CSE 564
Visualization & Visual Analytcs

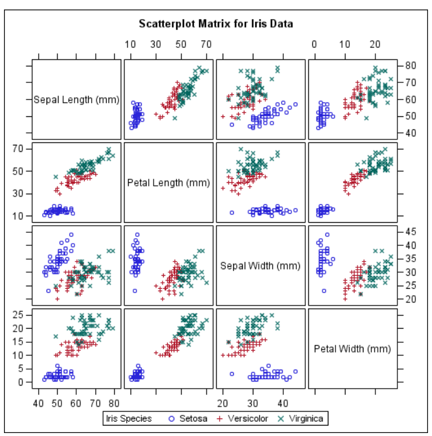
Evaluation of User Studies

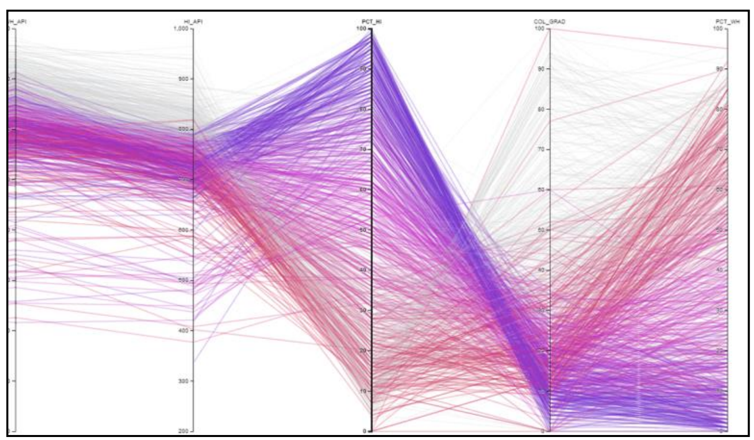
Klaus Mueller

Computer Science Department
Stony Brook University

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

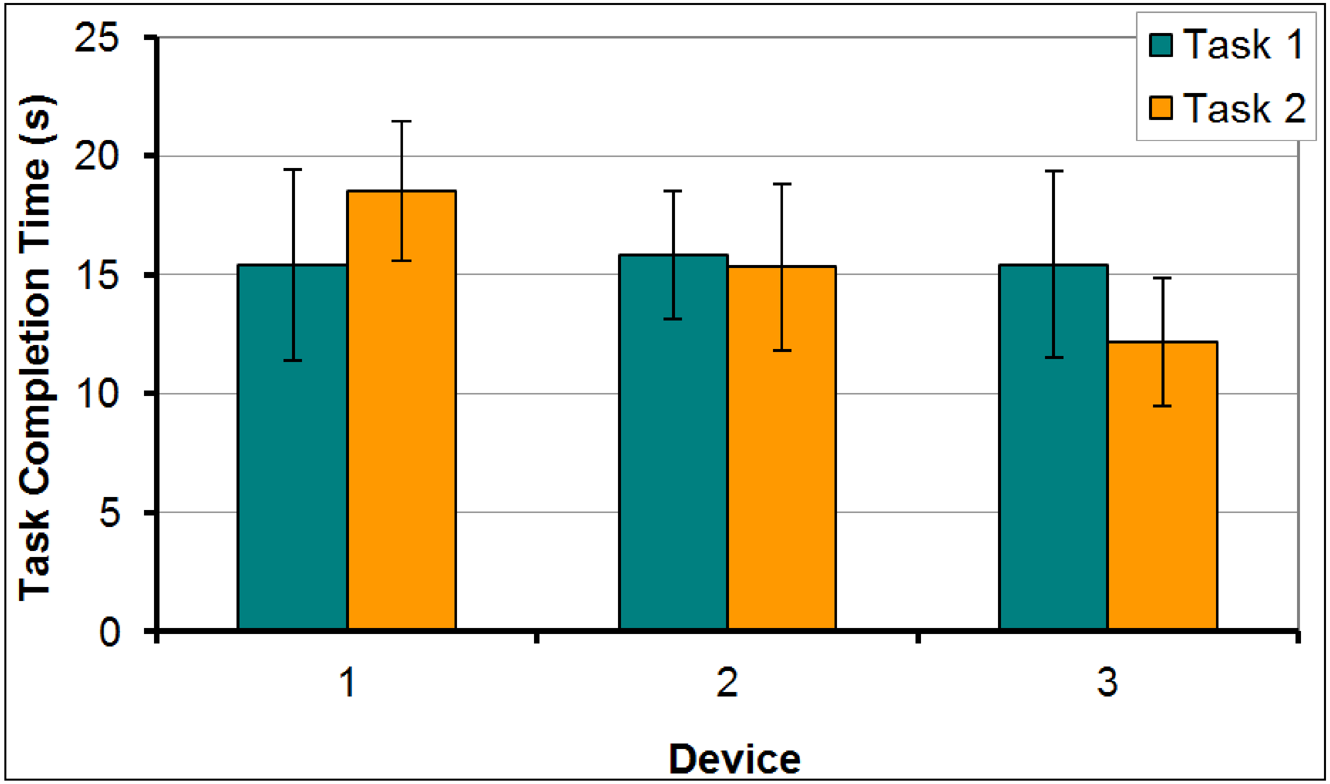
- Measure time and accuracy
- Do this for each of the three visualizations

You Get a Result Like This

Participant	Device 1		Device 2		Device 3	
	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2
1	11	18	15	13	20	14
2	10	14	17	15	11	13
3	10	23	13	20	20	16
4	18	18	11	12	11	10
5	20	21	19	14	19	8
6	14	21	20	11	17	13
7	14	16	15	20	16	12
8	20	21	18	20	14	12
9	14	15	13	17	16	14
10	20	15	18	10	11	16
11	14	20	15	16	10	9
12	20	20	16	16	20	9
Mean	15.4	18.5	15.8	15.3	15.4	12.2
SD	4.01	2.94	2.69	3.50	3.92	2.69

You Get a Result Like This

- Which visualization is best (1, 2, or 3)?



Suppose Also…

- Assume your boss is female
- Your cohort has male and female users
- Do females prefer one visualization over another
  - you probably want to use the one the females preferred
  - now you’re curious, is there a gender bias?

Next Some Basics

Standard Deviation

$$\sigma = \sqrt{\frac{\sum (x - \overline{x})^2}{n}}$$

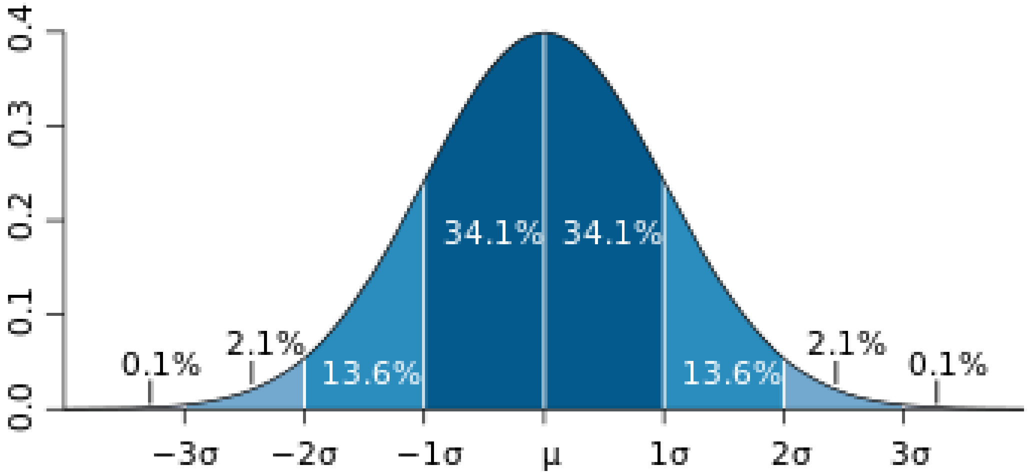
σ = standard deviation

$\sum$ = sum of

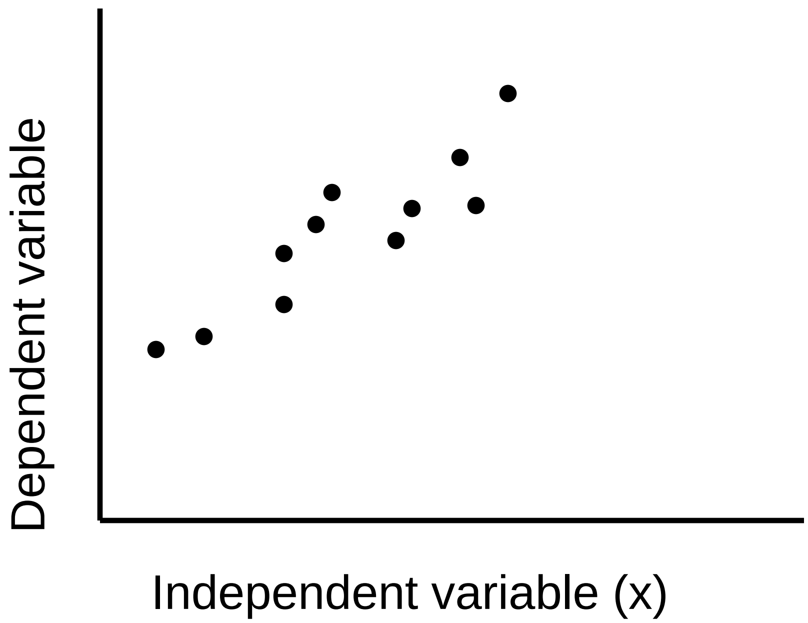
x = each value in the data set

$\overline{x}$ = mean of all values in the data set

n = number of value in the data set



Regression

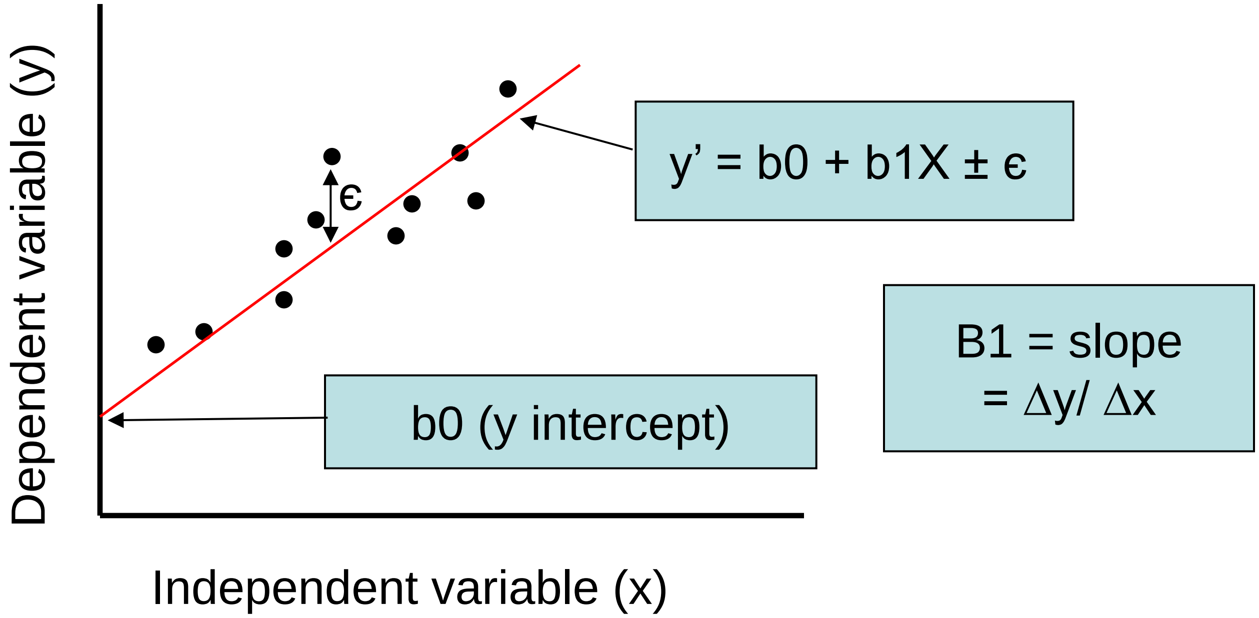


Regression is the attempt to explain the variation in a dependent variable using the variation in independent variables.

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.

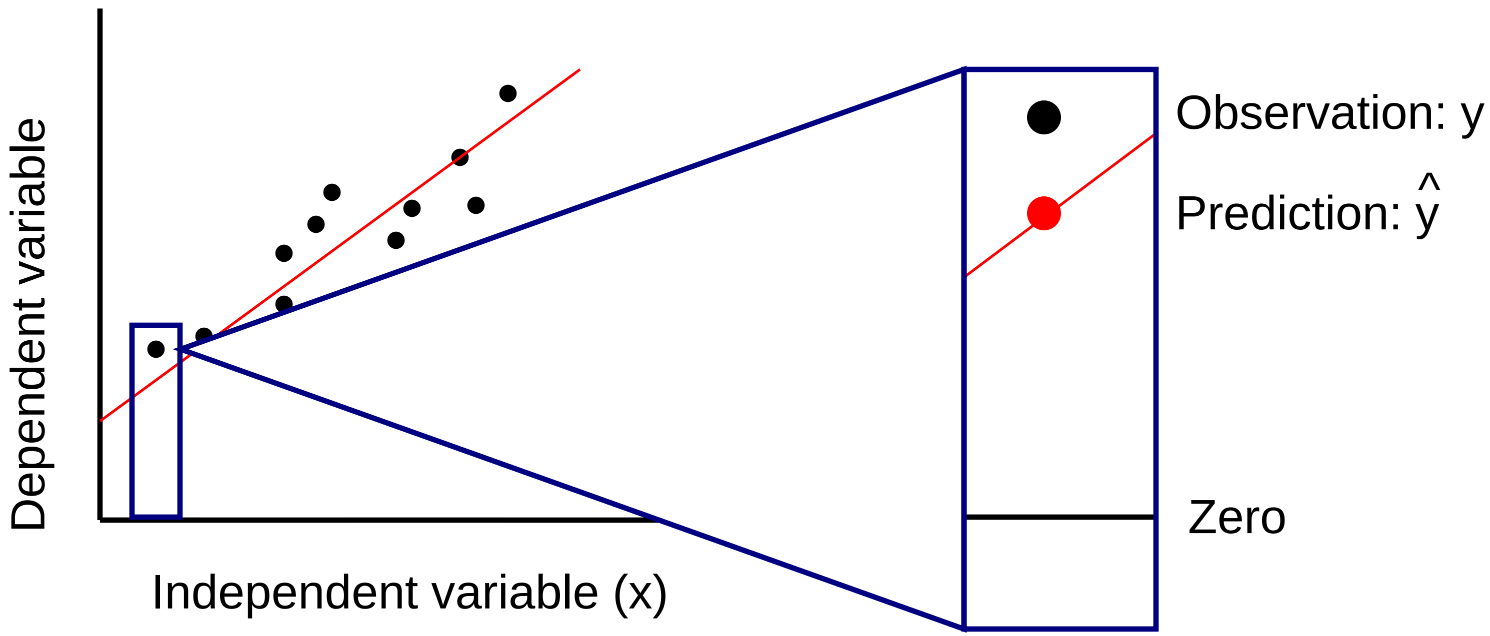
Simple Linear Regression



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.

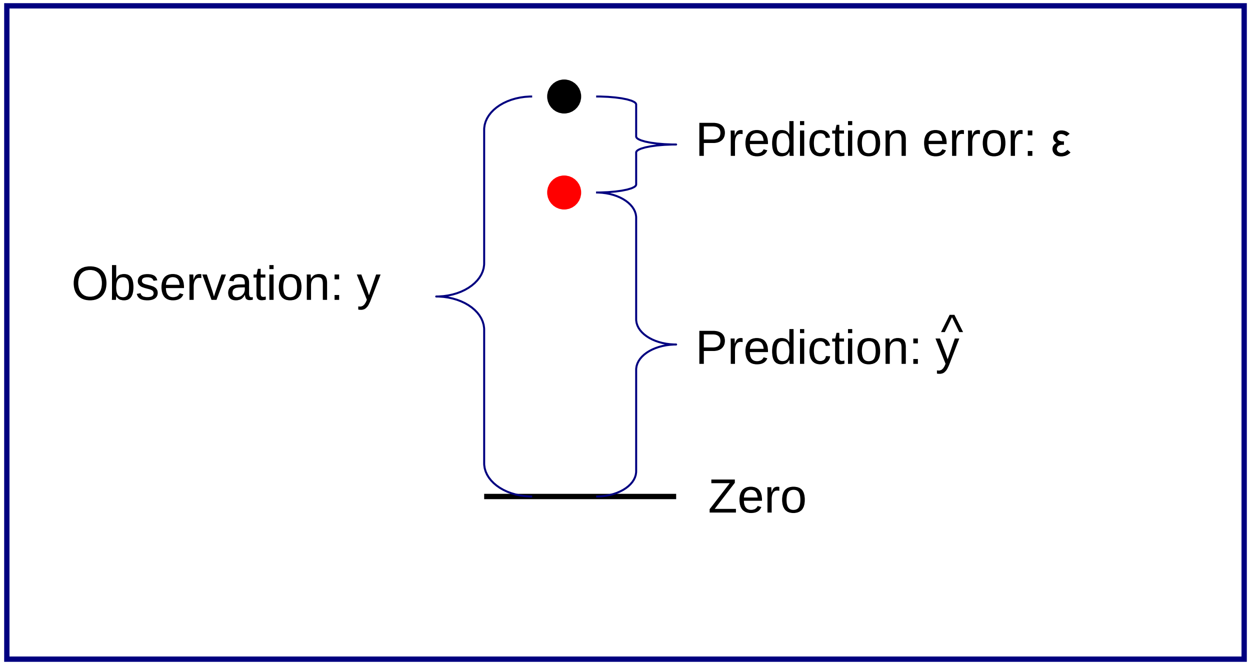
Simple Linear Regression



The function will make a prediction for each observed data point.

The observation is denoted by y and the prediction is denoted by $\hat{y}$.

Simple Linear Regression

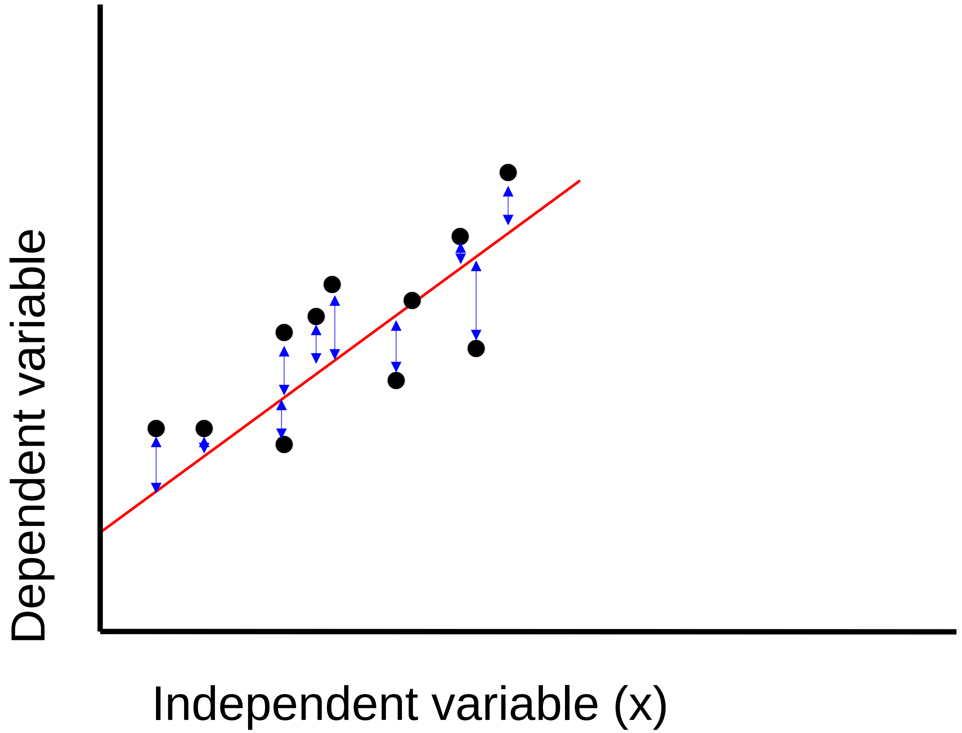


For each observation, the variation can be described as:

$$y = \hat{y} + \varepsilon$$

Actual = Explained + Error

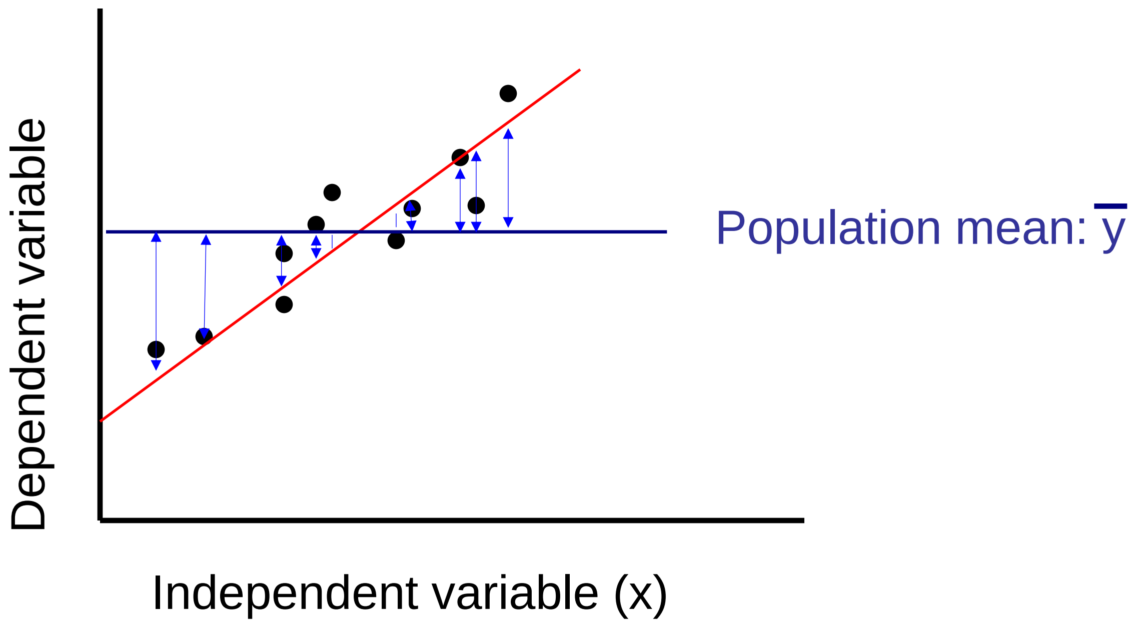
Regression



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.

Calculating SSR



The Sum of Squares Regression (SSR) is the sum of the squared differences between the prediction for each observation and the population mean.

Regression Formulas

The Total Sum of Squares (SST) is equal to SSR + SSE.

Mathematically,

SSR = $\sum (\hat{y} - \bar{y})^2$ **(measure of explained variation)**

SSE = $\sum (y - \hat{y})^2$ **(measure of unexplained variation)**

SST = SSR + SSE = $\sum (y - \bar{y})^2$ **(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)

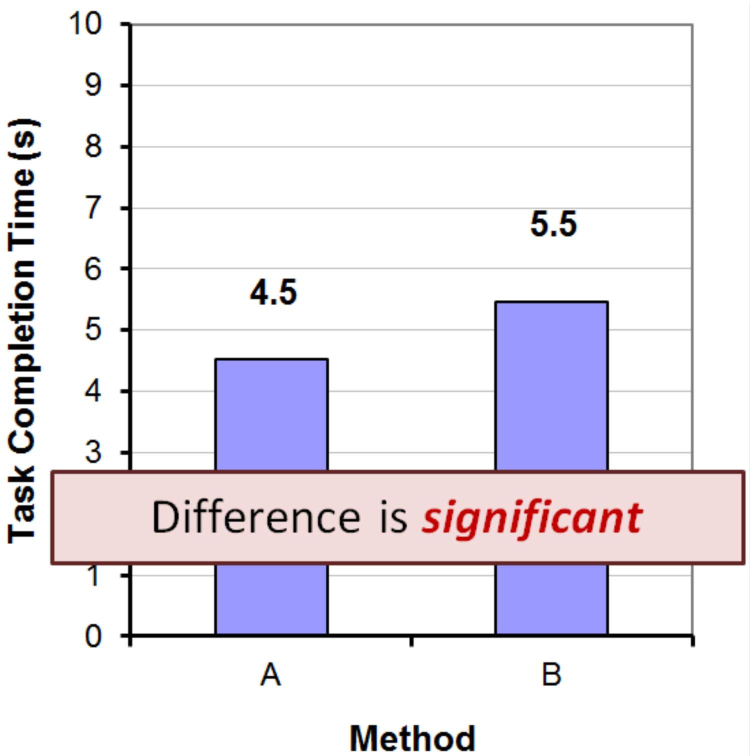
Why Analyze the Variance?

- 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?

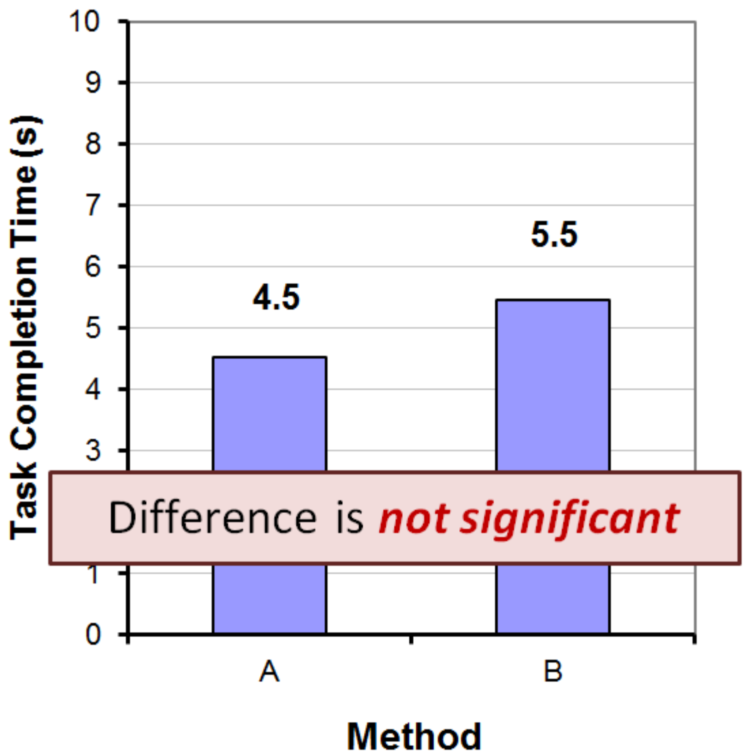
- Let's explain through two simple examples (next slide)

Example #1



“Significant” implies that in all likelihood the difference observed is due to the test conditions (Method A vs. Method B).

Example #2

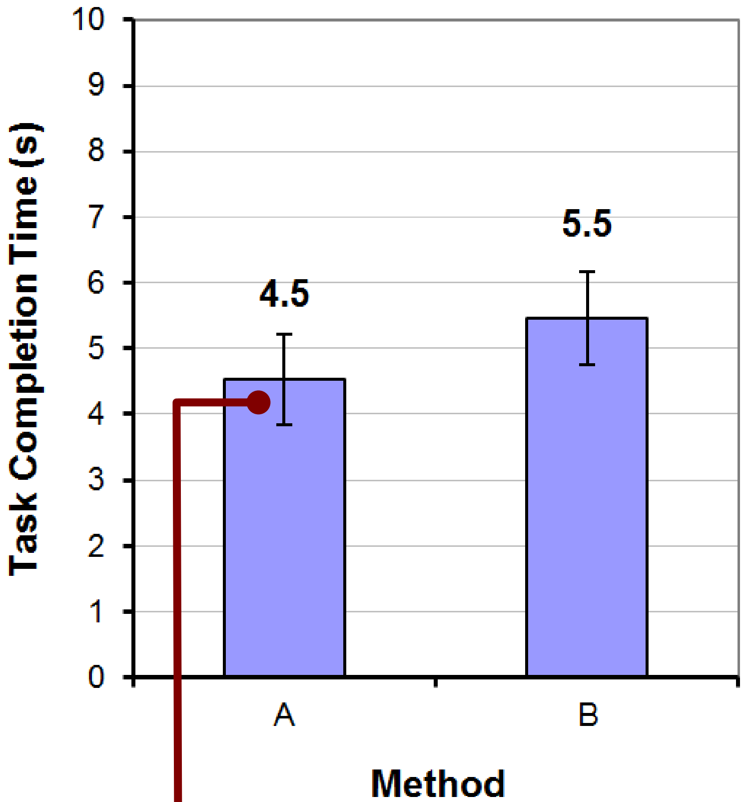


“Not significant” implies that the difference observed is likely due to chance.

File: `06-AnovaDemo.xlsx`

Example #1 - Details

Note: Within-subjects design

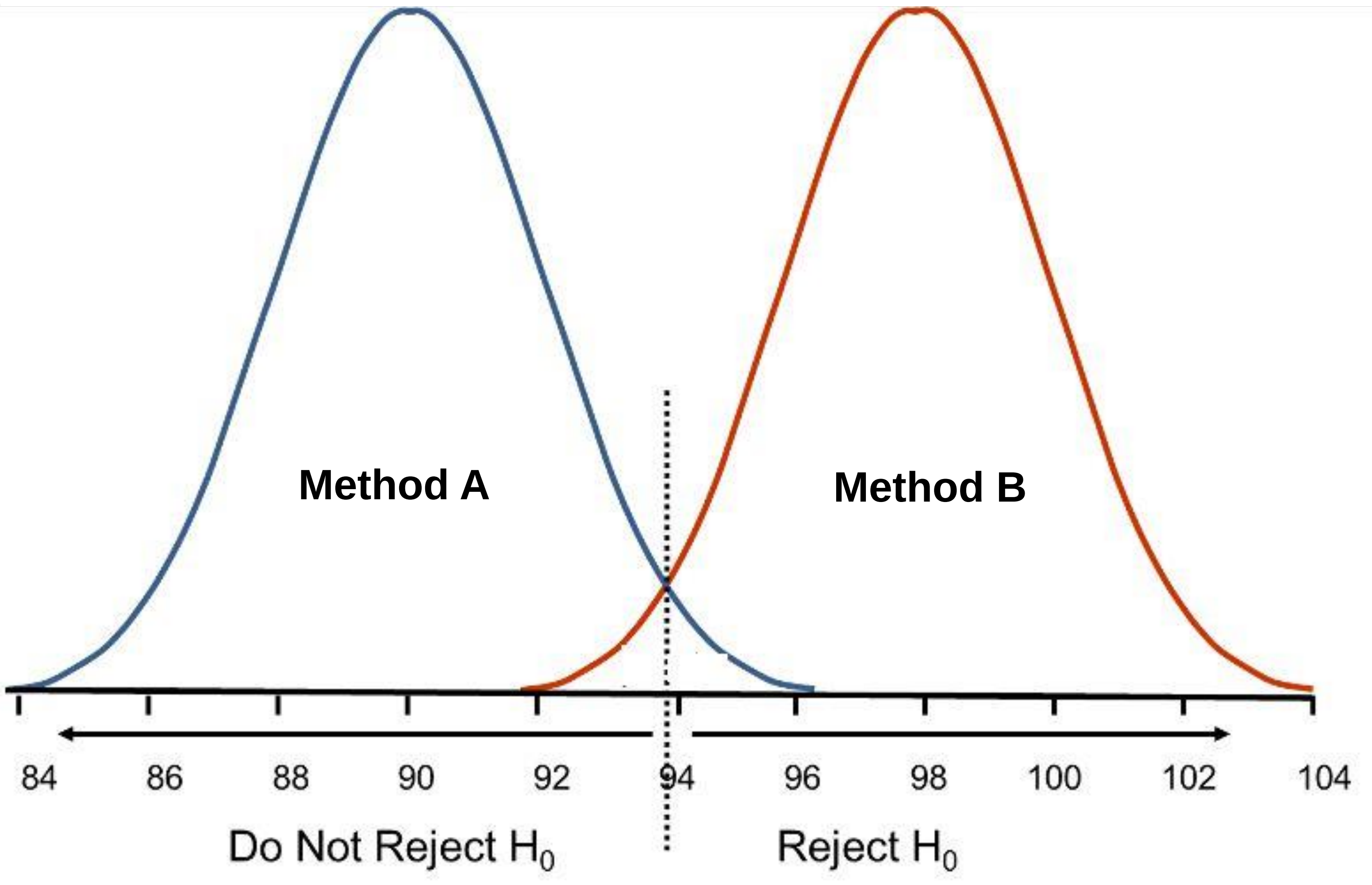


Participant	Method	
	A	B
1	5.3	5.7
2	3.6	4.8
3	5.2	5.1
4	3.6	4.5
5	4.6	6.0
6	4.1	6.8
7	4.0	6.0
8	4.8	4.6
9	5.2	5.5
10	5.1	5.6
Mean	4.5	5.5
SD	0.68	0.72

Error bars show
±1 standard deviation

Note: *SD* is the square root of the variance

Reject or Not Reject – That's the Question



Example #1 – ANOVA[1]

ANOVA Table for Task Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	9	5.080	.564				
Method	1	4.232	4.232	9.796	.0121	9.796	.804
Method * Subject	9	3.888	.432				

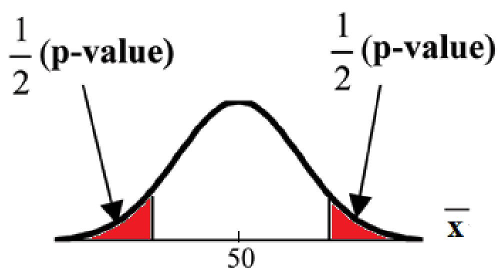
Probability of obtaining the observed data if the null hypothesis is true

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[1]

SS within method groups

MS=SS/df

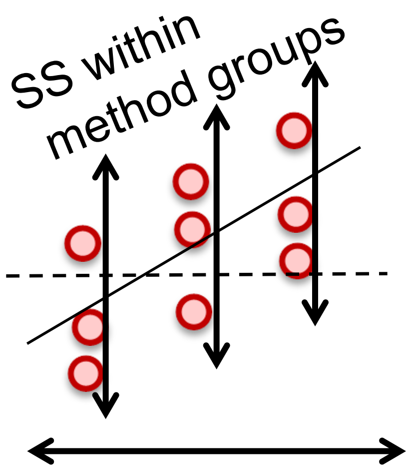
MS between/within here: 4.232/0.564 = 9.796

ANOVA Table for Task Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	9	5.080	.564				
Method	1	4.232	4.232	9.796	.0121	9.796	.804
Method * Subject	9	3.888	.432				

SS between method groups

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



SS between method groups

Reported as…

$F_{1,9} = 9.80, p < .05$

Thresholds for "p"

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

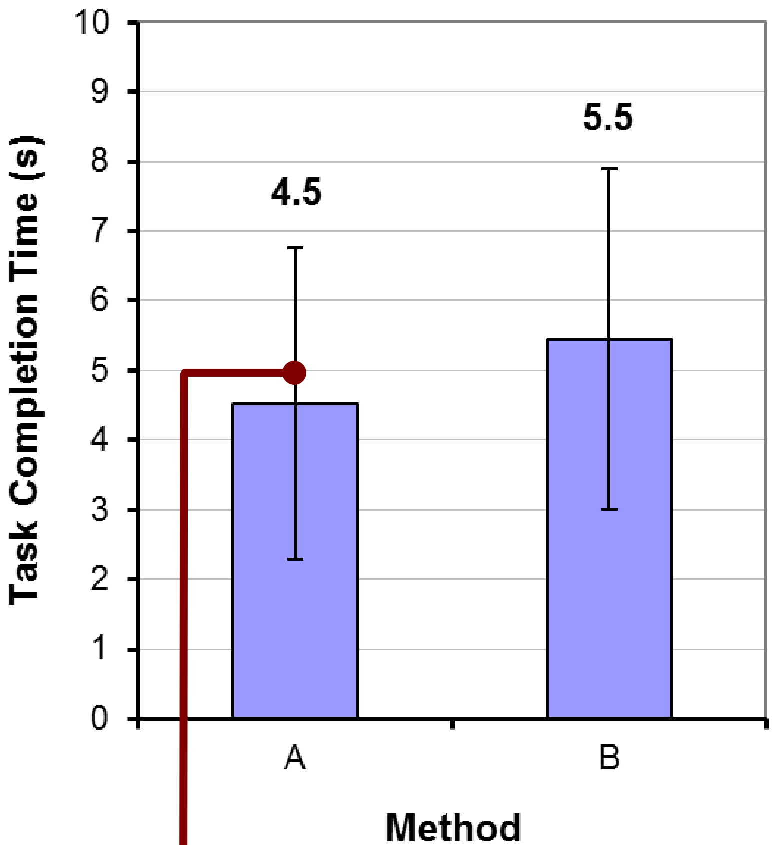
[1] ANOVA table created by *StatView* (now marketed as *JMP*, a product of SAS; www.sas.com)

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



Participant	Method	
	A	B
1	2.4	6.9
2	2.7	7.2
3	3.4	2.6
4	6.1	1.8
5	6.4	7.8
6	5.4	9.2
7	7.9	4.4
8	1.2	6.6
9	3.0	4.8
10	6.6	3.1
Mean	4.5	5.5
SD	2.23	2.45

Error bars show
±1 standard deviation

Example #2 – ANOVA

ANOVA Table for Task Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	9	37.372	4.152				
Method	1	4.324	4.324	.626	.4491	.626	.107
Method * Subject	9	62.140	6.904				

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

Reported as…

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

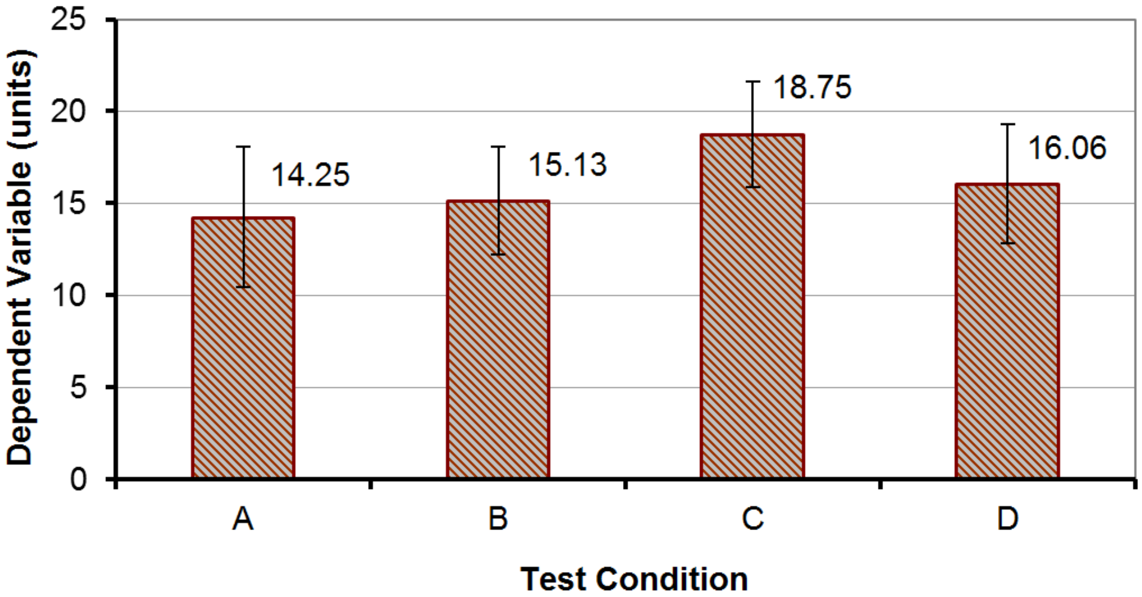
Note: For non-significant effects, use "ns" if $F < 1.0$, or "$p > .05$" if $F > 1.0$.

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

Participant	Test Condition			
	A	B	C	D
1	11	11	21	16
2	18	11	22	15
3	17	10	18	13
4	19	15	21	20
5	13	17	23	10
6	10	15	15	20
7	14	14	15	13
8	13	14	19	18
9	19	18	16	12
10	10	17	21	18
11	10	19	22	13
12	16	14	18	20
13	10	20	17	19
14	10	13	21	18
15	20	17	14	18
16	18	17	17	14
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)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	15	81.109	5.407				
Test Condition	3	182.172	60.724	4.954	.0047	14.862	.896
Test Condition * Subject	45	551.578	12.257				

- 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 → $(n – 1)$
  - Residual → $(n – 1)(m – 1)$
  - 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 %

	Mean Diff.	Crit. Diff.	P-Value	
A, B	-.875	3.302	.9003	
A, C	-4.500	3.302	.0032	S
A, D	-1.813	3.302	.4822	
B, C	-3.625	3.302	.0256	S
B, D	-.938	3.302	.8806	
C, D	2.688	3.302	.1520	

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

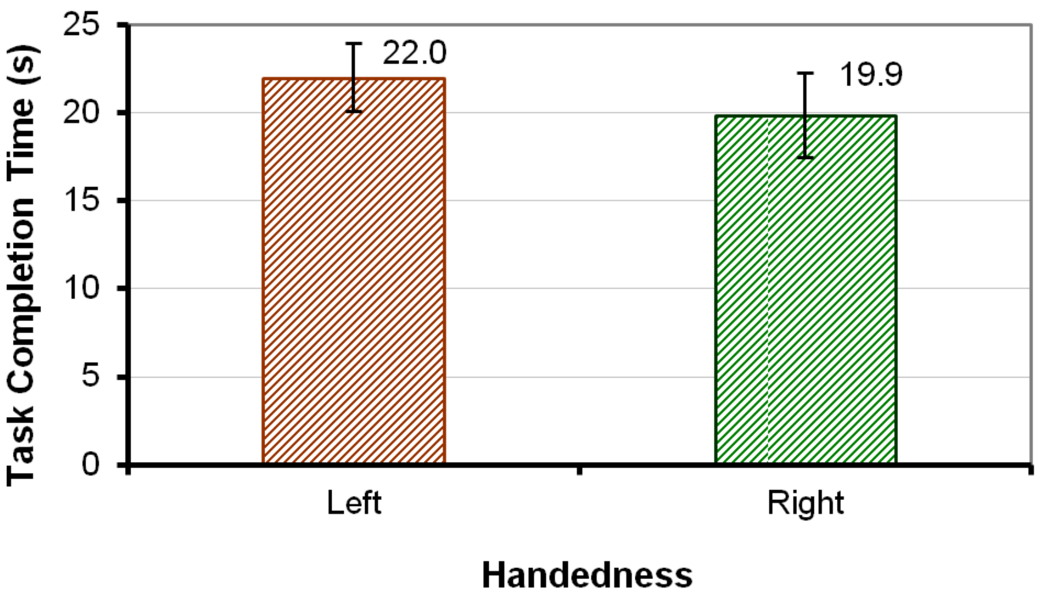
Between-subjects Designs

- 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 →

Participant	Task Completion Time (s)	Handedness
1	23	L
2	19	L
3	22	L
4	21	L
5	23	L
6	20	L
7	25	L
8	23	L
9	17	R
10	19	R
11	16	R
12	21	R
13	23	R
14	20	R
15	22	R
16	21	R
Mean	20.9	
SD	2.38	

Summary Data and Chart

Handedness	Task Completion Time (s)	
	Mean	*SD*
Left	22.0	1.93
Right	19.9	2.42



ANOVA

ANOVA Table for Task Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Handedness	1	18.063	18.063	3.781	.0722	3.781	.429
Residual	14	66.875	4.777				

- The difference was not statistically significant ($F_{1,14}$ = 3.78, p > .05)
- Degrees of freedom:
  - Effect → (n – 1)
  - Residual → (m – n)
  - Note: single-factor, between-subjects design

Two-way ANOVA

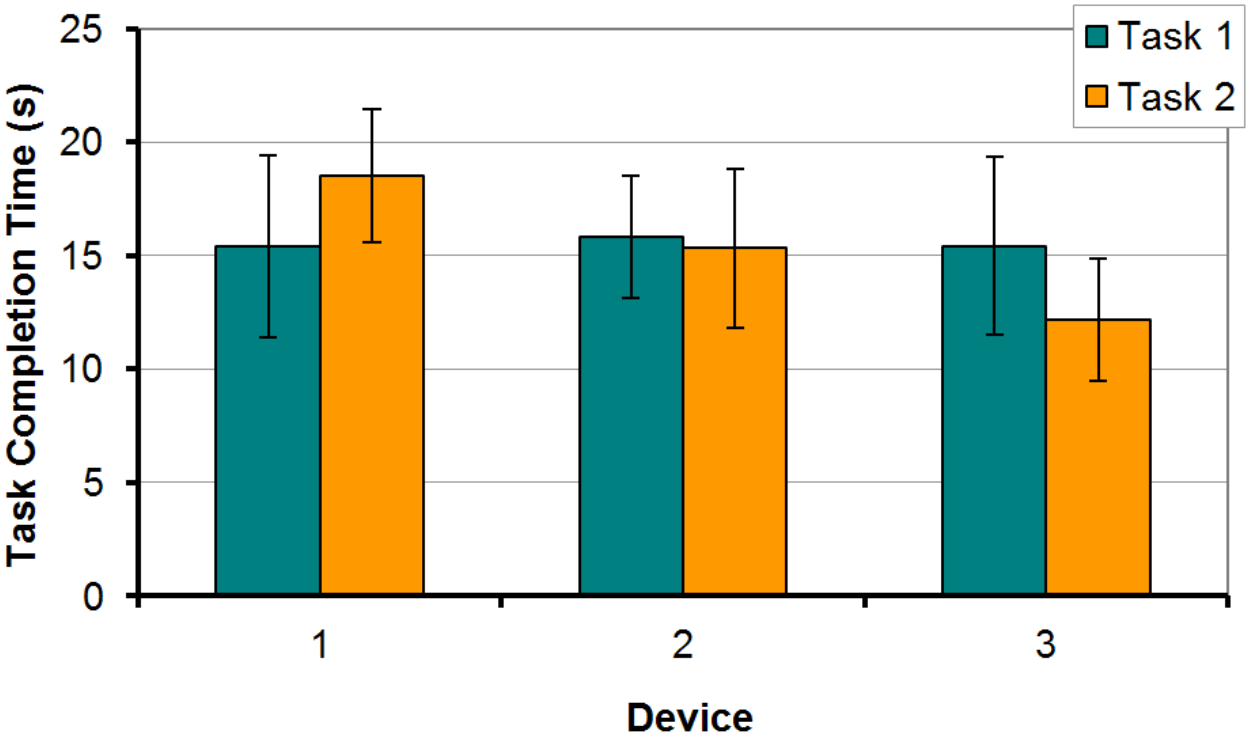
- An experiment with two independent variables is a *two-way design*
- ANOVA tests for
  - Two main effects + one interaction effect
- Example
  - Independent variables
    - Device → D1, D2, D3 (e.g., mouse, stylus, touchpad)
    - Task → T1, T2 (e.g., point-select, drag-select)
  - Dependent variable
    - Task completion time (or something, this isn't important here)
  - Both IVs assigned within-subjects
  - Participants: 12
  - Data set (next slide)

Data Set

Participant	Device 1		Device 2		Device 3	
	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2
1	11	18	15	13	20	14
2	10	14	17	15	11	13
3	10	23	13	20	20	16
4	18	18	11	12	11	10
5	20	21	19	14	19	8
6	14	21	20	11	17	13
7	14	16	15	20	16	12
8	20	21	18	20	14	12
9	14	15	13	17	16	14
10	20	15	18	10	11	16
11	14	20	15	16	10	9
12	20	20	16	16	20	9
Mean	15.4	18.5	15.8	15.3	15.4	12.2
SD	4.01	2.94	2.69	3.50	3.92	2.69

Summary Data and Chart

	Task 1	Task 2	*Mean*
Device 1	15.4	18.5	17.0
Device 2	15.8	15.3	15.6
Device 3	15.4	12.2	13.8
Mean	15.6	15.3	15.4



ANOVA

ANOVA Table for Task Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	11	134.778	12.253				
Device	2	121.028	60.514	5.865	.0091	11.731	.831
Device * Subject	22	226.972	10.317				
Task	1	.889	.889	.076	.7875	.076	.057
Task * Subject	11	128.111	11.646				
Device * Task	2	121.028	60.514	5.435	.0121	10.869	.798
Device * Task * Subject	22	244.972	11.135				

Can you pull the relevant statistics from this chart and craft statements indicating the outcome of the ANOVA?

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 ($F_{1,11} = 0.076$, 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.

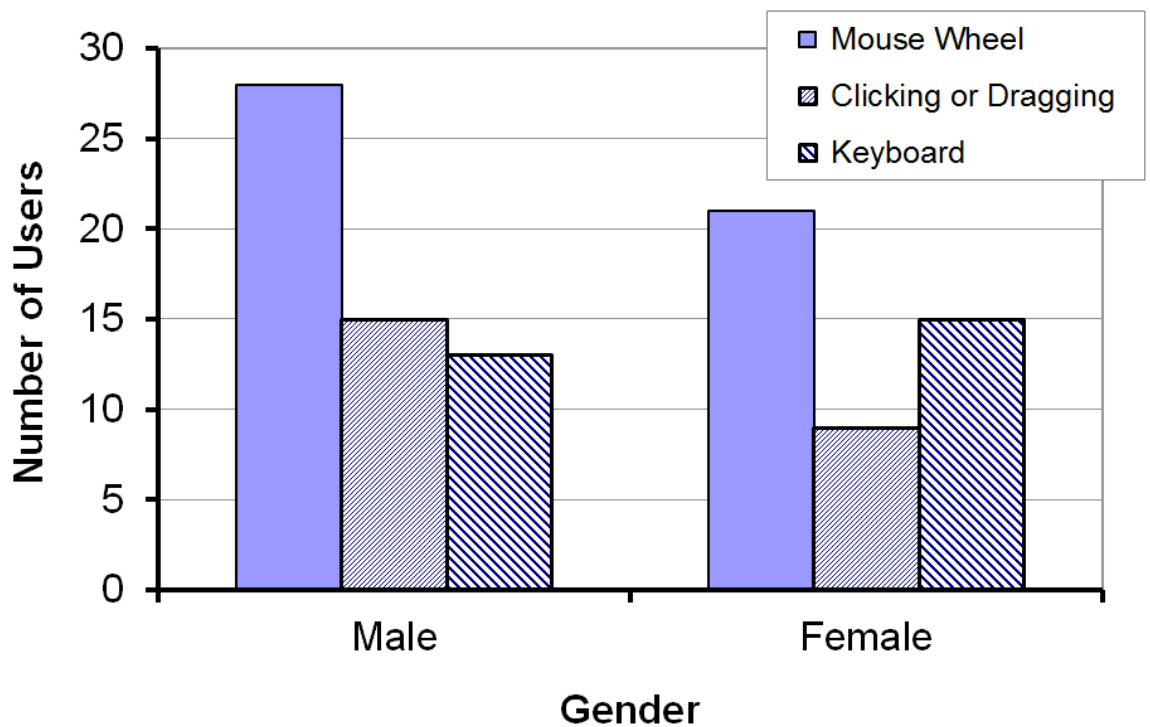
Chi-square Test (Nominal Data)

- 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

Observed Number of Users				
Gender	Scrolling Method			Total
	MW	CD	KB	
Male	28	15	13	56
Female	21	9	15	45
Total	49	24	28	101

MW = mouse wheel
CD = clicking, dragging
KB = keyboard



Chi-square – Example #1

56.0·49.0/101=27.2

Expected Number of Users				
Gender	Scrolling Method			Total
	MW	CD	KB	
Male	27.2	13.3	15.5	56.0
Female	21.8	10.7	12.5	45.0
Total	49.0	24.0	28.0	101

(Expected-Observed)^2/Expected=(28-27.2)^2/27.2

Chi Squares				
Gender	Scrolling Method			Total
	MW	CD	KB	
Male	0.025	0.215	0.411	0.651
Female	0.032	0.268	0.511	0.811
Total	0.057	0.483	0.922	**1.462**

Significant if it exceeds critical value (next slide)

$$\chi^2 = 1.462$$

(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 = number of rows, c = number of columns

Significance Threshold (α)	Degrees of Freedom							
	1	2	3	4	5	6	7	8
.1	2.71	4.61	6.25	7.78	9.24	10.65	12.02	13.36
.05	3.84	5.99	7.82	9.49	11.07	12.59	14.07	15.51
.01	6.64	9.21	11.35	13.28	15.09	16.81	18.48	20.09
.001	10.83	13.82	16.27	18.47	20.52	22.46	24.32	26.13

$\chi^2 = 1.462$ (< 5.99 ∴ not significant)