CSE 590
Data Science Fundamentals

Statistics Foundations

Klaus Mueller

Computer Science Department
Stony Brook University and SUNY Korea

Lecture	Topic	Projects
1	Intro, schedule, and logistics	
2	Data Science components and tasks	
3	Data types	Project #1 out
4	Introduction to R, statistics foundations	
5	Introduction to D3, visual analytics	
6	Data preparation and reduction	
7	Data preparation and reduction	Project #1 due
8	Similarity and distances	Project #2 out
9	Similarity and distances	
10	Cluster analysis	
11	Cluster analysis	
12	Pattern miming	Project #2 due
13	Pattern mining	
14	Outlier analysis	
15	Outlier analysis	Final Project proposal due
16	Classifiers	
17	Midterm	
18	Classifiers	
19	Optimization and model fitting	
20	Optimization and model fitting	
21	Causal modeling	
22	Streaming data	Final Project preliminary report due
23	Text data	
24	Time series data	
25	Graph data	
26	Scalability and data engineering	
27	Data journalism	
	Final project presentation	Final Project slides and final report due

Normal Distribution

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

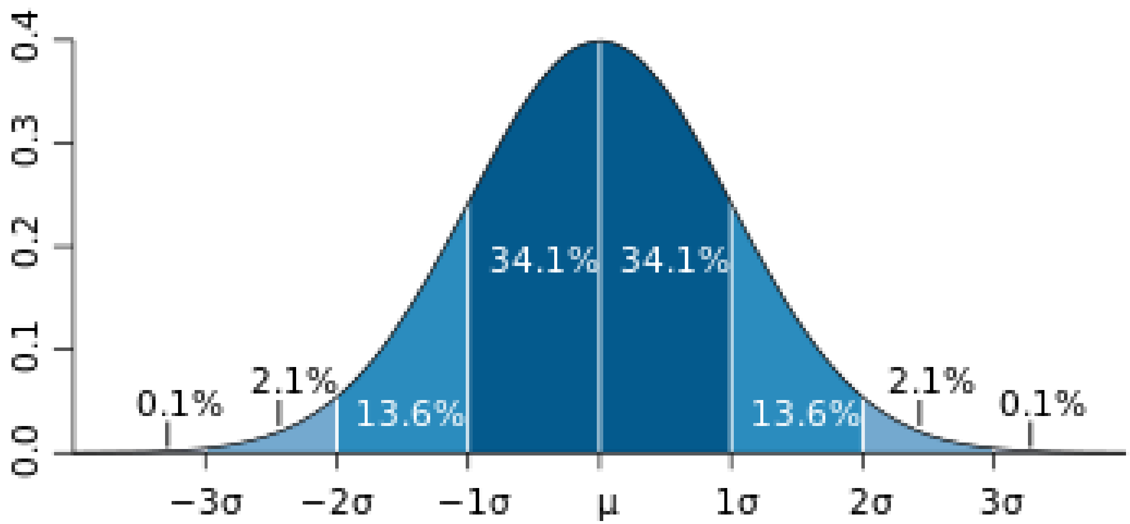
σ = standard deviation

$\sum$ = 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



Central Limit Theorem

Important relationship

- sample mean is normal distributed, too
- the standard error of these means is

$$s = \sqrt{s^2} \propto \frac{\sim 1}{\sqrt{n}} \quad \text{(for large } n\text{)}$$

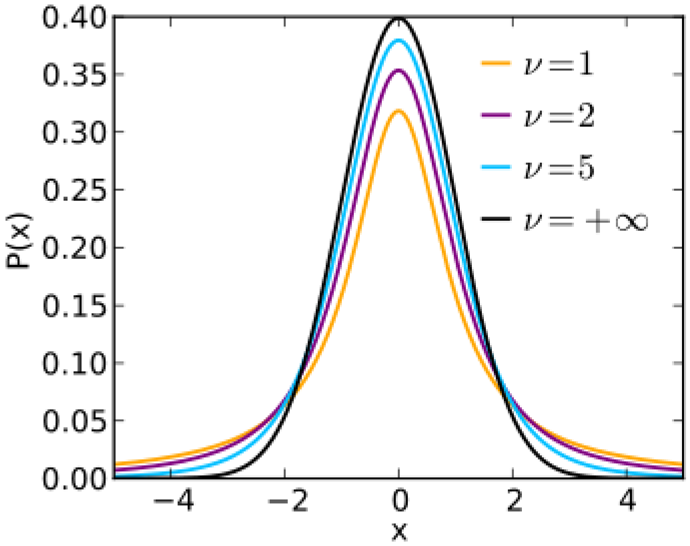
So the more data samples you have

- the smaller the sample deviation
- the closer the sample mean $\bar{x}$ is to the true mean μ

t-Distribution

When n is small the distribution tails are more expressed

- this is captured by the t-distribution
- once n gets large the t-distribution resembles the normal distribution



υ = degree of freedom = n-1

Chi Square Distribution

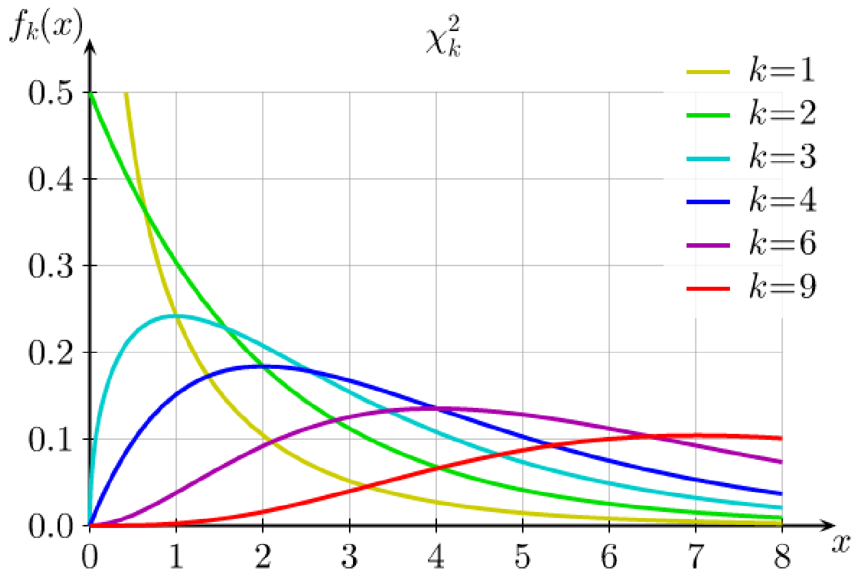
Written as χ2

$$Q = \sum_{i=1}^{k} Z_i^2,$$

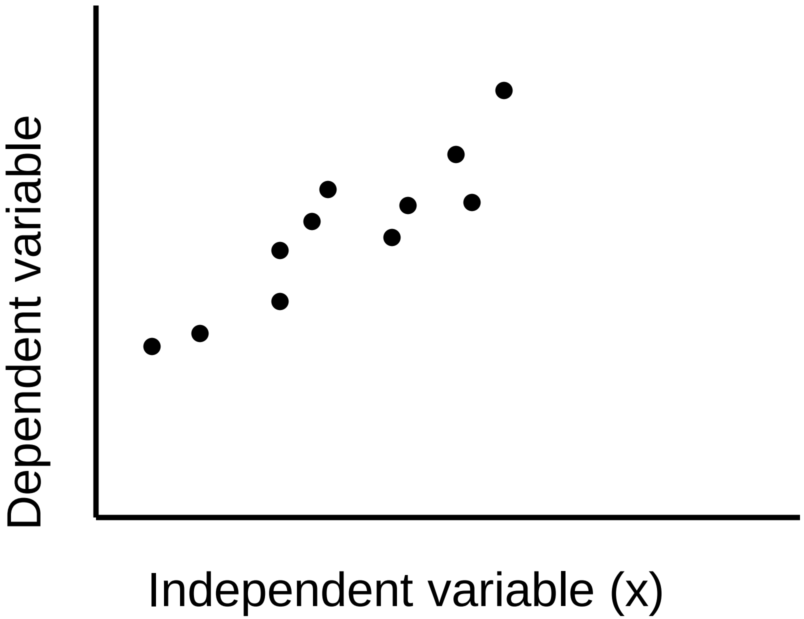
Special case of the Γ (gamma) distribution

Cumulative distribution function Q of the square of a random variable Z

- k is the degree of freedom = number of samples
- probability density function



Regression

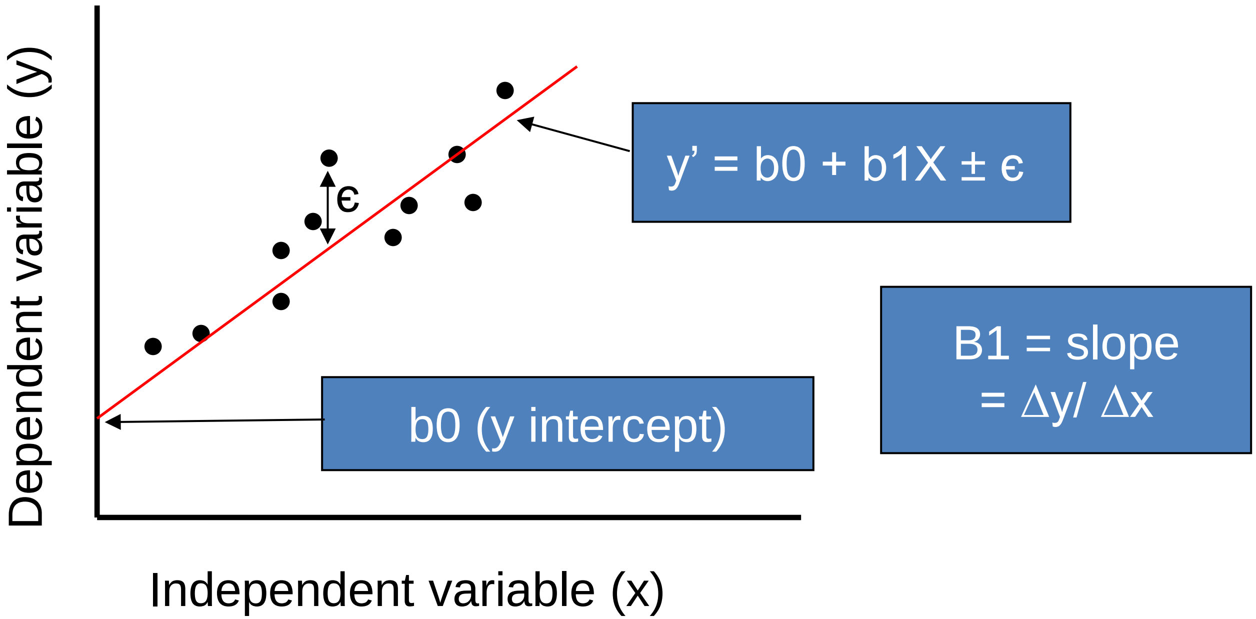


Regression is the attempt to explain the variation in a dependent variable using the variation in independent variables.

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.

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.

Calculations

Consider the linear model:

$$\sum_{j=1}^{n} X_{ij}\beta_j = y_i,\ (i = 1, 2, \dots, m),$$

In matrix form:

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{y},$$

where

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mn} \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}.$$

Error of the model:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\arg\min}\, S(\boldsymbol{\beta}),$$

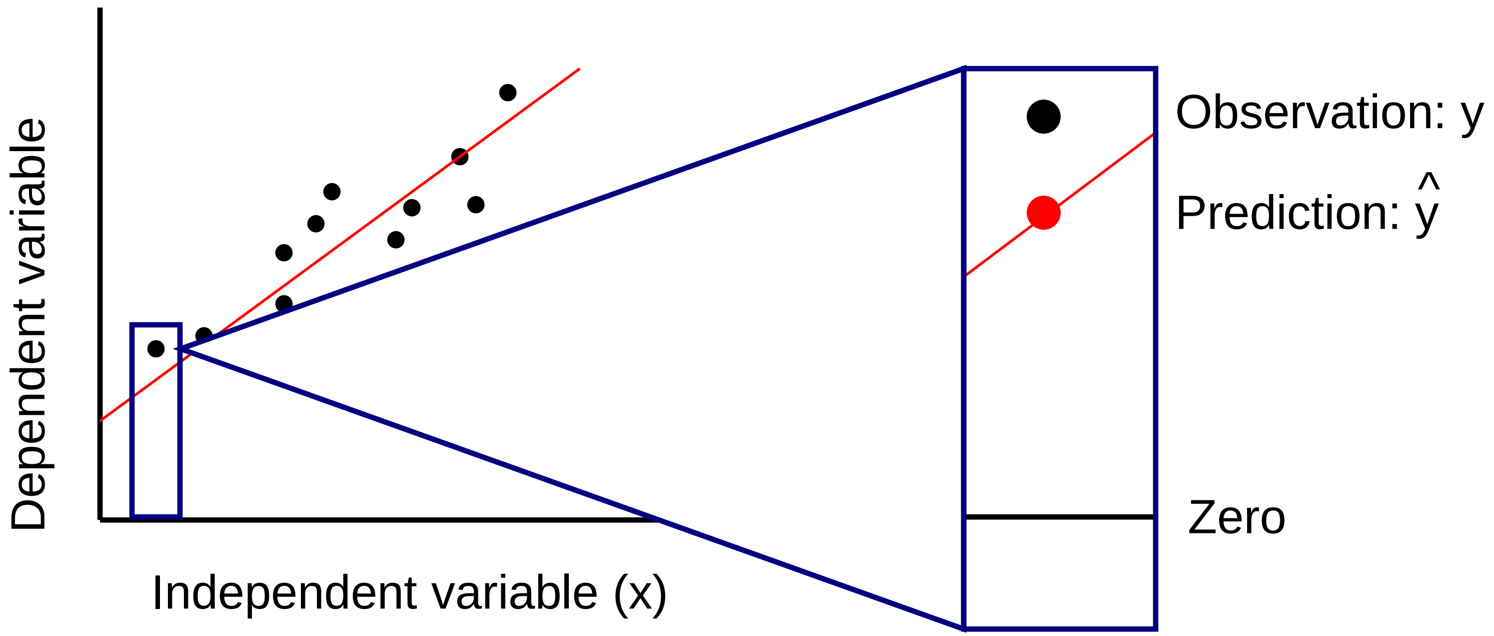
where the objective function S is given by

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{m} \left| y_i - \sum_{j=1}^{n} X_{ij}\beta_j \right|^2 = \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right\|^2.$$

Find the coefficients using least squares optimization

$$\widehat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \text{ and } \hat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

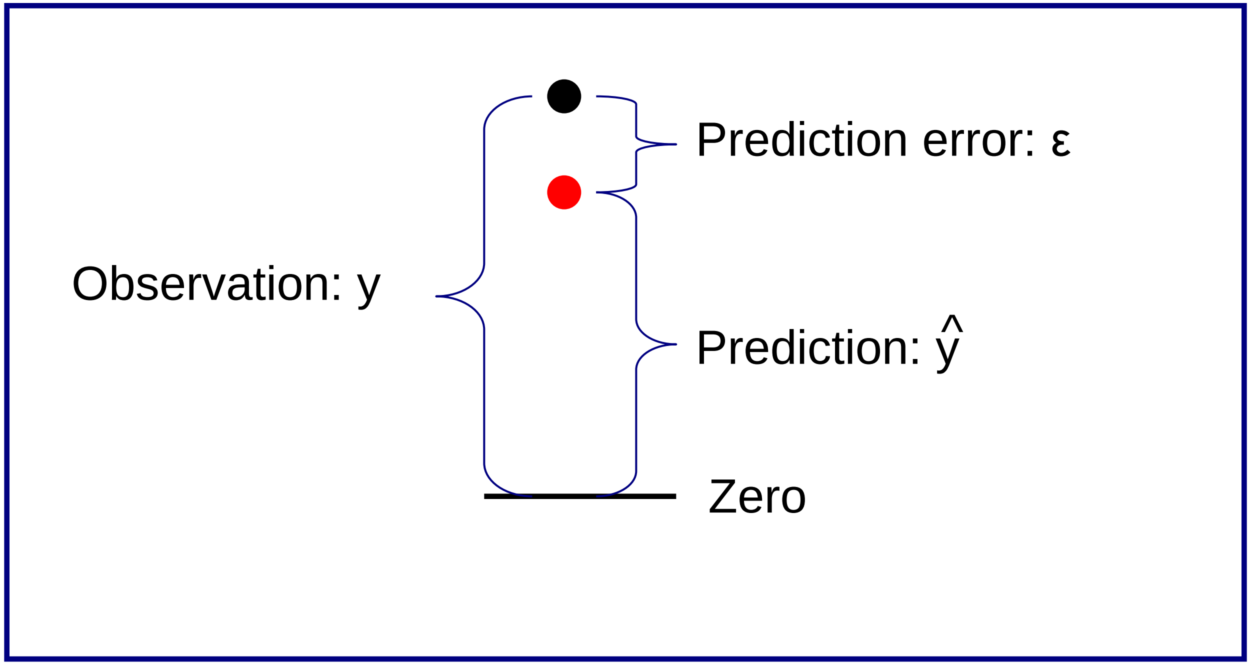
Linear Regression



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

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

Linear Regression

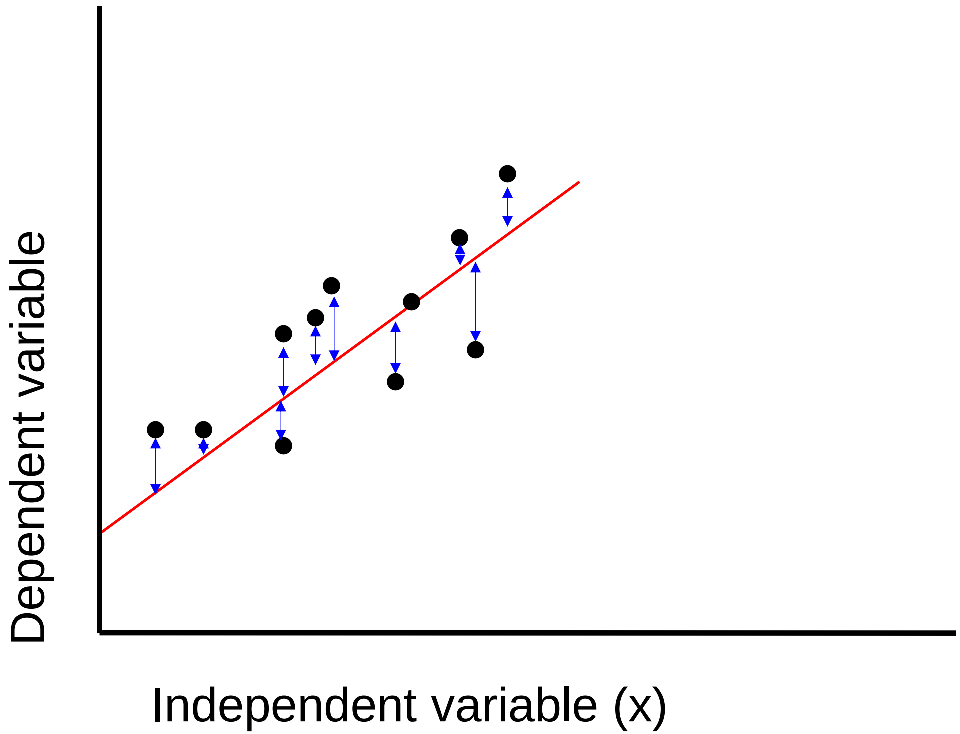


For each observation, the variation can be described as:

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

Actual = Explained + Error

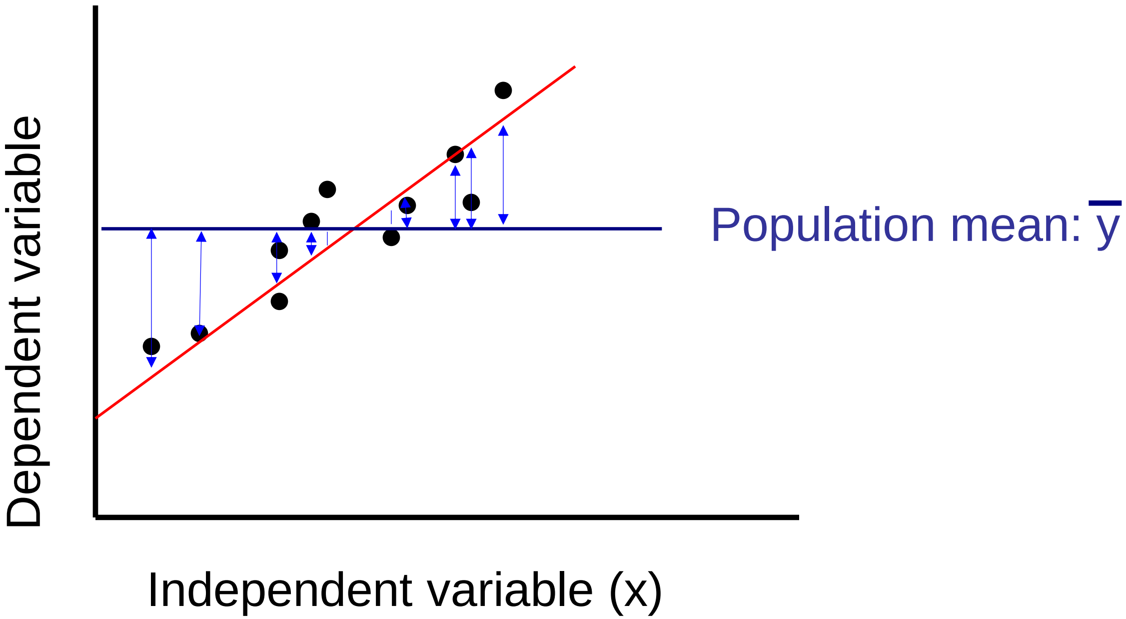
Regression



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)

Coefficient of determination (0…1)

- goodness of fit

$$R^2 \equiv 1 - \frac{SS_{\rm res}}{SS_{\rm tot}}.$$

Hypothesis Testing

next 17 slides adapted from L. Scott MacKenzie
Human Computer Interaction

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)

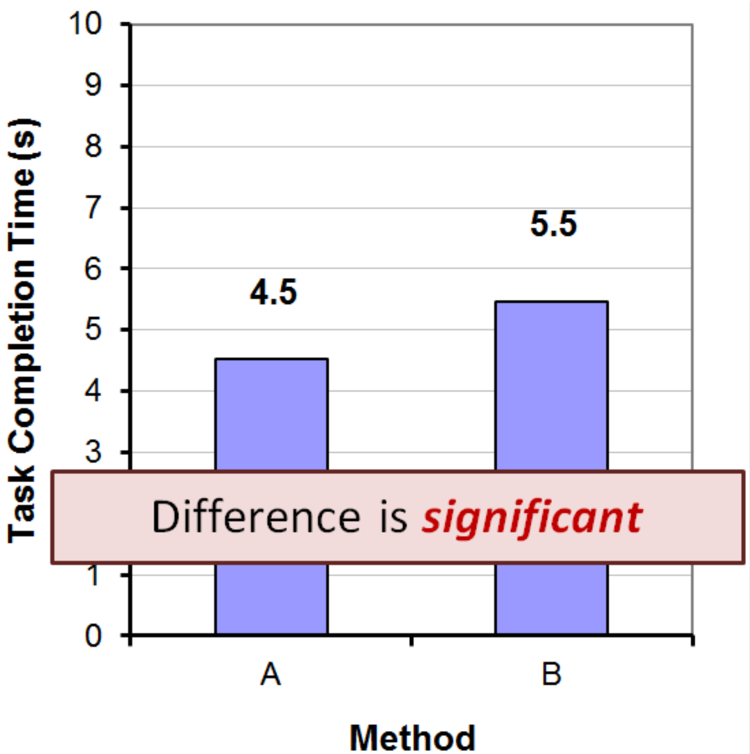
Why Analyse the Variance?

- Seems odd that we analyse the variance, but the research question is concerned with the overall means:

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

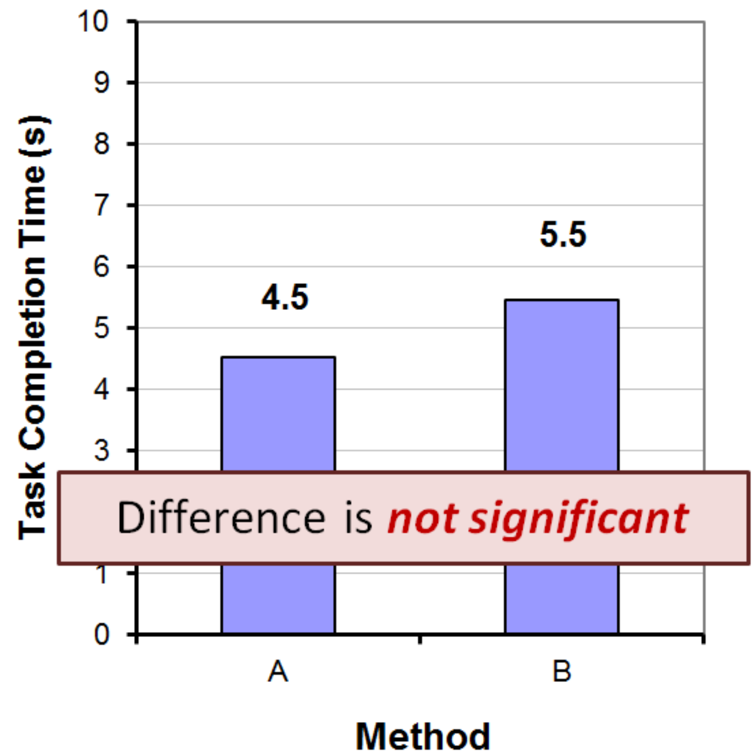
- Let's explain through two simple examples (next slide)

Example #1



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

Example #2

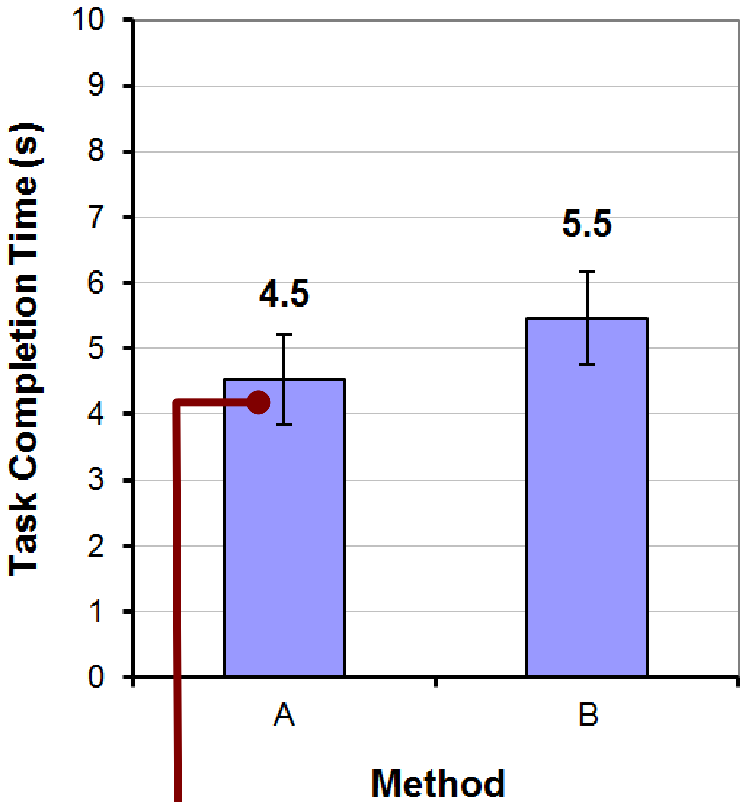


"Not significant" implies that the difference observed is likely due to chance.

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

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

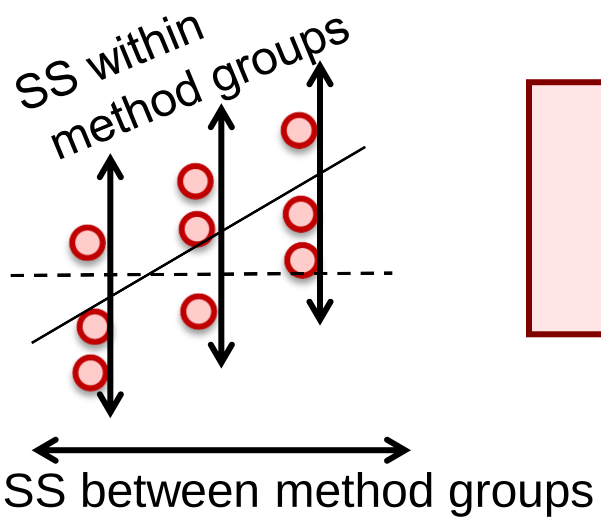
MS between/within

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



Reported as…

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

Thresholds for "p"

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

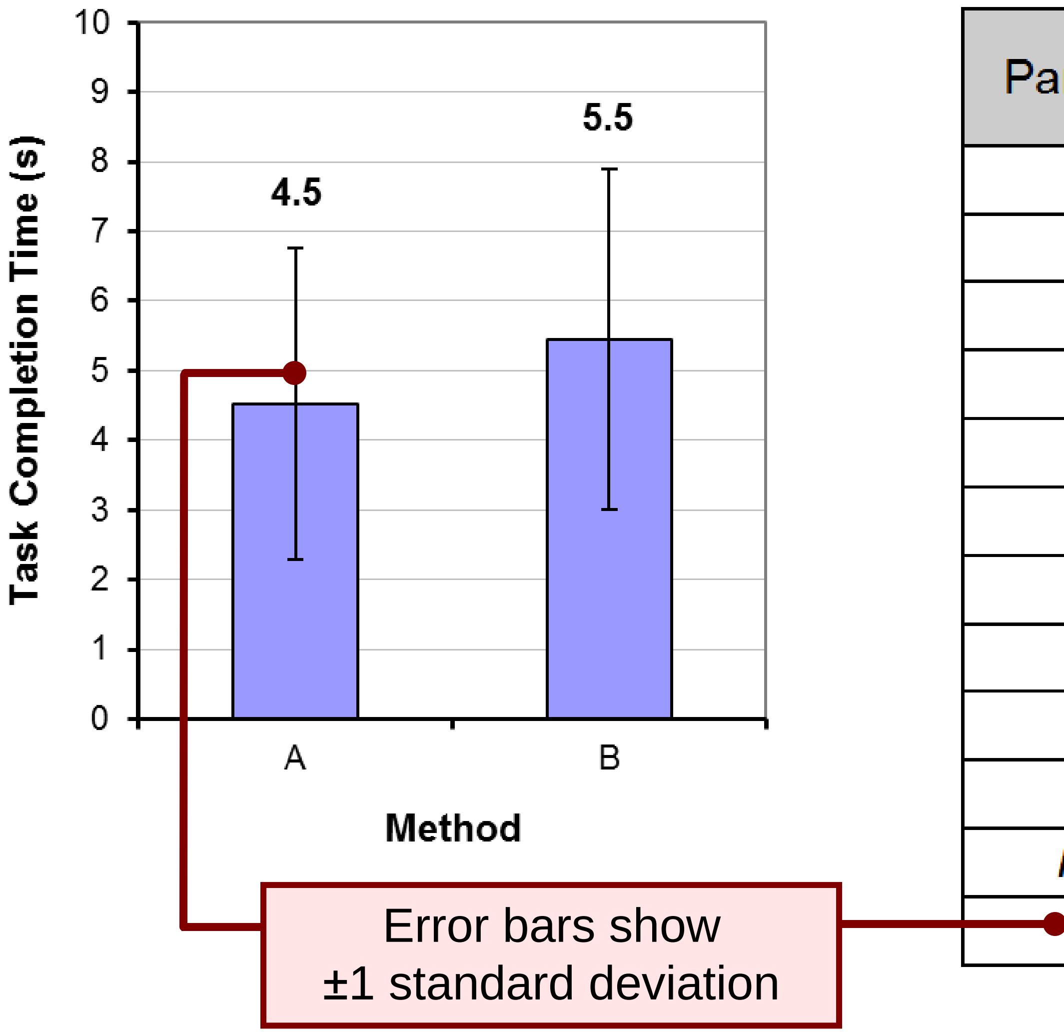
[1] ANOVA table created by *StatView* (now marketed as *JMP*, a product of SAS; www.sas.com)

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

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).

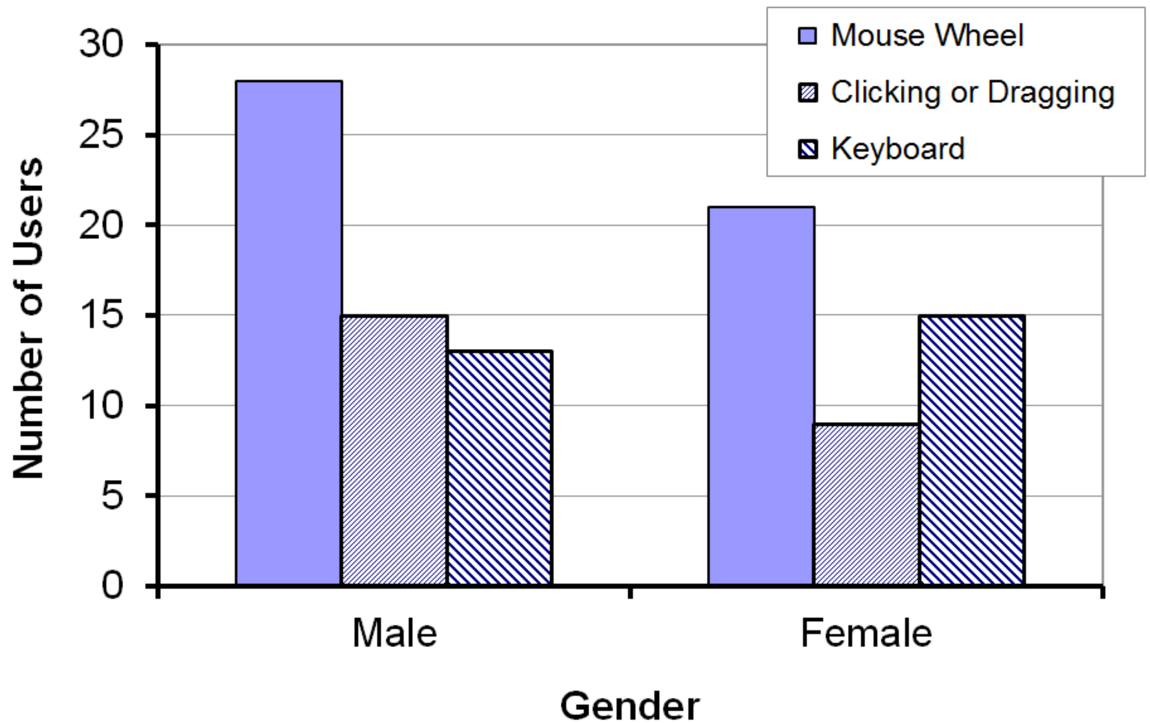
Chi-square Test (Nominal Data)

- 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

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

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 $\therefore$ not significant)

Chi-square – Example #2

- Research question:
  - *Do students, professors, and parents differ in their responses to the question: Students should be allowed to use mobile phones during classroom lectures?*
- Data:

Observed Number of People				
Opinion	Category			Total
	Student	Professor	Parent	
Agree	10	12	98	120
Disagree	30	48	102	180
Total	40	60	200	300

Chi-square – Example #2

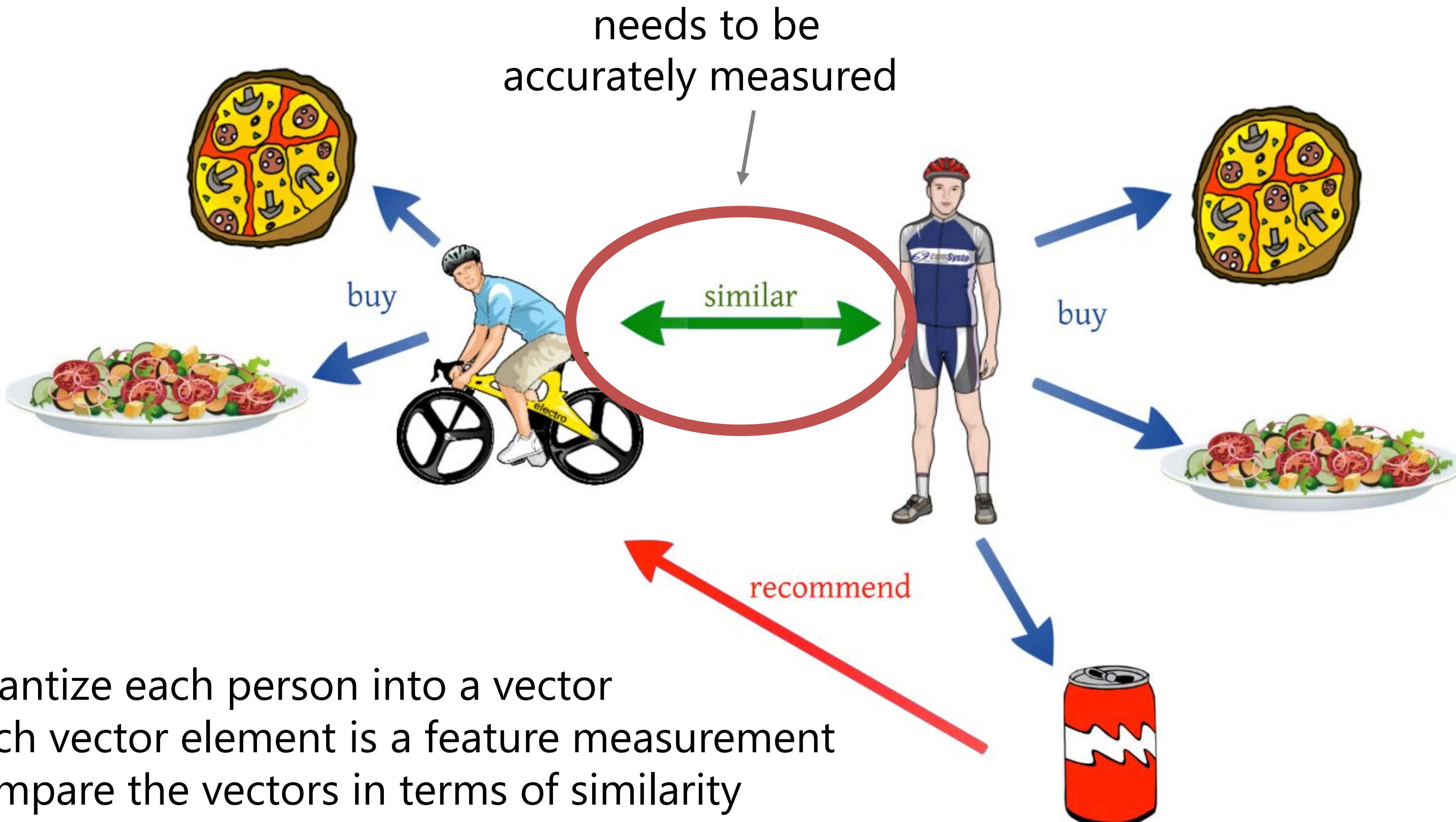
- Result: significant difference in responses ($\chi^2 = 20.5$, $p < .0001$)
- Post hoc comparisons reveal that opinions differ between students:parents and professors:parents (students:professors do not differ significantly in their responses)

```
text>type chisquare-ex2.txt
10   12   98
30   48  102

text>java ChiSquare chisquare-ex2.txt -ph
Chi-square(2) = 20.500
p = 0.0000

----------------------------------------------------
------ Pairwise Comparisons (using contrasts) ------
----------------------------------------------------
Pair 1:2   --->   Chi-square(2) =  0.340, p = 0.8437
Pair 1:3   --->   Chi-square(2) =  9.702, p = 0.0078
Pair 2:3   --->   Chi-square(2) = 21.475, p = 0.0000
----------------------------------------------------

text>
```

1 = students, 2 = professors, 3 = parents

How to Measure Similarity

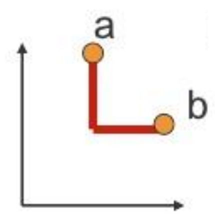
Similarity Functions



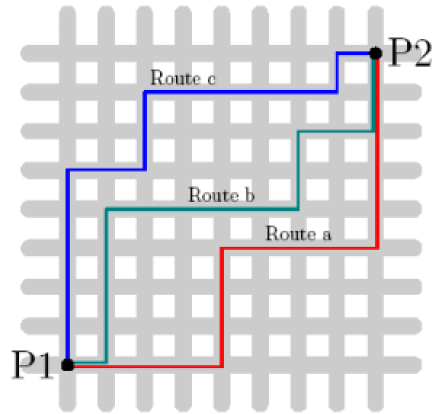
quantize each person into a vector
each vector element is a feature measurement
compare the vectors in terms of similarity
similarity also called distance functions

Metric Distances

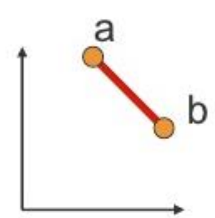
Manhattan distance



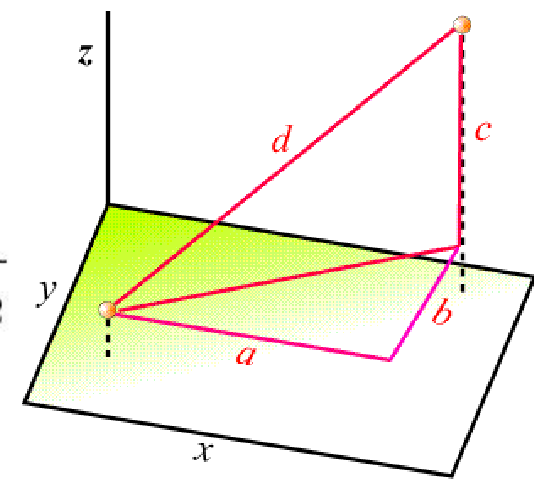
$$\text{dist}(a,b) = \|a - b\|_1 = \sum_i |a_i - b_i|$$



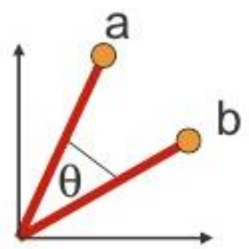
Euclidian distance



$$\text{dist}(a,b) = \|a - b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$$



Cosine Similarity



dist(a,b) = $cos^{-1} \frac{\langle a, b \rangle}{||a|| \, ||b||}$

how is this related to correlation?

Pearson's Correlation = correlation similarity

$$r = r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

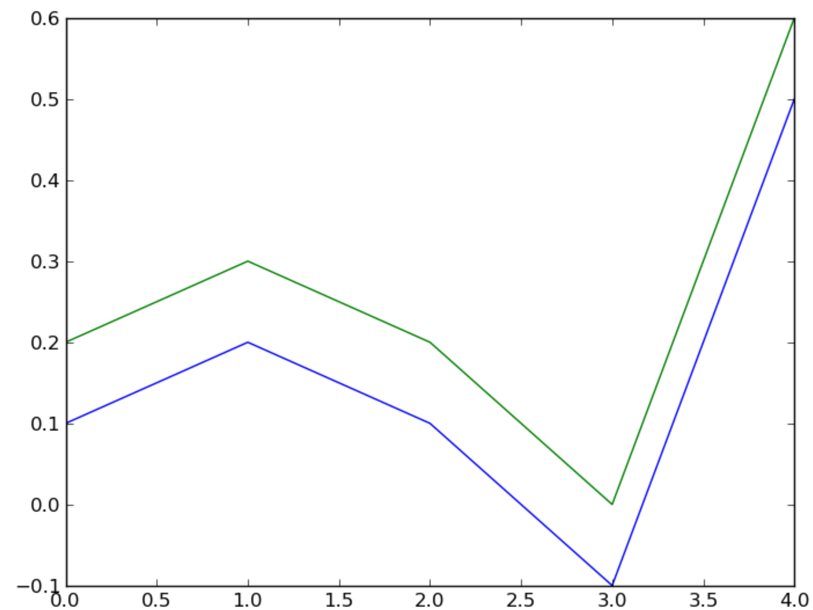
Correlation vs. Cosine Distance

Correlation distance is invariant to addition of a constant

- subtracts out by construction
- green and blue curve have correlation of 1
- but cosine similarity is < 1
- correlated vectors just vary in the same way
- cosine similarity is stricter

Both correlation and cosine similarity are invariant to multiplication with a constant

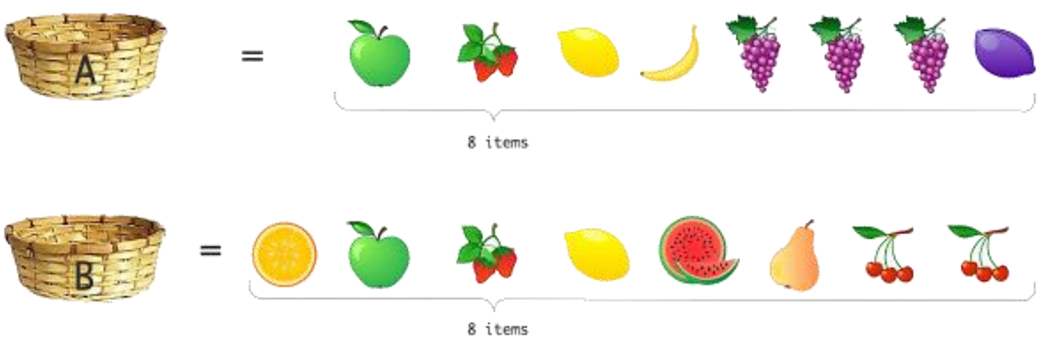
- invariant to scaling

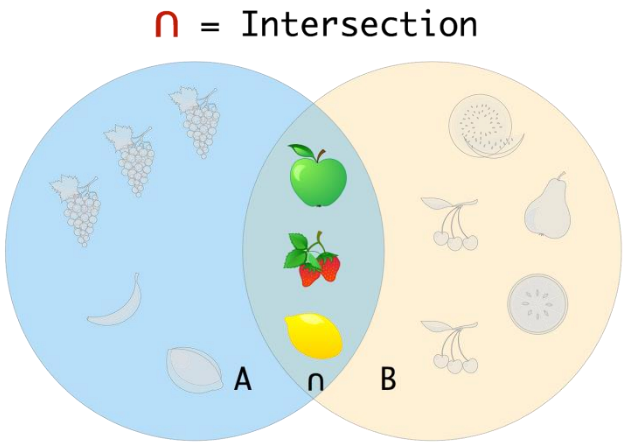


green = blue + 0.1

Jaccard Distance

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$



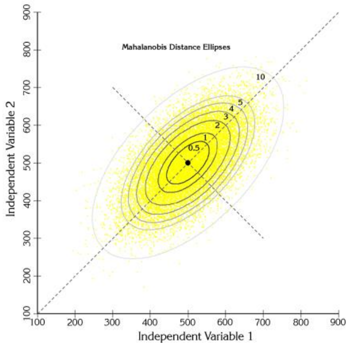


What's the Jaccard similarity of the two baskets A and B?

Mahalanobis Distance

The distance between a point P and a distribution D

- measures how many standard deviations P is away from the mean of D
- S is the covariance matrix of the distribution D



$$D_M(x) = \sqrt{(x-\mu)^T S^{-1} (x-\mu)}.$$

- when the covariance matrix is diagonal then the Mahalanobis distance reduces to the normalized Euclidian distance

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{N} \frac{(x_i - y_i)^2}{s_i^2}},$$

- what happens when the covariance matrix is the identity matrix?