CSE 332
Introduction to Visualization

System Design and Evaluation

Klaus Mueller

Computer Science Department
Stony Brook University
<table>
<thead>
<tr>
<th>Lecture</th>
<th>Topic</th>
<th>Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intro, schedule, and logistics</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Applications of visual analytics, data, and basic tasks</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Data preparation and reduction</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Data preparation and reduction</td>
<td>Project 1 out</td>
</tr>
<tr>
<td>5</td>
<td>Data reduction and similarity metrics</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Dimension reduction</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Introduction to D3</td>
<td>Project 2 out</td>
</tr>
<tr>
<td>8</td>
<td>Bias in visualization</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Perception and cognition</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Visual design and aesthetics</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Cluster and pattern analysis</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>High-Dimensional data visualization: linear methods</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>High-D data vis.: non-linear methods, categorical data</td>
<td>Project 3 out</td>
</tr>
<tr>
<td>14</td>
<td>Principles of interaction</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Visual analytics and the visual sense making process</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>VA design and evaluation</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Visualization of graphs and hierarchies</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Visualization of time-varying and time-series data</td>
<td>Project 4 out</td>
</tr>
<tr>
<td>19</td>
<td>Midterm</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Maps and geo-vis</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Computer graphics and volume rendering</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Techniques to visualize spatial (3D) data</td>
<td>Project 4 halfway report due</td>
</tr>
<tr>
<td>23</td>
<td>Scientific and medical visualization</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Scientific and medical visualization</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Non-photorealistic rendering</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Memorable visualizations, visual embellishments</td>
<td>Project 5 out</td>
</tr>
<tr>
<td>27</td>
<td>Infographics design</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Projects Hall of Fame demos</td>
<td></td>
</tr>
</tbody>
</table>
Case Study: What Causes Low MPG
Consider the salient features of a car (not really big data):

- miles per gallon (MPG)
- top speed
- acceleration (time to 60 mph)
- number of cylinders
- horsepower
- weight
- country origin

400 cars from the 1980s
<table>
<thead>
<tr>
<th>Country</th>
<th>Urban population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghanistan</td>
<td>785,030</td>
</tr>
<tr>
<td>Algeria</td>
<td>329,393</td>
</tr>
<tr>
<td>Angola</td>
<td>521,205</td>
</tr>
<tr>
<td>Antigua and Barbuda</td>
<td>16,999</td>
</tr>
<tr>
<td>Argentina</td>
<td>31,254,906</td>
</tr>
<tr>
<td>Armenia</td>
<td>15,242,096</td>
</tr>
<tr>
<td>Australia</td>
<td>32,375,328</td>
</tr>
<tr>
<td>Austria</td>
<td>6,458,057</td>
</tr>
<tr>
<td>Azerbaijan</td>
<td>15,876,732</td>
</tr>
<tr>
<td>Bahamas</td>
<td>6,944,760</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>7,581,519</td>
</tr>
<tr>
<td>Barbados</td>
<td>8,084,969</td>
</tr>
<tr>
<td>Belarus</td>
<td>26,565,152</td>
</tr>
<tr>
<td>Belgium</td>
<td>8,457,959</td>
</tr>
<tr>
<td>Belize</td>
<td>4,516,355</td>
</tr>
<tr>
<td>Benin</td>
<td>2,103,433</td>
</tr>
<tr>
<td>Bermuda</td>
<td>4,400,450</td>
</tr>
<tr>
<td>Bhutan</td>
<td>8,875,864</td>
</tr>
<tr>
<td>Bolivia</td>
<td>12,333,398</td>
</tr>
<tr>
<td>Bosnia and Herzegovina</td>
<td>94,420</td>
</tr>
<tr>
<td>Botswana</td>
<td>16,320,170</td>
</tr>
<tr>
<td>Brazil</td>
<td>36,821,818</td>
</tr>
<tr>
<td>Brunei</td>
<td>36,501,387</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>29,188,559</td>
</tr>
<tr>
<td>Burkina Faso</td>
<td>22,187,219</td>
</tr>
<tr>
<td>Burundi</td>
<td>5,881,105</td>
</tr>
<tr>
<td>Cambodia</td>
<td>6,966,311</td>
</tr>
<tr>
<td>Cameroon</td>
<td>7,511,710</td>
</tr>
<tr>
<td>Canada</td>
<td>13,765,152</td>
</tr>
<tr>
<td>Cape Verde</td>
<td>3,279,342</td>
</tr>
<tr>
<td>Cayman Islands</td>
<td>-</td>
</tr>
<tr>
<td>Central African Rep.</td>
<td>392,157</td>
</tr>
<tr>
<td>Chad</td>
<td>2,877,227</td>
</tr>
<tr>
<td>Chagos</td>
<td>4,265,455</td>
</tr>
</tbody>
</table>
Global Layout of The Car Data

Random
Seeking the Cause of Low MPG

Isolating MPG
Use the nested model

- devised by Tamara Munzner (UBC)
Define the tasks, data, workflow of target users

- the tasks are usually described in domain terms
- finding and eliciting the requirements is notoriously hard
- observe how domain users work and perform their tasks
- observe the pains they are having
- what are the limitations?
- what is currently impossible, slow, or tedious?

**domain problem characterization**
Map from domain vocabulary/concerns to abstraction

- may require some sort of transformation
- data and types are described in abstract terms
- numeric tables, relational/network, spatial, ...
- tasks and operations described in abstract terms
- generic activities: sort, filter, correlate, find trends/outliers...

**domain problem characterization**

**data/operation abstraction design**

**Step 2: Abstract Into a Design**
Visual encoding

- how to best show the data (also pay tribute to aesthetics)
- bar/pie/line charts, parallel coordinates, MDS plot, scatterplot, tree map, network, etc.

Interaction design

- how to best support the intent a user may have
- select, navigate, order, brush, ...

domain problem characterization

data/operation abstraction design

encoding/interaction technique design
Match Visualizations to Tasks

Saket, Endert, Demiralp, "Task-based effectiveness of basic visualizations." *IEEE TVCG*, 2018
Well-studied computer science problem
- create efficient algorithms
- should support human interaction
- else it would not comply with key principle of visual analytics
Let use the causality analyzer framework just presented
- use the car design example

Domain problem characterization
- how to design a faster car without elevating gas consumption

Data/operation abstraction design
- determine how the different car parameters depend on one another (specifically, how do speed/acceleration and mpg relate with respect to design)
- collect data of different car models and compute a causal network

Encoding/interaction technique design
- draw graph where parameters are nodes and causal links are edges
- provide interactions that allows users to test causal links and compute a score

Algorithm design
- Partial correlation followed by causal inferencing/conditioning
- Bayesian Information Criterion (BIC) to model Occam’s Razor
How the iPhone came about

- domain problem characterization (define need)
- data/operation abstraction design
- encoding/interaction technique design
- algorithm design

June 29, 2007
threat: wrong problem
validate: observe and interview target users

threat: bad data/operation abstraction
threat: ineffective encoding/interaction technique
validate: justify encoding/interaction design

threat: slow algorithm
validate: analyze computational complexity
implement system
validate: measure system time/memory
validate: qualitative/quantitative result image analysis
[test on any users, informal usability study]
validate: lab study, measure human time/errors for operation
validate: test on target users, collect anecdotal evidence of utility
validate: field study, document human usage of deployed system
validate: observe adoption rates
Validate along the way and refine
  ▪ formative user study

Extend to general user studies of the final design
  ▪ summative user study
  ▪ laboratory study
  ▪ smaller number of subjects but can use speak aloud protocol
  ▪ crowd-sourced via internet
  ▪ potentially greater number of subjects to yield better statistics but can be superficial

Let’s discuss evaluation studies next
Suppose…

• You boss asks you to come up with a visualization that can show 4 variables
• This reminds you of the great times at CSE 564
• You also remember these three visualizations
Which One Will You Implement?
Let’s Ask

• Your best friend
  – but will he/she be an unbiased judge?
• Ask more people
Testing with Users

• You will need
  – implementations
  – some users
  – a few tasks they can solve

• Ask each user to
  – find a certain relationship in the data
  – find certain data elements
  – and so on

• Measure time and accuracy
• Do this for each of the three visualizations
You Get a Result Like This

<table>
<thead>
<tr>
<th>Participant</th>
<th>Device 1</th>
<th>Device 2</th>
<th>Device 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Task 1</td>
<td>Task 2</td>
<td>Task 1</td>
</tr>
<tr>
<td>1</td>
<td>11</td>
<td>18</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>14</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>23</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>18</td>
<td>11</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>21</td>
<td>19</td>
</tr>
<tr>
<td>6</td>
<td>14</td>
<td>21</td>
<td>20</td>
</tr>
<tr>
<td>7</td>
<td>14</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>21</td>
<td>18</td>
</tr>
<tr>
<td>9</td>
<td>14</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>15</td>
<td>18</td>
</tr>
<tr>
<td>11</td>
<td>14</td>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td>20</td>
<td>16</td>
</tr>
<tr>
<td>Mean</td>
<td>15.4</td>
<td>18.5</td>
<td>15.8</td>
</tr>
<tr>
<td>SD</td>
<td>4.01</td>
<td>2.94</td>
<td>2.69</td>
</tr>
</tbody>
</table>
You Get a Result Like This

• Which visualization is best (1, 2, or 3)?
Next Some Basics
Standard Deviation

\[ \sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} \]

- \( \sigma \): standard deviation
- \( \sum \): sum of
- \( x \): each value in the data set
- \( \bar{x} \): mean of all values in the data set
- \( n \): number of value in the data set
Regression is the attempt to explain the variation in a dependent variable using the variation in independent variables. Regression is thus an explanation of causation.

If the independent variable(s) sufficiently explain the variation in the dependent variable, the model can be used for prediction.
Simple Linear Regression

The output of a regression is a function that predicts the dependent variable based upon values of the independent variables.

Simple regression fits a straight line to the data.

\[ y' = b_0 + b_1X \pm \epsilon \]

\[ b_0 \text{ (y intercept)} \]

\[ b_1 = \text{slope} = \frac{\Delta y}{\Delta x} \]
Simple Linear Regression

The function will make a prediction for each observed data point. The observation is denoted by $y$ and the prediction is denoted by $\hat{y}$. 
For each observation, the variation can be described as:

\[ y = \hat{y} + \epsilon \]

Actual = Explained + Error
A least squares regression selects the line with the lowest total sum of squared prediction errors.

This value is called the Sum of Squares of Error, or SSE.
The Sum of Squares Regression (SSR) is the sum of the squared differences between the prediction for each observation and the population mean.
The Total Sum of Squares (SST) is equal to SSR + SSE.

Mathematically,

\[ \text{SSR} = \sum (\hat{y} - \bar{y})^2 \text{ (measure of explained variation)} \]

\[ \text{SSE} = \sum (y - \hat{y})^2 \text{ (measure of unexplained variation)} \]

\[ \text{SST} = \text{SSR} + \text{SSE} = \sum (y - \bar{y})^2 \text{ (measure of total variation in } y) \]
remaining slides courtesy of Scott MacKenzie (York University)
“Human-Computer Interaction: An Empirical Research Perspective”
What is Hypothesis Testing?

• ... the use of statistical procedures to answer research questions

• Typical research question (generic):

  Is the time to complete a task less using Method A than using Method B?

• For hypothesis testing, research questions are statements:

  There is no difference in the mean time to complete a task using Method A vs. Method B.

• This is the null hypothesis (assumption of “no difference”)

• Statistical procedures seek to reject or accept the null hypothesis (details to follow)
Analysis of Variance

• The *analysis of variance* (ANOVA) is the most widely used statistical test for hypothesis testing in factorial experiments.

• Goal → determine if an independent variable has a significant effect on a dependent variable.

• Remember, an independent variable has at least two levels (test conditions).

• Goal (put another way) → determine if the test conditions yield different outcomes on the dependent variable (e.g., one of the test conditions is faster/slower than the other).
Why Analyze the Variance?

• Seems odd that we analyse the variance when the research question is concerned with the overall means:

Is the time to complete a task less using Method A than using Method B?

• Let’s explain through two simple examples (next slide)
“Significant” implies that in all likelihood the difference observed is due to the test conditions (Method A vs. Method B).

“Not significant” implies that the difference observed is likely due to chance.

File: 06-AnovaDemo.xlsx
Example #1 - Details

Note: Within-subjects design

<table>
<thead>
<tr>
<th>Participant</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td>1</td>
<td>5.3</td>
</tr>
<tr>
<td>2</td>
<td>3.6</td>
</tr>
<tr>
<td>3</td>
<td>5.2</td>
</tr>
<tr>
<td>4</td>
<td>3.6</td>
</tr>
<tr>
<td>5</td>
<td>4.6</td>
</tr>
<tr>
<td>6</td>
<td>4.1</td>
</tr>
<tr>
<td>7</td>
<td>4.0</td>
</tr>
<tr>
<td>8</td>
<td>4.8</td>
</tr>
<tr>
<td>9</td>
<td>5.2</td>
</tr>
<tr>
<td>10</td>
<td>5.1</td>
</tr>
</tbody>
</table>

Mean

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>4.5</td>
<td>5.5</td>
</tr>
</tbody>
</table>

Error bars show ±1 standard deviation

Note: SD is the square root of the variance
Make Sure to Randomize

• Eliminate any effect than the one you’re after
• Randomize the order in which the subjects run method A and B
  – else may get learning effects of the overall problem
  – method B may turn out better just because users learnt about the problem with method A
• Randomize the data sets or tasks they are asked to use when running method A and B
  – one dataset may be easier than the other
  – method B may turn out better just because the data or tasks was easier
Reject or Not Reject – That’s the Question

Method A

- Do Not Reject $H_0$

Method B

- Reject $H_0$
Example #1 – ANOVA

<table>
<thead>
<tr>
<th></th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F-Value</th>
<th>P-Value</th>
<th>Lambda</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>9</td>
<td>5.080</td>
<td>.564</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method * Subject</td>
<td>9</td>
<td>3.888</td>
<td>.432</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Probability of obtaining the observed data if the null hypothesis is true

Reported as...

\[ F_{1,9} = 9.80, \ p < .05 \]

Thresholds for “p”
- .05
- .01
- .005
- .001
- .0005
- .0001

1 ANOVA table created by StatView (now marketed as JMP, a product of SAS; www.sas.com)
Example #1 – ANOVA

ANOVA Table for Task Completion Time (s)

<table>
<thead>
<tr>
<th></th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F-Value</th>
<th>P-Value</th>
<th>Lambda</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>9</td>
<td>5.080</td>
<td>.564</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method * Subject</td>
<td>9</td>
<td>3.888</td>
<td>.432</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SS between method groups (difference of average treatment effect across groups)

SS within method groups (variation of subjects w/r to each treatment mean)

Reported as...

\[ F_{1,9} = 9.80, \, p < .05 \]

Probability of obtaining the observed data if the null hypothesis is true

Thresholds for “p”

- .05
- .01
- .005
- .001
- .0005
- .0001

SS between method groups

1 ANOVA table created by StatView (now marketed as JMP, a product of SAS; www.sas.com)
How to Report an $F$-statistic

The mean task completion time for Method A was 4.5 s. This was 20.1% less than the mean of 5.5 s observed for Method B. The difference was statistically significant ($F_{1,9} = 9.80$, $p < .05$).

- Notice in the parentheses
  - Uppercase for $F$
  - Lowercase for $p$
  - Italics for $F$ and $p$
  - Space both sides of equal sign
  - Space after comma
  - Space on both sides of less-than sign
  - Degrees of freedom are subscript, plain, smaller font
  - Three significant figures for $F$ statistic
  - No zero before the decimal point in the $p$ statistic (except in Europe)
Example #2 - Details

Error bars show ±1 standard deviation

<table>
<thead>
<tr>
<th>Participant</th>
<th>Method</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>1</td>
<td>2.4</td>
<td>6.9</td>
</tr>
<tr>
<td>2</td>
<td>2.7</td>
<td>7.2</td>
</tr>
<tr>
<td>3</td>
<td>3.4</td>
<td>2.6</td>
</tr>
<tr>
<td>4</td>
<td>6.1</td>
<td>1.8</td>
</tr>
<tr>
<td>5</td>
<td>6.4</td>
<td>7.8</td>
</tr>
<tr>
<td>6</td>
<td>5.4</td>
<td>9.2</td>
</tr>
<tr>
<td>7</td>
<td>7.9</td>
<td>4.4</td>
</tr>
<tr>
<td>8</td>
<td>1.2</td>
<td>6.6</td>
</tr>
<tr>
<td>9</td>
<td>3.0</td>
<td>4.8</td>
</tr>
<tr>
<td>10</td>
<td>6.6</td>
<td>3.1</td>
</tr>
</tbody>
</table>

Mean

SD

4.5
5.5

2.23
2.45
Example #2 – ANOVA

ANCOVA Table for Task Completion Time (s)

<table>
<thead>
<tr>
<th></th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F-Value</th>
<th>P-Value</th>
<th>Lambda</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>9</td>
<td>37.372</td>
<td>4.152</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>1</td>
<td>4.324</td>
<td>4.324</td>
<td>.626</td>
<td>.4491</td>
<td>.626</td>
<td>.107</td>
</tr>
<tr>
<td>Method * Subject</td>
<td>9</td>
<td>62.140</td>
<td>6.904</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Reported as...

$F_{1,9} = 0.626, \text{ ns}$

Note: For non-significant effects, use “ns” if $F < 1.0$, or “p > .05” if $F > 1.0$.  

Probability of obtaining the observed data if the null hypothesis is true.
Example #2 - Reporting

The mean task completion times were 4.5 s for Method A and 5.5 s for Method B. As there was substantial variation in the observations across participants, the difference was not statistically significant as revealed in an analysis of variance ($F_{1,9} = 0.626$, ns).
# More Than Two Test Conditions

<table>
<thead>
<tr>
<th>Participant</th>
<th>Test Condition</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>11</td>
<td>11</td>
<td>21</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>18</td>
<td>11</td>
<td>22</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>17</td>
<td>10</td>
<td>18</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>19</td>
<td>15</td>
<td>21</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>13</td>
<td>17</td>
<td>23</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>10</td>
<td>15</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>14</td>
<td>14</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>13</td>
<td>14</td>
<td>19</td>
<td>18</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>19</td>
<td>18</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>10</td>
<td>17</td>
<td>21</td>
<td>18</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>10</td>
<td>19</td>
<td>22</td>
<td>13</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>16</td>
<td>14</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>10</td>
<td>20</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>10</td>
<td>13</td>
<td>21</td>
<td>18</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>20</td>
<td>17</td>
<td>14</td>
<td>18</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>18</td>
<td>17</td>
<td>17</td>
<td>14</td>
</tr>
</tbody>
</table>

| Mean | 14.25 | 15.13 | 18.75 | 16.06 |
| SD   | 3.84  | 2.94  | 2.89  | 3.23  |
ANOVA

ANOVA Table for Dependent Variable (units)

<table>
<thead>
<tr>
<th></th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F-Value</th>
<th>P-Value</th>
<th>Lambda</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>15</td>
<td>81.109</td>
<td>5.407</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test Condition</td>
<td>3</td>
<td>182.172</td>
<td>60.724</td>
<td>4.954</td>
<td>.0047</td>
<td>14.862</td>
<td>.896</td>
</tr>
<tr>
<td>Test Condition * Subject</td>
<td>45</td>
<td>551.578</td>
<td>12.257</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• There was a significant effect of Test Condition on the dependent variable \(F_{3,45} = 4.95, p < .005\)

• Degrees of freedom
  – If \(n\) is the number of test conditions and \(m\) is the number of participants, the degrees of freedom are…
  – Effect  \(\rightarrow (n - 1)\)
  – Residual  \(\rightarrow (n - 1)(m - 1)\)
  – Note: single-factor, within-subjects design
Post Hoc Comparisons Tests

• A significant $F$-test means that at least one of the test conditions differed significantly from one other test condition

• Does not indicate which test conditions differed significantly from one another

• To determine which pairs differ significantly, a post hoc comparisons tests is used

• Examples:
  – Fisher PLSD, Bonferroni/Dunn, Dunnett, Tukey/Kramer, Games/Howell, Student-Newman-Keuls, orthogonal contrasts, Scheffé

• Scheffé test on next slide
### Scheffé Post Hoc Comparisons

#### Scheffe for Dependent Variable (units)

**Effect: Test Condition**

**Significance Level: 5 %**

<table>
<thead>
<tr>
<th></th>
<th>Mean Diff.</th>
<th>Crit. Diff.</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, B</td>
<td>-.875</td>
<td>3.302</td>
<td>.9003</td>
</tr>
<tr>
<td>A, C</td>
<td>-4.500</td>
<td>3.302</td>
<td>.0032</td>
</tr>
<tr>
<td>A, D</td>
<td>-1.813</td>
<td>3.302</td>
<td>.4822</td>
</tr>
<tr>
<td>B, C</td>
<td>-3.625</td>
<td>3.302</td>
<td>.0256</td>
</tr>
<tr>
<td>B, D</td>
<td>-.938</td>
<td>3.302</td>
<td>.8806</td>
</tr>
<tr>
<td>C, D</td>
<td>2.688</td>
<td>3.302</td>
<td>.1520</td>
</tr>
</tbody>
</table>

- Test conditions A:C and B:C differ significantly (see chart three slides back)
Between-subjects Designs

• Research question:
  – Do left-handed users and right-handed users differ in the time to complete an interaction task?

• The independent variable (handedness) must be assigned between-subjects

• Example data set →

<table>
<thead>
<tr>
<th>Participant</th>
<th>Task Completion Time (s)</th>
<th>Handedness</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23</td>
<td>L</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>L</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>L</td>
</tr>
<tr>
<td>4</td>
<td>21</td>
<td>L</td>
</tr>
<tr>
<td>5</td>
<td>23</td>
<td>L</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>L</td>
</tr>
<tr>
<td>7</td>
<td>25</td>
<td>L</td>
</tr>
<tr>
<td>8</td>
<td>23</td>
<td>L</td>
</tr>
<tr>
<td>9</td>
<td>17</td>
<td>R</td>
</tr>
<tr>
<td>10</td>
<td>19</td>
<td>R</td>
</tr>
<tr>
<td>11</td>
<td>16</td>
<td>R</td>
</tr>
<tr>
<td>12</td>
<td>21</td>
<td>R</td>
</tr>
<tr>
<td>13</td>
<td>23</td>
<td>R</td>
</tr>
<tr>
<td>14</td>
<td>20</td>
<td>R</td>
</tr>
<tr>
<td>15</td>
<td>22</td>
<td>R</td>
</tr>
<tr>
<td>16</td>
<td>21</td>
<td>R</td>
</tr>
</tbody>
</table>

Mean 20.9
SD 2.38
Summary Data and Chart

<table>
<thead>
<tr>
<th>Handedness</th>
<th>Task Completion Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>Left</td>
<td>22.0</td>
</tr>
<tr>
<td>Right</td>
<td>19.9</td>
</tr>
</tbody>
</table>

![Bar chart showing task completion times for left and right handedness]
ANOVA Table for Task Completion Time (s)

<table>
<thead>
<tr>
<th></th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F-Value</th>
<th>P-Value</th>
<th>Lambda</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Handedness</td>
<td>1</td>
<td>18.063</td>
<td>18.063</td>
<td>3.781</td>
<td>.0722</td>
<td>3.781</td>
<td>.429</td>
</tr>
<tr>
<td>Residual</td>
<td>14</td>
<td>66.875</td>
<td>4.777</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• The difference was not statistically significant ($F_{1,14} = 3.78, p > .05$)

• Degrees of freedom:
  – Effect $\rightarrow (n - 1)$
  – Residual $\rightarrow (m - n)$
  – Note: single-factor, between-subjects design
Two-way ANOVA

• An experiment with two independent variables is a two-way design

• ANOVA tests for
  – Two main effects + one interaction effect

• Example
  – Independent variables
    • Device $\rightarrow$ D1, D2, D3 (e.g., mouse, stylus, touchpad)
    • Task $\rightarrow$ T1, T2 (e.g., point-select, drag-select)
  – Dependent variable
    • Task completion time (or something, this isn’t important here)
  – Both IVs assigned within-subjects
  – Participants: 12
  – Data set (next slide)
## Data Set

<table>
<thead>
<tr>
<th>Participant</th>
<th>Device 1</th>
<th></th>
<th>Device 2</th>
<th></th>
<th>Device 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Task 1</td>
<td>Task 2</td>
<td>Task 1</td>
<td>Task 2</td>
<td>Task 1</td>
<td>Task 2</td>
</tr>
<tr>
<td>1</td>
<td>11</td>
<td>18</td>
<td>15</td>
<td>13</td>
<td>20</td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>14</td>
<td>17</td>
<td>15</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>23</td>
<td>13</td>
<td>20</td>
<td>20</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>18</td>
<td>11</td>
<td>12</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>21</td>
<td>19</td>
<td>14</td>
<td>19</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>14</td>
<td>21</td>
<td>20</td>
<td>11</td>
<td>17</td>
<td>13</td>
</tr>
<tr>
<td>7</td>
<td>14</td>
<td>16</td>
<td>15</td>
<td>20</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>21</td>
<td>18</td>
<td>20</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>9</td>
<td>14</td>
<td>15</td>
<td>13</td>
<td>17</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>15</td>
<td>18</td>
<td>10</td>
<td>11</td>
<td>16</td>
</tr>
<tr>
<td>11</td>
<td>14</td>
<td>20</td>
<td>15</td>
<td>16</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td>20</td>
<td>16</td>
<td>16</td>
<td>20</td>
<td>9</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>15.4</td>
<td>18.5</td>
<td>15.8</td>
<td>15.3</td>
<td>15.4</td>
<td>12.2</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td>4.01</td>
<td>2.94</td>
<td>2.69</td>
<td>3.50</td>
<td>3.92</td>
<td>2.69</td>
</tr>
</tbody>
</table>
Summary Data and Chart

<table>
<thead>
<tr>
<th></th>
<th>Task 1</th>
<th>Task 2</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device 1</td>
<td>15.4</td>
<td>18.5</td>
<td>17.0</td>
</tr>
<tr>
<td>Device 2</td>
<td>15.8</td>
<td>15.3</td>
<td>15.6</td>
</tr>
<tr>
<td>Device 3</td>
<td>15.4</td>
<td>12.2</td>
<td>13.8</td>
</tr>
<tr>
<td>Mean</td>
<td>15.6</td>
<td>15.3</td>
<td>15.4</td>
</tr>
</tbody>
</table>

![Chart showing task completion times for different devices.](chart.png)
Can you pull the relevant statistics from this chart and craft statements indicating the outcome of the ANOVA?

**ANOVA Table for Task Completion Time (s)**

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F-Value</th>
<th>P-Value</th>
<th>Lambda</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>11</td>
<td>134.778</td>
<td>12.253</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Device</td>
<td>2</td>
<td>121.028</td>
<td>60.514</td>
<td>5.865</td>
<td>.0091</td>
<td>11.731</td>
<td>.831</td>
</tr>
<tr>
<td>Device * Subject</td>
<td>22</td>
<td>226.972</td>
<td>10.317</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task</td>
<td>1</td>
<td>.889</td>
<td>.889</td>
<td>.076</td>
<td>.7875</td>
<td>.076</td>
<td>.057</td>
</tr>
<tr>
<td>Task * Subject</td>
<td>11</td>
<td>128.111</td>
<td>11.646</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Device * Task</td>
<td>2</td>
<td>121.028</td>
<td>60.514</td>
<td>5.435</td>
<td>.0121</td>
<td>10.869</td>
<td>.798</td>
</tr>
<tr>
<td>Device * Task * Subject</td>
<td>22</td>
<td>244.972</td>
<td>11.135</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
ANOVA - Reporting

The grand mean for task completion time was 15.4 seconds. Device 3 was the fastest at 13.8 seconds, while device 1 was the slowest at 17.0 seconds. The main effect of device on task completion time was statistically significant ($F_{2,22} = 5.865, p < .01$). The task effect was modest, however. Task completion time was 15.6 seconds for task 1. Task 2 was slightly faster at 15.3 seconds; however, the difference was not statistically significant ($F_{1,11} = 0.076, \text{ ns}$). The results by device and task are shown in Figure x. There was a significant Device $\times$ Task interaction effect ($F_{2,22} = 5.435, p < .05$), which was due solely to the difference between device 1 task 2 and device 3 task 2, as determined by a Scheffé post hoc analysis.
Chi-square Test (Nominal Data)

- A *chi-square test* is used to investigate relationships
- Relationships between categorical, or nominal-scale, variables representing attributes of people, interaction techniques, systems, etc.
- Data organized in a *contingency table* – cross tabulation containing counts (frequency data) for number of observations in each category
- A chi-square test compares the *observed values* against *expected values*
- Expected values assume “no difference”
- Research question:
  - *Do males and females differ in their method of scrolling on desktop systems?* (next slide)
Chi-square – Example #1

<table>
<thead>
<tr>
<th>Gender</th>
<th>Scrolling Method</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MW</td>
<td>CD</td>
</tr>
<tr>
<td>Male</td>
<td>28</td>
<td>15</td>
</tr>
<tr>
<td>Female</td>
<td>21</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>49</td>
<td>24</td>
</tr>
</tbody>
</table>

MW = mouse wheel
CD = clicking, dragging
KB = keyboard
Chi-square – Example #1

56.0\cdot49.0/101=27.2

<table>
<thead>
<tr>
<th>Gender</th>
<th>Expected Number of Users</th>
<th>Scrolling Method</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MW</td>
<td>CD</td>
</tr>
<tr>
<td>Male</td>
<td>27.2</td>
<td>13.3</td>
<td>15.5</td>
</tr>
<tr>
<td>Female</td>
<td>21.8</td>
<td>10.7</td>
<td>12.5</td>
</tr>
<tr>
<td>Total</td>
<td>49.0</td>
<td>24.0</td>
<td>28.0</td>
</tr>
</tbody>
</table>

\frac{(\text{Expected}-\text{Observed})^2}{\text{Expected}}=\frac{(28-27.2)^2}{27.2}

<table>
<thead>
<tr>
<th>Gender</th>
<th>Chi Squares</th>
<th>Scrolling Method</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MW</td>
<td>CD</td>
</tr>
<tr>
<td>Male</td>
<td>0.025</td>
<td>0.215</td>
<td>0.411</td>
</tr>
<tr>
<td>Female</td>
<td>0.032</td>
<td>0.268</td>
<td>0.511</td>
</tr>
<tr>
<td>Total</td>
<td>0.057</td>
<td>0.483</td>
<td>0.922</td>
</tr>
</tbody>
</table>

\chi^2 = 1.462

(See HCI:ERP for calculations)
Chi-square Critical Values

- Decide in advance on *alpha* (typically .05)
- Degrees of freedom
  - $df = (r - 1)(c - 1) = (2 - 1)(3 - 1) = 2$
  - $r = \text{number of rows, } c = \text{number of columns}$

<table>
<thead>
<tr>
<th>Significance Threshold ($\alpha$)</th>
<th>Degrees of Freedom</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>.1</td>
<td>2.71</td>
</tr>
<tr>
<td>.05</td>
<td>3.84</td>
</tr>
<tr>
<td>.01</td>
<td>6.64</td>
</tr>
<tr>
<td>.001</td>
<td>10.83</td>
</tr>
</tbody>
</table>

$\chi^2 = 1.462 \ (< 5.99 \therefore \text{not significant})$