max planck institut
informatik

Introduction to Statistics and Statistical Learning

Alejandro Pironti

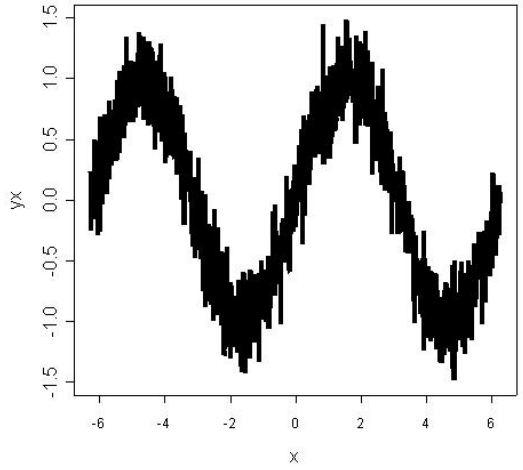
Dar es Salaam

November 16th, 2011

euresist

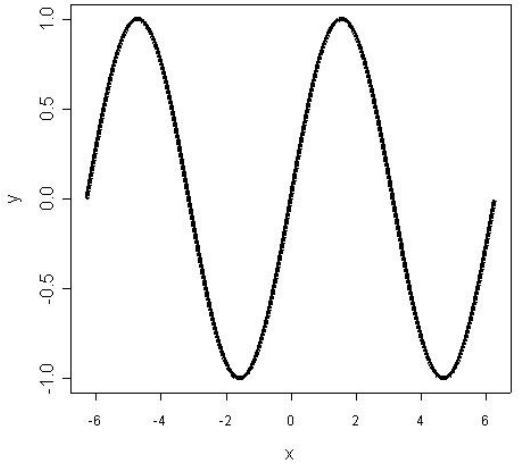
Dealing with Randomness

Measurement with noise



Statistics can help

De-noised measurement



- Scientific experiments are almost always influenced by randomness
- Measurement results are not precisely predictable due to:
  - Measurement errors
  - Biological variability
  - Sample selection
- We need methods for dealing with randomness
- By randomness we **do not** always mean absolute lack of pattern

Random Events

- Events that are subject to randomness are called **random events**
- The set of all possible outcomes (random events) of an experiment subject to randomness is called sample space
- Random events may be elementary or compound

Sample space Ω = {1,2,3,4,5,6}
Elementary event A = {1}
Compound event B= {2,4,6}

Sample Space

- By defining the sample space, we define the possible outcomes of our experiment
- This might be idealized, such that we ignore certain possible outcomes
- Sample spaces might be:
  - Discrete
  - Continuous

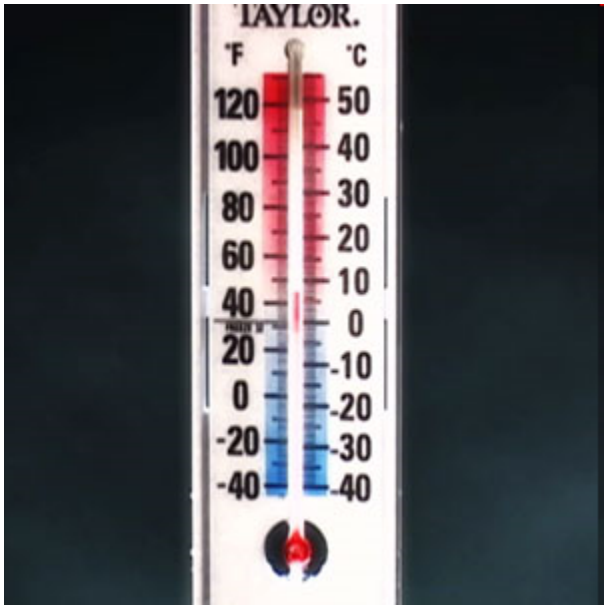


Image: http://www.learner.org/courses/essential/physicalsci/session7/closer1.html

The sample space of a temperature-measuring experiment could be defined to be discrete or continuous:
Ω = [-40;50]
Ω = {cold, warm, hot}

Probability

- What is the probability of obtaining a 6 when throwing a die?
- Classic probability:

$$P(\{6\}) = \frac{\#\text{ outcomes with a six}}{\#\text{ possible outcomes}} = \frac{1}{6}$$

- Sex of newborns

  Ω = {male, female}

  P({female}) = 0.5???????
- Statistic probability:

$$\hat{P}(\{female\}) = f(\{female\}) = \frac{\#\text{ females born}}{\#\text{ children born}}$$

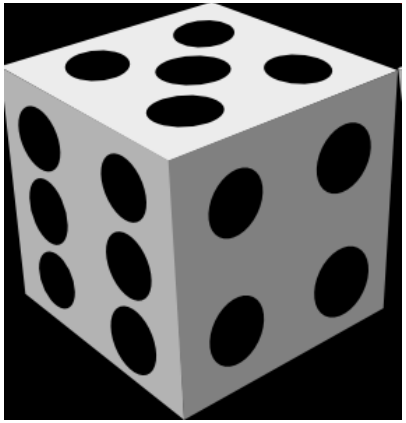
Random sample

n	f({female})
10	0.4
100	0.63
5 000	0.47
50 000	0.478
300 000	0.482

Axioms of Probability

Let A and B be events of the sample space Ω and P(A), P(B) their probabilities.

1. A probability P(A) is assigned to each event A from Ω s.t. $0 \leq P(A) \leq 1$
2. The probabilies of Ω (certain event) and Ø (impossible event) are $P(\Omega) = 1$ and $P(\emptyset) = 0$
3. For two mutually exclusive events $P(A \cup B) = P(A) + P(B)$



1. P({1}) = 1/6; P({1,2}) = 1/3; P({1,2,3,}) = 1/2
2. P({1,2,3,4,5,6}) = 1; P({}) = 0
3. P({1,2}) = 1/3
   P({3,4,5}) = 1/2
   P({1,2} U {3,4,5}) =
   P({1,2,3,4,5}) =
   P({1,2}) + P({3,4,5}) =
   1/2 + 1/3 =
   5/6

Random Variable

- A random variable assigns each event a number
- Experiment:
  - Throw two dice
  - Define a random variable X that assigns each result of this experiment the number represented by the dice.
- This induces a new sample space:

  {2,3,4,5,6,7,8,9,10,11,12}

Image: http://www.freefoto.com/images/11/12/11_12_64---Dice_web.jpg

Random Variable

- Original sample space:
  - 36 events
  - Each probability 1/36
- New sample space has other probabilities:
  - P(X=2) = P(X=12) = 1/36
  - P(X=3) = P(X=11) = 2/36
  - P(X=4) = P(X=10) = 3/36
  - P(X=5) = P(X=9) = 4/36
  - P(X=6) = P(X=8) = 5/36
  - P(X=7) = 6/36
- The sum of all probabilities equals 1

Image: http://www.freefoto.com/images/11/12/11_12_64---Dice_web.jpg

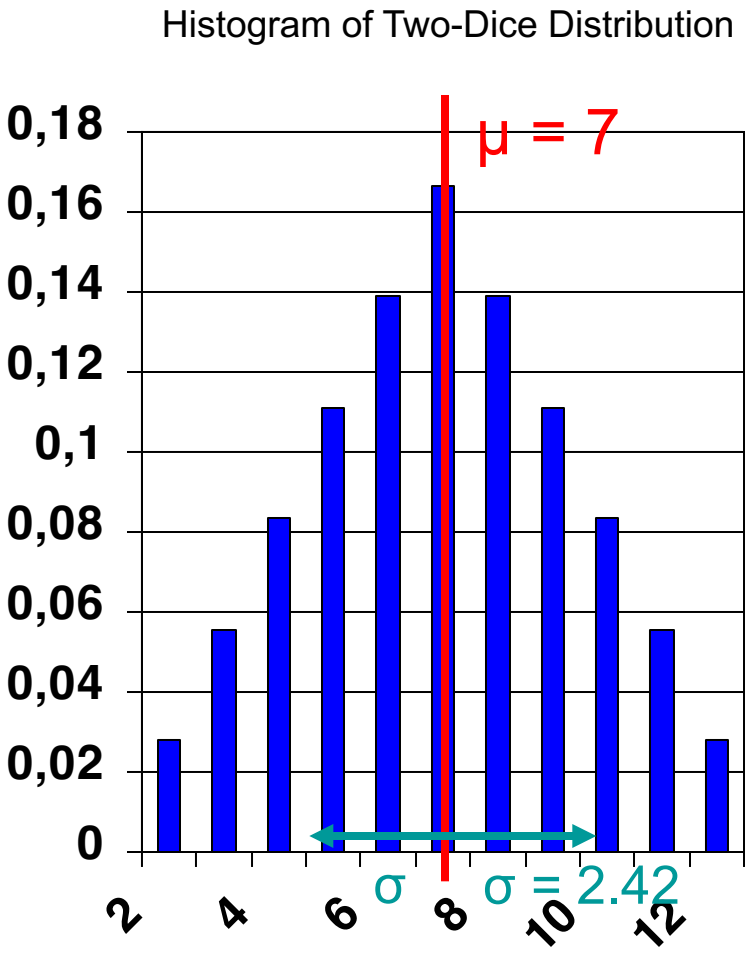
Probability Distribution

- The set of probabilities of a random variable is called probability distribution.
- It can be characterized by its expected value

$$E(X) = \mu = \sum_{i \in X} i \cdot P(X = i)$$

and its variance

$$Var(X) = \sigma^2 = \sum_{i \in X} P(X = i) \cdot (i - \mu)^2$$



Continuous Random Variable

- Some measurements imply continuous values, e.g. height and weight of a person
- Nonetheless subject to randomness
- A certain distribution of the measurement values exists
- A random variable that operates on continuous values is continuous itself

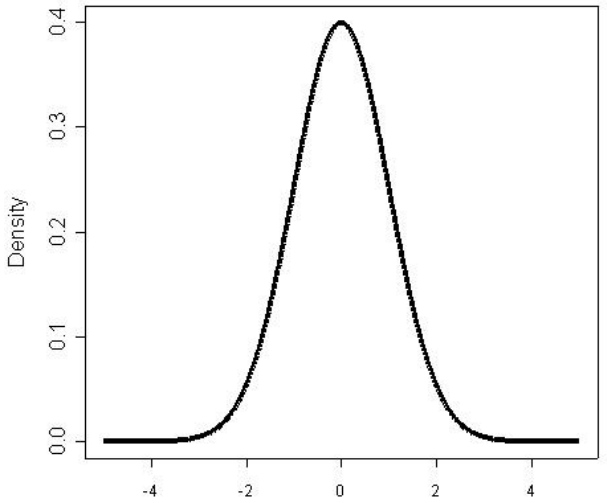
Probability Density

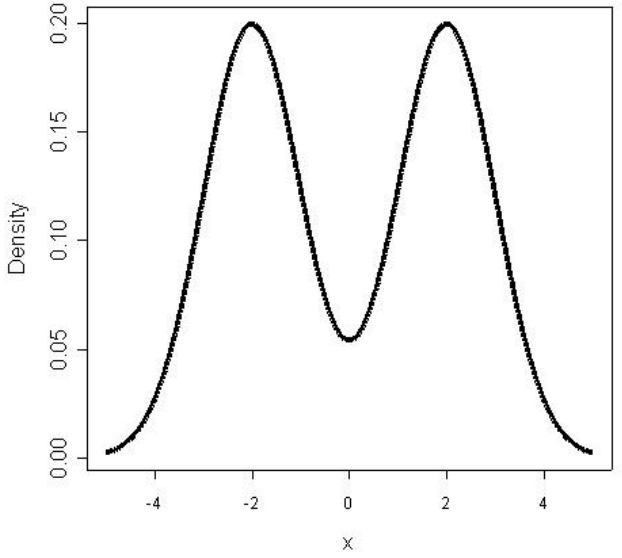
- The probability distribution of a continuous variable is given by its **probability density function**

$$P(a \le X \le b) = \int_a^b f(x)dx$$

$$P(X = a) = \int_a^a f(x)dx = 0$$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$



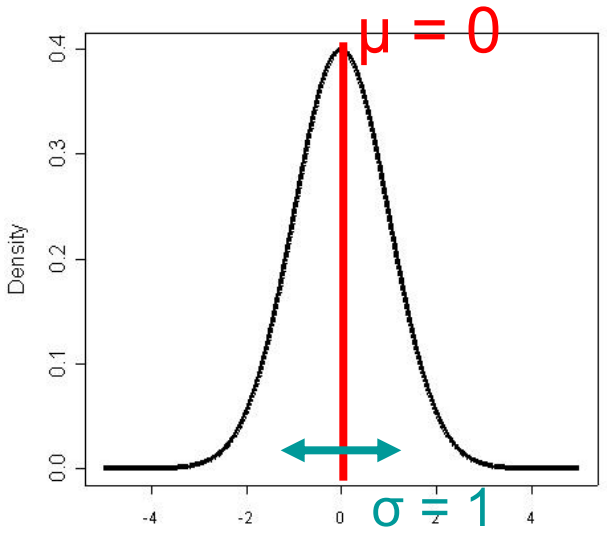


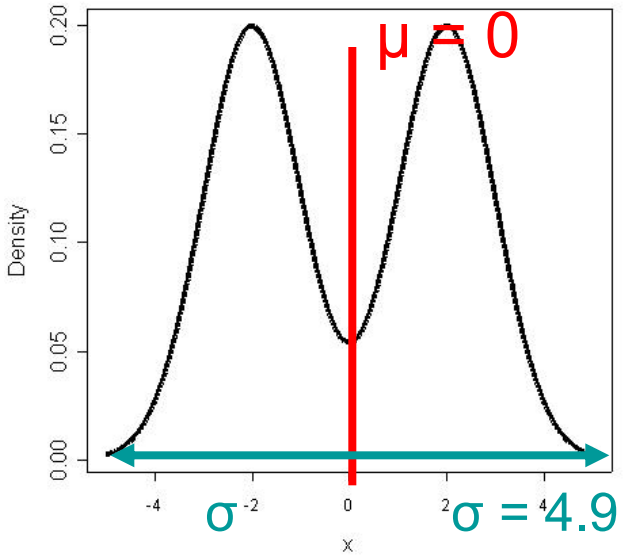
Mean and Variance of Continuous Distributions

- For continuous distributions, the expected value and variance are:

$$E(X) = \mu = \int_{-\infty}^{\infty} f(x) \cdot x dx$$

$$Var(X) = \sigma^2 = \int_{-\infty}^{\infty} f(x) \cdot (x - \mu)^2 dx$$





Sample Mean and Variance

- $\mu_{STUDY} = 6$
- $\sigma^2_{STUDY} = 10$
- $\mu_{SLEEP} = 8$
- $\sigma^2_{SLEEP} = 2.5$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \mu)^2$$

# Study Hours	# Sleep Hours
2	10
4	9
6	8
8	7
10	6

Important Distributions: The Binomial Distribution

- A soccer player has a 0.25 chance of scoring a goal. If he attempts 4 shots in a match, how can we model the probability that he scores 0, 1, 2, 3 or 4 goals?

Image: http://www.sxc.hu/photo/1155825

Important Distributions: The Binomial Distribution

$$P(0\ \text{goals}) = 1 \cdot (1-0.25)^4 \approx 0.32$$

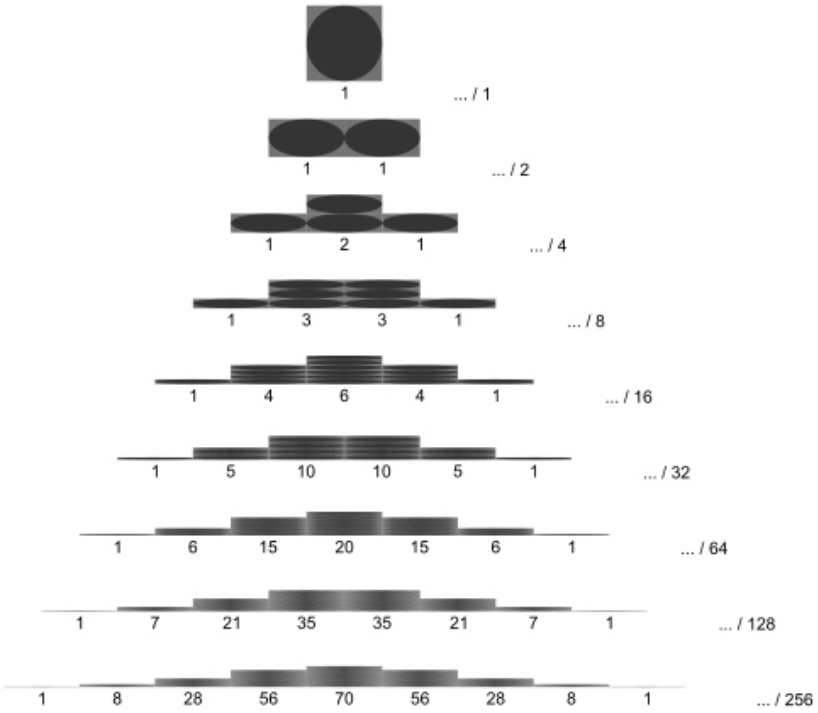
$$P(1\ \text{goal}) = 4 \cdot 0.25 \cdot (1-0.25)^3 \approx 0.42$$

$$P(2\ \text{goals}) = 6 \cdot 0.25^2 \cdot (1-0.25)^2 \approx 0.21$$

$$P(3\ \text{goals}) = 4 \cdot 0.25^1 \cdot (1-0.25)^3 \approx 0.05$$

$$P(4\ \text{goals}) = 1 \cdot 0.25^4 \approx 0.004$$

Important Distributions: The Binomial Distribution



- The probability of getting k "successes" in n trials with success probability p

$$f(k,n,p) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\mu = n \cdot p$$

$$\sigma^2 = n \cdot p \cdot (1-p)$$

Important Distributions: The Normal Distribution

- The normal distribution is used for modeling measurement errors
- It is a simple model for the variability produced by the complex interplay of several factors

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

Bean machine by Francis Galton

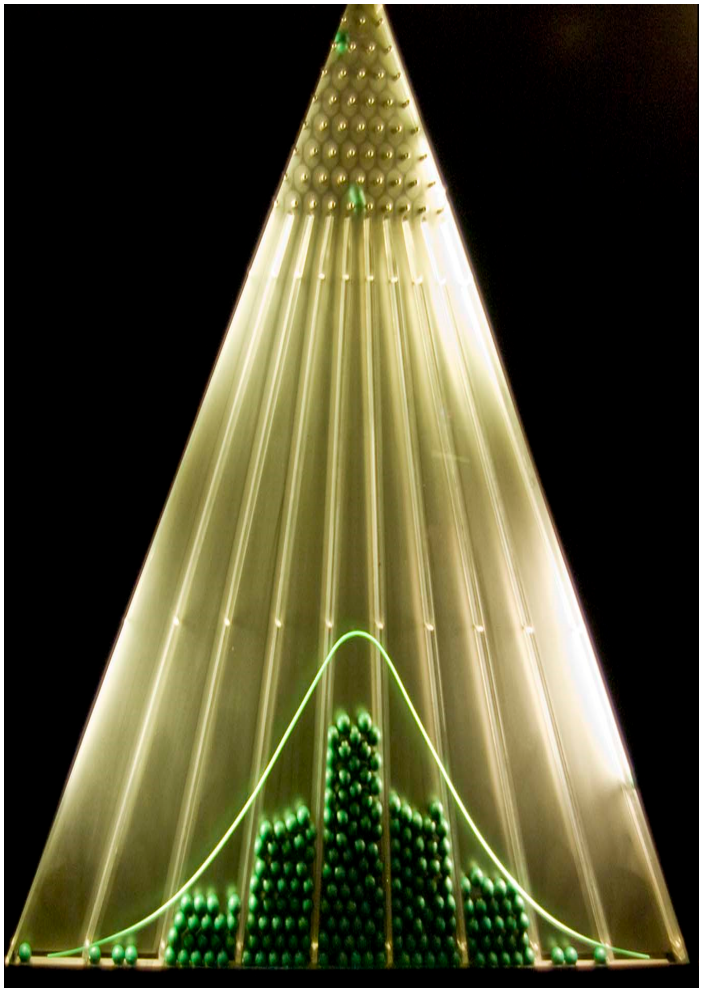


Image: http://en.wikipedia.org/wiki/Bean_machine

Conditional Probability

- What is the probability of being HIV-infected in this sample?
- What is the probability of being HIV-infected, given that the test was reactive?
- What is the probability of being healthy, given that the test was reactive?

	HIV infection	Healthy	Total
Reactive test	99	3	102
Non-reactive test	1	897	898
Total	100	900	1000

Conditional Probability

Conditional probability formula:

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

If you multiply by P(B):

$$P(A \cap B) = P(A \mid B) \cdot P(B)$$

If A and B are independent:

$$P(A \mid B) = P(A)$$

$$P(A \cap B) = P(A) \cdot P(B)$$

Joint Probability Distribution

	HIV infection	Healthy	Total
Reactive test	0.099	0.003	0.102
Non-reactive test	0.001	0.897	0.898
Total	0.100	0.900	1

Two variables with two possible outcomes each

Statistical Independence

$$P(A \mid B) = P(A)$$

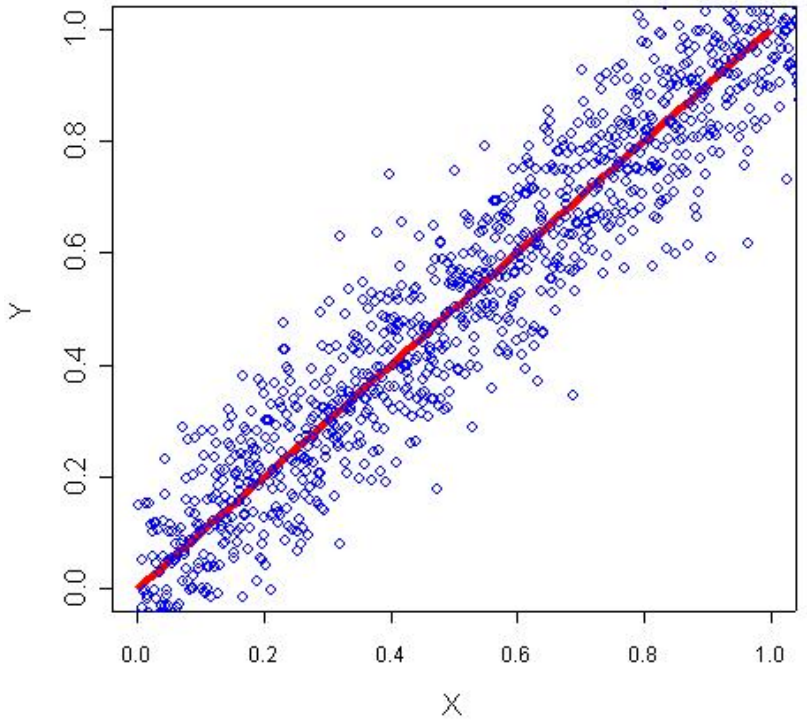
$$P(A \cap B) = P(A) \cdot P(B)$$

	Tasty	Bad	Total
Red Apple	50	50	100
Yellow Apple	50	50	100
Total	100	100	200

The variables apple color and taste are statistically independent

Correlation

- A statistical relationship between two random variables or sets of data is called **dependence**.
- A **correlation** in a statistical relationship involving dependence.
- Correlation is a prerequisite for causation but does not imply it.



Pearson Correlation Coefficient

- A number between -1 and 1 that measures linear dependency between variables on the interval scale

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \cong \frac{\sum_i (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_i (x_i - \mu_x)^2 \sum_i (y_i - \mu_y)^2}}$$

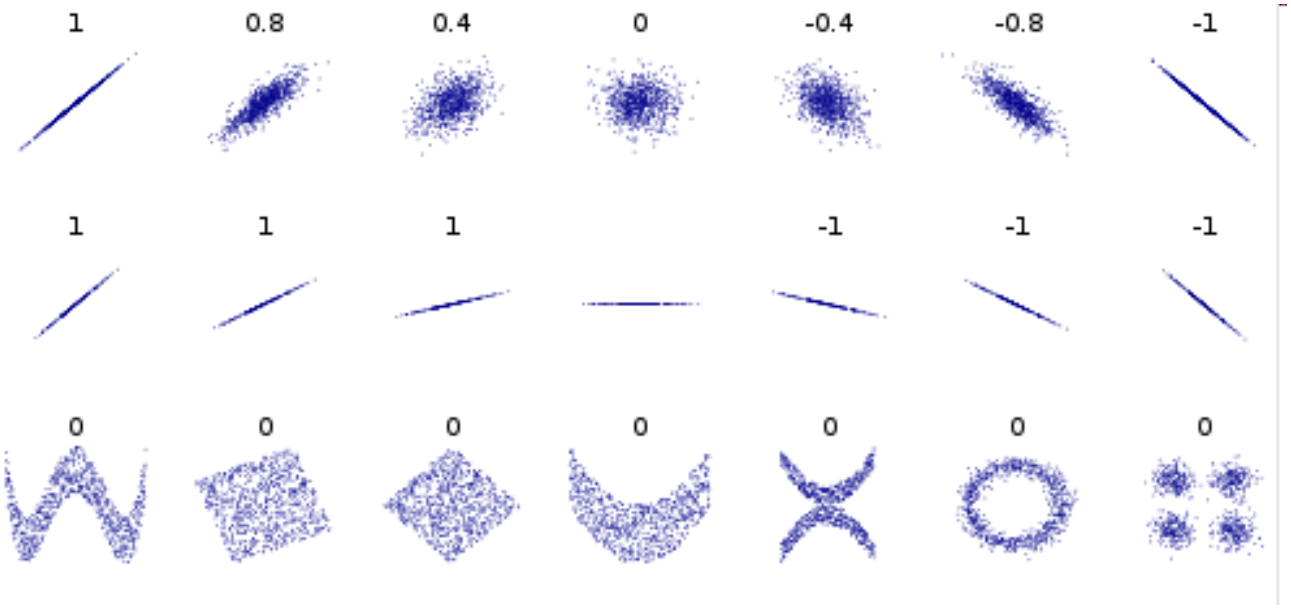
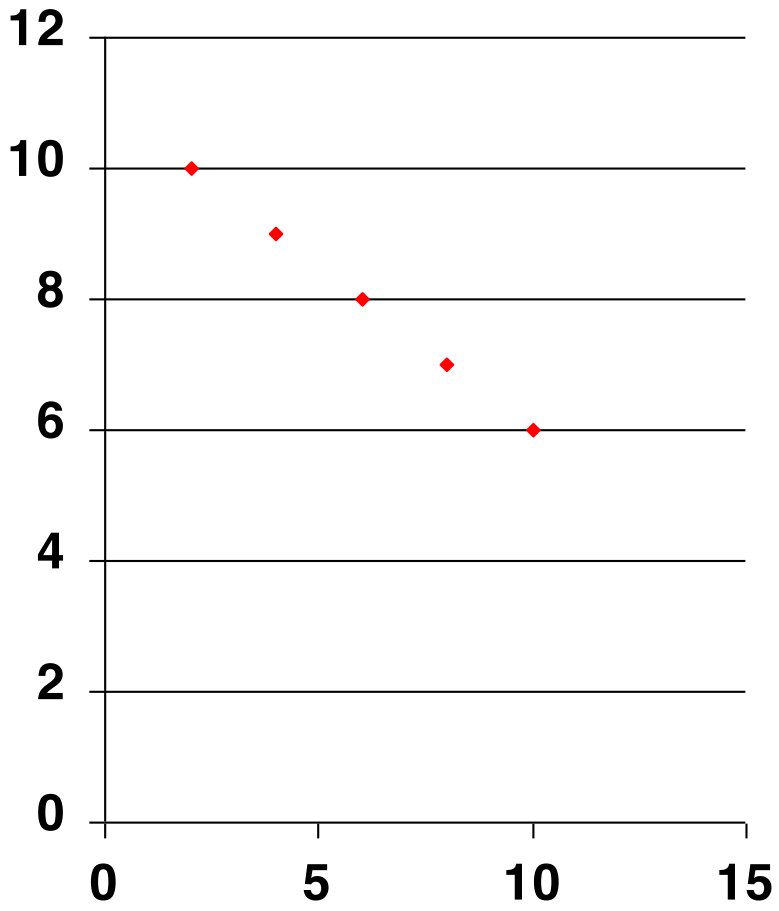


Image: http://en.wikipedia.org/wiki/Correlation

Pearson Correlation Coefficient

# Study Hours	# Sleep Hours
2	10
4	9
6	8
8	7
10	6



Pearson Correlation Coefficient

X	Y	$(X-\mu_X)$	$(Y-\mu_Y)$	$(X-\mu_X)(Y-\mu_Y)$	$(X-\mu_X)^2$	$(Y-\mu_Y)^2$
2	10	-4	2	-8	16	4
4	9	-2	1	-2	4	1
6	8	0	0	0	0	0
8	7	2	-1	-2	4	1
10	6	4	-2	-8	16	4
30	40	0	0	**-20**	**40**	**10**

Pearson Correlation Coefficient

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \cong \frac{\sum_i (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_i (x_i - \mu_x)^2 \sum_i (y_i - \mu_y)^2}} =$$

$$\frac{-20}{\sqrt{40 \cdot 10}} = \frac{-20}{\sqrt{400}} = \frac{-20}{20} = -1$$

Statistical Test Theory

- During experimentation, measurements are performed at a series of conditions
- Since our measurements are subject to randomness, we would like to know how likely it is for a difference to be due to randomness

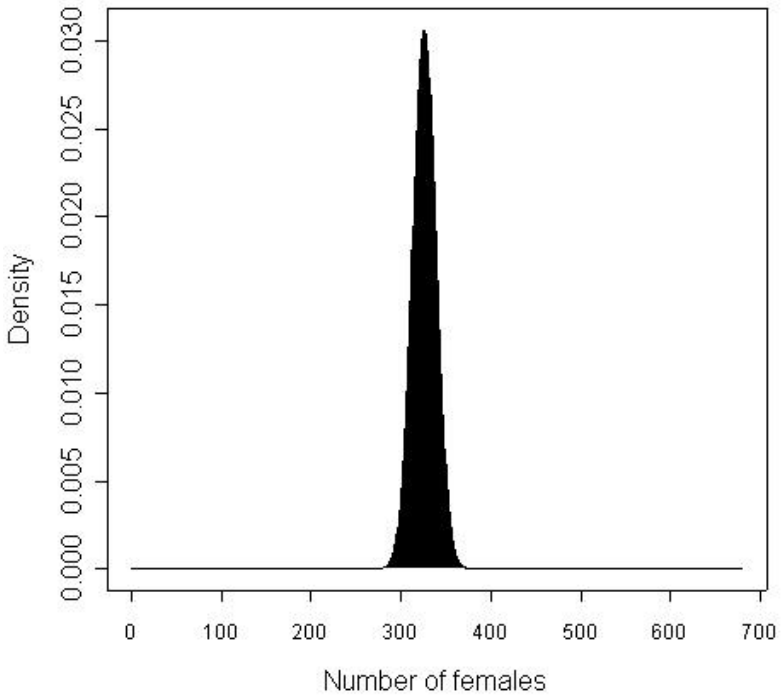
Statistical Test Theory

- To assess the possibility of randomness, we need a measure of what happens at random
- Example:
  Over the years, it has been empirically determined that approx. 48% of all newborns are females. However, a survey in 3 hospitals reported 680 newborns of which 51% were females. In this context, is 3% a strong, significant difference?
- How do we model what happens at random in this case?

Statistical Test Theory

- We could use a binomial distribution with n = 680 and p = 0.48!!!!
- With it, we can assess how probable it is, to get 347 female births by chance.
- f(347,680,0.48)= 0.008774934
- Is this number comparable?

$$f(k,n,p) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$



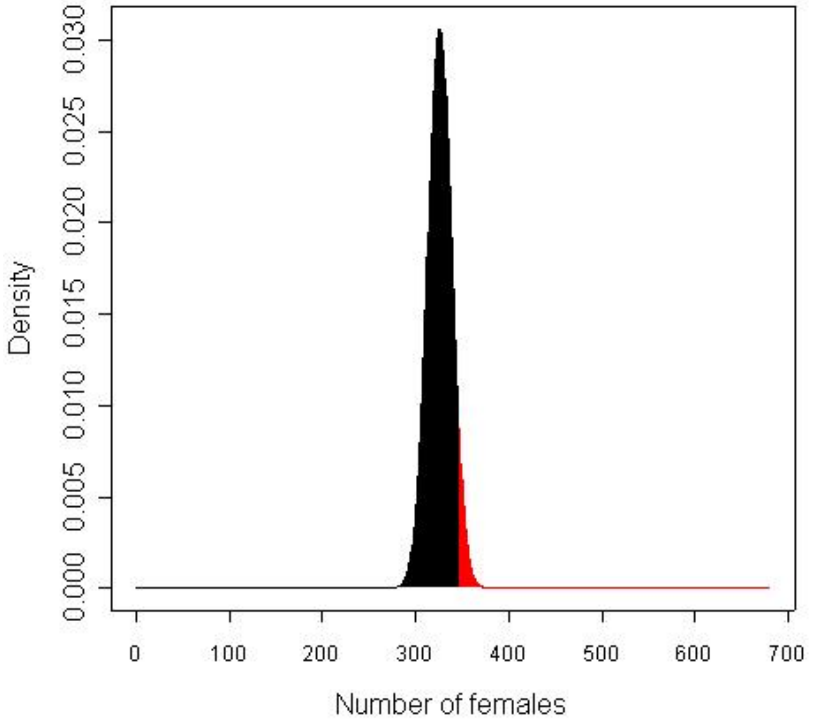
Statistical Test Theory

- What is the probability of getting 347 *or more* female births by chance?

$$P(X \geq 347) = \sum_{i=347}^{680} f(i,680,0.48) =$$

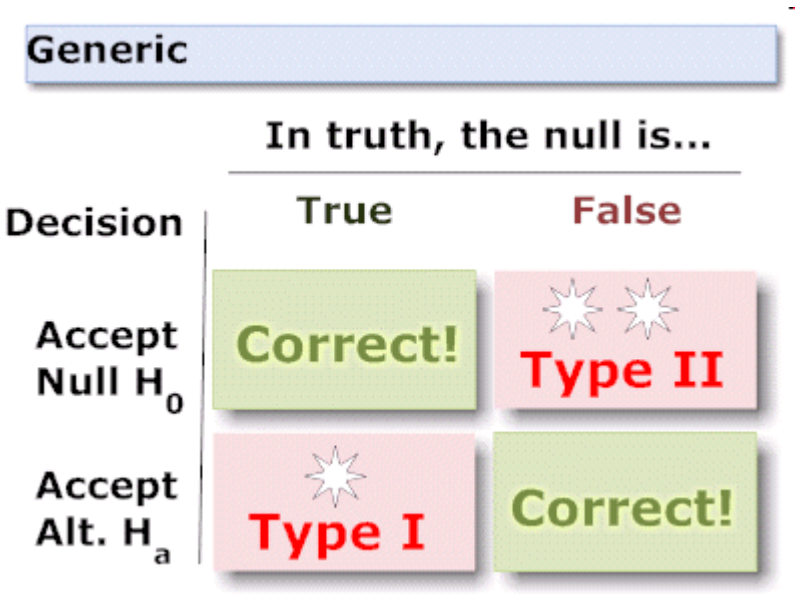
$$0.05271453$$

- This number is comparable.



Statistical Test Theory

- A statistical test helps us decide between two hypotheses:
  - H_0: Differences are due to randomness
  - H_1: Differences are significant
- There is a risk of error when using the test
- α errors happen, when we erroneously chose H_1
  - This error is bound by the α risk or α error probability
  - More commonly called significance level
- β errors happen, we erroneously chose H_0



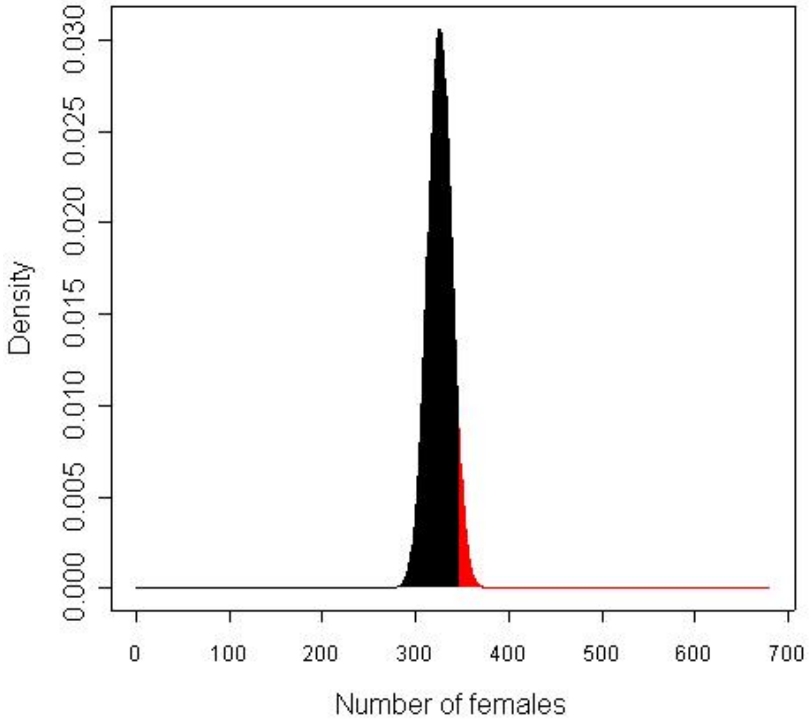
Statistical Test Theory

- What is the probability of getting 347 *or more* female births by chance?

$$P(X \geq 347) = \sum_{i=347}^{680} f(i,680,0.48) =$$

$$0.05271453$$

- We choose H_0 if the p-value is above our significance level, otherwise, we choose H_1



Fisher's Exact Test

- Measure of correlation between variables with categorical values
- Example:

  We draw a random sample of 24 teenagers and we record their gender and whether they are dieting.

  In this random sample, are women more propense to diet than men?

	Men	Women	Total
Dieting	1	9	10
Not Dieting	11	3	14
Total	12	12	24

Fisher's Exact Test

- Fisher's exact test gives us a measure of association by using a hypergeometric distribution.

	x_1	x_2	Total
y_1	a	b	a+b
y_2	c	d	c+d
Total	a+c	b+d	a+b+c+d = n

Number of ways we can get a out ouf a+b

Number of ways we could get c out of c+d

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

Number of ways we could get a+c out off all

Fisher's Exact Test

- What is the probability of getting such an association or a more extreme one by chance?
- p=0.002759

	Men	Women	Total
Dieting	1	9	10
Not Dieting	11	3	14
Total	12	12	24

Statistical Learning

- Use statistical methods to learn from data
- Examples:
  - Predict whether a patient who had a heart attack will have a second one
  - Identify the risk factors for prostate cancer
  - Estimate the amount of sugar in a person's blood from its infrared absorption spectrum
  - Analyse the mutations in an HIV sequence to propose a suitable therapy

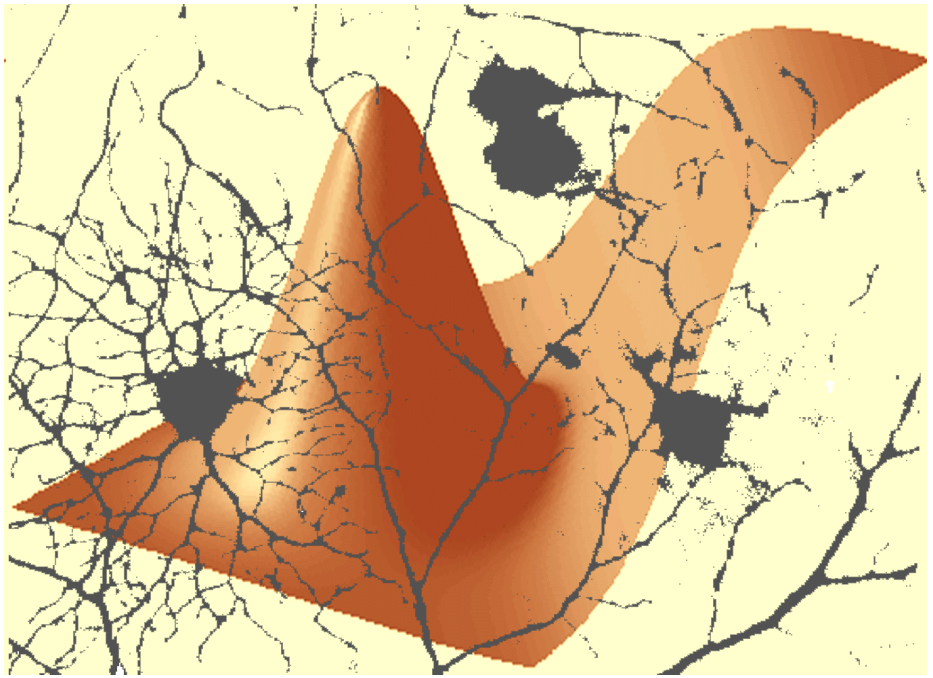


Image: http://www.netral.com/formation/formation-espci-en.html

Statistical Learning Algorithms

- Statistical learning algorithms (SLA) have the capability of learning from data
- Need a training set of data with
  - Outcomes
  - Feature measurements
- Use data to build a prediction model
- Use model to predict new, unseen data

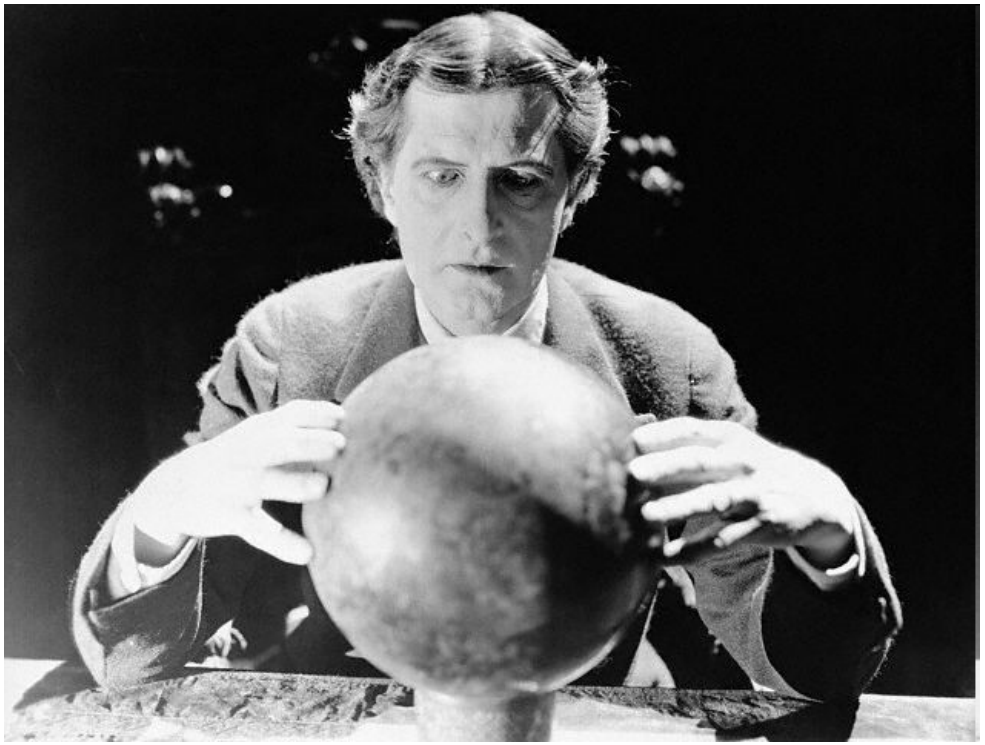
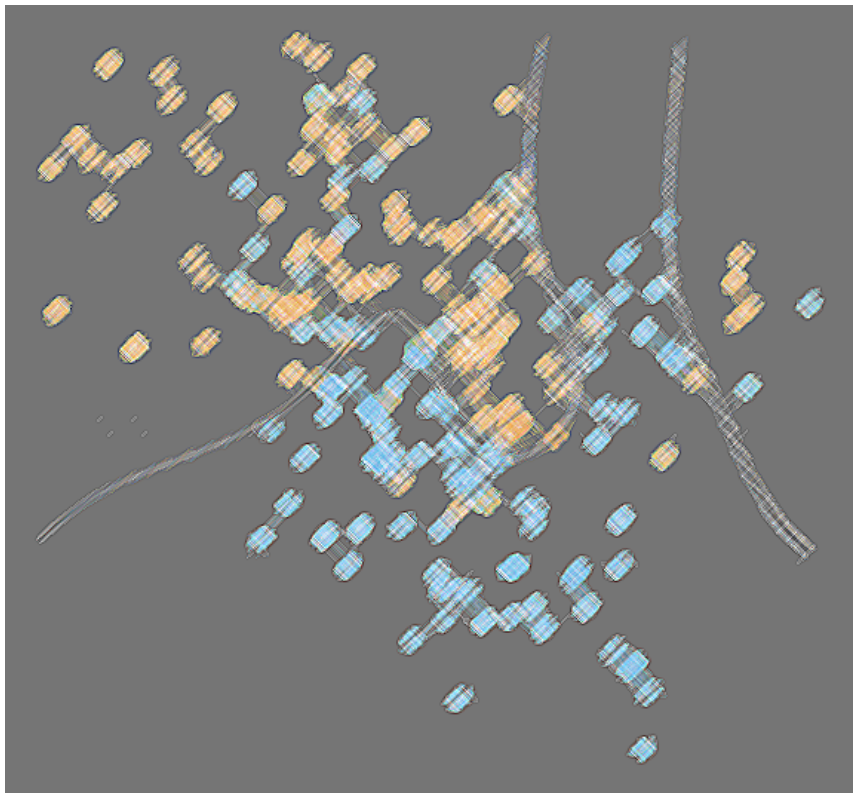


Image: http://sidschwab.blogspot.com/2011/09/prediction.html

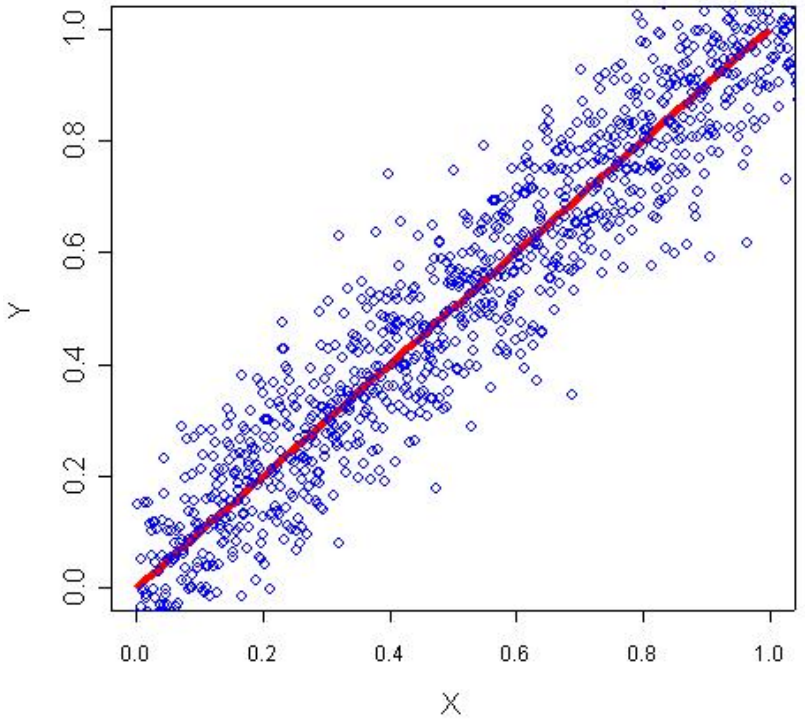
Training Set

- SLA learn from their training set by example
- Recognize trends and infer rules
- Outcomes can be
  - Quantitative: numbers
  - Qualitative: categories
- Estimating numbers: regression
- Categorizing: classification



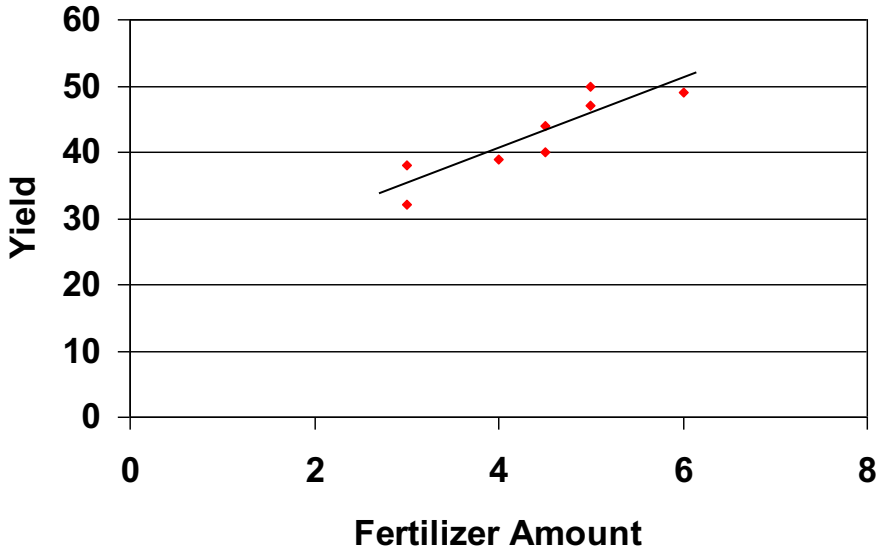
Linear Regression

- Measurement with noise
- Linear relationship between measurements
- Dependent and independent variables
- Represent point cloud with a line
- y = mx + b



Example

- In an agricultural experiment, different quantities of fertilizer are tested. The resulting yield is measured



Index	Fertilizer	Yield
1	3.0	32
2	3.0	38
3	4.0	39
4	4.5	40
5	4.5	44
6	5.0	47
7	5.0	50
8	6.0	49

Fitting Linear Models

- Out of the data, we need to fit the m and the b in y=mx+b
- Least squares method

Index	Fertilizer	Yield
1	3.0	32
2	3.0	38
3	4.0	39
4	4.5	40
5	4.5	44
6	5.0	47
7	5.0	50
8	6.0	49

$$SS_{RES}(m,b) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (mx_i + b))^2$$

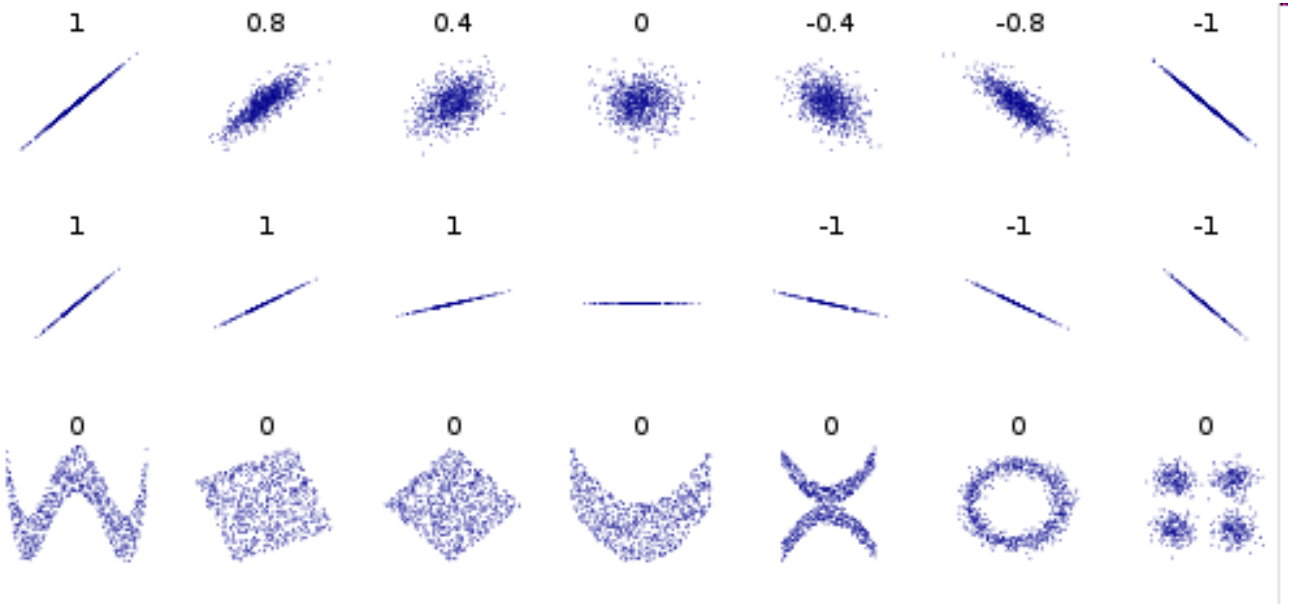
$$\hat{m} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{b} = \bar{y} - \hat{m}\bar{x}$$

m = 5.4, b = 18.7

Assessing Linear Models

- Linear models perform the better, the more the point cloud resembles a line



Assessing Linear Models

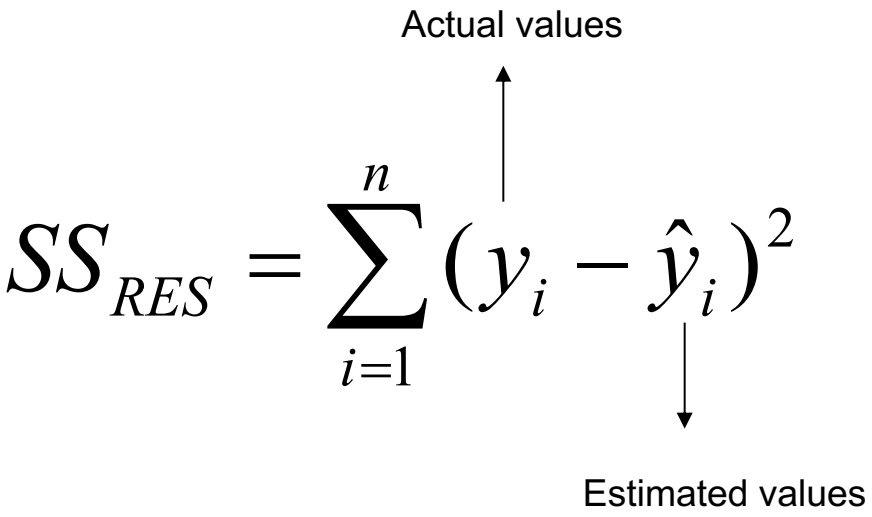
- Four numbers help us assess the goodness of fit
  - The residual sum of squares
  - The total sum of squares
  - The regression sum of squares
  - The coefficient of determination

$$SS_{RES} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad SS_{REG} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SS_{TOTAL} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad r^2 = \frac{SS_{REG}}{SS_{TOTAL}}$$

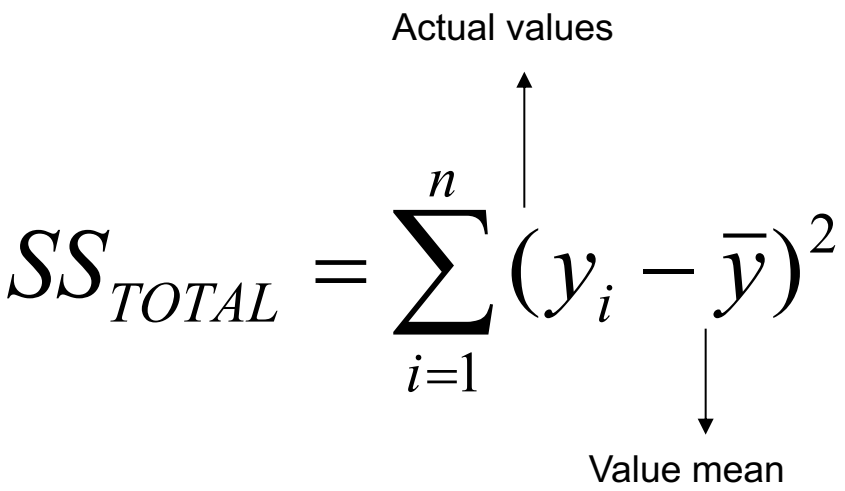
The Residual Sum of Squares

- Tells us how well the minimization went. The smaller the value, the better



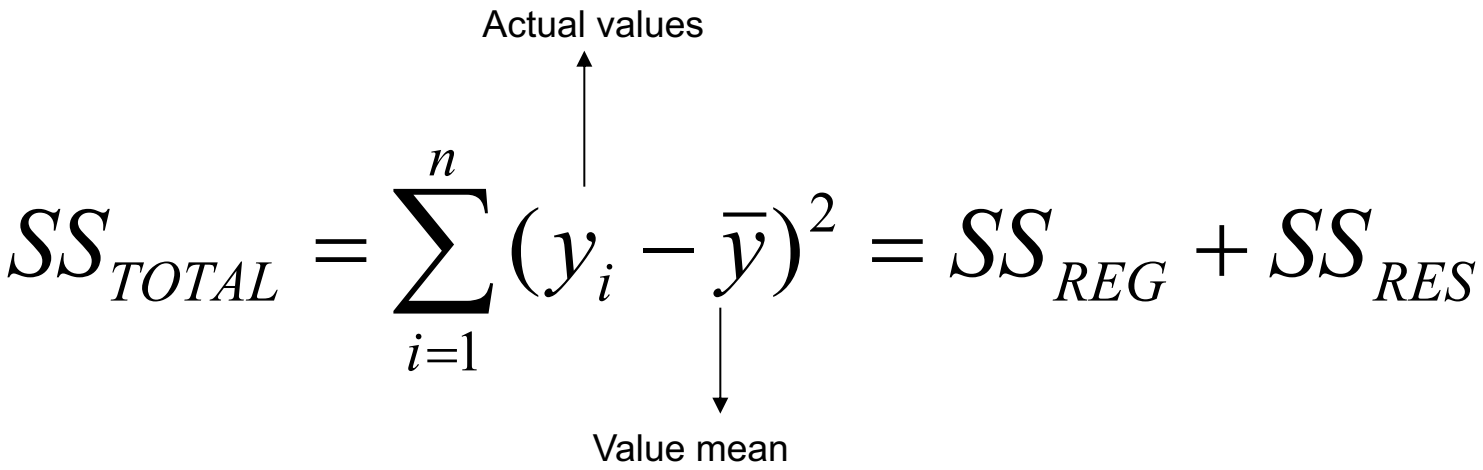
The Total Sum of Squares

- This value tells us how much the actual values deviate from their mean. This is a measure of variability of the training data



The Total Sum of Squares

- This number tells us how much the estimated values deviate from their means. This is a measure of variability of the estimated values



The Coefficient of Determination

- The coefficient of determination is a number between 0 and 1
- It tells us how well does the model explain the variability in the training data; the larger the number the better

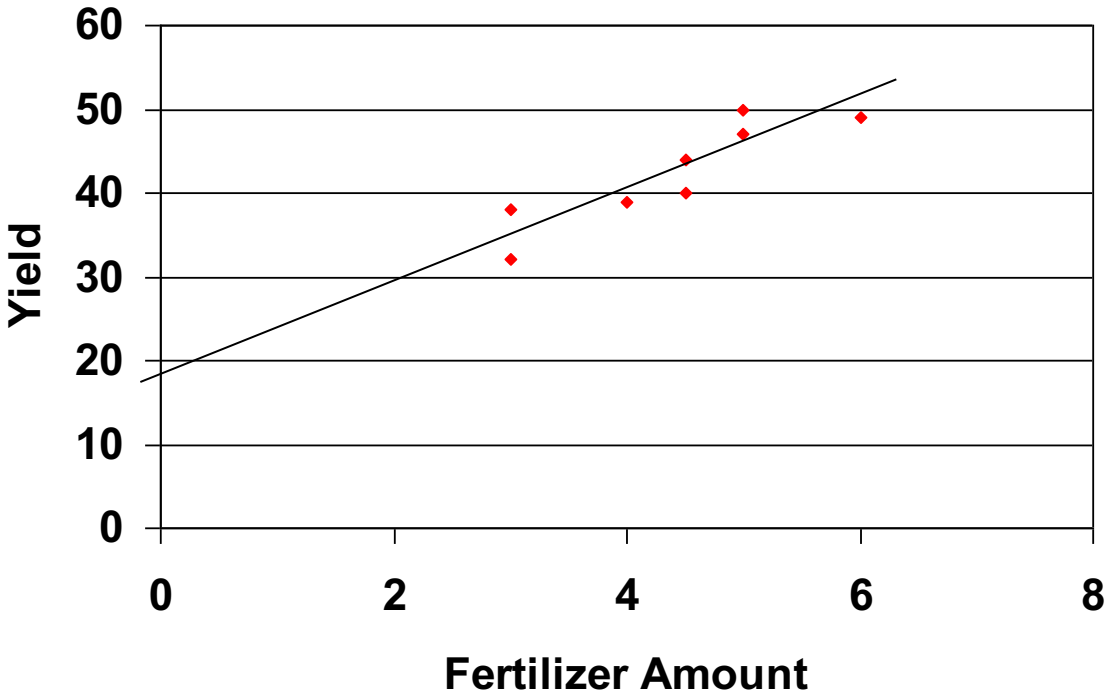
$$r^2 = \frac{SS_{REG}}{SS_{TOTAL}}$$

← Variability of the estimated values

← Variability of the data

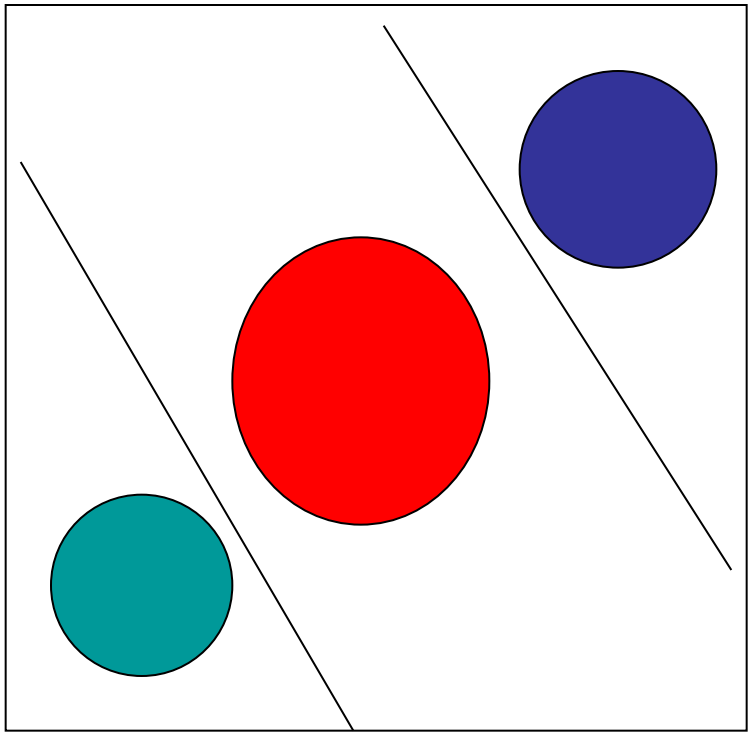
Our Fertilizer Example

- m = 5.4
- b = 18.7
- SS_{RES} = 54
- SS_{REG} = 214
- SS_{TOTAL} = 268
- r^2 = 0.79



Logistic Regression

- Linear method for classification: linear class boundaries are constructed
- Uses conditional probabilities to classify according to a score between 0 and 1
- All scores sum up to one, however, the scores are not *true* probabilities



Logistic Regression

- The model tries to find a linear boundary between two classes by setting the probability ratio of the two classes equal to a line.

Let $\{1,2,...,K\}$ be a set of classes and x a measurement.

$$\log \frac{P(G=1 \mid X=x)}{P(G=K \mid X=x)} = m_1 x + b_1$$

$$\log \frac{P(G=2 \mid X=x)}{P(G=K \mid X=x)} = m_2 x + b_2$$

$$\vdots$$

$$\log \frac{P(G=K-1 \mid X=x)}{P(G=K \mid X=x)} = m_{k-1} x + b_{k-1}$$

Logistic Regression

Solving for

$$P(G = k \mid X = x)$$

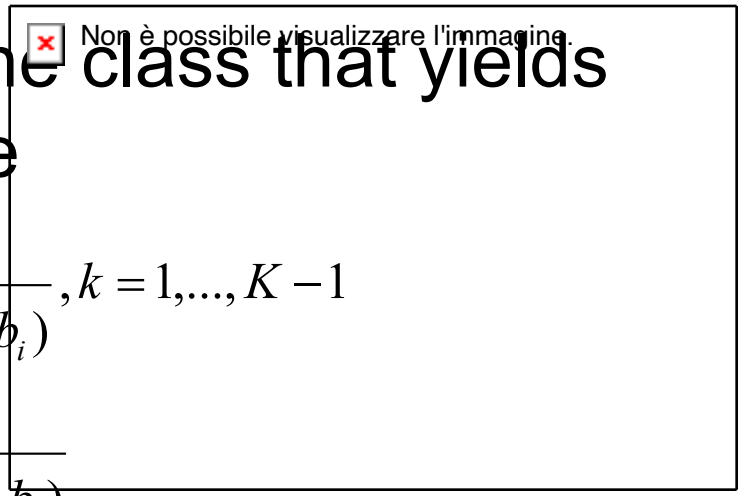
we get

$$P(G = k \mid X = x) = \frac{\exp(m_k x + b_k)}{1 + \sum_{i=1}^{K-1} \exp(m_i x + b_i)}, k = 1, \ldots, K-1$$

$$P(G = K \mid X = x) = \frac{1}{1 + \sum_{i=1}^{K-1} \exp(m_i x + b_i)}$$

Logistic Regression

- A new point x is classified to the class that yields the maximum probability score



$$P(G=k \mid X=x)=\frac{\exp(m_k x+b_k)}{1+\sum_{i=1}^{K}\exp(m_i x+b_i)}, k=1,\ldots,K-1$$

$$P(G=K \mid X=x)=\frac{1}{1+\sum_{i=1}^{K}\exp(m_i x+b_i)}$$

- Finding out the m's and the b's implies a numerical minimization
- I just want you to understand the principle

Summary of Logistic Regression

- Linear classification method, i.e. linear boundaries
- Works with conditional probabilities
- Yields a probabilistic score for each class
- The accuracy of a classifier is measured by the misclassification error

$$\frac{\#\,\text{incorrectly classified samples}}{\#\,\text{samples}}$$

Survival Analysis

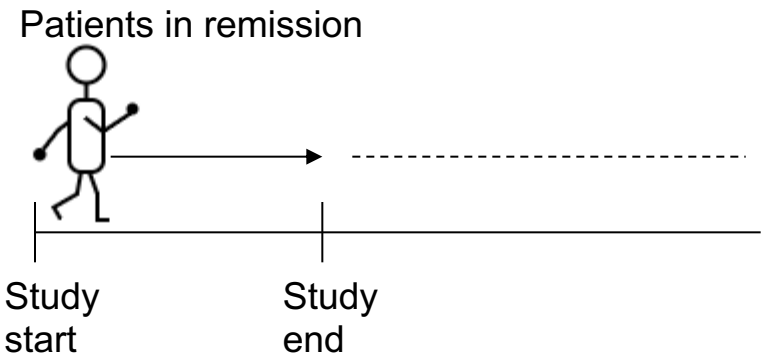
- We want to analyze and predict the time until a certain event occurs
  E.g.
  - Death
  - Disease
  - Relapse
  - Recovery
- The time until the event occurs is called **survival time** and the event itself can be referred to as **failure**

Image: http://personal.psu.edu/cdc5064/Tools/.resource/speed/stopwatch1.png

For the survival-analysis part of the course, the following book was used: Kleinbaum D. and Klein M. Survival Analysis. A Self-Learning-Text. 2nd Edition. 2005. Springer.

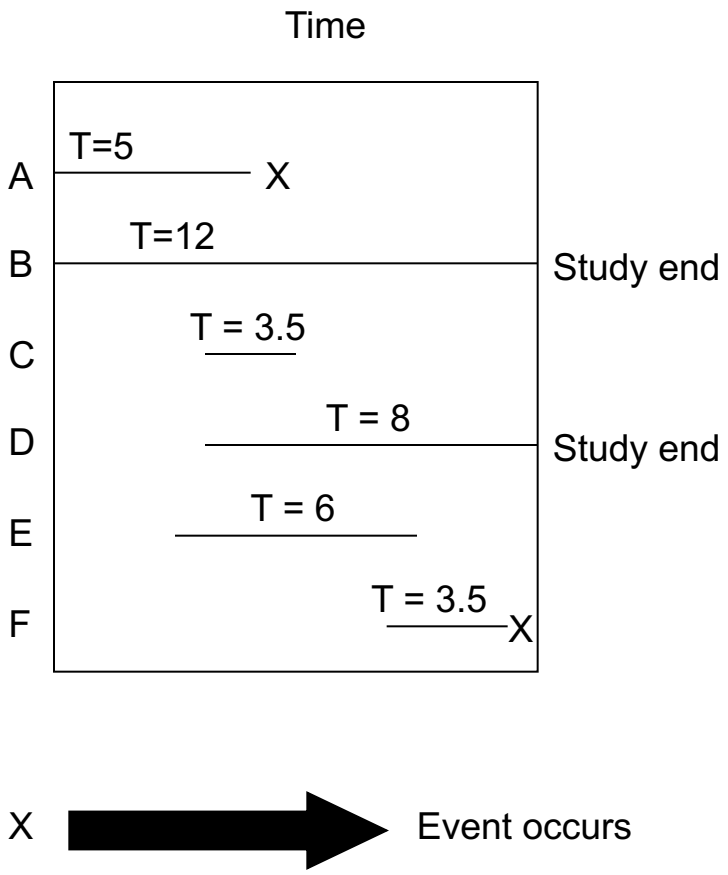
Censoring

- In survival analysis we deal with the problem of censoring
- We have some information, but we **don't know the survival time exactly**
- E.g. Leukemia patients who are followed until they go out of remission. If the study ends while a patient is still in remission, then we say that his survival time is censored



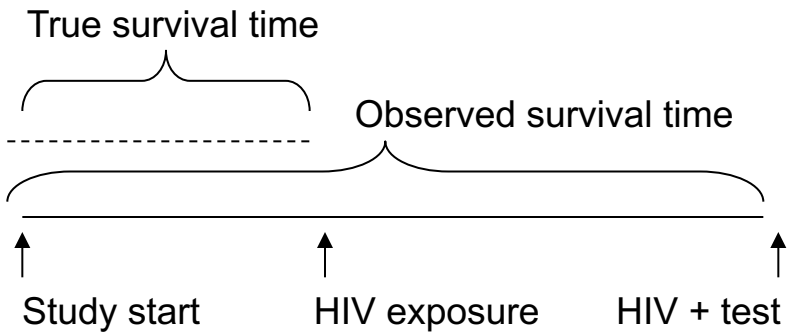
Censoring is Common

- Reasons for censoring are:
  - No event has happened at the time the study ends
  - Loss to follow-up
  - Withdrawal from study
- To the right:
  - We observe an event in patients A and F
  - Patients B, C, D, and E are censored



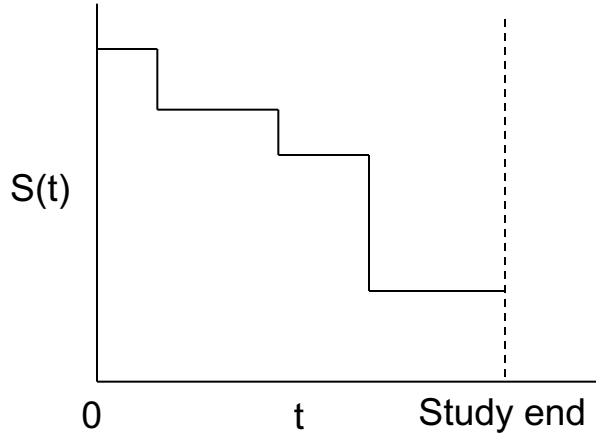
Right- and Left-Censored Data

- In the previous example, the patients were right-censored
- The patients' survival time is greater or equal to the observed survival time
- In left-censored data, the situation is mirrored. Real survival time is less or equal to the observed survival time



Survival Analysis Terminology

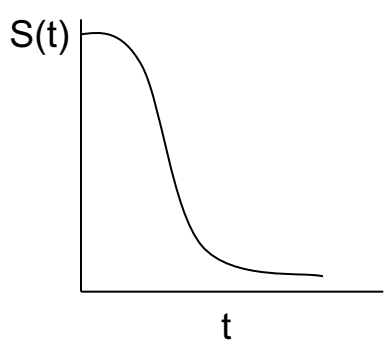
- Random variable T: survival time of a person
- Survivor function **S(t):** probability that a certain person survives for at least a specified time **t**
- Hazard function **h(t):** instantaneous risk of failure at time t, given that the patient has survived up to that time point

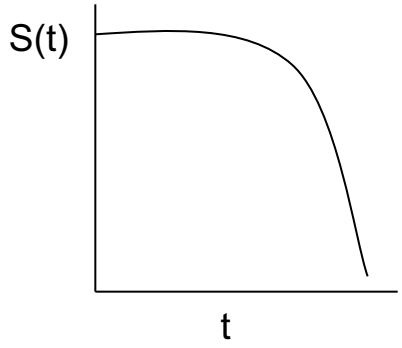


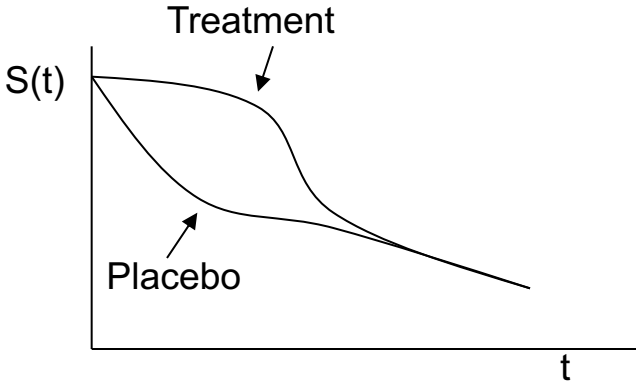
$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t}$$

Goals of Survival Analysis

1. Estimate and interpret survivor and/or hazard functions from survival data
2. To compare survivor and/or hazard functions
3. Assess the relationship of explanatory variables to survival time. This goal requires additional mathematical modelling.





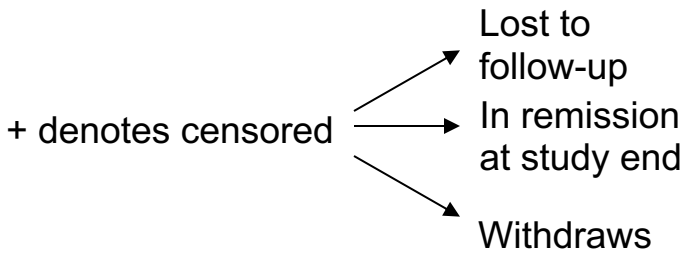


Example: Leukemia Patients

Remission times (weeks) for two groups of leukemia patients

Group 1 (Treatment) n = 21	Group 2 (Placebo) n = 21
6,6,6,7,10	1,1,2,2,3
13,16,22,23	4,4,5,5,
6+,9+,10+,11+	8,8,8,8,
17+,19+,20+	11,11,12,12
25+,32+,32+,	15,17,22,23
34+,35+	

	# failed	# censored	Total
Group 1	9	12	21
Group 2	21	0	21



Kaplan-Meier Estimates

- The Kaplan-Meier method can be used to estimate the survival function S(t) from survival data
- Computations are based on conditional probabilities

$$\text{Fraction at } t_j : P(T > t_j \mid T \geq t_j)$$

t: time point
n: number of patients surviving at time point
m: number of patients who failed at time point
q: number of censored patients at time point

Group 1 (treatment)

t_j	n_j	m_j	q_j	$S(t_j)$
0	21	0	0	
6	21	3	1	.8571
7	17	1	1	.8067
10	15	1	2	.7529
13	12	1	0	.6902
16	11	1	3	.6275
22	7	1	0	.5378
23	6	1	5	.4482

Kaplan Meier Estimates

$t_{(j)}$	n_j	m_j	q_j	$S(t_j)$
0	21	0	0	
1	21	2	0	
2	19	2	0	
3	17	1	0	
4	16	2	0	
5	14	2	0	
8	12	4	0	
11	8	2	0	
12	6	2	0	
15	4	1	0	
17	3	1	0	
22	2	1	0	
23	1	1	0	

Kaplan-Meier Curve