 3(n+1)$

$= \frac{12}{16(17)} \left( \frac{31^2 + 35^2 + 15^2 + 55^2}{4} \right) - 3(17) = 8.96$

Rejection region: Use Table 5. For a right-tailed chi-square test with $\alpha = .05$ and $df = 4-1 = 3$, reject $H₀$ if $7.81 < \chi^2 < 8.96$.

Reject $H₀$. There is sufficient evidence to indicate that there is a difference in test scores for the four teaching techniques.

The Friedman $F_r$ Test

- The Friedman $F_r$ Test is the nonparametric equivalent of the randomized block design with $k$ treatments and $b$ blocks.
- All $k$ measurements within a block are ranked from 1 to $b$.
- We use the sums of the ranks of the $k$ treatment observations to compare the $k$ treatment distributions.
- Model: $X_{i,j} = \mu + \alpha_i + \beta_j + \varepsilon_{i,j}$. $1 \leq i \leq a \leq ak$ (treatment effects), $1 \leq j \leq b$ (block)

$T_i$ is the nonparametric equivalent of the randomized block design with $k$ treatments and $b$ blocks.

H₀: the $k$ treatments are identical versus $H_a$: at least one distribution is different

Test statistic: Friedman $F_r$

When $H₀$ is true, the test statistic $F_r$ has an approximate $\chi^2$ distribution with $df = k - 1$.

Use a right-tailed rejection region or $p$-value based on the Chi-square distribution.

Example

A student is subjected to a stimulus and we measure the time until the student reacts by pressing a button. Four students are used in the experiment, each is subjected to three stimuli, and their reaction times are measured.

Do the distributions of reaction times differ for the three stimuli?

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>Subject 1</th>
<th>Subject 2</th>
<th>Subject 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>(1)</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>7 (2.5)</td>
<td>1.1</td>
<td>(1.0)</td>
</tr>
<tr>
<td>3</td>
<td>9 (1.0)</td>
<td>1.3</td>
<td>10 (1.2)</td>
</tr>
<tr>
<td>4</td>
<td>5 (1.5)</td>
<td>7 (1.2)</td>
<td>8 (1.3)</td>
</tr>
</tbody>
</table>

Rank the three measurements, within each block, from 1 to 3. Tied observations are assigned average of the ranks they would have gotten if not tied.

Calculate

$T_i =$ rank sum for the $i$th treatment $i = 1, 2, \ldots, k$

and the test statistic

$F_r = \frac{12}{b \times k \times (k+1)} \left( \sum_{i=1}^{k} T_i^2 \right) - 3 \times b \times (k + 1)$

Test statistic: $F_r = \frac{12}{6 \times 2 \times 3} \left( \sum T_i^2 - 36(k+1) \right)$

$= \frac{12}{12(4)} \left( 4.5^2 + 11^2 + 8.5^2 - 3(4)(4) \right) = 5.375$

$F_r = 5.375 < 7.81$

$F_r$ is less than the critical value of $7.81$, so we fail to reject $H₀$. There is not sufficient evidence to indicate that the three treatments differ in location.
**Reaction Times**

H₀: the distributions of reaction times are the same
H₁: the distributions differ in location

Test statistic: \( F = \frac{12}{bk(k+1)} \sum Y_i^2 - 3b(k+1) \)
\[ \frac{12}{12} \cdot (4.5^2 + 11^2 + 8.5^2) - 3(4)(4) = 5.375 \]

Rejection region: Use Table 5.

For a right-tailed chi-square test with \( \alpha = .05 \) and \( df = 3-1 = 2 \), reject H₀ if \( H \geq 5.99 \).

Do not reject H₀. There is insufficient evidence to indicate that there is a difference in reaction times for the three stimuli.

**Rank Correlation Coefficient**

The rank correlation coefficient, Spearman \( r_s \) is the nonparametric equivalent of the Pearson correlation coefficient \( r \).

The two variables are each ranked from smallest to largest and the ranks are denoted as \( x \) and \( y \).

We are interested in the strength of the relationship (correlation) between the two variables.

**Example**

Five elementary school science teachers have been ranked by a judge according to their teaching ability. They have also taken a national “teacher’s exam”. Is there agreement between the judge’s rank and the exam score?

<table>
<thead>
<tr>
<th>Teacher</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judge’s Rank</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Exam score</td>
<td>72</td>
<td>69</td>
<td>62</td>
<td>93</td>
<td>80</td>
</tr>
</tbody>
</table>

If the judge’s rank is low (best teacher), we might expect the teacher’s score to be high. We look for a negative association between the ranked measurements.

\[ r_s = \frac{6 \sum d^2}{n(n^2-1)} \]

For a one-tailed test with \( \alpha = .05 \) and \( n = 5 \), reject H₀ if \( r_s \geq .900 \). We do not reject H₀. Not enough evidence to indicate a negative association.
Summary

The nonparametric analogues of the parametric procedures presented in Chapters 10–14 are straightforward and fairly simple to implement.

The Wilcoxon rank sum test is the nonparametric analogue of the two-sample t test.

The sign test and the Wilcoxon signed-rank test are the nonparametric analogues of the paired-sample t test.

The Kruskal-Wallis H test is the rank equivalent of the one-way analysis of variance F test.

The Friedman F test is the rank equivalent of the randomized block design two-way analysis of variance F test.

Spearman’s rank correlation $r_s$ is the rank equivalent of Pearson’s correlation coefficient.

Key Concepts

I. Nonparametric Methods
1. These methods can be used when the data cannot be measured on a quantitative scale, or when
2. The numerical scale of measurement is arbitrarily set by the researcher, or when
3. The parametric assumptions such as normality or constant variance are seriously violated.

II. Wilcoxon Rank Sum Test: Independent Random Samples
1. Jointly rank the two samples: Designate the smaller sample as sample 1. Then
\[ T = \frac{\sum_{i=1}^{n_1} R_i - \frac{n_1(n_1 + 1)}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} \]
where $R_i$ is the rank of observation $i$ in sample 1.

III. Sign Test for a Paired Experiment
1. Find $x$, the number of times that observation A exceeds observation B for a given pair.
2. To test for a difference in two populations, test $H_0 : p = 0.5$ versus a one- or two-tailed alternative.
3. Use Table 1 of Appendix I to calculate the $p$-value for the test.

IV. Wilcoxon Signed-Rank Test: Paired Experiment
1. Calculate the differences in the paired observations. Rank the absolute values of the differences. Calculate the rank sums $T^+$ and $T^-$ for the positive and negative differences, respectively. The test statistic $T$ is the smaller of the two rank sums.
2. Table 8 of Appendix I has critical values for the rejection of $H_0$.
3. When the sample sizes are large, use the normal approximation:
\[ z = \frac{T - \mu}{\sigma_T} \]
\[ \mu_T = \frac{n(n + 1)}{4} \]
\[ \sigma_T = \sqrt{\frac{n(n + 1)(2n + 1)}{24}} \]

V. Kruskal-Wallis H Test: Completely Randomized Design
1. Jointly rank the $n$ observations in the $k$ samples. Calculate the rank sums, $T_i$ = rank sum of sample $i$, and the test statistic
\[ H = \frac{12}{n(n + 1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n + 1) \]
where $n_i$ is the sample size of sample $i$.
2. If the null hypothesis of equality of distributions is false, $H$ will be unusually large, resulting in a one-tailed test.
3. For sample sizes of five or greater, the rejection region for $H$ is based on the chi-square distribution with $(k - 1)$ degrees of freedom.
VI. The Friedman $F_i$ Test: Randomized Block Design

1. Rank the responses within each block from 1 to $k$. Calculate the rank sums $T_1, T_2, \ldots, T_k$, and the test statistic $F_i = \frac{k(k+1)}{4} \sum_{j=1}^{k} T_j^2 - 3b(k+1)$

2. If the null hypothesis of equality of treatment distributions is false, $F_i$ will be unusually large, resulting in a one-tailed test.

3. For block sizes of five or greater, the rejection region for $F_i$ is based on the chi-square distribution with $(k-1)$ degrees of freedom.

VII. Spearman’s Rank Correlation Coefficient

1. Rank the responses for the two variables from smallest to largest.

2. Calculate the correlation coefficient for the ranked observations:

$$r_s = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \quad \text{or} \quad r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

3. Table 9 in Appendix I gives critical values for rank correlations significantly different from 0.

4. The rank correlation coefficient detects not only significant linear correlation but also any other monotonic relationship between the two variables.