Tutorial on Statistics, Probability and Information Theory for Language Engineers

Prof. Ibrahim F. Imam

Full Professor and Assistant Dean,
College of Computing and Information Technology
Arab Academy for Science, Technology & Maritime Transport, Cairo

Adjunct Professor, Computer Science Department,
College of Engineering, Virginia Tech. University, VA, USA

Email: ifi05@yahoo.com Phone: 012-2242929

Contents of the Tutorial

OUTLINE

BASIC MATHEMATICS

Part 0

Basic Concepts

BASIC MATHEMATICS

$$\sum_{i=1}^{n} i = 1 + 2 + \ldots + n$$

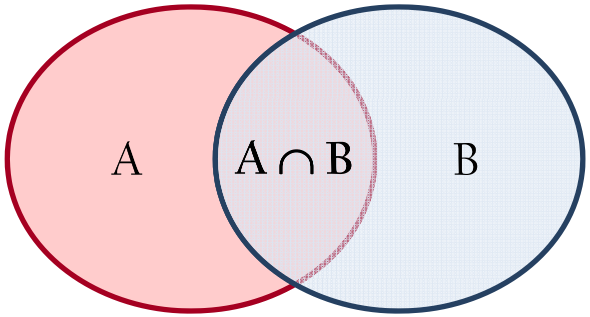
$$\prod_{i=1}^{n} i = 1 * 2 * \ldots * n$$

$$\sum_{i=1}^{n} ki = k\sum_{i=1}^{n} i$$

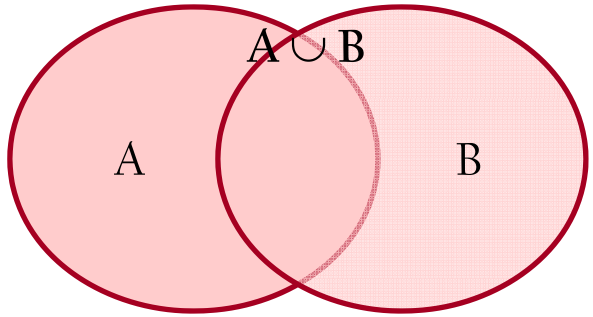
$$\prod_{i=1}^{n} ki = k\prod_{i=1}^{n} i$$

Introduction to Set Theory

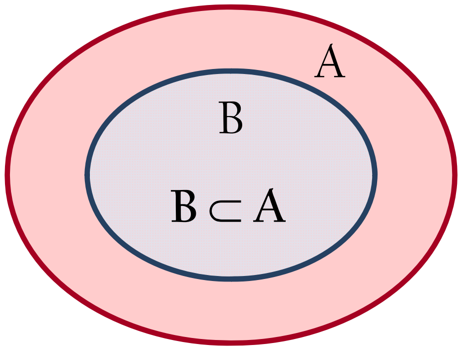
- A set is a collection of distinct items (Example: A = {1, 2, 3, 4, 5})



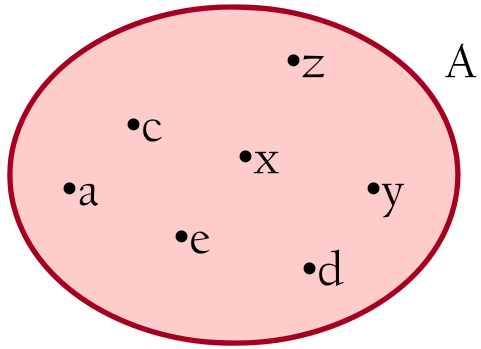
Intersection



Union



Sub-set & Super-set



$x \in A$; $a \in A$; $d \in A$; ...

Introduction to Set Theory

- $A = \{a, c, e, d, x, y, z\}$ $B = \{b, c, d, y, m, n\}$ $C = \{c, d\}$

$A \cap B = \{c, d, y\}$

Intersection

$A \cup B = \{a, b, c, d, e, m, n, x, y, z\}$

Union

$A \not\subset B$ $C \subset B$ $C \subset A$

Sub-set & Super-set

$x \in A$; $x \notin B$; $x \notin C$

Belong Relationship

Φ/ϕ is the empty set $\cap \cup \subset \not\subset \in \notin \neg \wedge \vee$

Introduction to Set Theory

- $A \cap (B \cap C) = (A \cap B) \cap C$ & $A \cup (B \cup C) = (A \cup B) \cup C$
- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
- $\neg(\neg A) = A$
- $\neg(A \cap B) = \neg A \cup \neg B$

Introduction to Propositional Logic

- It is also called the Zero Order Logic
- A sentence X can be either true or false (1 or 0)

X
0
1

Y
0
1

X	Y	$X \wedge Y$
0	0	0
0	1	0
1	0	0
1	1	1

X	Y	$X \vee Y$
0	0	0
0	1	1
1	0	1
1	1	1

X	Y	$X \rightarrow Y$
0	0	1
0	1	1
1	0	0
1	1	1

X	Y	X XOR Y
0	0	1
0	1	0
1	0	0
1	1	1

$X \rightarrow Y = \neg X \vee Y$
$\neg(X \wedge Y) = \neg X \vee \neg Y$
$X \wedge X = X$ & $X \vee X = X$
$X \vee (Y \wedge Z) = (X \vee Y) \wedge (X \vee Z)$
$\neg(\neg X) = X$

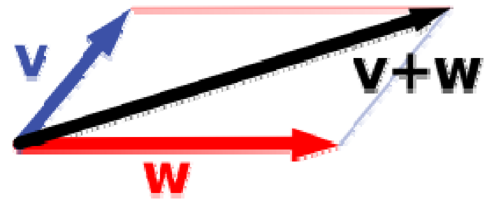
Introduction to Vectors

Part 1

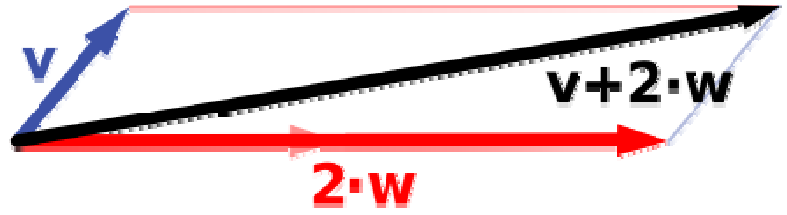
Representing Documents
As Vectors

Introduction to Vectors

Adding two vectors
$(x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1 + y_2)$

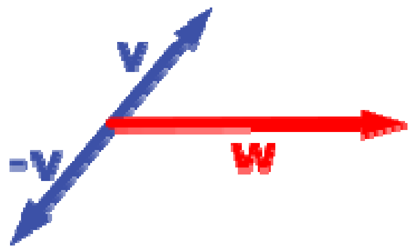


Multiplying a vector by a constant and adding it to another vector
$(x_1, y_1) + (2.x_2, 2.\ y_2) = (x_1 + 2x_2, y_1 + 2y_2)$



Multiplying a vector by -1
$-(x_1, y_1) = (-x_1, -y_1)$

Multiplying a vector by a constant
$2\ .\ (x_2, y_2) = (2x_2, 2y_2)$



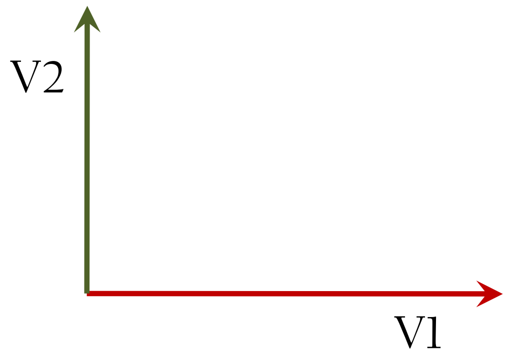
Introduction to Vectors

Multiplying two orthogonal vectors equal to zero.

Examples:

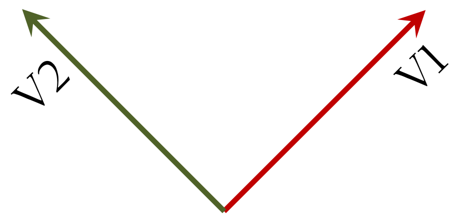
V1 =(5, 0) & V2=(0, 4)

V1 . V2 = 0



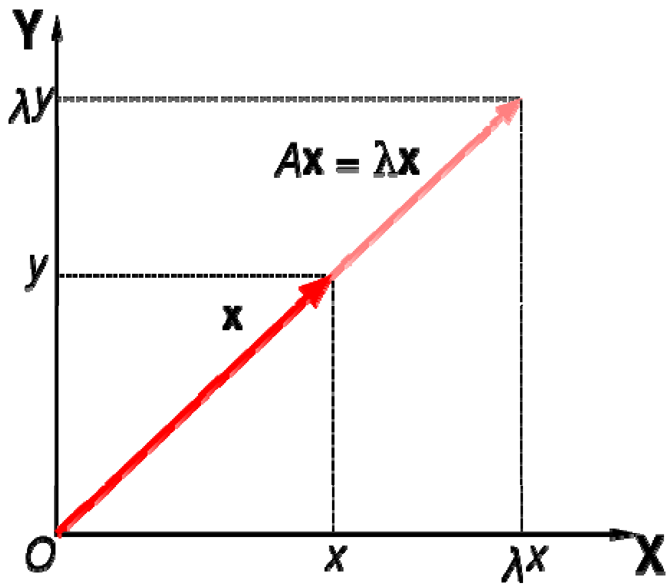
V1 =(5, 4) & V2=(-4, 5)

V1 . V2 = 0



Eigen Values & Eigen Vectors

- An eigenvector of a matrix *A* is a nonzero vector *x*, where *A.x* is similar to applying a linear transformation λ to *x* which, may change in length, but not direction
- *A* acts to stretch the vector *x*, not change its direction, so *x* is an eigenvector of *A*



$$Ax - \lambda Ix = 0$$

$$(A - \lambda I)x = 0$$

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \lambda \begin{bmatrix} x \\ y \end{bmatrix}$$

$if\ \ there\ exist\ an\ inverse\ (A - \lambda I)^{-1},\ then\ x = 0$

$we\ need\ \det(A - \lambda I) = 0\ to\ avoid\ the\ trevial\ solution\ x = 0$

$$\det(A - \lambda I) = 0$$

Example on Eigen Values & Eigen Vectors

- Suppose $\underline{A}$ is 2x2 matrix

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

$$\det\begin{bmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{bmatrix} = (2-\lambda)^2 - 1 = 0$$

$$\lambda = 1 \quad or \quad \lambda = 3$$

$$for\ \lambda = 3, \quad \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = 3\begin{bmatrix} x \\ y \end{bmatrix}$$

$$for\ \lambda = 1, \quad \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = 1\begin{bmatrix} x \\ y \end{bmatrix}$$

$$\begin{bmatrix} 2x+y \\ x+2y \end{bmatrix} = \begin{bmatrix} 3x \\ 3y \end{bmatrix}$$

$$2x + y = 3x$$

$$\boxed{x = y}$$

$$\begin{bmatrix} 2x+y \\ x+2y \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix}$$

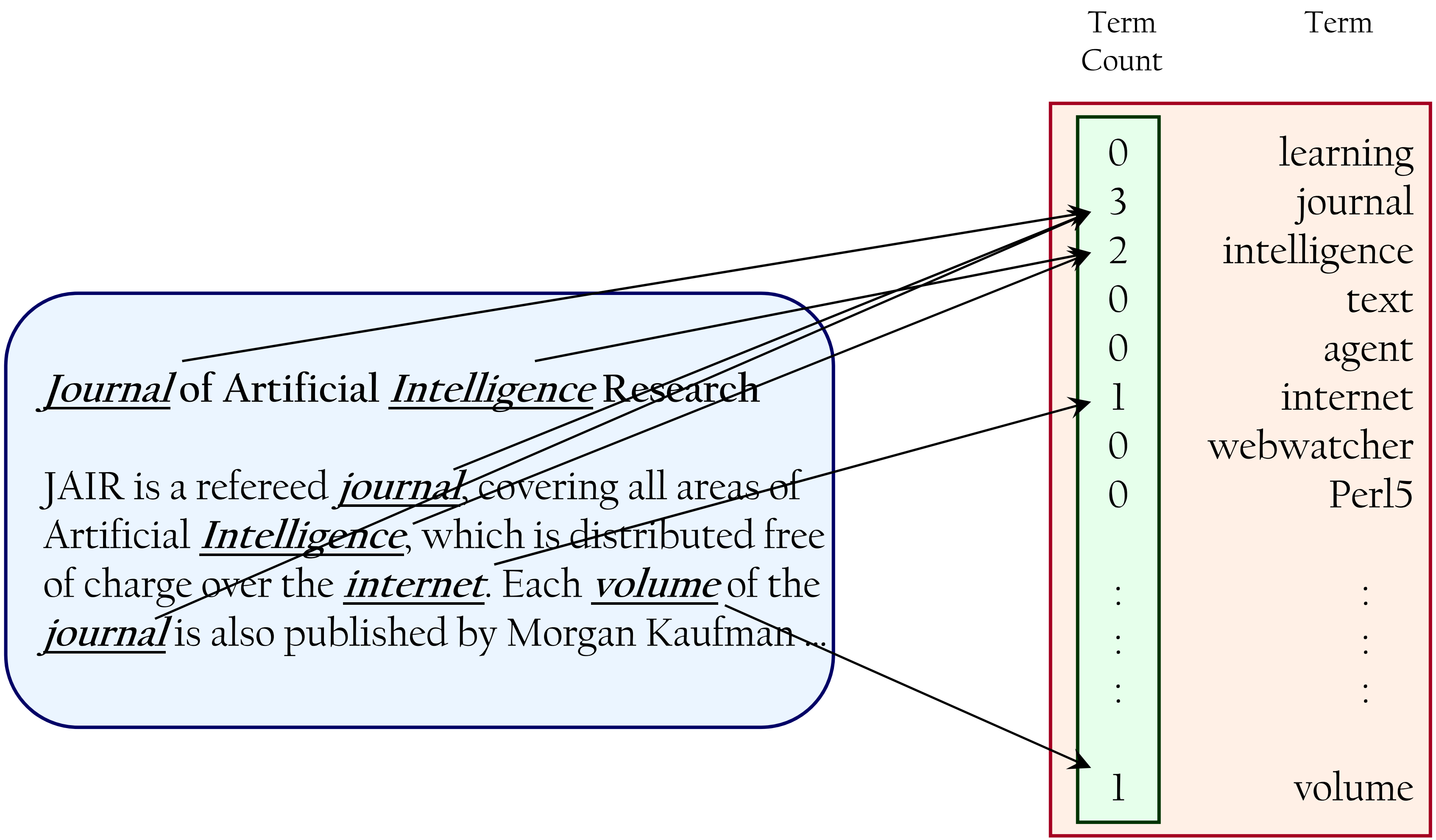
$$2x + y = x$$

$$\boxed{x = -y}$$

The eigenvectors are:

$$\boxed{\begin{bmatrix} 1 \\ 1 \end{bmatrix}} \qquad \boxed{\begin{bmatrix} 1 \\ -1 \end{bmatrix}}$$

Representing Documents as Vectors

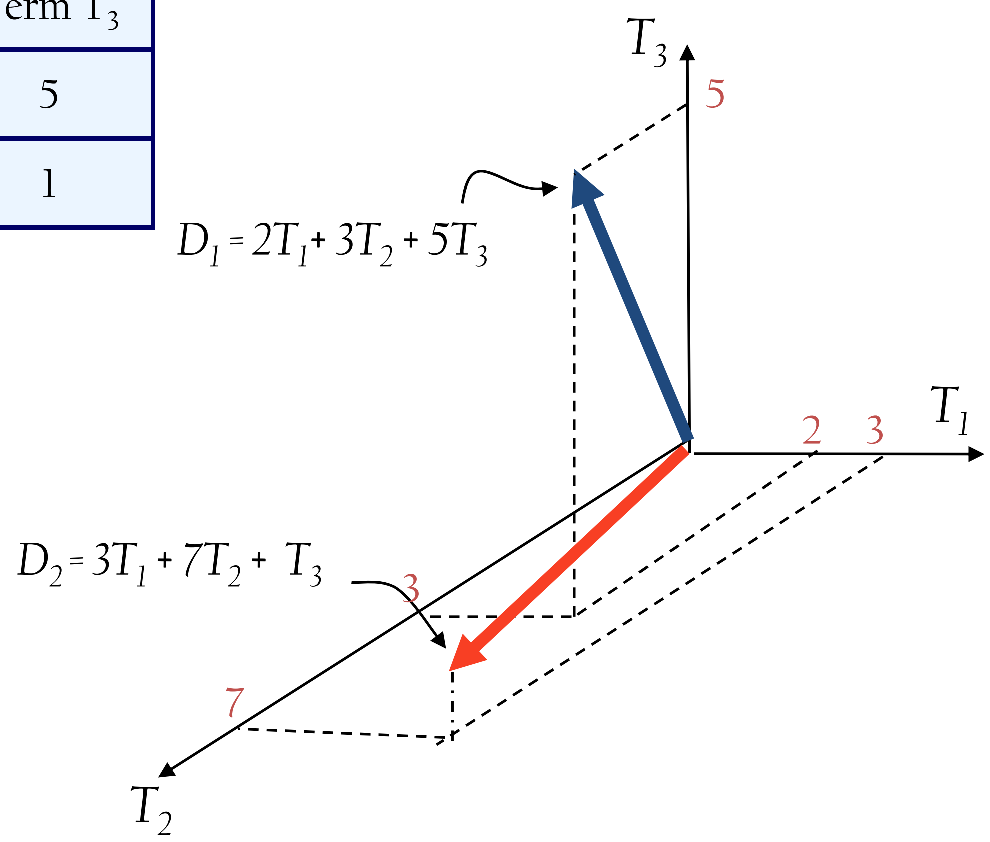


Documents as Vectors

Suppose we have two documents containing three nouns only

	Term T_1	Term T_2	Term T_3
Document D_1	2	3	5
Document D_2	3	7	1

$$D_1 = \begin{bmatrix} 2 \\ 3 \\ 5 \end{bmatrix} \qquad D_2 = \begin{bmatrix} 3 \\ 7 \\ 1 \end{bmatrix}$$



Dimensionality Reduction

Term Count	Term
34	Home
32	Garden
15	Room
14	Window
11	Furniture
11	Restroom
6	Floor
5	Kitchen
5	Balcony
1	Chimney
1	Street
1	City
1	Dog
1	Lake

Dimensionality Reduction

- Term Count
- tfidf
- Chi-Square
- Information Gain
- Gain Ratio

Term Count	Term
15	Room
14	Window
11	Furniture
11	Restroom
6	Floor
5	Kitchen
5	Balcony

PROBABILITY

Part 2

-Introduction
-Terminology

What Is Probability?

- A priori probability ***P(e)***: The chance that e happens
- Conditional probability ***P(f / e)***: The chance of f given e
- Joint probability ***P(e, f)***: The chance of e and f both happening; If e and f are independent, then P(e, f) = P(e) * P(f); If e and f are dependent then P(e, f) = P(e) * P(f | e)

  For example, if e stands for "the first roll of the die comes up 5" and f stands for "the second roll of the die comes up 3," then P(e,f) = P(e) * P(f) = 1/6 * 1/6 = 1/36.

$$\sum_e P(e) = 1 \qquad \sum_e P(e \mid f) = 1$$

BASIC Probabilities

$$P(A \cup B) = \begin{cases} P(A) + P(B) & A \,\&\, B \text{ are not dependant} \\ P(A) + P(B) - P(A, B) & A \,\&\, B \text{ are dependant} \end{cases}$$

- For example, when drawing a single card at random from a regular deck of cards, the chance of getting a heart or a face card (J,Q,K) (or one that is both) is

$$\frac{13}{52} + \frac{12}{52} - \frac{3}{52} = \frac{22}{52}$$

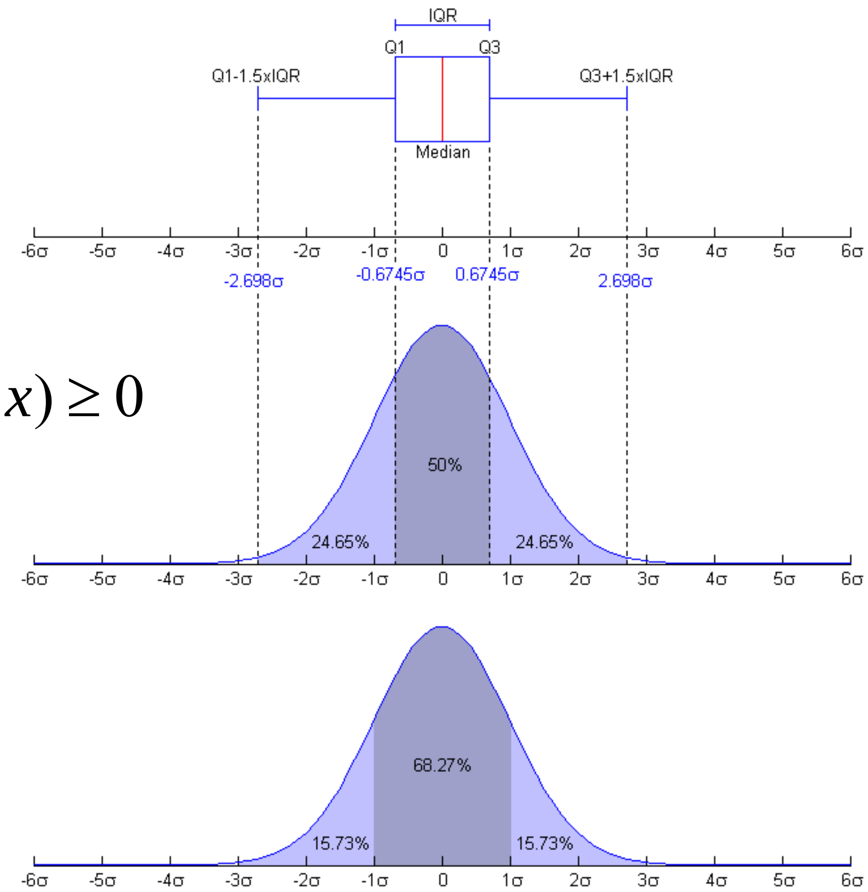
A	$P(A) \in [0, 1]$
not A	$P(A') = 1 - P(A)$
A or B	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ $= P(A) + P(B) \quad$ if A and B are mutually exclusive
A and B	$P(A \cap B) = P(A\|B)P(B)$ $= P(A)P(B) \quad$ if A and B are independent
A given B	$P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$

Probability Density Function PDF

- Probability density function (pdf) is a function that represents a probability distribution in terms of integrals

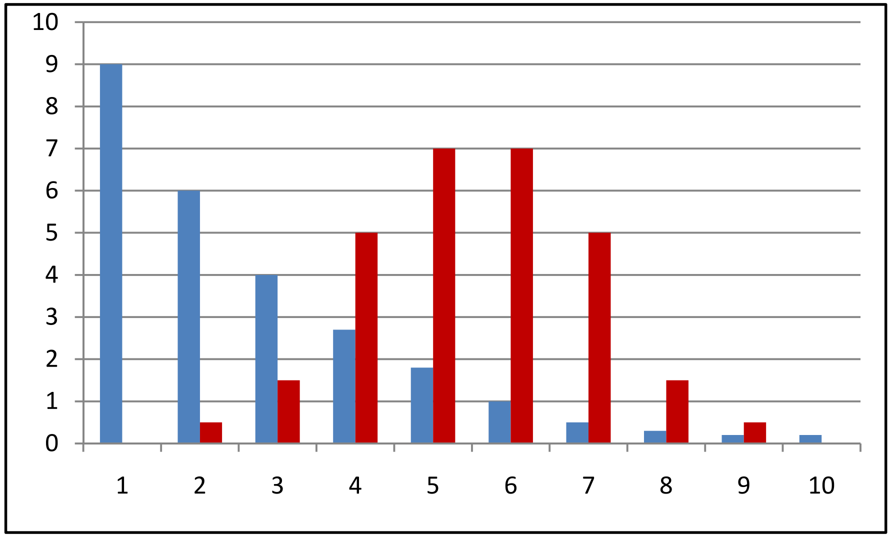
$$\int_{a}^{b} f(x)\,dx$$

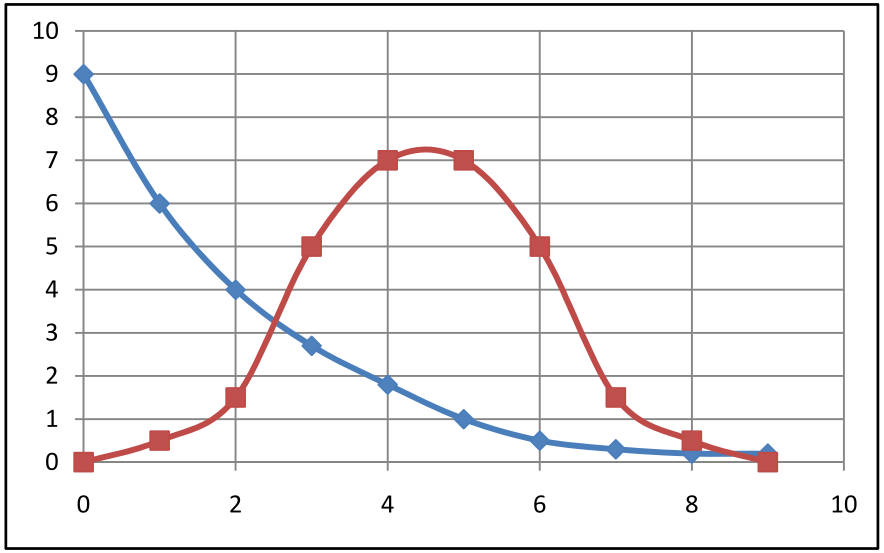
$$\int_{-\infty}^{\infty} f(x)dx = 1 \qquad \& \qquad f(x) \geq 0$$



Probability Density Function PDF

- The Summation is used with Discrete Data





Conditional & Bayesian Probability

- **Conditional probability** is the probability of some event A, given the occurrence of some other event B
- Conditional probability is written $P(A|B)$, and is read "the probability of A, given B"

$$P(A \mid B) = \frac{P(A, B)}{P(B)}$$

- Bayesian probability, the probability of a hypothesis given the data (the *posterior*), is proportional to the product of the likelihood times the prior probability (often just called the *prior*)
- The likelihood brings in the effect of the data, while the prior specifies the belief in the hypothesis before the data was observed

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}$$

STATISTICS

Part 3

Introduction

Statistics

- Statistics is a Mathematical Science pertaining to the ***collection***, ***analysis***, ***interpretation** **or*** ***explanation**, **and*** ***presentation*** of data

Statistical Terminologies

- Measures of Central Tendency ***(Mean***, Median, Mode)

$$\bar{x} = (1/n)\sum_{i=1}^{n} x_i$$

- ***Population Variance*** measures statistical dispersion of data points from the expected value (mean)

$$Var(X) = E\left[(X - E(X))^2\right]$$
$$= (1/n)\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sigma^2$$

- ***Standard Deviation*** is a measure of the variability or dispersion of a population; Low SD indicates very close data points to the mean; High SD indicates spread out data points

$$sd(X) = \sqrt{\sigma^2}$$

- ***Covariance*** measures how much two variables change together

$$Cov(X,Y) = E\left[(X - E(X))(Y - E(Y))\right]$$

- ***Correlation*** (coefficient) indicates the strength and direction of a *linear* relationship between two random variables

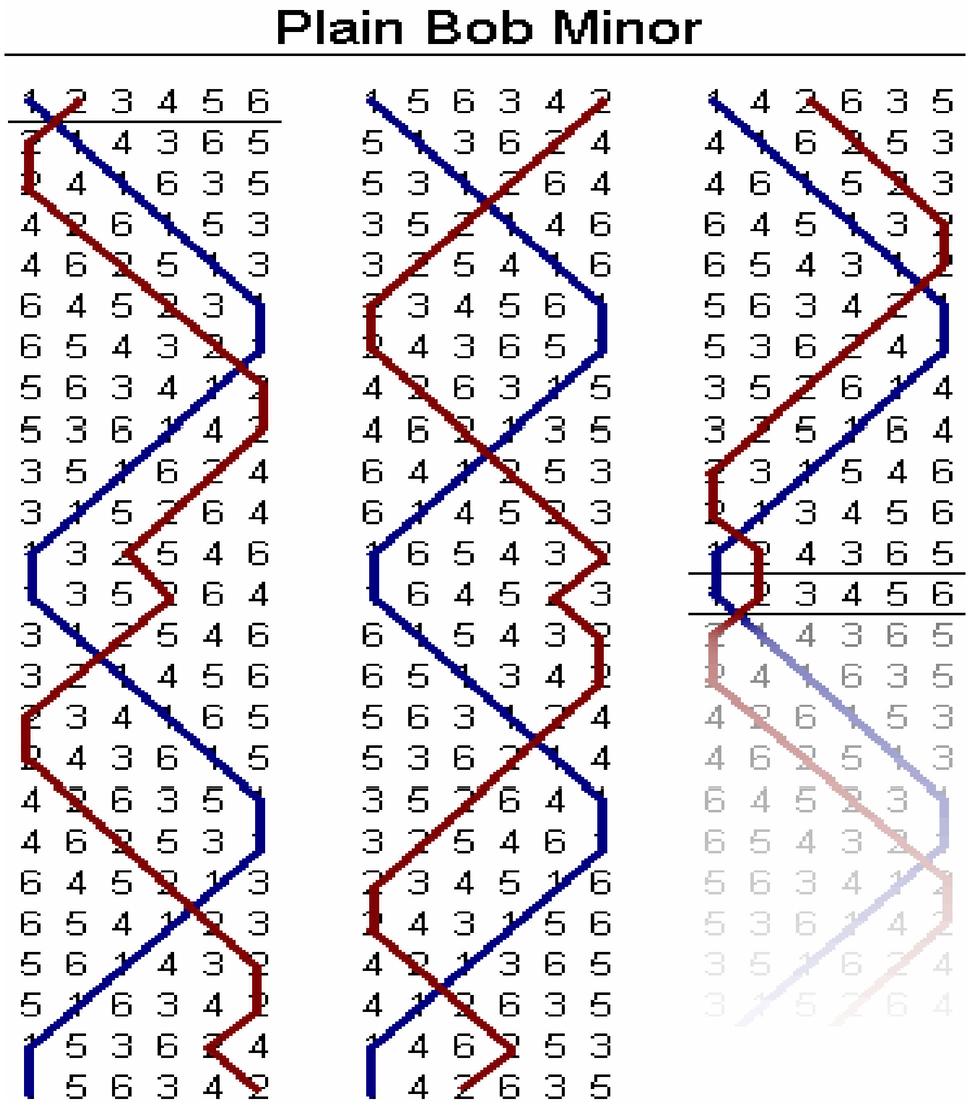
$$Corr(X,Y) = \frac{Cov(X,Y)}{sd(X) * sd(Y)} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

STATISTICS

Part 4

Permutations & Computations

Introduction
to Permutations & Computations



Permutations

- Suppose an ordered set of n different objects
- For ***ordered*** selection of r objects from a set of n (n≥r) different objects, the number of permutations of r from n, *i.e.* the number of different possible ordered selections, is usually denoted by P_r^n.

$$P_r^n = \frac{n!}{(n-r)!}$$

لدينا ثلاثة أرقام ا، ب، ج. يتم إختيار أول رقم وضربه فى 10، ويتم ضرب الرقم الثانى فى 100، ويتم ضرف الرقم الثالث فى 1000، ثم يتم جمع الثلاثة أرقام الجديدة. كم رقم يمكن إستنتاجه من هذه الأرقام الثلاثة.

مثال: 1، 2، 3   (3210، 3120، 2130، ...)

الحل: ؟

$P_0^n = 1$	$P_1^n = n$	$P_n^n = n!$

Permutations

Example:

r	g	b	y

Suppose we have 4 elements and need to select 3 elements in order; there are 24 different combinations

$$P_3^4 = \frac{4!}{(4-3)!} = \frac{4!}{1!} = 4*3*2 = 24$$

r g b	r b g	g r b	g b r
b g r	b r g	r g y	r y g
g r y	g y r	y r g	y g r
r b y	r y b	b r y	b y r
y r b	y b r	g b y	g y b
b g y	b y g	y g b	y b g

Permutations

- Suppose a set {A, B, C}, we have 6 (=3!) permutations of {A, B, C} are *ABC, ACB, BAC, BCA, CAB and CBA*
- Suppose a set {A, B, C, D}, there are 24 = $P^4{}_3$ = (4 × 3 × 2) permutations of 3 letters from {A, B, C, D}
- If the n objects are not all different, and there are n_r objects of type 1, n_2 objects of type 2, ..., n_k objects of type k, where $n_1+n_2+\ldots+n_k=n$, then the number of different ordered arrangements is

$$\frac{n!}{n_1!n_2!n_3!\ldots n_k!}$$

a	a	a	b	b	b	c	c	c	c	d	d	d	d

$$\frac{14!}{3!*3!*4!*4!}$$

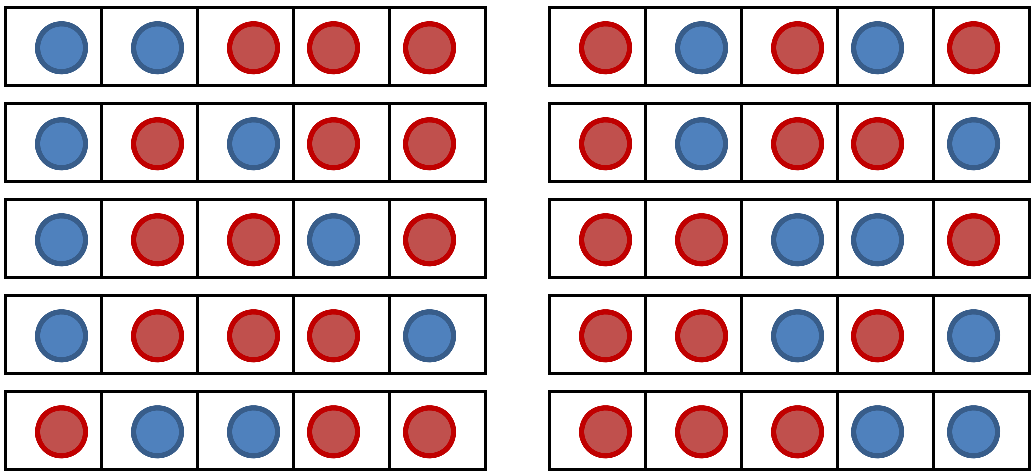
Computations

The number of ways of picking k ***unordered*** outcomes from n possibilities. Also known as the binomial coefficient or choice number and read “n choose k,“

$$C_k^n = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

لدينا ثلاثة كرات حمراء و كرتان زرقاء. كم طريقة يمكن بها ترتيب الخمس كرات.

مثال: (ح،ح،ح،ز،ز)، (ح،ح،ز،ح،ز)
الحل:



Computations

For example: suppose we have the set {1, 2, 3, 4}, we need to calculate the number of combinations of selecting two elements out of the set

$$C_2^4 = \binom{4}{2} = \frac{4!}{2!*2!} = 6$$

namely {1,2}, {1,3}, {1,4}, {2,3}, {2,4}, and {3,4}.

Suppose we have 4 places and filled only 2 of them. The combination to fill the other two cells with the other two numbers equal to 1. Muir (1960) uses the nonstandard notations

$$\overline{C}_k^n = \binom{n-k}{k} \qquad \overline{C}_2^4 = \binom{2}{2} = \frac{2!}{2!*0!} = 1$$

$C_0^n = 1$	$C_1^n = n$	$C_n^n = 1$

STATISTICS

Part 5

Popular Distributions

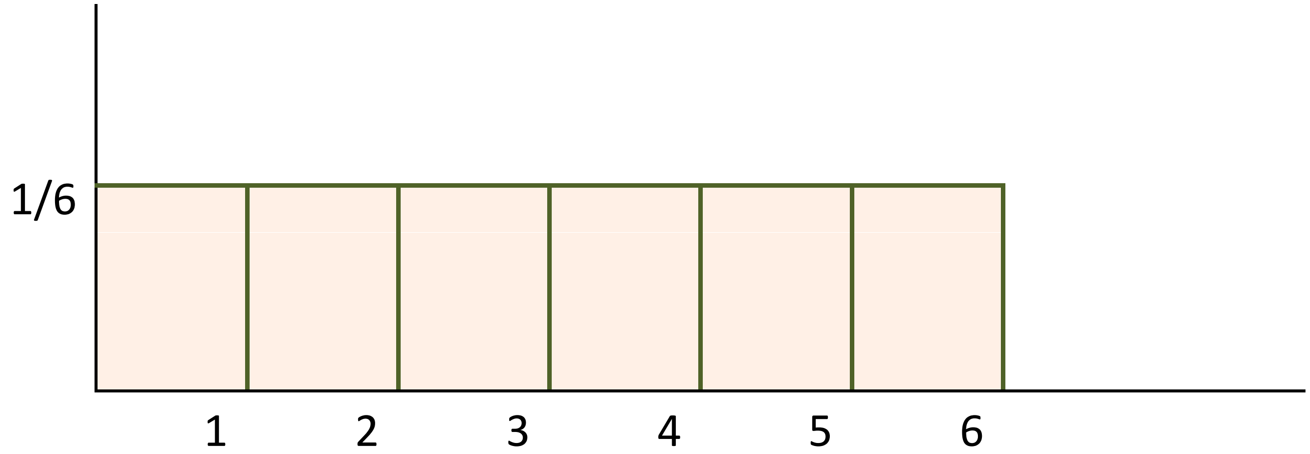
Popular Distributions

Probability Distribution identifies the probability of each value of an unidentified random variable

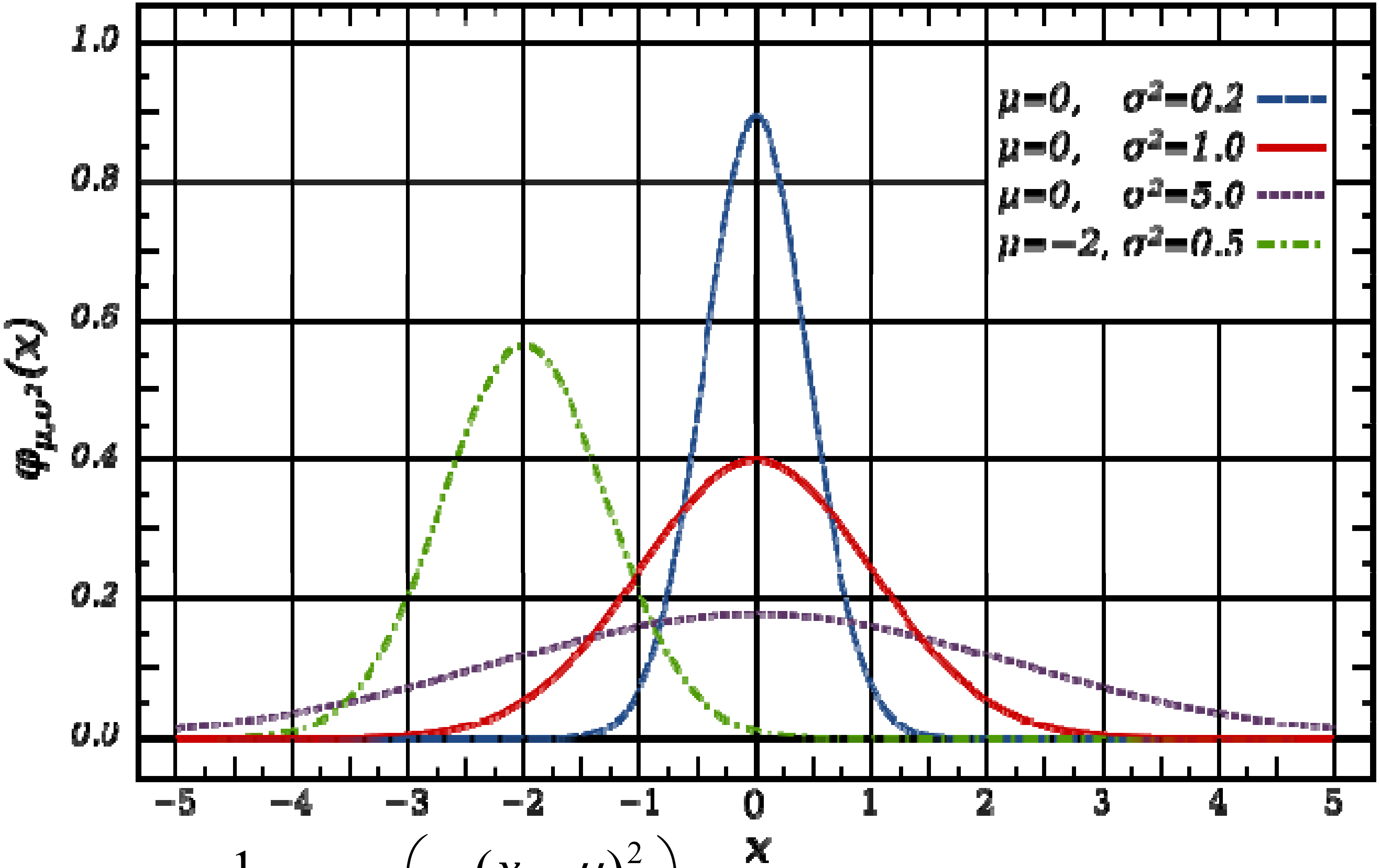
- *Uniform Distribution*
- *Normal (Gaussian) Distribution*
- *Chi-Square Distribution*
- *Exponential Distribution*
- *Poisson Distribution*
- *T Distribution*
- *F Distribution*

The Uniform Distribution

- The probability is equal for all outcomes
- Suppose a fair dice is thrown, the probability of getting any of its 6 faces equal to 1/6
- The area under the line equal to 1

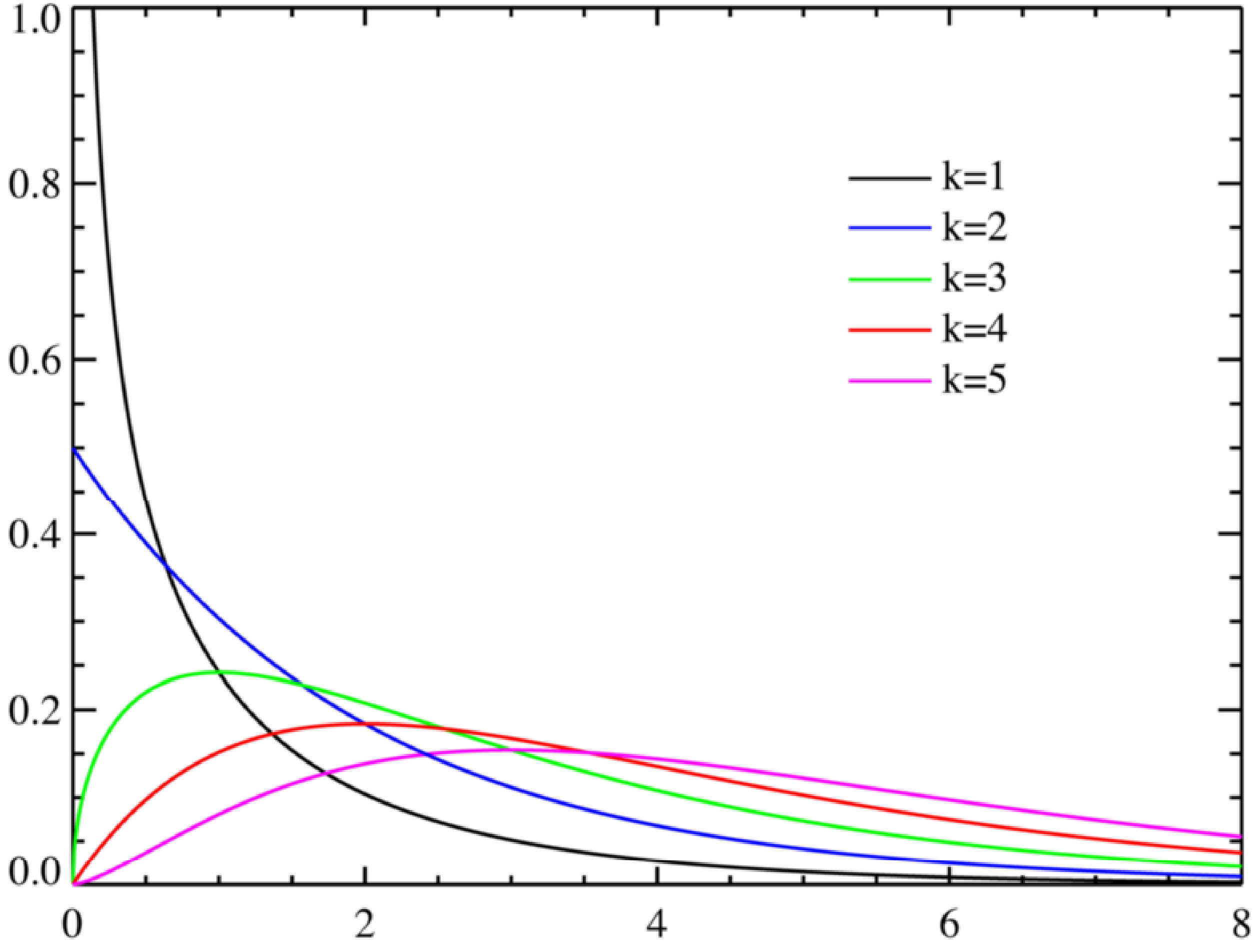


The Normal/Gaussian Distribution



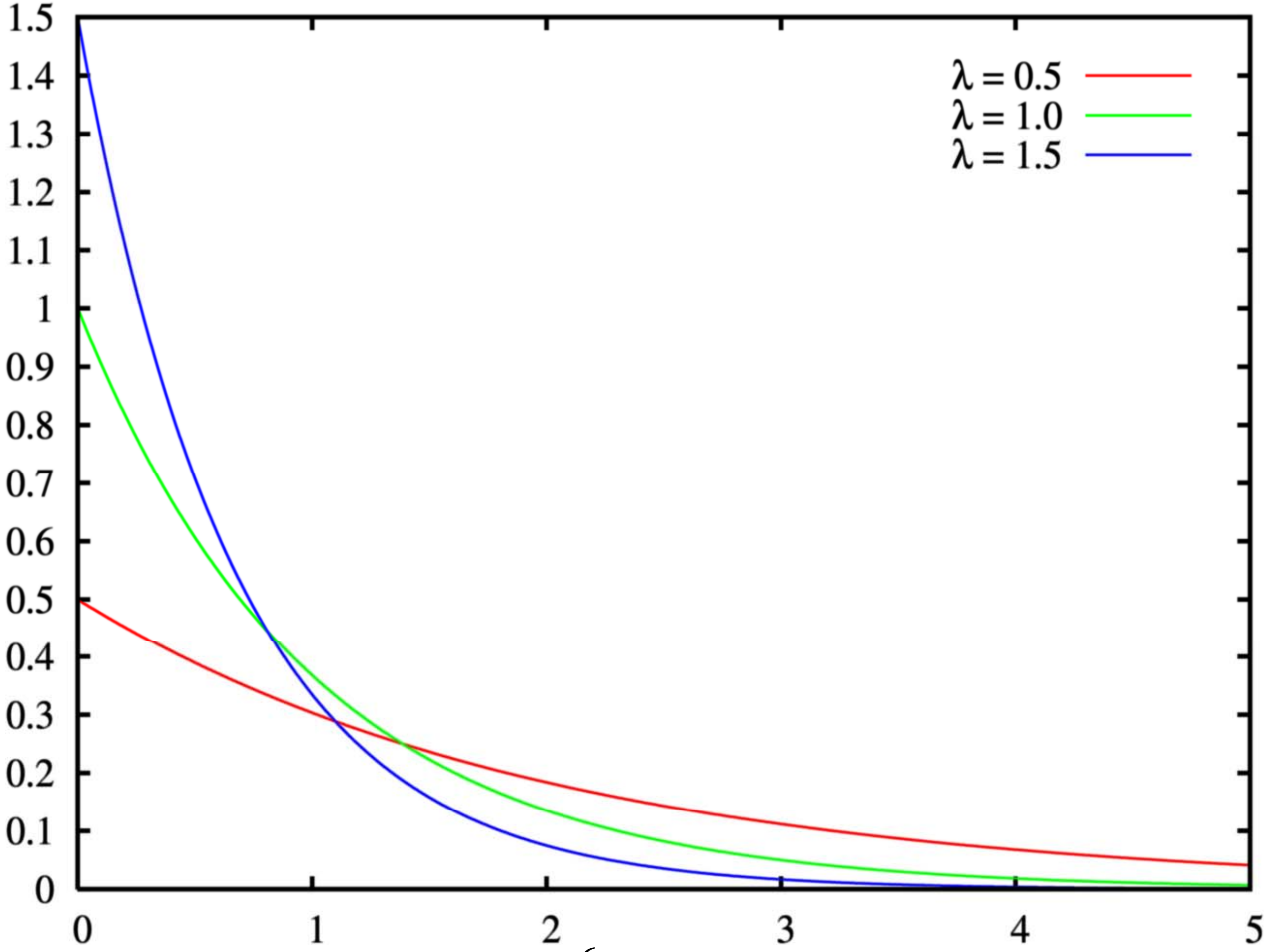
$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

The Chi-Square Distribution



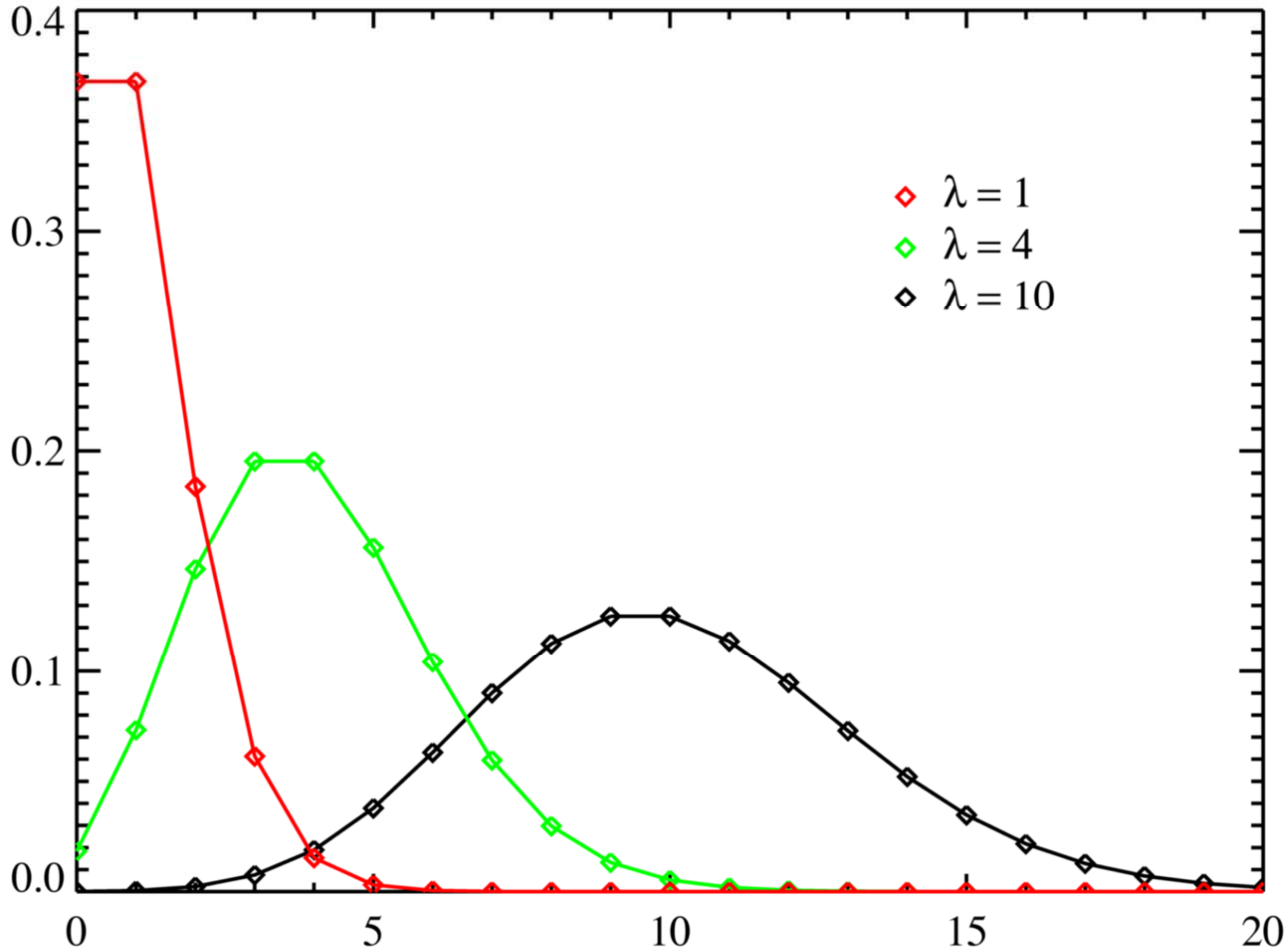
$$f(x;k) = \begin{cases} \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2} & for\ x > 0 \\ 0 & for\ x \leq 0 \end{cases}$$

The Exponential Distribution



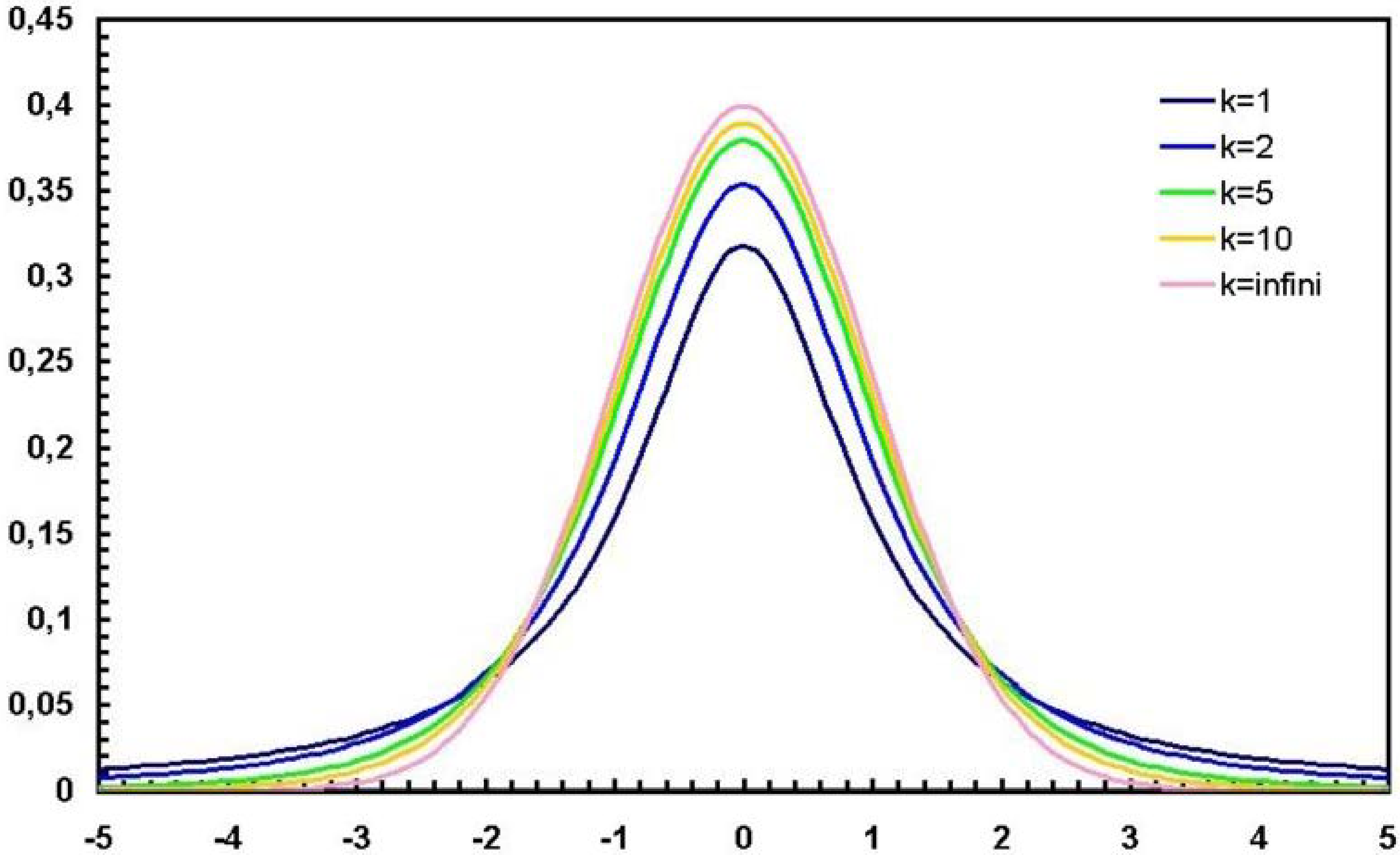
$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & for\ x > 0 \\ 0 & for\ x \le 0 \end{cases}$$

The Poisson Distribution



$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

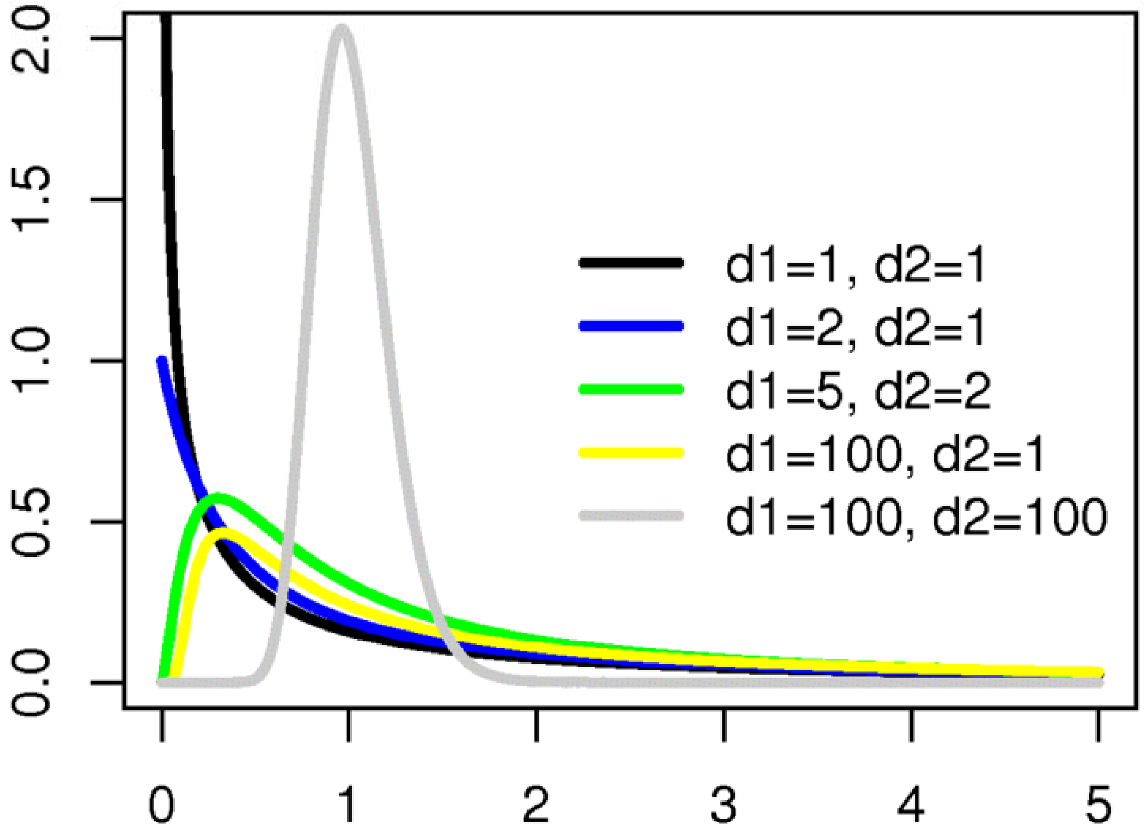
The T Distribution



$$f(t) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{v\pi}\,\Gamma(\frac{v}{2})}\left(1+\frac{t^2}{v}\right)^{-(v+1)/2}$$

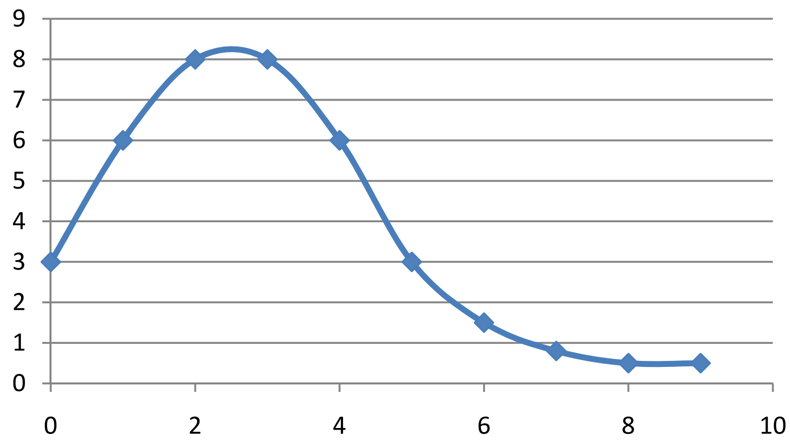
t-**distribution** arises in the problem of estimating the mean of a normally distributed population when the sample size is small

The F Distribution



$$f(x) = \frac{\sqrt{\frac{(d_1\,x)^{d_1}\,\,d_2^{d_2}}{(d_1\,x+d_2)^{d_1+d_2}}}}{x\,\mathrm{B}\!\left(\frac{d_1}{2},\frac{d_2}{2}\right)}$$

Fitting Chi-Square



Vector a

15
14
11
11
6
5
5

$$\max \quad \chi^2 = \sum_{i=1}^{n} \frac{(a_i - E_i)^2}{E_i}$$

$$E_{ij} = (15+14+11+11+6+5+5)/7 = 9.57$$

$$\chi^2 = (1/9.57) * ((15-9.57)^2 + (14-9.57)^2 + (11-9.57)^2 + (11-9.57)^2 + (6-9.57)^2 + (5-9.57)^2 + (5-9.57)^2) = 107.71/9.57 = 11.26$$

Measuring Term-Category Correlation

$$\chi^2(t_k, c_i) = \frac{[P(t_k, c_i)P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i)P(\bar{t}_k, c_i)]^2}{P(t_k)P(\bar{t}_k)P(c_i)P(\bar{c}_i)}$$

$P(t_k, c_i)$ ➔probability document x contains term t and belongs to category c.

$P(\bar{t}_k, c_i)$ ➔probability document x does not contain term t and belongs to category c.

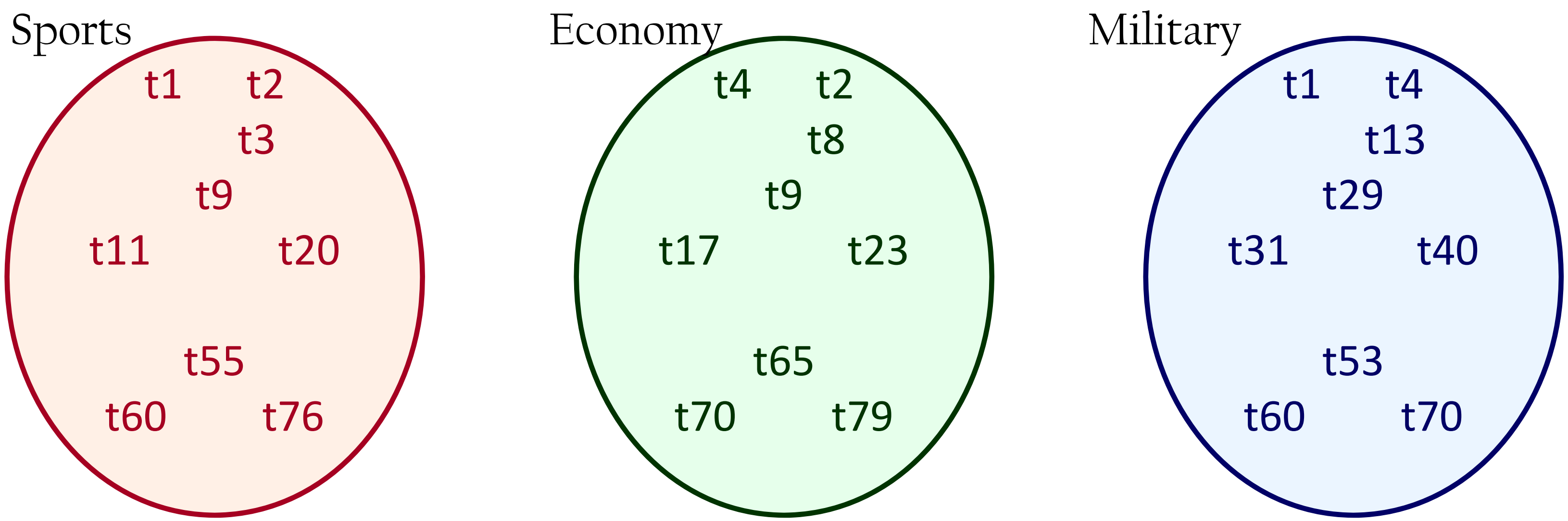
$P(t_k, \bar{c}_i)$ ➔probability document x contains term t and does not belong to category c.

$P(\bar{t}_k, \bar{c}_i)$ ➔probability document x does not contain term t and does not belong to category c.

$P(t)$ ➔probability of term t

$P(c)$ ➔probability of category c

Testing The Membership



$$\chi^2(t_k, c_i) = \frac{[P(t_k, c_i)P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i)P(\bar{t}_k, c_i)]^2}{P(t_k)P(\bar{t}_k)P(c_i)P(\bar{c}_i)}$$

$$\chi^2(t_1, Sports\ ) = \frac{\left[\frac{1}{9} * \frac{14}{16} - \frac{1}{16} * \frac{8}{9}\right]^2}{\frac{2}{27} * \frac{25}{27} * \frac{9}{27} * \frac{18}{27}}$$

Using Chi-Square for Categorization

Another Example:

Term	**Frequency per Category**				**Total**
	Communication	Phone	Business	Army	
Link	15	6	2	12	35
Wire	10	12	0	8	30
Total	25	18	2	20	**65**

$$\chi^2(link\,,\,phone) = \frac{[6/65)*(18/65)-(29/65)*(12/65)]^2}{(35/65)*(30/65)*(18/65)*(47/65)}$$

Using Chi-Square for Multiple sets of Terms

Group 1	Category		Total
	0	1	
Term 1	3	2	5
Term 2	0	4	4
Term 3	2	3	5
Total	5	9	14

Group 2	Category		Total
	0	1	
Term 5	1	3	4
Term 7	4	6	10
Total	5	9	14

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(a_{ij}-E_{ij})^2}{E_{ij}}$$

$$E_{ij} = (T_{ci} * T_{vj}) \big/ T$$

$$\chi^2(Group\,1) = (3-1.78)^2/1.78 + (2-3.21)^2/3.21 + (0-1.42)^2/1.42 + (4-2.57)^2/2.57 + (2-1.78)^2/1.78 + (3-3.21)^2/3.21 = 3.62$$

$$\chi^2(Group\ 2) = (1-1.42)^2/1.42 + (3-2.57)^2/2.57 + (4-3.57)^2/3.57 + (6-6.43)^2/6.43 =$$

Mingers, J., (1989a). "An Empirical Comparison of selection Measures for Decision-Tree Induction", *Machine Learning*, Vol. 3, No. 3, (pp. 319-342), Kluwer Academic Publishers.

Attribute Selection Criteria: Chi-Square

Example

- T2 is quantized into two intervals 21 (T2<=21) and (T2>21)
- T3 is quantized into two intervals 15 (T3<=15) and (T3>15)

T1	T2	T3	T4	D
1	25	10	A	1
1	30	30	A	0
1	35	25	B	0
1	22	35	B	0
1	19	10	B	1
2	22	30	A	1
2	33	18	B	1
2	14	5	A	1
2	31	15	B	1
3	21	20	A	0
3	15	10	A	0
3	25	20	B	1
3	18	20	B	1
3	20	36	B	1

T2	Decision D		Total
	0	1	
<=21	1	3	4
>21	4	6	10
Total	5	9	14

T1	Decision D		Total
	0	1	
1	3	2	5
2	0	4	4
3	2	3	5
Total	5	9	14

T3	Decision D		Total
	0	1	
<=15	1	4	5
>15	4	5	9
Total	5	9	14

T4	Decision D		Total
	0	1	
A	3	3	6
B	2	6	8
Total	5	9	14

Attribute Selection Criteria: Chi-Square

$$\chi^2(A) = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(a_{ij}-E_{ij})^2}{E_{ij}}$$

where A is the attribute to be evaluated against the decision attribute, n is the number of distinct values of A, m is the number of distinct values of the decision attribute, a_{ij} is the correlation frequency of value number i from A and value number j from the decision attribute;

$$E_{ij} = {(T_{ci} * T_{vj})}/{T}$$

where T_{ci} is the total number of examples belonging to class ci, T_{vj} is the number of examples containing the value vj of the given attribute

$$\chi^2(X1) = (3-1.78)^2/1.78 + (2-3.21)^2/3.21 + (0-1.42)^2/1.42 + (4-2.57)^2/2.57 + (2-1.78)^2/1.78 + (3-3.21)^2/3.21 = 3.62$$

$$\chi^2(X4) = (3-3.9)^2/3.9 + (3-2.1)^2/2.1 + (6-5.1)^2/5.1 + (2-2.9)^2/2.9 = 1.1$$

D1	Decision D5		Total
	0	1	
1	3	2	5
2	0	4	4
3	2	3	5
Total	5	9	14

D2	Decision D5		Total
	0	1	
<=21	1	3	4
>21	4	6	10
Total	5	9	14

D3	Decision D5		Total
	0	1	
<=15	1	4	5
>15	4	5	9
Total	5	9	14

D4	Decision D5		Total
	0	1	
A	3	3	6
B	2	6	8
Total	5	9	14

Mingers, J., (1989a). "An Empirical Comparison of selection Measures for Decision-Tree Induction", *Machine Learning*, Vol. 3, No. 3, (pp. 319-342), Kluwer Academic Publishers.

STATISTICS

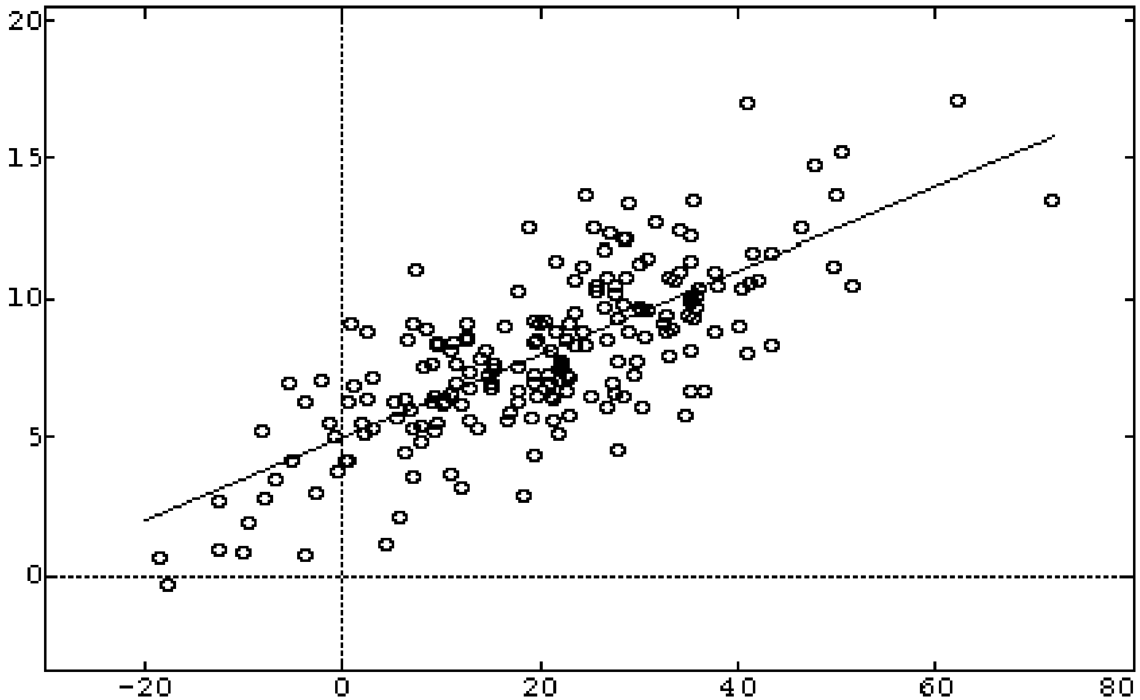
Part 6

Regression

Linear Regression

- The linear model states that the dependent variable is ***directly proportional*** to the value of the independent variable
- Thus if a theory implies that Y increases in direct proportion to an increase in X, it implies a specific mathematical model of behavior

$$y = ax + b$$



In case of two dimensions

$$a = slope = \frac{(y_2 - y_1)}{(x_2 - x_1)}$$

$$b = y_2 - slope * x_2$$

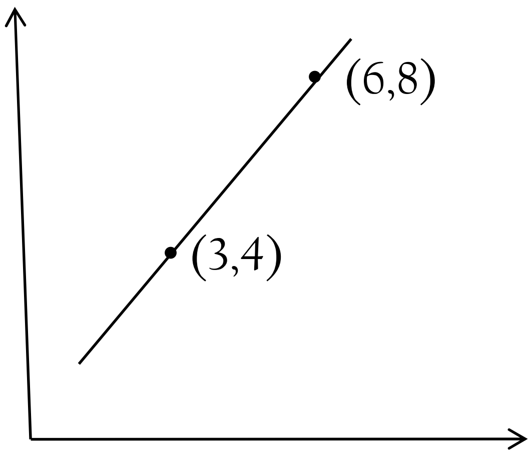
Linear Regression

$$y = ax + b$$

$$8 = 6a + b \quad \& \quad 4 = 3a + b$$

$$\frac{8-b}{6} = a \qquad \& \qquad 4 = 3 * \frac{8-b}{6} + b$$

$$b = 0 \qquad \& \qquad a = \frac{4}{3} = 1.333$$



$$Slope = \frac{8-4}{6-3} = 1.333$$

$$b = 4 - \frac{4}{3} * 3 = 0$$

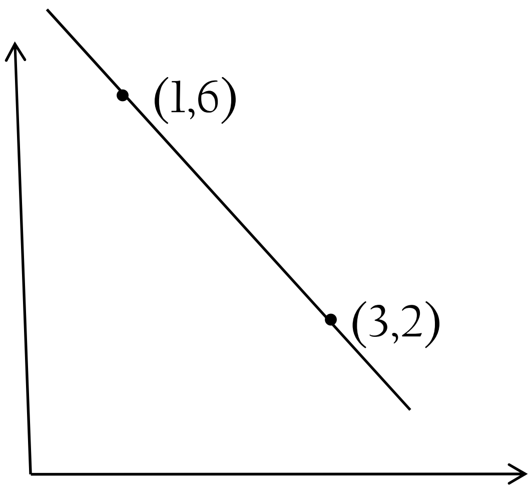
Linear Regression

$$y = ax + b$$

$$6 = a + b \quad \& \quad 2 = 3a + b$$

$$6 - b = a \qquad \& \qquad 2 = 3*(6-b) + b$$

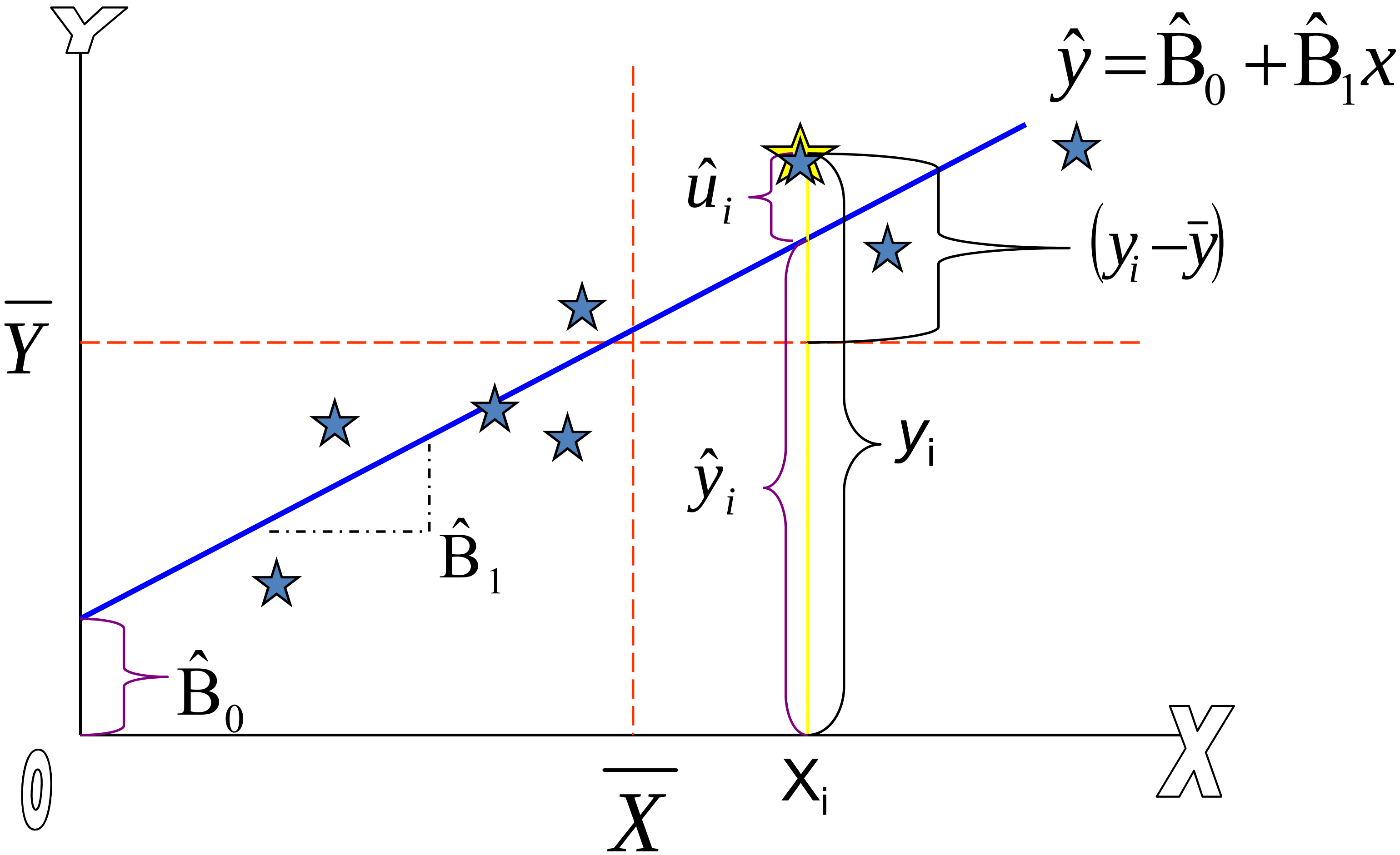
$$b = 8 \qquad \& \qquad a = 6 - 8 = -2$$



$$Slope = \frac{6-2}{1-3} = \frac{4}{-2} = -2$$

$$b = 2 + 2*3 = 8$$

Linear Regression

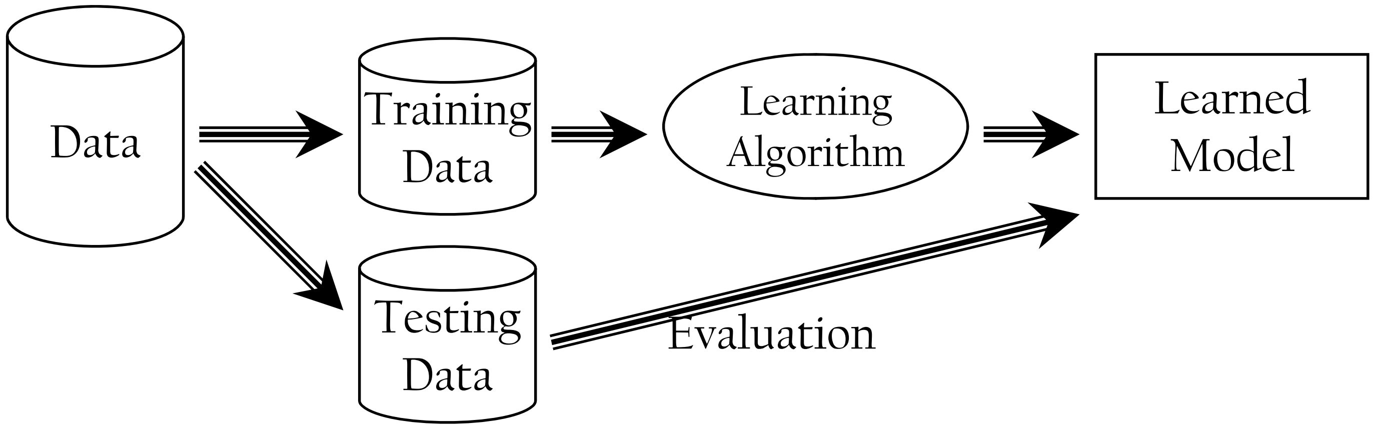


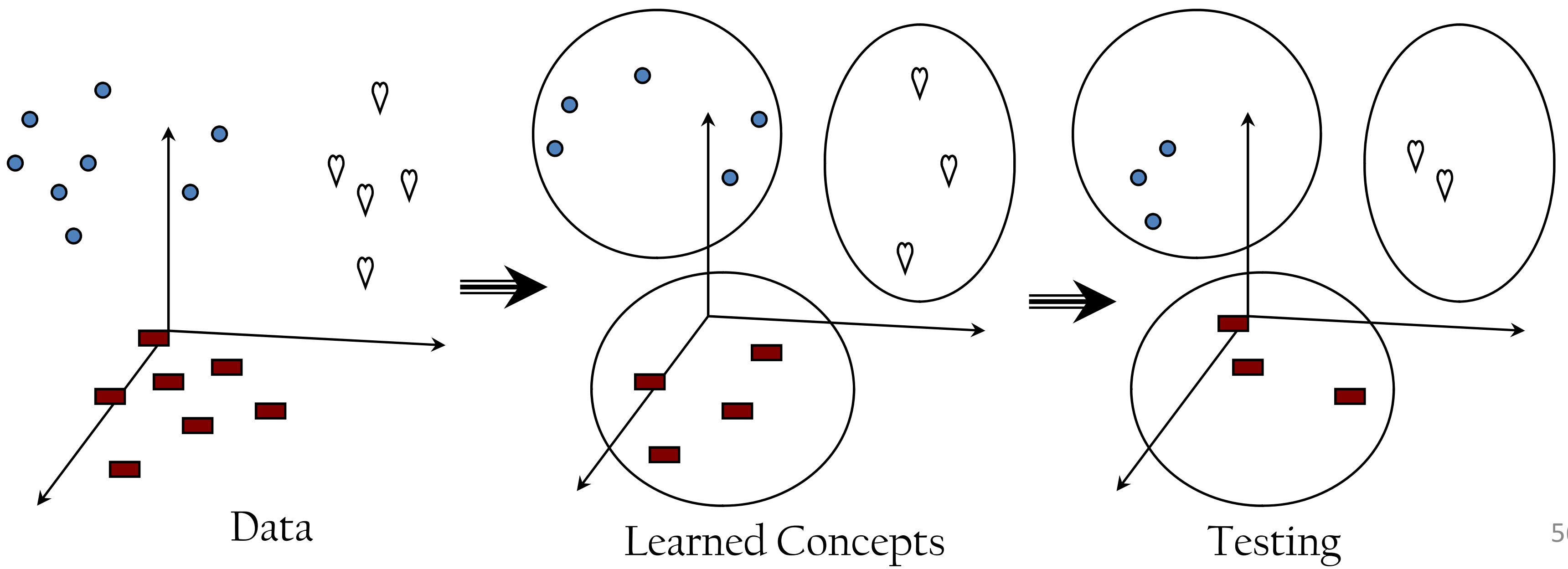
Statistics and Testing

Part 7

Testing Samples & Calculating Accuracy

Training & Testing



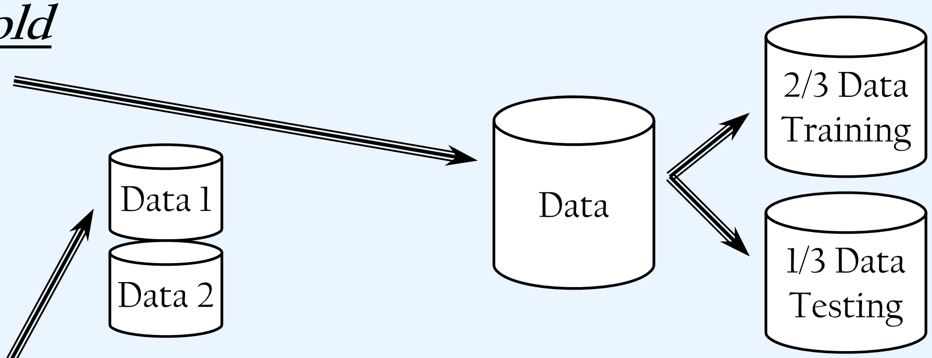


Testing Approaches

- ***Two-Cross-Fold***

  Train on 2/3rd

  Test on 1/3rd

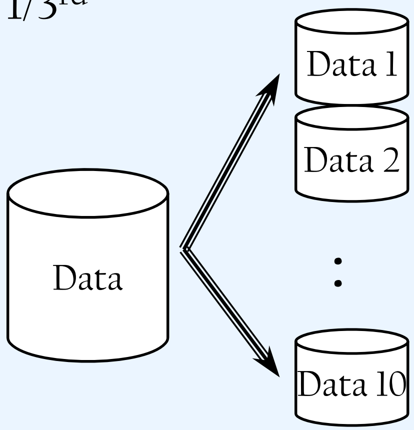


- ***Ten-Cross-Fold***

  Train on 9/10th

  Test on 1/10th

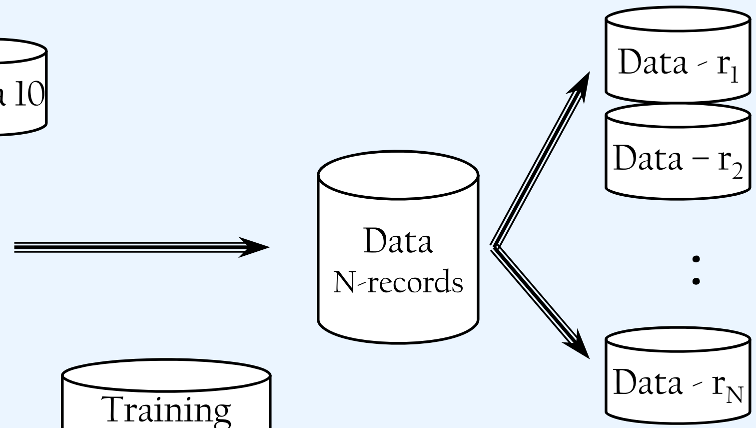
  Repeat 10 times



- ***Hold-One-Out***

  Train on all data but one

  Test on the selected one

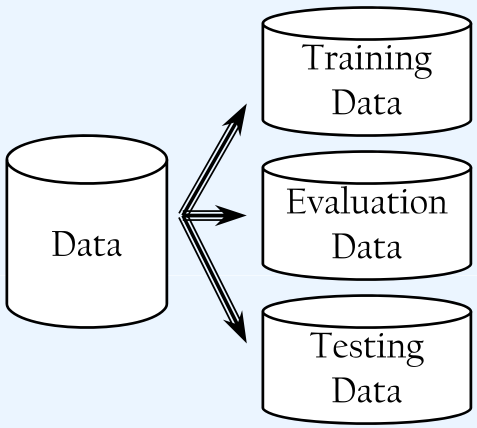


- ***Learning Evaluation vs. Testing***

  Train on Training Data

  Evaluate on Evaluation Data

  Test on Testing Data



Accuracy & Error

Example: Suppose you have a classification model C, and 100 testing records from two classes (P & N). Suppose the following are the classification results:

		Actual	
		P	N
Obtained	P	TP	FP
	N	FN	TN

		Actual	
		P	N
Obtained	P	40	10
	N	5	45

- Accuracy vs. Error Rate
  - ***Accuracy*** = (40+45)/100 = 85%
  - ***Error Rate*** = (10+5)/100 = 15%
- True vs. False Classification
  - ***True Positive:*** = 88.88%
  - ***True Negative:*** = 81.82%
  - ***False Positive:*** = 11.12%
  - ***False Negative:*** = 18.18%
- Flexible Matching
  - *Using Nearest Neighbors (e.g., majority of nearest 3 neighbors)*
  - Using Fuzzy rules (assigning probability for each decision and taking it into consideration when calculating the accuracy)
  - Assigning small weights for the false positive and false negative results (not zero)
- Testing for Multiple Classes ????

Precision, Recall, and F-Measure

Accuracy: is the percentage of correct results

Error: is the percentage of wrong results

Accuracy only reacts to real errors, and doesn't show how many correct results have been found as such

Precision:

Precision shows the percentage of correct results within an answer:

$$Precision = (tp) / (tp + fp)$$

Recall:

Recall is the percentage of the correct system results over all correct results:

$$Recall = (tp) / (tp + fn)$$

Makhoul, John; Francis Kubala; Richard Schwartz; Ralph Weischedel: Performance measures for information extraction. In: Proceedings of DARPA Broadcast News Workshop, Herndon, VA, February 1999

Precision, Recall, and F-Measure

Precision and Recall can be defined differently for different tasks

For example: In Information Retrieval,

- Recall = |{relevant documents} ∩ {documents retrieved}| /

  / |{relevant documents}|

- Precision = |{relevant documents} ∩ {documents retrieved}| /

  / |{documents retrieved}|

Christopher D. Manning and Hinrich Sch¨utze, Foundations of Statistical Natural Language Processing, MIT Press, 1999.

Precision, Recall, and F-Measure

F-Measure (harmonic mean):

F_β "measures the effectiveness of β times as much importance to recall as precision". The general form of F-Measure:

$$F_\beta = (1+\beta^2) * (\text{precision} * \text{recall}) / (\beta^2 * \text{precision} + \text{recall})$$

when β=1,

$$F_1 = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$$

STATISTICS

Part 8

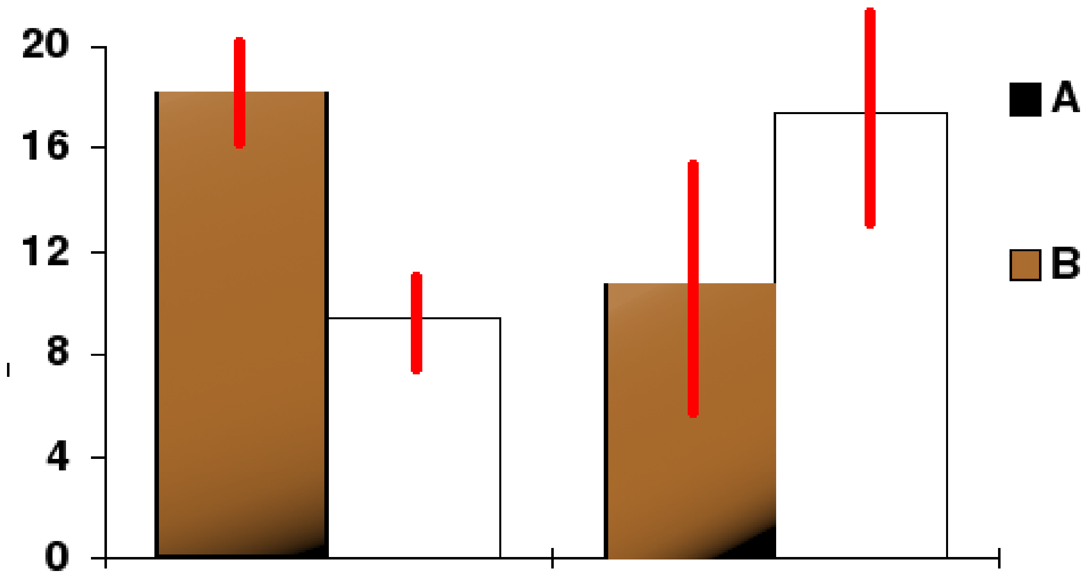
Test of Significance

Test of Significance (1/5)

- The probability that a result is not due to chance; or Is the observed value differs enough from a hypothesized value?
- The hypothesized value is called the null hypothesis
- If this probability is sufficiently low, then the difference between the parameter and the statistic is said to be "statistically significant"
- Just how low is sufficiently low? The choice of 0.05 and 0.01 are most commonly used

- Suppose your algorithm produced error rate of 1.5 and another algorithm produced an error of 2.1 on the same data set; are the two algorithms similar?

Test of Significance (2/5)



- The top ends of the bars indicate observation means
- The red line segments represent the confidence intervals surrounding them
- The difference between the two populations on the left is significant
- However, it is a common misconception to suppose that two parameters whose 95% confidence intervals fail to overlap are significantly different at the 5% level

Test of Significance (3/5)

- The system you are comparing against reported results of 250; the value reported is considered as a random variable X; the distribution of X is assumed as normal distribution with unknown mean and standard deviation σ=2.5; You ran your system 25 times; it reported values (x1, x2, … , x25); the average of these values is 250.2.

$$\hat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{25} x_i = 250.2$$

Sample Mean

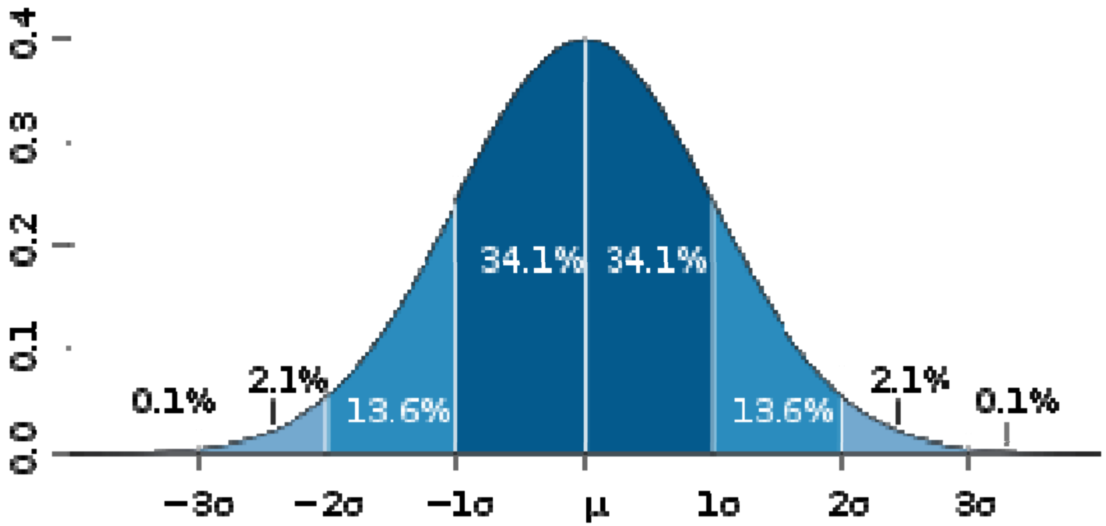
$$\text{Standard Error} = \sigma / \sqrt{n} = 2.5 / \sqrt{25} = 0.5$$

n is the sample size

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} = \frac{\overline{X} - \mu}{0.5}$$

μ is not known

Test of Significance (4/5)



$$P(-z \le Z \le z) = 1 - \alpha = 0.95$$

$$\Phi(z) = P(Z \le z) = 1 - \frac{\alpha}{2} = 0.975$$

From Tables

$$z = \Phi^{-1}(\Phi(z)) = \Phi^{-1}(0.975) = 1.96$$

$$0.95 = 1 - \alpha = P(-z \le Z \le z) = P(-1.96 \le \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \le 1.96)$$

Test of Significance (5/5)

$$P(-z \le Z \le z) = P(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + 1.96\frac{\sigma}{\sqrt{n}})$$

$$P(-z \le Z \le z) = P(\overline{X} - 1.96 * 0.5 \le \mu \le \overline{X} + 1.96 * 0.5)$$

$$P(-z \le Z \le z) = P(\overline{X} - 0.98 \le \mu \le \overline{X} + 0.98)$$

$$Our\ Interval \quad = (250.2 - 0.98;\ 250.2 + 0.98)$$

$$Our\ Interval \quad = (249.22;\ 251.0)$$

- Any value within this interval is not significant

The Information Theory

Part 9

Introduction
Entropy

The Information Theory

The information conveyed by a message can be measured in bits by its probability

The Information Theory: Given Data

Attributes:
D1, D2, D3, D4

Domain(D1)={1,2,3}

Domain(D2)={1,2}

Domain(D3)={1,2}

Domain(D4)={A,B}

D1	D2	D3	D4	D5
1	2	1	A	1
1	2	2	A	0
1	2	2	B	0
1	2	2	B	0
1	1	1	B	1
2	2	2	A	1
2	2	2	B	1
2	1	1	A	1
2	2	1	B	1
3	1	2	A	0
3	1	1	A	0
3	2	2	B	1
3	1	2	B	1
3	1	2	B	1

Decision Attributes: D5

Domain(D5)={0,1}

Two Decisions: 0, 1

The Information Theory: Given Data

D1		1		2		3	
D4	D3\D2	1	2	1	2	1	2
A	1		1	1		0	
	2		0		1	0	
B	1	1	1		1	1	
	2		0		1	1	1

D1	D2	D3	D4	D5
1	2	1	A	1
1	2	2	A	0
1	2	1	B	0
1	2	2	B	0
1	1	1	B	1
2	2	2	A	1
2	2	2	B	1
2	1	1	A	1
2	2	1	B	1
3	1	2	A	0
3	1	1	A	0
3	2	2	B	1
3	1	1	B	1
3	1	2	B	1

The Information Theory: Entropy

THE INFORMATION THEORY: information conveyed by a message depends on its probability and can be measured in bits as minus the logarithm (base 2) of that probability

suppose $D_1, \ldots, D_m$ are m attributes and $C_1, \ldots, C_n$ are n decision classes in a given data. Suppose S is any set of cases, and T is the initial set of training cases $S \subset T$. The **frequency of class C_i in the set S** is:

$$freq(C_i, S) = Number\ of\ examples\ in\ S\ belonging\ to\ C_i$$

If |S| is the total number of examples in S, ***the probability that an example selected at random from S belongs to class C_i*** is

$$freq(C_i, S) / |S|$$

The information conveyed by the message that "**a selected example belongs to a given decision class, C_i**", is determined by

$$-\log_2(freq(C_i, S) / |S|) \quad bits$$

The Information Theory: Entropy

The information conveyed by the message "***a selected example belongs to a given decision class, C_i***"

$$-\log_2(freq(C_i, S)/|S|) \quad bits$$

The Entropy: The expected information from a message stating class membership is given by

$$Info(S) = -\sum_{i=1}^{k}(freq(C_i, S)/|S|)*\log_2(freq(C_i, S)/|S|) \quad bits$$

info(S) is known as the ***entropy*** of the set S. When S is the initial set of training examples, *info(S) determines the average amount of information needed to identify the class of an example in S*.

The Information Theory: The Gain Ratio

Example

$$freq(0,S)=5 \qquad freq(1,S)=9$$

$$freq(0,S)/|S|=5/14 \qquad freq(1,S)/|S|=9/14$$

S

D1	D2	D3	D4	D5
1	2	1	A	1
1	2	2	A	0
1	2	2	B	0
1	2	2	B	0
1	1	1	B	1
2	2	2	A	1
2	2	2	B	1
2	1	1	A	1
2	2	1	B	1
3	1	2	A	0
3	1	1	A	0
3	2	2	B	1
3	1	2	B	1
3	1	2	B	1

The Entropy: the average amount of information needed to identify the class of an example in S

$$Info(S)=-9/14*\log_2(9/14)-5/14*\log_2(5/14)=0.94bits$$

Using D_1 to Split the data provide 3 subsets of data

$$Info_{D_1}(S_1)=-3/5*\log_2(3/5)-2/5*\log_2(2/5)=0.94$$

$$Info_{D_1}(S_2)=-4/4*\log_2(4/4)=0.94$$

$$Info_{D_1}(S_3)=-2/5*\log_2(2/5)-3/5*\log_2(3/5)=0.94$$

$$Info_{D_1}(S)=(5/14)*Info_{D_1}(S_1)+(4/14)*Info_{D_1}(S_2)+(5/14)*Info_{D_1}(S_3)=0.694$$

The Information Theory: The Gain Ratio

Suppose attribute $\underline{D_i}$ is selected to be the root and it has $\underline{k}$ possible values. The expected information of selecting D to partition the training set S, $\text{info}_{D_i}(S)$, can be calculated as follows:

$$Info_{D_i}(S) = \sum_{i=1}^{k} \left({|S_i|}/{|S|} \right) * Info(S_i)$$

S_i is the subset number i of the data; k is the number of values of D_i

The information gained by partitioning the training examples S into subset using the attribute D_1 is given by

$$Gain(X_i) = Info(S) - Info_{D_i}(S)$$

The Information Theory: The Gain Ratio

The attribute to be selected is the attribute with maximum gain value. Quinlan found out that a key attribute will have the maximum gain. This is not good!

$$Split\_Info(S) = -\sum_{i=1}^{k} (|S_i|/|S|) * \log_2 (|S_i|/|S|)$$

The gain ratio is given by:

$$Gain\_Ratio(D_i) = Gain(D_i) / Split\_Info(D_i)$$

Quinlan, J.R., (1993). "C4.5: Programs for Machine Learning", Morgan Kaufmann, Los Altos, California.

The Information Theory: The Gain Ratio

Example Cont.

$$Info_{D_1}(S) = (5/14) * Info_{D_1}(S_1) + (4/14) * Info_{D_1}(S_2) + (5/14) * Info_{D_1}(S_3) = 0.694$$

$$Gain(D_1) = 0.94 - 0.694 = 0.246$$

$$Split\_Info(S) = -5/14 * \log_2(5/14) - 4/14 * \log_2(4/14) - 5/14 \log_2(5/14) = 1.577 \quad bits$$

$$Gain\_Ratio(D_1) = 0.246 / 1.577 = 0.156$$

S

D1	D2	D3	D4	D5
1	2	1	A	1
1	2	2	A	0
1	2	2	B	0
1	2	2	B	0
1	1	1	B	1
2	2	2	A	1
2	2	2	B	1
2	1	1	A	1
2	2	1	B	1
3	1	2	A	0
3	1	1	A	0
3	2	2	B	1
3	1	2	B	1
3	1	2	B	1

Information Gain: Term vs. Category

It measures the classification power of a term

$$IG(t_k, c_i) = \sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} P(t,c) \log_2 \frac{P(t,c)}{P(t)P(c)}$$

$P(t_k, c_i)$ ➔probability document x contains term t and belongs to category c.

$P(\bar{t}_k, c_i)$ ➔probability document x does not contain term t and belongs to category c.

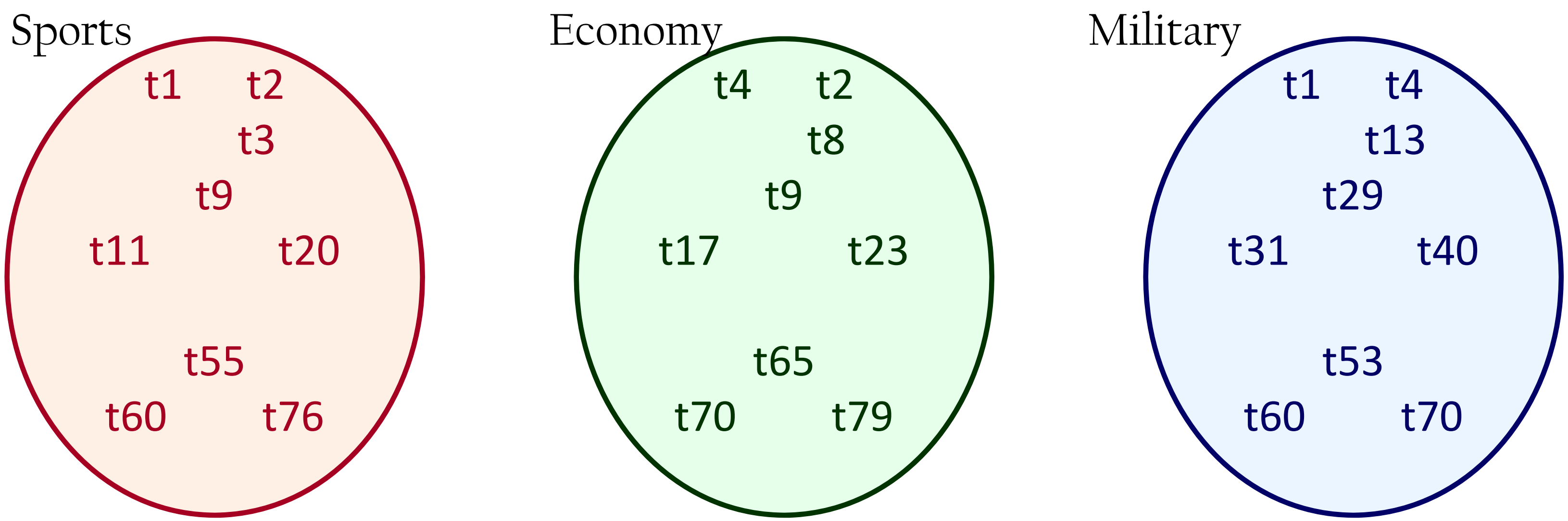
$P(t_k, \bar{c}_i)$ ➔probability document x contains term t and does not belong to category c.

$P(\bar{t}_k, \bar{c}_i)$ ➔probability document x does not contain term t and does not belong to category c.

$P(t)$ ➔probability of term t.

$P(c)$ ➔probability of category c.

Testing The Membership



$$IG(t_k, c_i) = \sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} P(t,c) \log_2 \frac{P(t,c)}{P(t)P(c)}$$

$$IG(t_1, sport) = \frac{1}{9} * \log_2 \frac{1/9}{(2/27)*(9/27)} + \frac{8}{9} * \log_2 \frac{8/9}{(25/27)*(9/27)}$$
$$+ \frac{1}{18} * \log_2 \frac{1/18}{(2/27)*(18/27)} + \frac{17}{27} * \log_2 \frac{17/27}{(25/27)*(18/27)}$$

The Gain Ratio

$$GR(t_k, c_i) = \frac{\sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} P(t, c) \log_2 \frac{P(t, c)}{P(t) P(c)}}{- \sum_{c \in \{c_i, \bar{c}_i\}} P(c) \log_2 P(c)}$$

$P(t_k, c_i)$ ➔probability document x contains term t and belongs to category c.

$P(\bar{t}_k, c_i)$ ➔probability document x does not contain term t and belongs to category c.

$P(t_k, \bar{c}_i)$ ➔probability document x contains term t and does not belong to category c.

$P(\bar{t}_k, \bar{c}_i)$ ➔probability document x does not contain term t and does not belong to category c.

$P(t)$ ➔probability of term t.

$P(c)$ ➔probability of category c.

Basics for Language Engineers

Part 10

Evaluating Documents

Term Frequency & Inverse Document Frequency

Usually a combination of the term frequency and the inverse document frequency

$$TFIDF = w_{ik} = tf_{ik} \times idf_{ik}$$

$$tf_{ik} = 1 + \log_2(tr_{ik}) \qquad \textit{and zero when} \log = 0$$

$$idf_{ik} = \log_2\left(\frac{N}{n_{ik}}\right) \qquad \textit{and zero when} \log = 0$$

tf_{ik} is the term frequency of term i in document k, tr_{ik} is the count of term i in document k, idf_{ik} is the inverse document frequency of term i in document k, N is the total number of documents in the collection, n_{ik} is the number of occurrence of term i in document k, w_{ik} is the weight of term i in document k. Logarithm has been used to reduces the difference between the weight of high and low frequency terms. Logarithm of base 2 is used when vectors are full of binary TFIDF weights 0 and 1. Logarithm of base 10 is used when vectors are full of TFIDF weights except binary ones. TFIDF weights values are not normalized.

The Magical Recipe

$$tf_{ik} = 1 + \log_2(tr_{ik}) \qquad and \quad zero \quad when \quad \log = 0$$

$$idf_{ik} = \log_2\left(\frac{N}{n_{ik}}\right) \qquad and \quad zero \quad when \quad \log = 0$$

$$\boxed{\log_2 x = \log_{10} x / \log_{10} 2}$$

Term Count			Term frequency	
D_1	D_2		D_1	D_2
$\begin{bmatrix} 2 \\ 3 \\ 5 \end{bmatrix}$	$\begin{bmatrix} 3 \\ 7 \\ 1 \end{bmatrix}$	$\Longrightarrow$	$\begin{bmatrix} 2 \\ 2.6 \\ 3.3 \end{bmatrix}$	$\begin{bmatrix} 2.6 \\ 3.8 \\ 1 \end{bmatrix}$

STATISTICAL ASSOCIATIONS

Part II

Association Rules

Learning Term-Association

T1	T2	T3	T4	T5	T6	T7	
1	1	1	1	1	1	1	D1
2	1	2	1	1	1	2	D2
1	2	3	1	1	1	3	D3
2	2	1	2	1	2	4	D4
1	1	2	2	1	1	5	D5
2	1	3	2	1	2	6	D6
1	2	1	3	2	2	7	D7
2	2	2	3	2	2	8	D8
1	1	3	3	2	2	9	D9
2	1	1	1	2	1	1	D10
1	2	2	1	2	2	2	D11
2	2	3	1	2	1	3	D12
1	1	1	2	3	1	4	D13
2	1	2	2	3	1	5	D14
1	2	3	2	3	1	6	D15
2	2	1	3	3	1	7	D16
1	1	2	3	3	2	8	D17
2	1	3	3	3	1	9	D18

D1	D2	D3	D4	D5	D6	D7	
1	1	1	1	1	1	1	T1
2	1	2	1	1	1	2	T2
1	2	3	1	1	1	3	T3
2	2	1	2	1	2	4	T4
1	1	2	2	1	1	5	T5
2	1	3	2	1	2	6	T6
1	2	1	3	2	2	7	T7
2	2	2	3	2	2	8	T8
1	1	3	3	2	2	9	T9
2	1	1	1	2	1	1	T10
1	2	2	1	2	2	2	T11
2	2	3	1	2	1	3	T12
1	1	1	2	3	1	4	T13
2	1	2	2	3	1	5	T14
1	2	3	2	3	1	6	T15
2	2	1	3	3	1	7	T16
1	1	2	3	3	2	8	T17
2	1	3	3	3	1	9	T18

Learning Term-Association

AR Syntax:

(condition 1) (condition 2) … (condition n) strength of association

Suppose we quantized the term weights

Drive two association rules with two Conditions and frequency greater than 0.25.

(T1 = 1) (T6 = 1) 5/18
(T1 = 2) (T2 = 1) 5/18

Question:
Drive association rules with two conditions and frequency greater than 0.38.

T1	T2	T3	T4	T5	T6	T7	T8
1	1	1	1	1	1	1	1
2	1	2	1	1	1	2	2
1	2	3	1	1	1	3	3
2	2	1	2	1	2	4	4
1	1	2	2	1	1	5	5
2	1	3	2	1	2	6	6
1	2	1	3	2	2	7	1
2	2	2	3	2	2	8	2
1	1	3	3	2	2	9	3
2	1	1	1	2	1	1	4
1	2	2	1	2	2	2	5
2	2	3	1	2	1	3	6
1	1	1	2	3	1	4	1
2	1	2	2	3	1	5	2
1	2	3	2	3	1	6	3
2	2	1	3	3	1	7	4
1	1	2	3	3	2	8	5
2	1	3	3	3	1	9	6

Learning Term-Association

The strength of an association rule can be measure by:

- Leverage
- Coverage
- Support
- Strength
- Lift

T1	T2	T3	T4	T5
1	1	1	1	1
2	1	2	1	1
1	2	3	1	1
2	2	1	2	1
1	1	2	2	1
2	1	3	2	1
1	2	1	3	2
2	2	2	3	2
1	1	3	3	2
2	1	1	1	2
1	2	2	1	2
2	2	3	1	2
1	1	1	2	3
2	1	2	2	3
1	2	3	2	3
2	1	1	3	3

1. Calculating LEVERAGE for the rule:

(T1 = 2) (T2 = 1)

- Number of records = 16
- Records having (T1 = 2) = 8
- Records having (T2 = 1) = 9
- Records having (T1 = 2) (T2 = 1) = **4**
- % of the cover (T1 = 2) (T2 = 1) = 4/16
- Records expected to be covered by (T1 = 2) (T2 = 1) if they were independent = (8 * 9) / 16 = **4.5**
- Leverage Count = 4.5 – 4 = 0.5
- Leverage Proportion = 0.5 / 16 = 1/32

Learning Term-Association

2. Calculating COVERAGE for the rule:

(T1 = 2) (T2 = 1)

- The coverage count for all conditions but the last one (T2=1) = 8
- The coverage proportional = 8/16 = 1/2

3. Calculating SUPPORT for the rule:

(T1 = 2) (T2 = 1)

- The support count for all conditions = 4
- The support proportional = 4/16 = 1/4

4. Calculating STRENGTH for the rule:

(T1 = 2) (T2 = 1)

- The strength count for all conditions but the last one (T2=1) = 8
- The last condition covers 4 out of those 8
- The strength proportional = 4/8 = 1/2

T1	T2	T3	T4	T5
1	1	1	1	1
2	1	2	1	1
1	2	3	1	1
2	2	1	2	1
1	1	2	2	1
2	1	3	2	1
1	2	1	3	2
2	2	2	3	2
1	1	3	3	2
2	1	1	1	2
1	2	2	1	2
2	2	3	1	2
1	1	1	2	3
2	1	2	2	3
1	2	3	2	3
2	1	1	3	3

Learning Term-Association

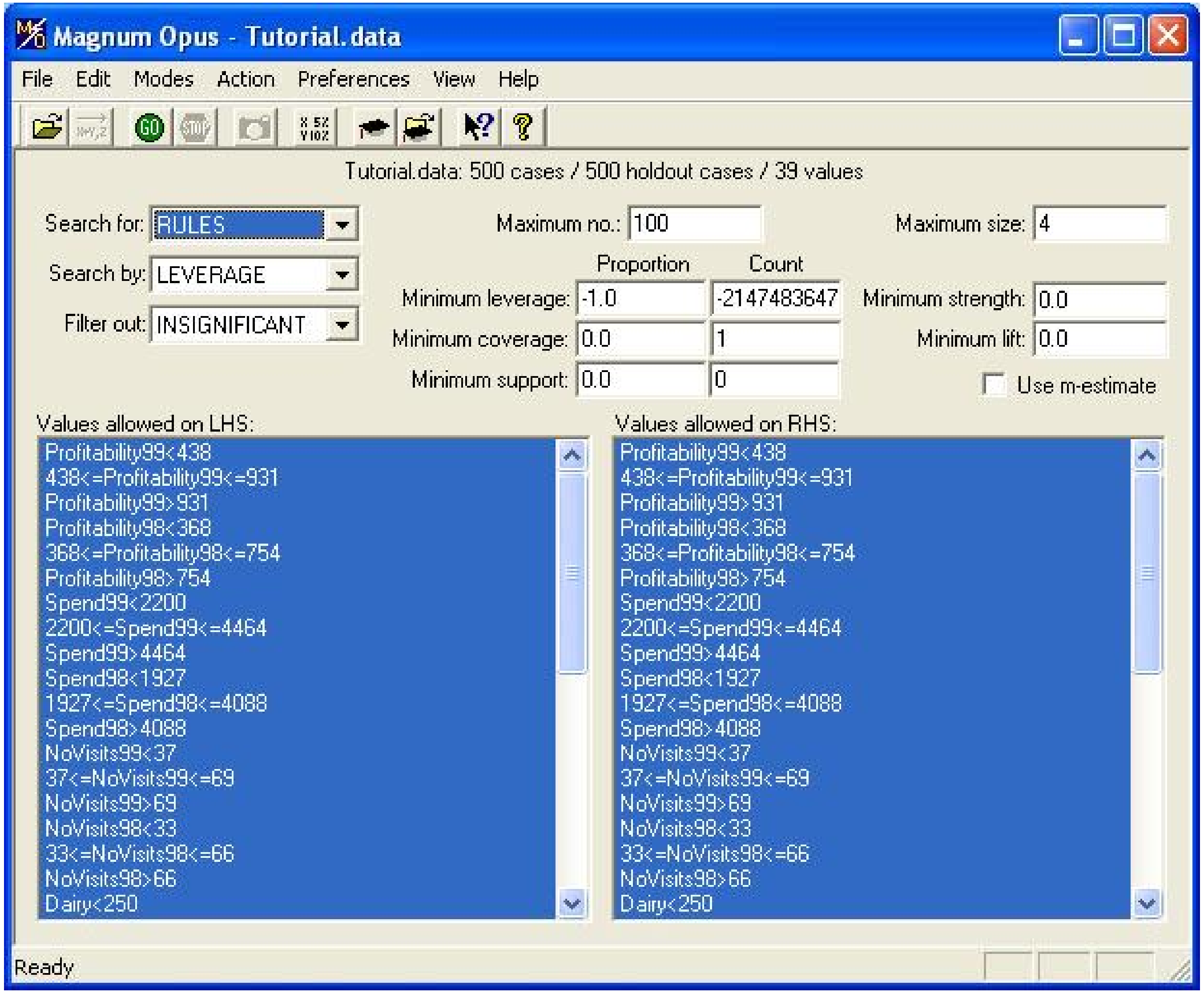
5. Calculating LIFT for the rule:

(T1 = 2) (T2 = 1)

- Total number of examples = 16
- Records covered by all conditions but the last condition (T2=1) = 8
- Records covered by the last condition = 8
- Records covered by all conditions = 4
- Strength = 4 / 8 = 1/2
- Cover proportion of all conditions but the last one (T2=1) = 8 / 16 = 1/2
- LIFT = strength / (cover proportion of all condition but the last) = (1/2) / (1/2) = 1

T1	T2	T3	T4	T5
1	1	1	1	1
2	1	2	1	1
1	2	3	1	1
2	2	1	2	1
1	1	2	2	1
2	1	3	2	1
1	2	1	3	2
2	2	2	3	2
1	1	3	3	2
2	1	1	1	2
1	2	2	1	2
2	2	3	1	2
1	1	1	2	3
2	1	2	2	3
1	2	3	2	3
2	1	1	3	3

The Magnum Opus System



Attributes and their values for the Tutorial database

- Profitability99: numeric 3
- Profitability98: numeric 3
- Spend99: numeric 3
- Spend98: numeric 3
- NoVisits99: numeric 3
- NoVisits98: numeric 3
- Dairy: numeric 3
- Deli: numeric 3
- Bakery: numeric 3
- Grocery: numeric 3
- SocioEconomicGroup: categorical
- Promotion1: t, f
- Promotion2: t, f

Statistical Association

Magnum Opus

DEMO

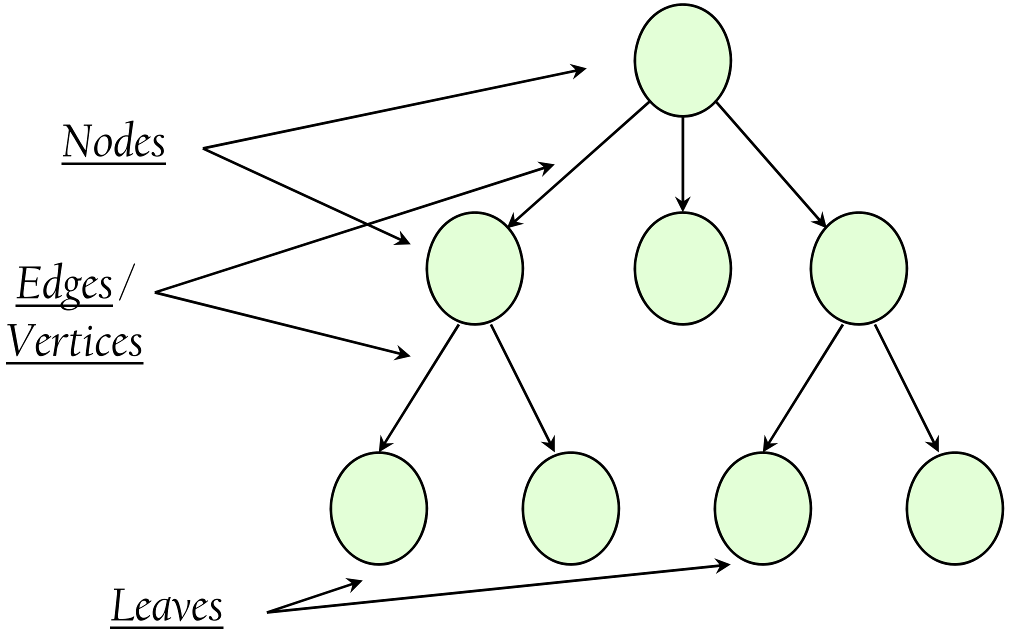
DECISION TREES

Part 12

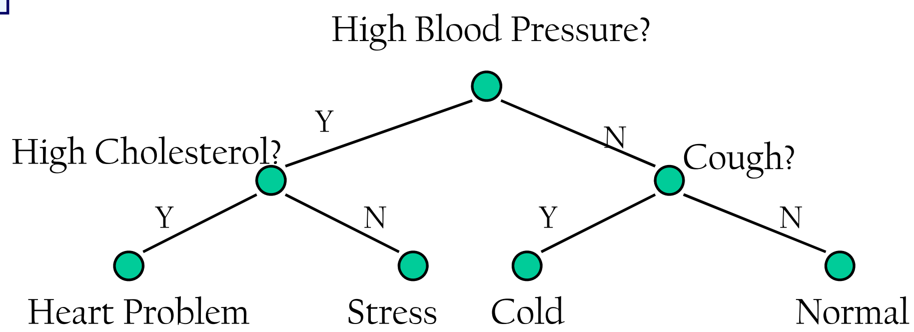
Using Statistical & Information Theory

Learning Decision Trees

- A *Tree* is a Directed Acyclic Graph *(DAG)* + each node has one parent at most
- A *Decision Tree* is a tree where nodes associated with attributes, edges associated with attribute values, and leaves associated with decisions



Example:



Learning Decision Trees

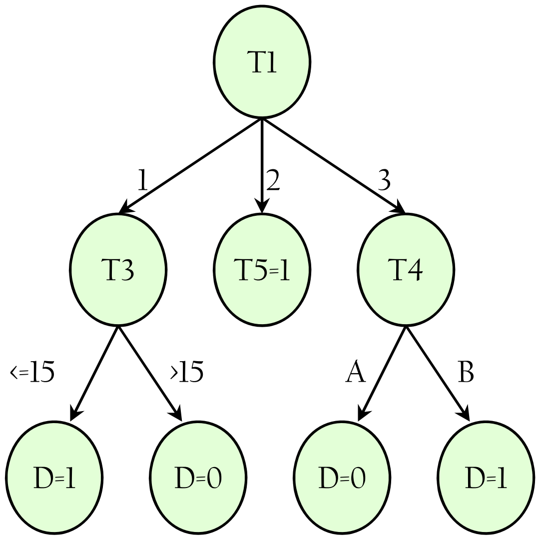
Attribute Selection Criteria

- Logical Based
- Information Based
- Statistical Based

Information Theory

Example

- T2 is quantized into two intervals at 21 (T2<=21) and (T2>21)
- T3 is quantized into two intervals at 15 (T3<=15) and (T3>15)



T1	T2	T3	T4	D
1	25	10	A	1
1	30	30	A	0
1	35	25	B	0
1	22	35	B	0
1	19	10	B	1
2	22	30	A	1
2	33	18	B	1
2	14	5	A	1
2	31	15	B	1
3	21	20	A	0
3	15	10	A	0
3	25	20	B	1
3	18	20	B	1
3	20	36	B	1

Decision Trees

C5

DEMO

NEURAL NETWORKS

Part 13

How It Works?

Learning Neural Networks

Supervised

In terms of Design

The user defines the number of nodes and levels in the hidden layer

As Learning Algorithm

The data is labeled and both input and output are given to the neural network

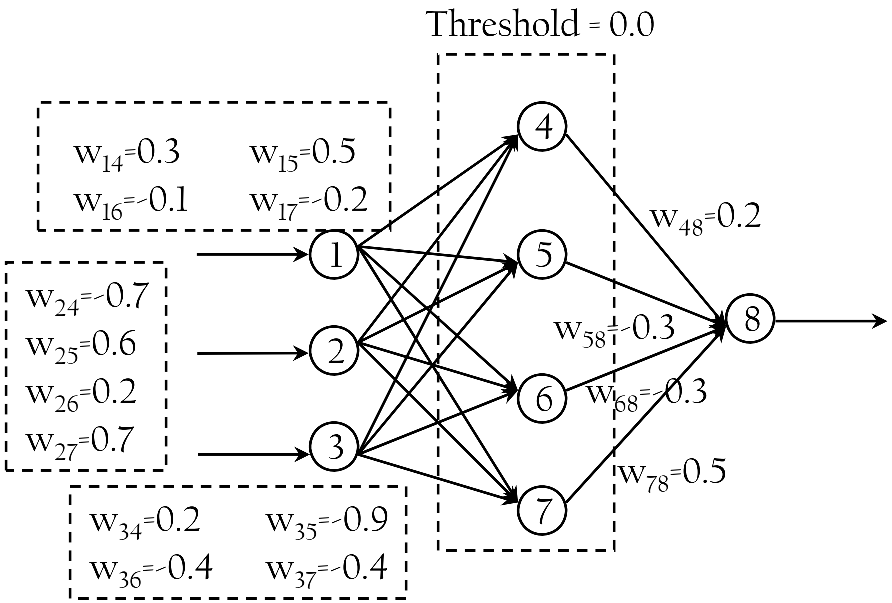
Unsupervised

In terms of Design

No. of nodes and levels in the hidden layer are defined automatically by the algorithm

As Learning Algorithm

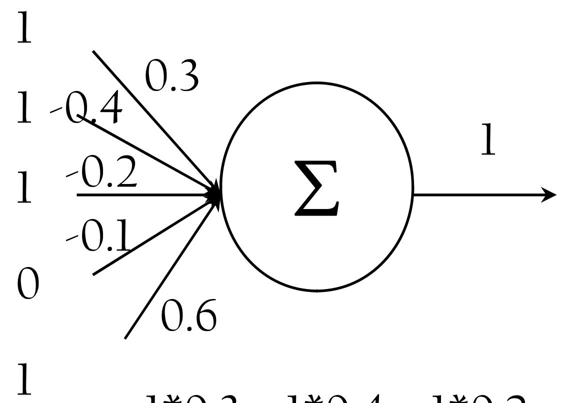
The data is not labeled. Only the input records are given to the neural network



Test Data

A	B	C	Decision
0	0	0	
0	0	1	
0	1	0	
0	1	1	1
1	0	0	
1	0	1	
1	1	0	
1	1	1	

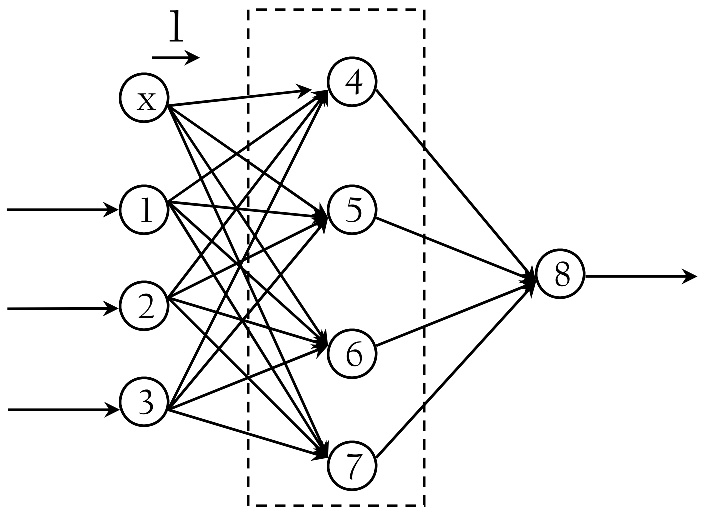
Learning Neural Networks



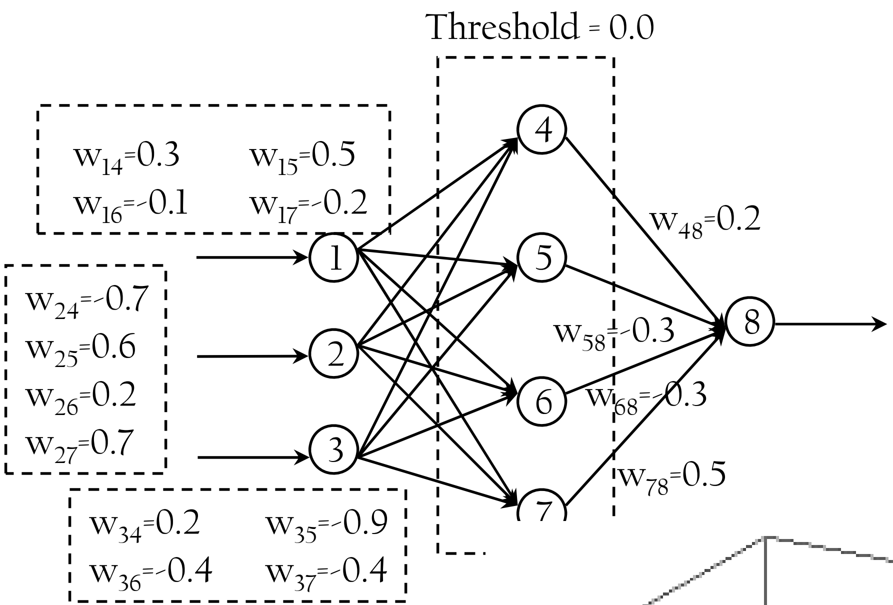
The Sigmoid Function

$=1*0.3 - 1*0.4 - 1*0.2 - 0*0.1 + 1*0.6 = 0.3 > 0.0$

To avoid setting the threshold:

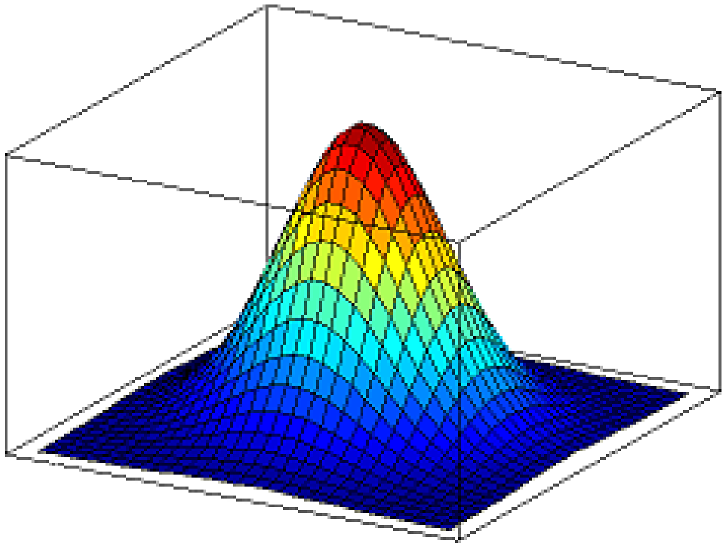


Learning Neural Networks



Test Data

A	B	C	Decision
0	0	0	
0	0	1	
0	1	0	
0	1	1	
1	0	0	
1	0	1	
1	1	0	
1	1	1	



MACHINE TRANSLATION

Part 14

Statistical Machine Translation

Statistical Machine Translation

- For each English sentence “e”, we need the Arabic sentence “a” which maximize $P(a|e)$

  $P(a|e)=P(a)*P(e|a)/P(e)$

English Document $\Longrightarrow$ Arabic Document

Language Model

- A statistical **language model** assigns a probability to a sequence of *m* words by means of a probability distribution
- Record every sentence that anyone ever says in Arabic; Suppose you record a database of one billion utterances; If the sentence "كيف حالك؟" appears 76,413 times in that database, then we say P(كيف حالك؟) = 76,413/1,000,000,000 = 0.000076413
- One big problem is that many perfectly good sentences will be assigned a P(e) of zero

Arabic Sentence	Probability
كيف حالك	0.000076413
الولد سعيد	0.000066392

N-Grams

- An n-word substring is called an n-gram
- If n=2, we say bigram. If n=3, we say trigram
- Let P(y | x) be the probability that word y follows word x

  P(y | x) = number-of-occurrences("xy") / number-of-occurrences("x")

  P(z | x y) = number-of-occurrences("xyz") / number-of-occurrences("xy")

➔ P(ذهب الولد إلى المدرسة) = P(ذهب | start-of-sentence) *
P(الولد | ذهب) * P(إلى | الولد) * P(المدرسة | إلى) *
P(end-of-sentence | المدرسة)

➔ P(ذهب الولد إلى المدرسة) = P(ذهب | start-of-sentence) *
P(الولد | start-of-sentence ,ذهب) * P(إلى | الولد ,ذهب) *
P(المدرسة | إلى ,الولد) * P(end-of-sentence | المدرسة ،إلى) *
P(end-of-sentence | end-of-sentence ,المدرسة)

N-Grams Language Model

$$P(w_1,\ldots,w_m) = \prod_{i=1}^{m} P(w_i \mid w_1,\ldots,w_{i-1}) \approx \prod_{i=1}^{m} P(w_i \mid w_{i-(n-1)},\ldots,w_{i-1})$$

$$P(w_i \mid w_{i-(n-1)},\ldots,w_{i-1}) = \frac{count(w_{i-(n-1)},\ldots,w_i)}{count(w_{i-(n-1)},\ldots,w_{i-1})}$$

Example:

In a bigram (n=2) language model, the approximation looks like

$$P(I, saw, the, red, house) \approx P(I)P(saw \mid I)P(the \mid saw)P(red \mid the)P(house \mid red)$$

In a trigram (n=3) language model, the approximation looks like

$$P(I, saw, the, red, house) \approx P(I)P(saw \mid I)P(the \mid I, saw)P(red \mid saw, the)P(house \mid the, red)$$

Translation Model

- P(a | e), the probability of an Arabic string "a" given an English string "e". This is called a translation model
- P(a | e) will be a module in overall English-to-Arabic machine translation system; When we see an actual English string e, we want to reason backwards ... What Arabic string a is (1) likely to be uttered, and (2) likely to subsequently translate to e? We're looking for the a that maximizes P(a) * P(e | a)

Arabic Sentence	English Sentence	P(a|e)
ذهب الولد إلى المدرسة	The boy went to School	0.0034
إنخفاض البورصة اليوم	Today, the stock market went down	0.00021
:	:	

Translation Model

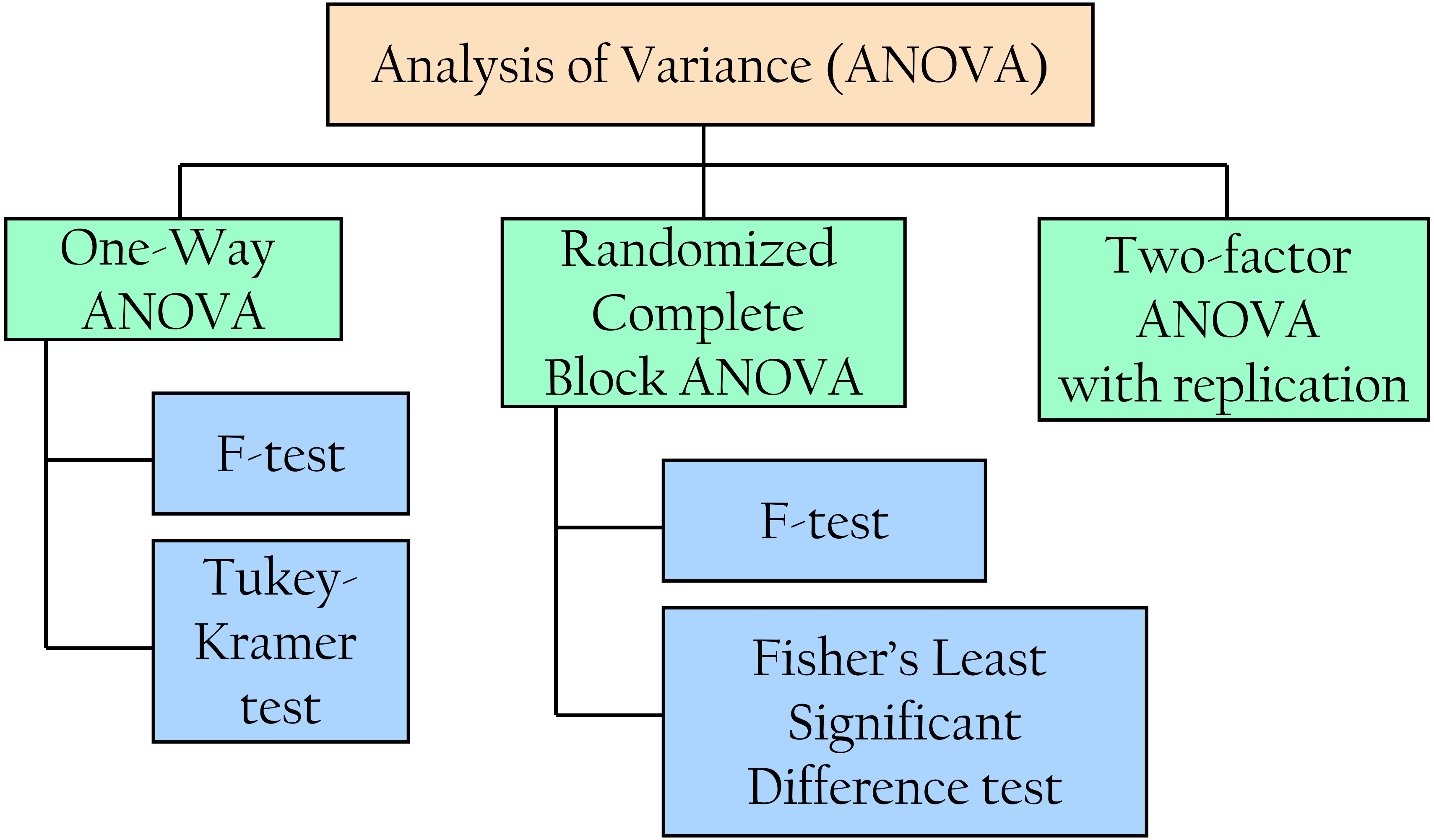
- For each word a_i in an Arabic sentence ($i = 1 \ldots l$), we choose a fertility ϕ_i. The choice of fertility depends on the Arabic word in question. It is not dependent on the other Arabic words in the Arabic sentence, or on their fertilities
- For each word a_i, we generate ϕ_i English words. The choice of English word depends on the Arabic word that generates it. It is not dependent on the Arabic context around the Arabic word. It is not dependent on other English words that have been generated from this or any other Arabic word
- All those English words are permuted. Each English word is assigned an absolute target "position slot." For example, one word may be assigned position 3, and another word may be assigned position 2 -- the latter word would then precede the former in the final English sentence. The choice of position for a English word is dependent solely on the absolute position of the Arabic word that generates it

STATISTICS

Part 15

Analysis of Variance
ANOVA

Analysis of Variance ANOVA

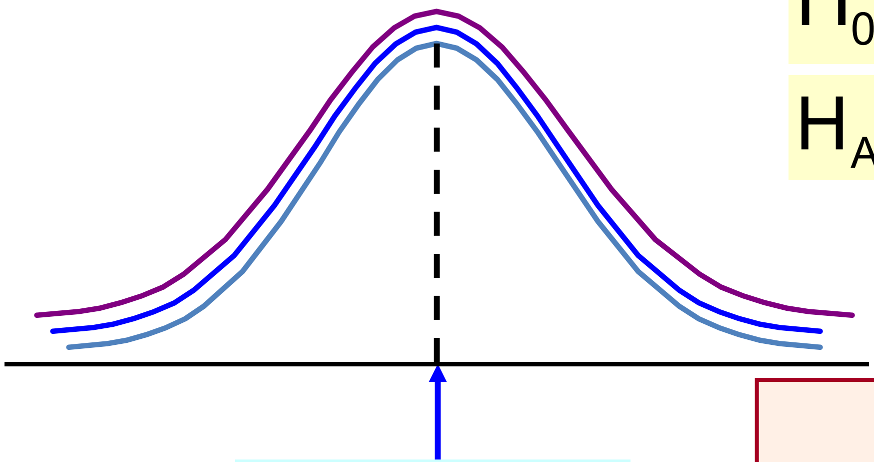


ONE WAY ANOVA

- Evaluate the difference among the means of three or more populations
- Assumptions
  - Populations are normally distributed
  - Populations have equal variances
  - Samples are randomly and independently drawn

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

$$H_A : \text{Not all } \mu_i \text{ are the same}$$



$$\mu_1 = \mu_2 = \mu_3$$

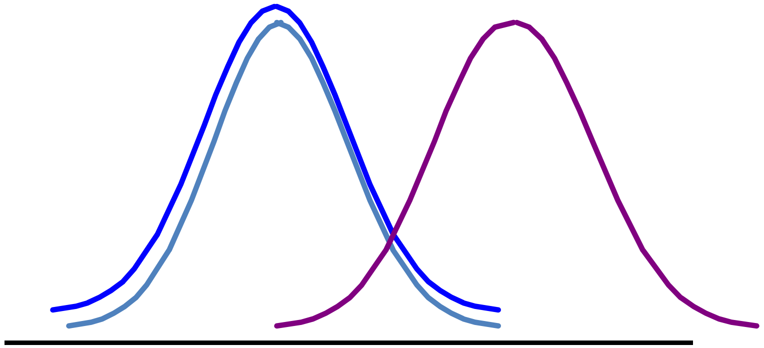
All Means are the same:
The Null Hypothesis is True

ONE WAY ANOVA

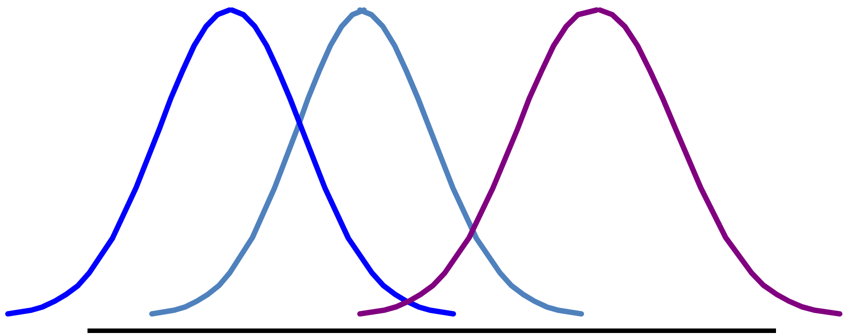
$$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$$

$$H_A : \text{Not all } \mu_i \text{ are the same}$$

At least one mean is different:
The Null Hypothesis is NOT true
(Treatment Effect is present)



or



$$\mu_1 = \mu_2 \neq \mu_3 \qquad \mu_1 \neq \mu_2 \neq \mu_3$$

Partitioning the Variations

$$SST = SSB + SSW$$

SST = Total Sum of Squares
SSB = Sum of Squares Between
SSW = Sum of Squares Within

Total Variation = the aggregate dispersion of the individual data values across the various factor levels (SST)

Between-Sample Variation = dispersion among the factor sample means (SSB)

Within-Sample Variation = dispersion that exists among the data values within a particular factor level (SSW)

Partition of Total Variation

Total Variation (SST)

= **Variation Due to Factor (SSB)** + **Variation Due to Random Sampling (SSW)**

Commonly referred to as:

- Sum of Squares Between
- Sum of Squares Among
- Sum of Squares Explained
- Among Groups Variation

Commonly referred to as:

- Sum of Squares Within
- Sum of Squares Error
- Sum of Squares Unexplained
- Within Groups Variation

Total Sum of Squares

$$\boxed{SST} = SSB + SSW$$

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{\bar{x}})^2$$

Where:

SST = Total sum of squares

k = number of populations (levels or treatments)

n_i = sample size from population i

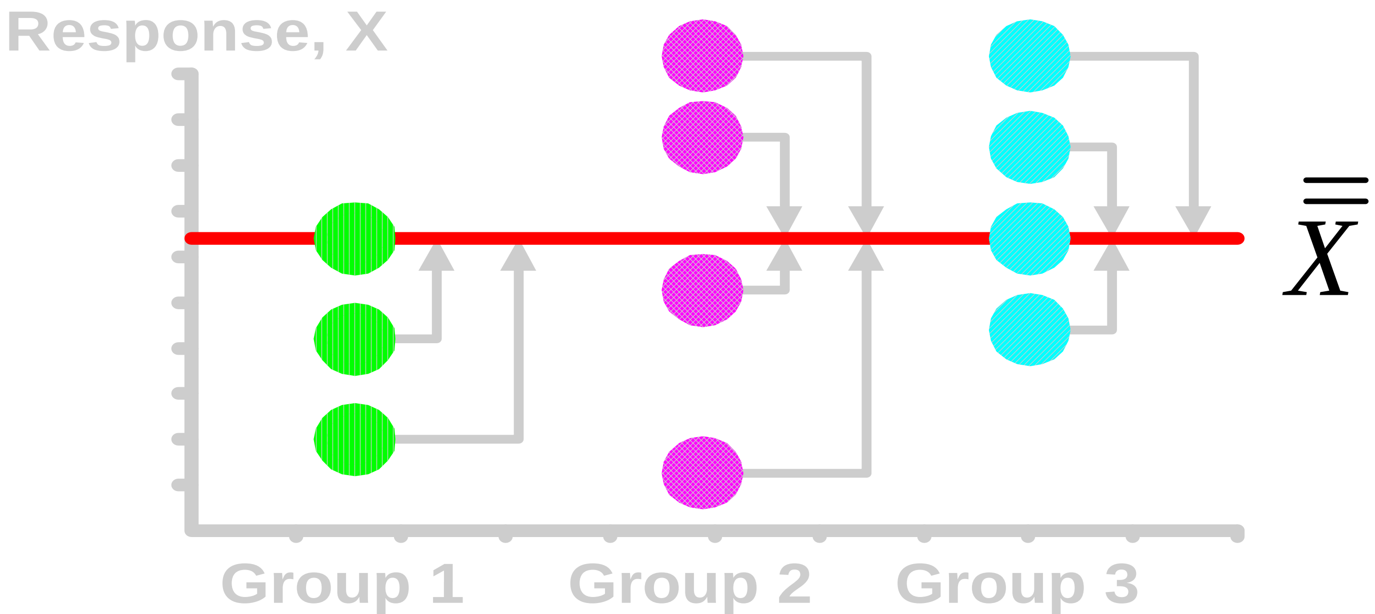
x_{ij} = j^{th} measurement from population i

$\bar{\bar{x}}$ = grand mean (mean of all data values)

Total Variation

(continued)

$$\boxed{SST = (x_{11} - \bar{\bar{x}})^2 + (x_{12} - \bar{\bar{x}})^2 + ... + (x_{kn_k} - \bar{\bar{x}})^2}$$



Sum of Squares Between

$$SST = \boxed{SSB} + SSW$$

$$\boxed{SSB = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{\bar{x}})^2}$$

Where:

SSB = Sum of squares between

k = number of populations

n_i = sample size from population i

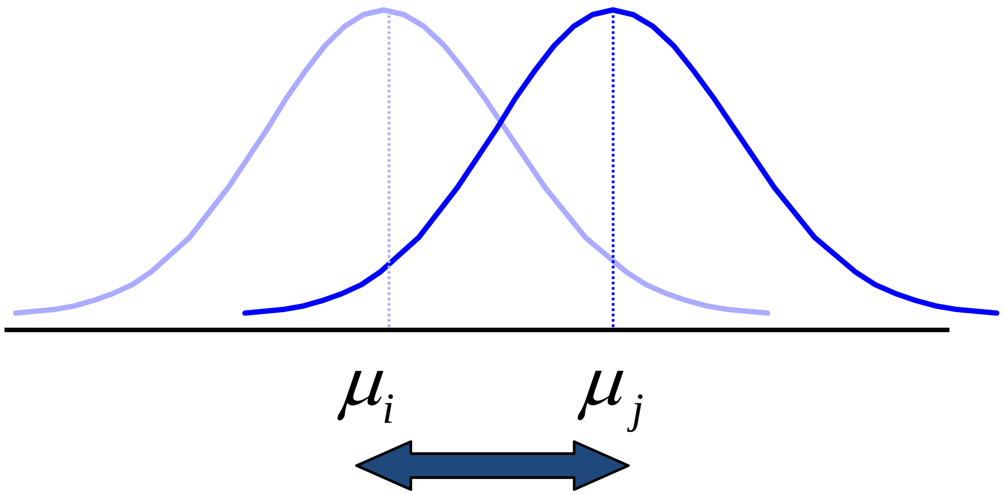
$\bar{x}_i$ = sample mean from population i

$\bar{\bar{x}}$ = grand mean (mean of all data values)

Between-Group Variation

$$SSB = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{\bar{x}})^2$$

Variation Due to
Differences Among Groups



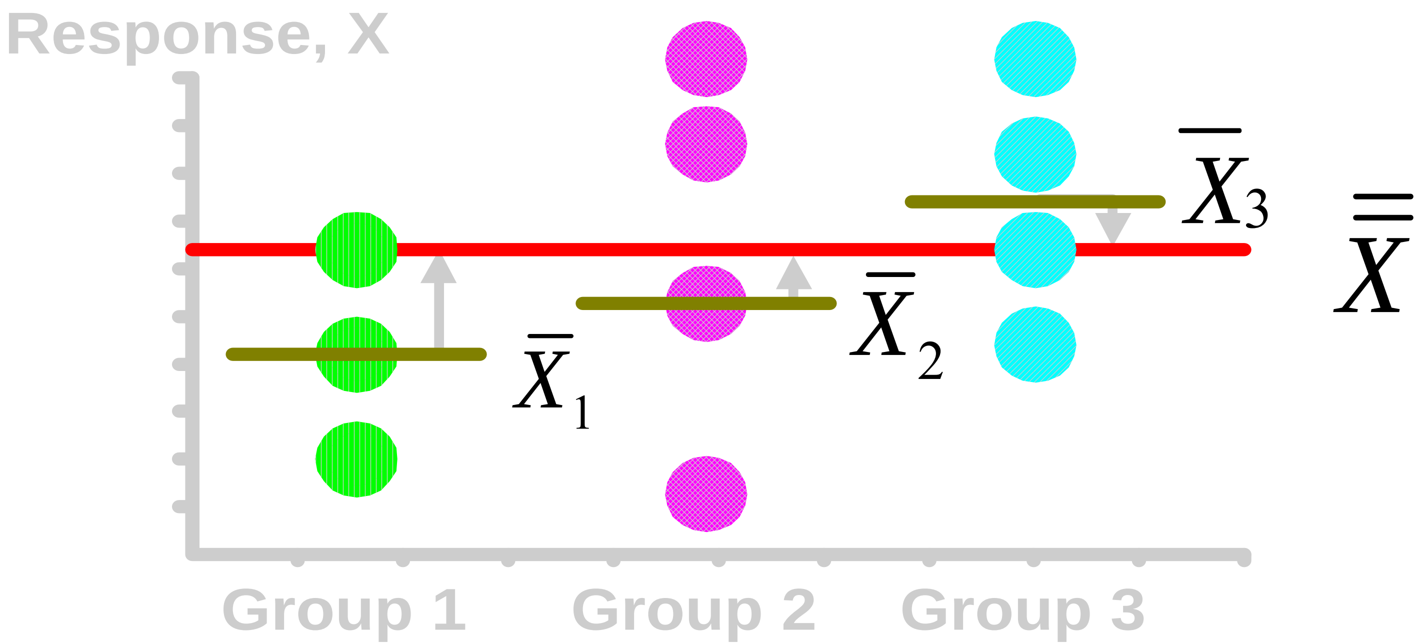
$$MSB = \frac{SSB}{k-1}$$

Mean Square Between =
SSB/degrees of freedom

Between-Group Variation

(continued)

$$SSB = n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + \dots + n_k(\bar{x}_k - \bar{\bar{x}})^2$$



Sum of Squares Within

$$SST = SSB + \boxed{SSW}$$

$$\boxed{SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_j} (x_{ij} - \bar{x}_i)^2}$$

Where:

SSW = Sum of squares within

k = number of populations

n_i = sample size from population i

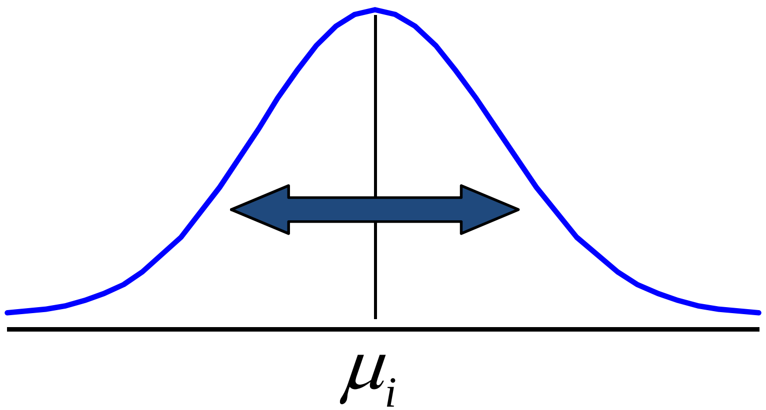
$\bar{x}_i$ = sample mean from population i

x_{ij} = j^{th} measurement from population i

Within-Group Variation

$$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_j} (x_{ij} - \bar{x}_i)^2$$

Summing the variation within each group and then adding over all groups



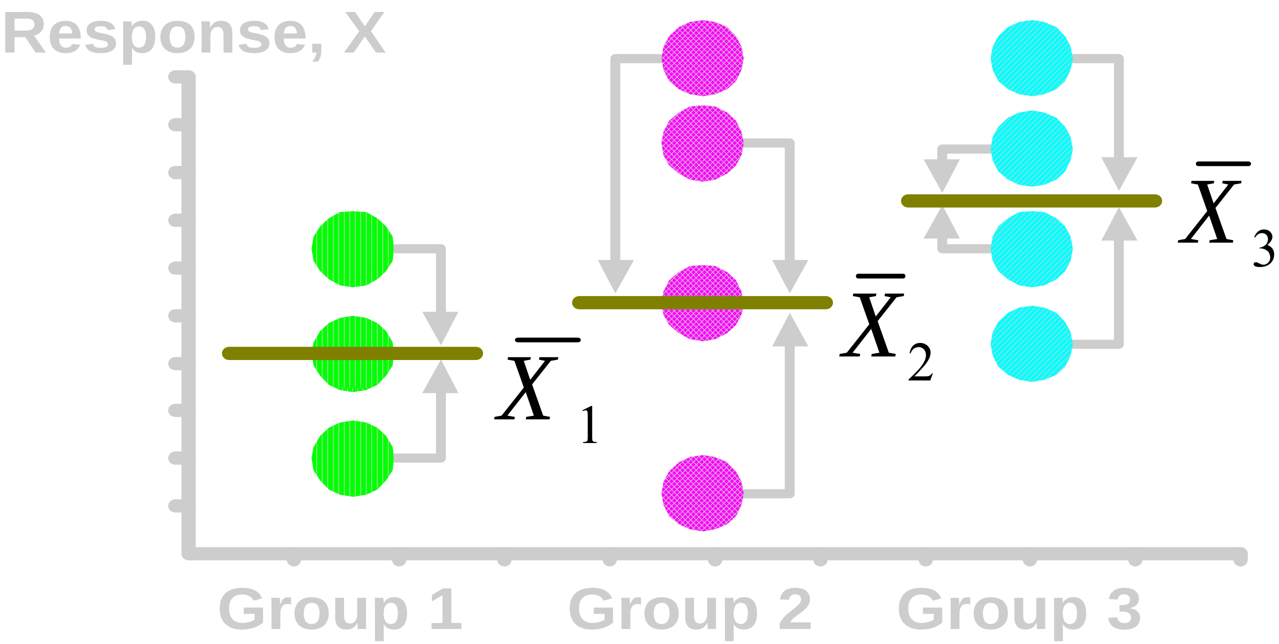
$$MSW = \frac{SSW}{N - k}$$

Mean Square Within = SSW/degrees of freedom

Within-Group Variation

(continued)

$$SSW = (x_{11} - \bar{x}_1)^2 + (x_{12} - \bar{x}_2)^2 + \ldots + (x_{kn_k} - \bar{x}_k)^2$$

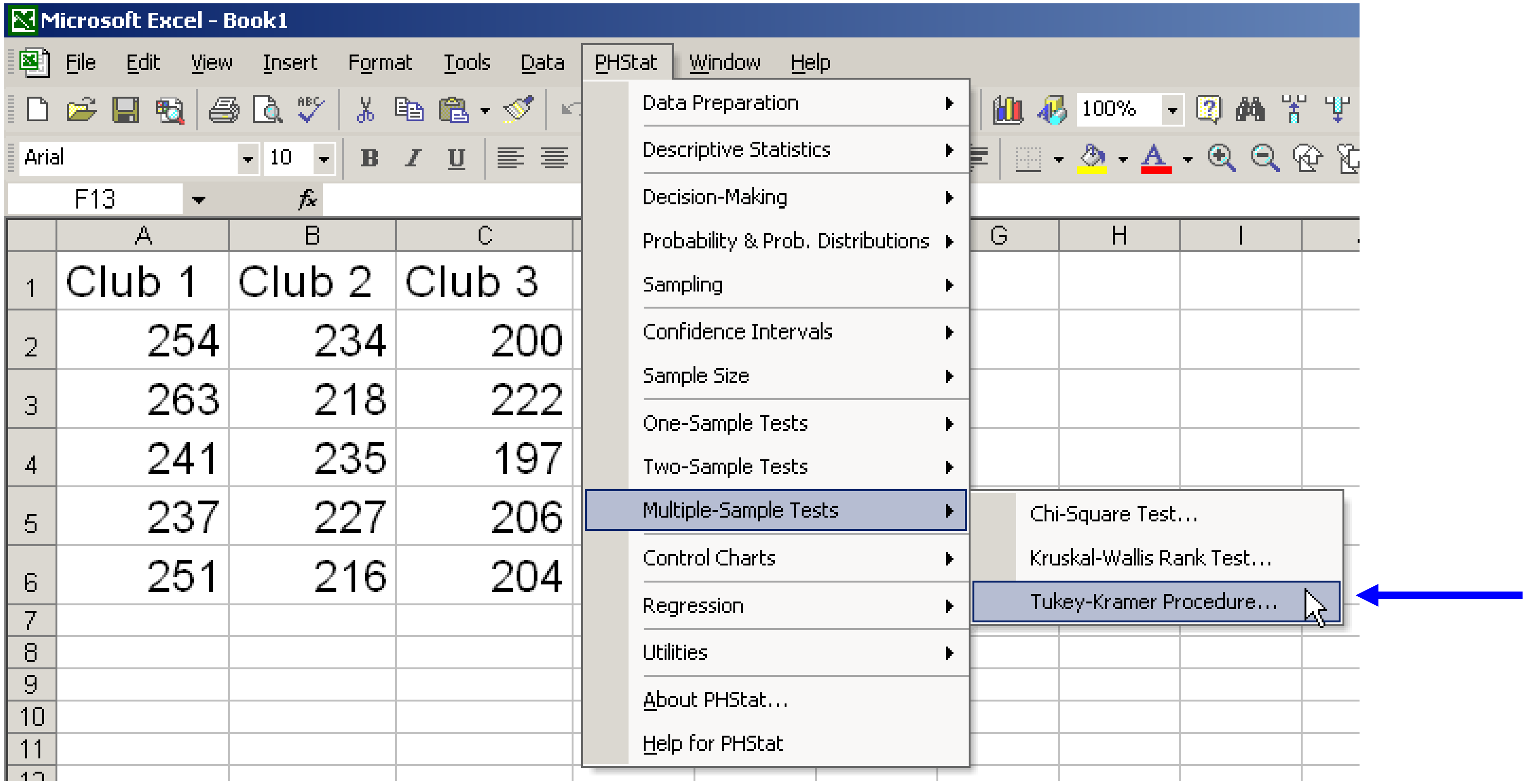


One-Way ANOVA Table

Source of Variation	SS	df	MS	F ratio
Between Samples	SSB	k - 1	$MSB = \frac{SSB}{k - 1}$	$F = \frac{MSB}{MSW}$
Within Samples	SSW	N - k	$MSW = \frac{SSW}{N - k}$	
Total	SST = SSB+SSW	N - 1		

k = number of populations
N = sum of the sample sizes from all populations
df = degrees of freedom

Tukey-Kramer in PHStat



Probability

Part 16

Bayesian Networks

Bayesian Networks (Watch Me!)

Conclusion

1- Basic Concepts

2- Introduction to Vectors

3- Probability

4- Statistics

5- Regression

6- Statistics & Testing

7- Test of Significance

8- Information Theory

9- Basics for Language Engineers

10- Statistical Association

11- Statistical Machine Translation

12- Analysis of Variance

13- Bayesian Networks

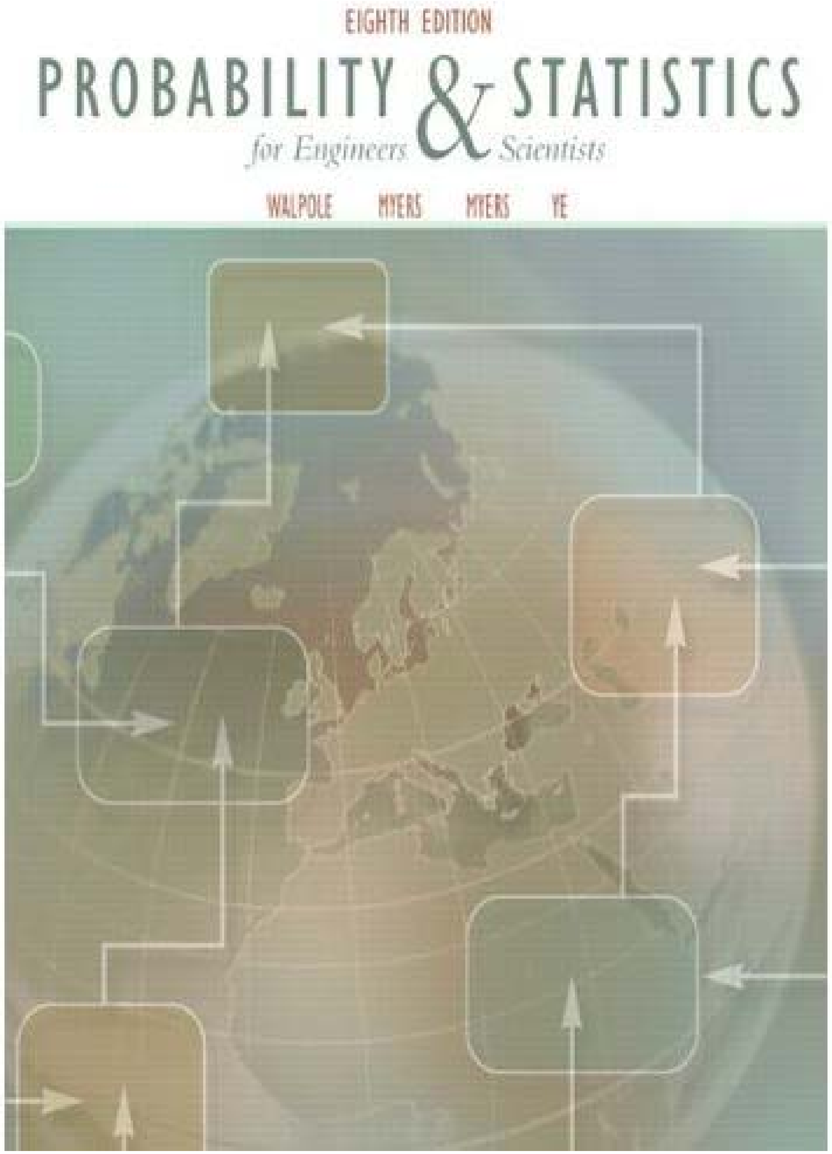
REFERENCES

- W. Weaver (1955). Translation (1949). In: *Machine Translation of Languages*, MIT Press, Cambridge, MA.
- P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, **19(2)**, 263-311.
- S. Vogel, H. Ney and C. Tillmann. 1996. HMM-based Word Alignment in StatisticalTranslation. In COLING '96: The 16th International Conference on Computational Linguistics, pp. 836-841, Copenhagen, Denmark.
- F. Och and H. Ney. (2003). A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, 29(1):19-51
- P. Koehn, F.J. Och, and D. Marcu (2003). Statistical phrase based translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL)*.
- D. Chiang (2005). A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*.
- F. Och and H. Ney. (2003). A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, 29(1):19-51
- P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. ACL 2007, Demonstration Session, Prague, Czech Republic
- Q. Gao, S. Vogel, "Parallel Implementations of Word Alignment Tool", Software Engineering, Testing, and Quality Assurance for Natural Language Processing, pp. 49-57, June, 2008
- W. J. Hutchens and H. Somers. (1992). An Introduction to Machine Translation, 18.3:322. ISBN 0-12-36280-X

REFERENCES

- W. The Sage Dictionary of Statistics, pg. 76, Duncan Cramer, Dennis Howitt, 2004, ISBN 076194138X
- E.L. Lehmann and Joseph P. Romano (2005). *Testing Statistical Hypotheses* (3E ed.). New York, NY: Springer. ISBN 0387988645
- D.R. Cox and D.V.Hinkley (1974). *Theoretical Statistics*. ISBN 0412124293.
- Fisher, Sir Ronald A. (1956) [1935]. "Mathematics of a Lady Tasting Tea". in James Roy Newman. *The World of Mathematics, volume 3*. http://books.google.com/books?id=oKZwtLQTmNAC&pg=PA1512&dq=%22mathematics+of+a+lady+tasting+tea%22&sig=8-NQlCLzrh-oV0wjfwa0EgspSNU
- R.A. Fisher, the Life of a Scientist, Box, 1978, p134
- Mccloskey, Deirdre (2008). *The Cult of Statistical Significance*. Ann Arbor: University of Michigan Press. ISBN 0472050079
- *What If There Were No Significance Tests?*, Harlow, Mulaik & Steiger, 1997, ISBN 978-0-8058-2634-0
- Rosnow, R.L. & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276-1284
- Loftus, G.R. 1991. On the tyranny of hypothesis testing in the social sciences. Contemporary Psychology 36: 102-105
- Cohen, J. 1990. Things I have learned (so far). American Psychologist 45: 1304-1312. ^ Introductory Statistics, Fifth Edition, 1999, pg. 521, Neil A. Weiss, ISBN 0-201-59877-9
- Ioannidis JP (July 2005). "Contradicted and initially stronger effects in highly cited clinical research". *JAMA* **294** (2): 218–28.

REFERENCES


EIGHTH EDITION
PROBABILITY & STATISTICS
for Engineers & Scientists
WALPOLE MYERS MYERS YE