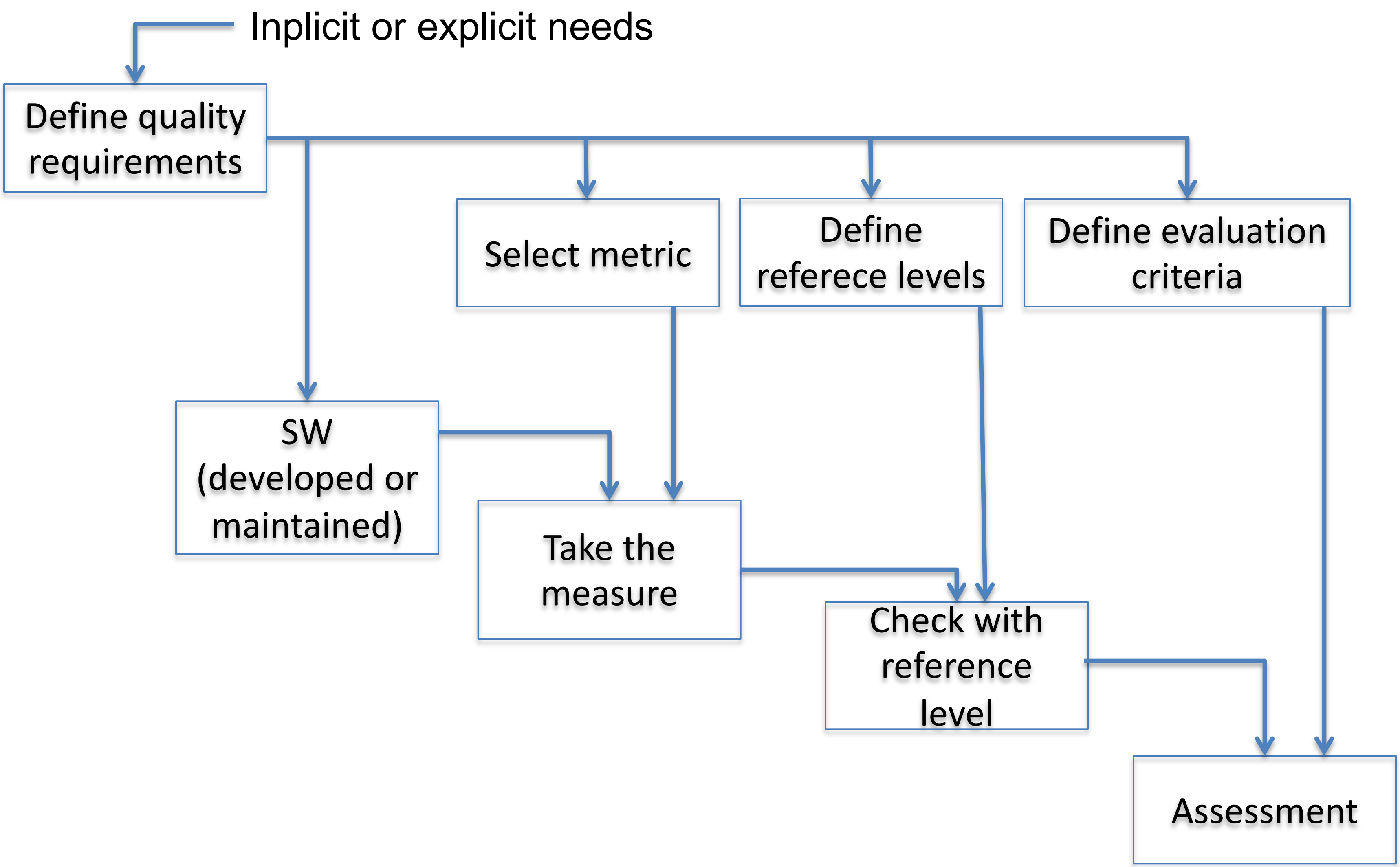
Introduction to the theory of measure and some concepts of descriptive and inferential statistics

(Some slides taken from slideshare.net)

Why do we measure?

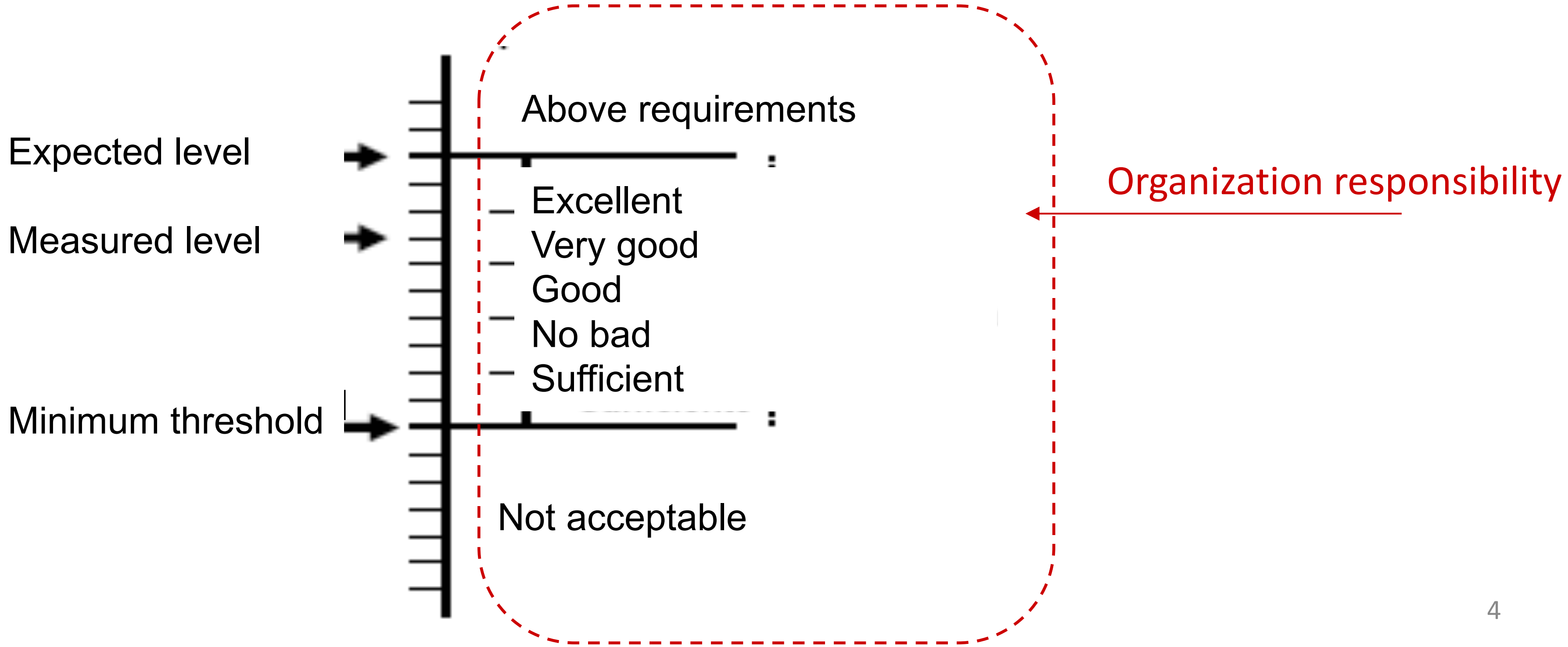
- We cannot govern what we cannot measure (De Marco, 1982)
- Measures, in the field of Software Engineering, are meant for:
  - Verifying how far quality parameters are from reference values
  - Identifying deviations from temporal and resource allocation planning
  - Identifying productivity indicators
  - **Validating the effect of strategies aimed at improving the development process (quality, productivity, planning, cost control)**
- We measure for monitoring and making decisions

Pragmatically



Measurement process

- Rating: definition of reference levels
- Metrics provide quantitative values which do not inherently correspond to quality judgments
- We have to map quantitative data to a qualitative scale



Basics of measure theory

Measurement Scales

- We consider five measurement scales:

- Nominal Scale
- Ordinal Scale
- Interval Scale
- Ratio Scale
- Absolute Scale

Nominal Scale

- Nominal scale classifies persons or objects into two or more categories
- Members of a category have a common set of characteristics, and each member may only belong to one category
- Other names: categorical, discontinuous, dichotomous (only two categories)

Nominal Scale

- A pre-defined non ordered set of distinct values
  - E.g., possible types of programming errors (syntactical, semantical, etc.) **without** defining an order of worseness
  - Possible operators {= , !=}
  - If we use this scale the average value makes sense only if we want to check the **frequency** by which certain measures fall into certain categories

Ordinal Scale

- Ordinal values allow us to rank the order of the items we measure in terms of which has less and which has more of the quality represented by the variable, but still they do **not allow** us to say "**how much more**“
- Possible operators {=, !=, >, <}
- Example: Student rank (A,B,C,D), CMMI levels

Ordinal Scale

- Ordinal scale **classifies** subjects and **rank** them in terms of how they possess the characteristic of interest
- Members are placed in terms of highest to lowest
- Ordinal scales do not state the difference between two adjacent ranks
- On some scales it is assumed that the distance between the ranks is equal but <u>we have to be careful</u> if we want to compute and use the <u>average</u>
- E.g., Likert scale of an experimental therapy
  - 1: recovered; 2: light complications; 3: medium complications; 4: hard complications; 5: death
  - 50 recovered and 50 death →
    - ... medium complications (!)

Interval Scale

- Interval scales allow us to rank the order of the items that are measured, and to quantify and compare the sizes of differences between them
- For example:
  - a student's exam performance: a score of 26 will be higher than 24 and lower than 28 and the difference between them is 2 points (**equal intervals**)
- Interval scales normally have an arbitrary minimum and maximum point
  - e.g. 17 to 31
- A score of 17 does not represent an absence of knowledge, nor does a score of 31 represent perfect knowledge

Interval Scale

- Interval scale requires a precise definition of the unit of measure to be used
- An example of interval scale is the temperature in C° or F°
- Integer or real values
- Possible operators {=, !=, <, >, +, -}
- The presence of an arbitral zero implies that you cannot **compare** the magnitude of two values, e.g., 80 °F is NOT four times hotter than 20° F, while you can say that there are 60° of difference

Ratio Scale

- Very similar to interval scale
- It has all the properties of interval scales, and it has an absolute (not arbitrary) zero point
- Height, weight, time, speed, and temperature in Kelvin degree are examples of ratio scales
- Possible operators {=, !=, <, >, +, -, *, /}
- For example, we can say that a person who runs a mile in 5 minutes is twice faster than a person who runs a mile in 10 minutes
- Ratio scales are often used in physical measurements (where absolute zero exists); conversely they are not often used in educational research and testing
- Ratio s. $\subset$ Interval s. $\subset$ Ordinal s. $\subset$ Nominal s.

Absolute scale

- It is a ratio scale and it ranges on non negative integers
- In this scale we count the actual occurrences of entities
  - E.g., Lines of Code (LOC) constituting a program
  - Number of errors
  - ...

Choosing a scale

- The choice of a scale depends on the attribute to be measured
- The chosen scale must correspond to a set of relations which are valid for the attribute
- For example, if it is not possible to determine if a product is reliable twice or three times of another we **have** to choose either an ordinal or a nominal scale

Synopsis of measure scales

Scale types	Admissible transformations	Basic empirical operation	Appropriate statistical indexes	Appropriate statistical tests	EXAMPLES
NOMINAL	any one-to-one transformation	equality test	Mode Frequency	not parametric	labeling classify
ORDINAL	$M(x) \geq M(Y)$ implies that $M'(x) \geq M'(Y)$	equality test and > <	Median Percentiles Spearman *r* Kendall W Kendall T	not parametric	preferences ordering di entità
INTERVALS	$M'=aM+n(a>0)$ [*positive, linear*]	equality test and > < + and -	Aritmetic mean Standard deviation Pearson correlation Multiple correlation	not parametric	Fahrenheit o Celsius date time
RATIO	$M'=aM$ $(a>0)$ [*similarity transformation*]	equality test and > < + and - * and /	Geometric mean Armonic mean Coefficiente di variazione Percentage variation Correlation index	not parametrico and parametric	time intervals Kelvin lenghts
ABSOLUTE	$M'=M$ [*identity*]				entity count

Types of measures

Ratio and Proportion

- Ratio: The result of a division between two values that come from two **different** and **disjoint** (logical) domains. The result is <u>typically</u> multiplied by 100 to avoid too little numbers. But a ratio is NOT a percentage
  - E.g. , (males/females)
  - It can have values above and below 1
  - E.g. , (lines of comments/LOC) * 100
- Proportion: The result of a division between two values where the dividend contributes to the divisor, e.g., a/(a + b)
  - E.g., number of satisfied users/number of users
  - It can assume values between 0 and 1
  - Often the divisor is composed of various elements for which we want to compute the proportions
    - E.g. a + b + c = N; a/N + b/N + c/N = 1
  - A fraction is a proportion between real values

Percentage

- A **proportion** or **fraction** expressed by normalizing the divisor to 100
  - E.g., defects in the requirements were 15%, in the design 25%, in the code 60%
  - The percentage must be used by indicating the involved values
  - The use of percentage must be avoided when values are less than 30-50
  - E.g., defects were 25% in the requirements, 15% in the design, and 60% in the code
  - E.g., defects in the project were 20, 5 in the requirements, 3 in the design, and 12 in the code

Rate

- Identifies a value associated with the dynamics of a phenomenon
- Typically it measures the change of a quantity (y) with respect to another quantity (x) on which it depends
- Usually x is the time
- E.g., crude rate of births in a certain year
  - (N/P)*k
    - N: births in the observed year
    - P: population (computed in the middle of the year)
    - K: a constant, typically 1000

Rate

- Elements of the divisor can become or produce elements of the dividend
  - This means there is a “risk exposure”
- It is **crude** because births are not produced by all the population but only by fertile women
- We can improve the previous formula by using P’ instead of P, where P’ is the number of women btw 15 and 44 years old (not crude)

$$(N/P')*k$$

Software Engineering example

- One year Defect rate = (D/OFE) *k dove
  - D=Defect observed within the **observation period of one year**
  - OFE =Opportunities For Error

- Software example: Defect rate: (D/KLOC)* 1000 where: D=Defect observed within the **observation period** (e.g., one year) and KLOC= thousands of Line Of Code

- It is a "crude" rate : KLOC do not coincide with OFE
  - One defect may rise from more than one LOC
  - One LOC may generate more than one defect

Definition, working definition, and measures

- Before to enter in the scope of the various measures and reference systems we recall some basic notions of the Theory of Measure
- Let’s use an example based on the following statement

  “the more rigorous the final part of the sw development process the higher the quality of the sw released to the customer”
- In order to accept/reject such statement we have to better define certain concepts:
  - **Development process**: requirements analysis, design,, …, integration,…., acceptance tests
  - **Final part of the development process**: integration and associated testing (after that the sw is quite stable)

Definition, working definition, and measures

- **Rigorous**: that adheres to the process documentation (quality manual)
  - This is still vague, we need some indicators:
    - E.g., if there is an inspection of the code we can use the working definition of: the percentage of code actually inspected
    - For the quality of the inspection we can use a working definition based on a Likert scale with 5 values
      - 1: low quality, … 5: high quality
    - Testing rigorousness could be associated with the working definition of the percentage of tested LOC
    - Testing effectiveness could be associated with the working definition of the number of removed defects per KLOC
  - **Quality of released software:** number errors per KLOC discovered during the system testing
  - Working definitions can be debatable, however they are
    - not ambiguous and
    - they can be measured

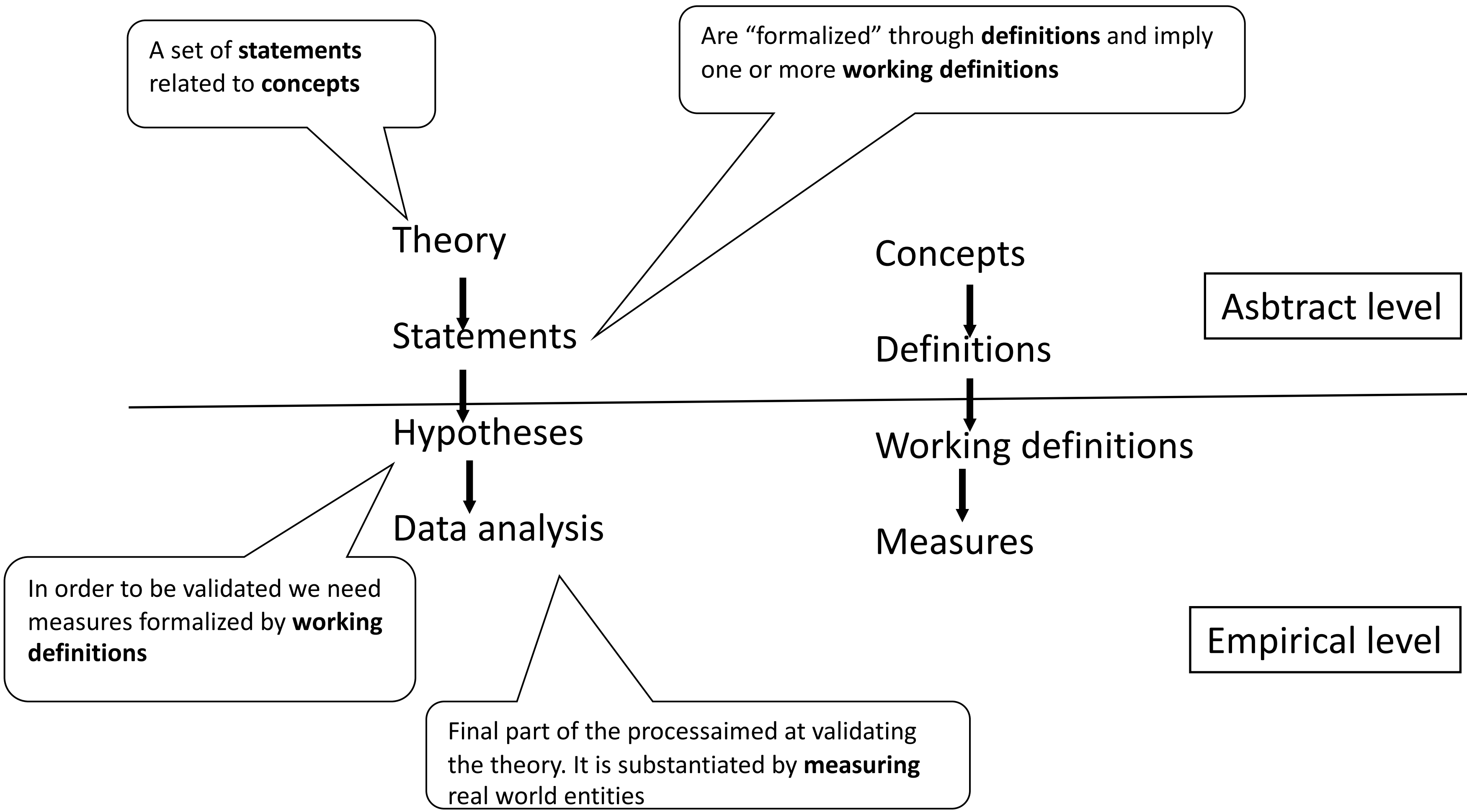
Definition, working definition, and measures

- Now we can rephrase the previous statement through the following hypotheses:

1. The greater the percentage of tested KLOC, the lesser the number of errors per KLOC discovered during the system testing
2. The greater the effectiveness of the inspection the lesser is the number of errors per KLOC discovered during the system testing
3. The greater the efficacy of tests, in terms of discovered errors, the lesser the number of errors per KLOC discovered during the system test

Definition, working definition, and measures

- The example shows the importance of measures and the need of different levels of abstraction



Definition, working definition, and measures

- In order to validate the theory previously formulated we need to:
  - introduce a unit of analysis e.g., component, project, etc.
  - validate the chosen indicators, i.e., means to collect and interpret measurements
  - perform statistical analysis in order to validate the hypotheses e.g., analysis of the variance

Basics of descriptive statistics (recall of main concepts)

Mean, variance, and standard deviation

- Consider a population of **n** known elements on which we want to perform a measurement
- E.g., the age of students in this class $\{x_1,\ldots,x_n\}$
- We define the following parameters:
  - Mean $\boldsymbol{\mu}= (x_1+ x_2+\ldots +x_n)/n$
  - Variance **var**$=[(x_1-\mu)^2+ (x_2-\mu)^2 +\ldots(x_n-\mu)^2]/n$
  - Standard deviation $\boldsymbol{\sigma}=\mathrm{var}^{1/2}$
- Usually the variance is indicated by $\boldsymbol{\sigma}^2$

Further observations on data (n known elements)

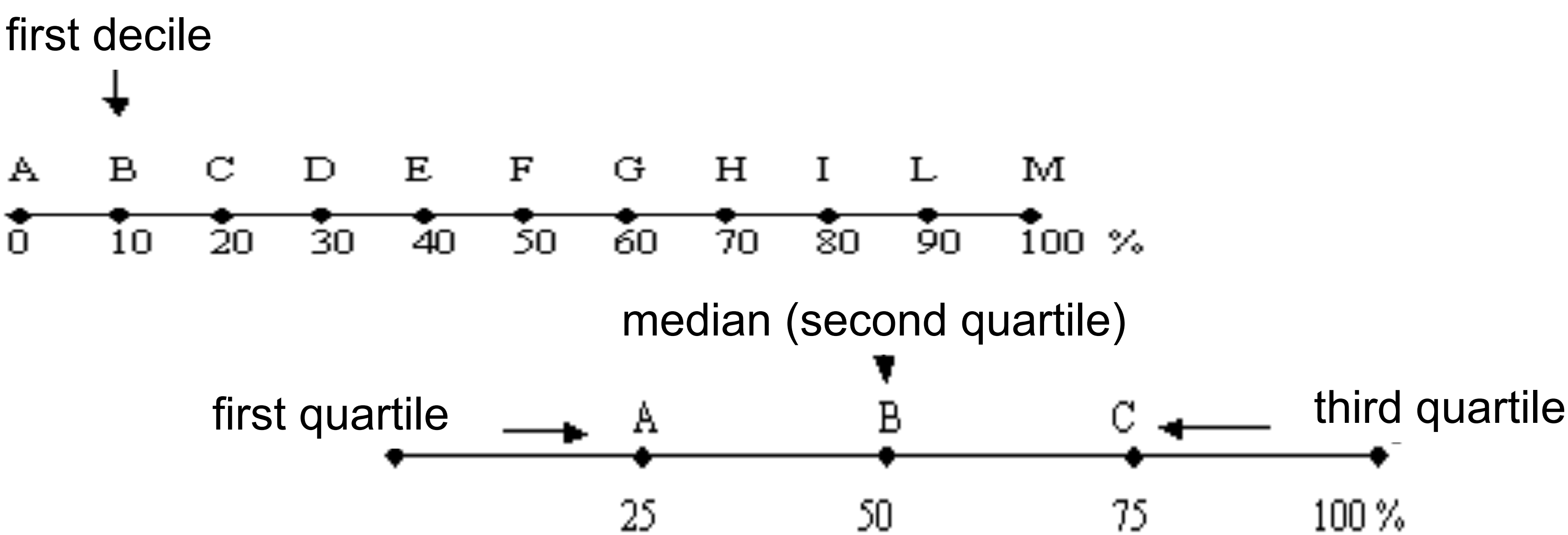
- Median
- Mode
- Percentile (Quartile)
- Frequency distribution

Median

- The median is the middle data point in an ordered set
- In order to compute the median we need a scale which is **at least an ordinal scale**
- To determine the median, sort the data from smallest to largest and find the middle data point
- It's the second quartile

Percentile, quartile

- A percentile (or centile) is the value of a variable below which a certain percent of observations fall
- So the 20th percentile is the value (or score) below which 20 percent of the observations may be found
- A quartile is any of the three values which divide the sorted data set into four equal parts, so that each part represents one fourth of the sampled population



Median

Sample data:	**Ordered Data:**
98cm	**54cm**
76cm	**76cm**
82cm	**82cm**
54cm	**90cm**
90cm	**98cm**

Median

Sample data:	Rearranged Data:
98cm	54cm
76cm	76cm
82cm	82cm
54cm	90cm
90cm	98cm

Median

- If there is an even number of data, there will be two middle points
- To find the median, take the average of those two points (that requires at least an interval scale!)

Median

Sample Data:

4ml

8ml

12ml

2ml

Ordered Data:

2ml

4ml

8ml

12ml

4 + 8 = 12ml

12/2 = 6ml

Mode

- The mode is the most frequently occurring data point
- To find the mode, arrange the data from smallest to largest, and then determine which amount occurs most often

Mode

Sample Data:		Rearranged Data:
20g	23g	20g 20g 20g
30g	30g	22g
22g	27g	23g 23g 23g 23g
25g	20g	24g
23g	24g	25g 25g
23g	25g	27g
20g	23g	30g 30g

Range

- The range is the distance between the smallest and largest data point.
- To calculate, determine the smallest data point and the largest data point, then subtract the smallest from the largest.

Range

Sample data:

98cm

76cm

82cm

54cm

90cm

Ordered Data:

54cm

76cm

82cm

90cm

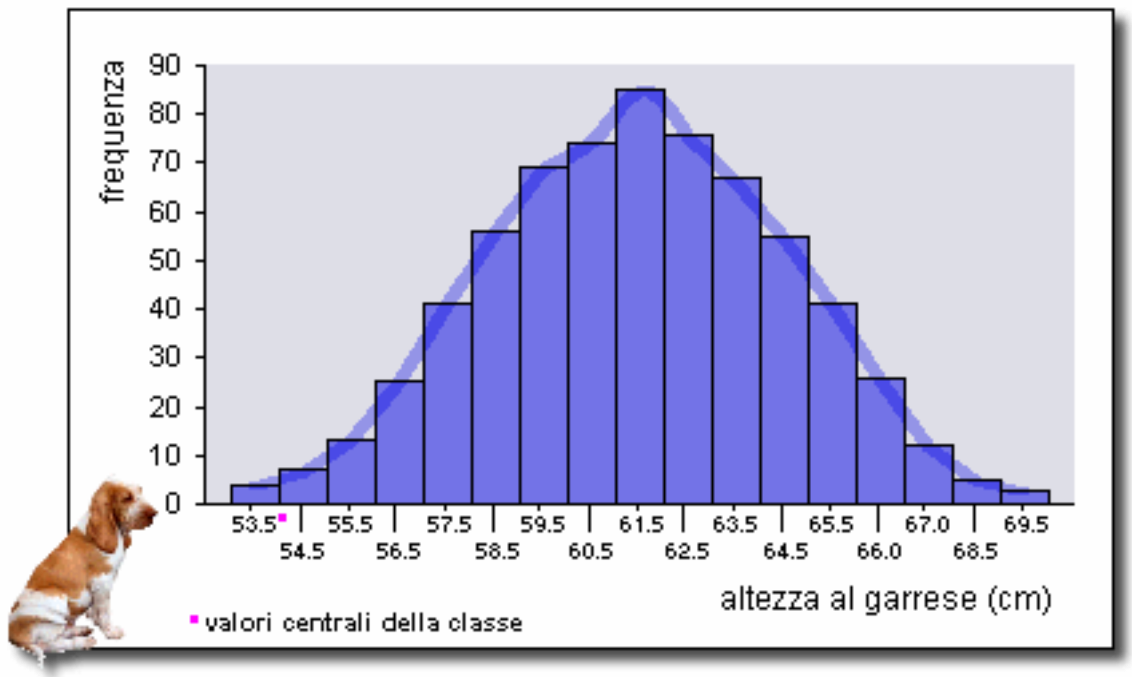
98cm

98cm – 54cm = 44cm

Frequency distribution

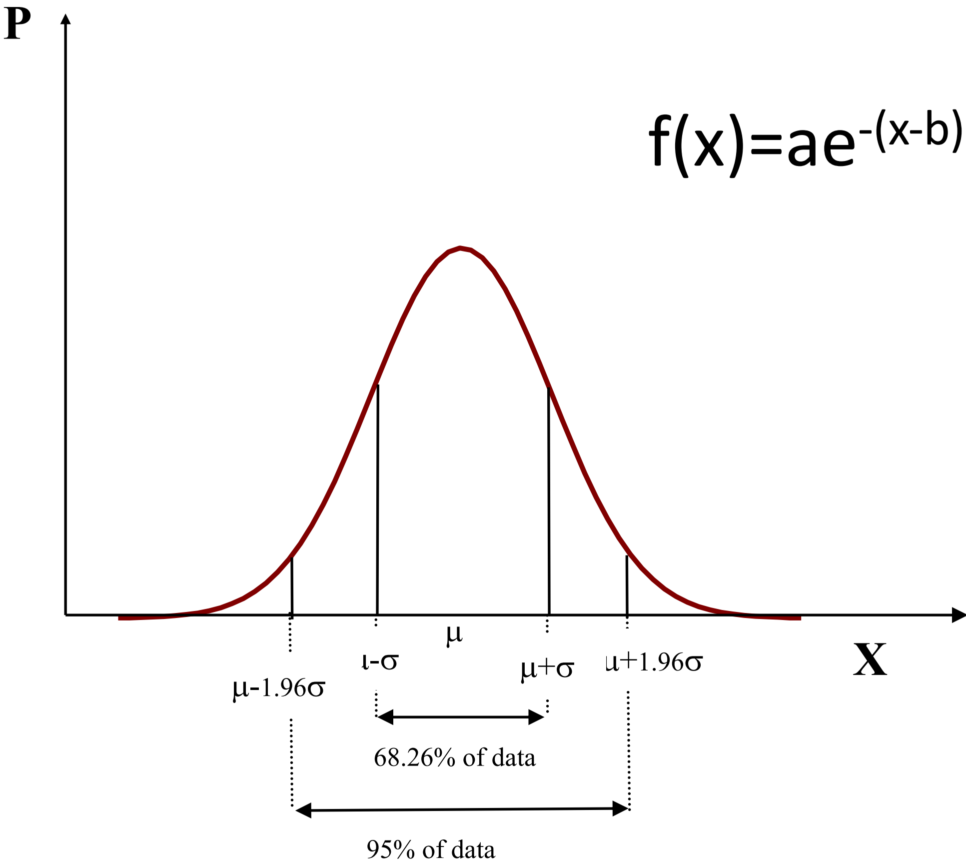
- It is obtained by splitting the values observed and by indicating, for each of them, the corresponding frequency
- As an alternative it is possible to divide the range of values in bins and counting all the elements in the same bin
- Typically the n elements follow a Normal distribution (or Gaussian)
- The analysis techniques for estimating the distribution of a sample is beyond the aims of this course

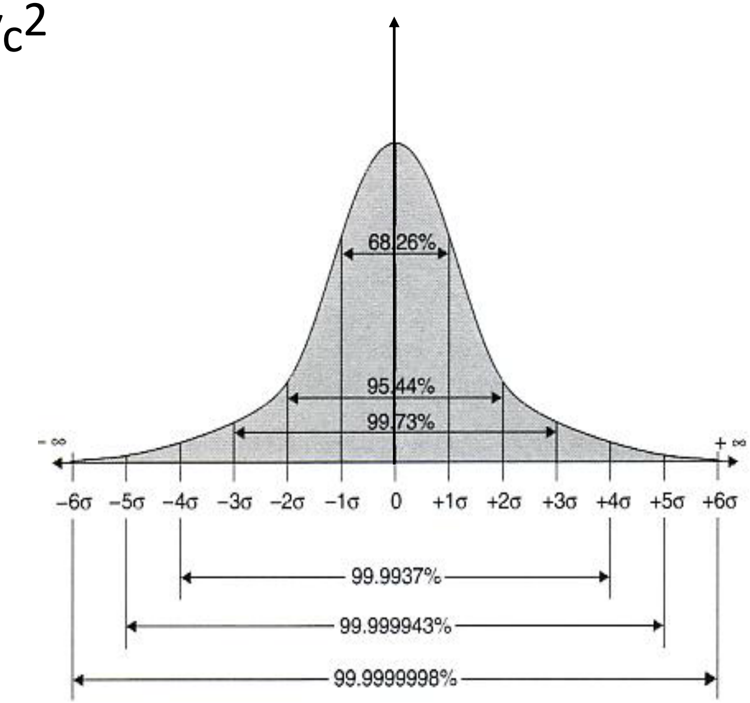
Altezza al garrese di 659 cani di razza "Bracco italiano". Istogramma.



Normal distribution

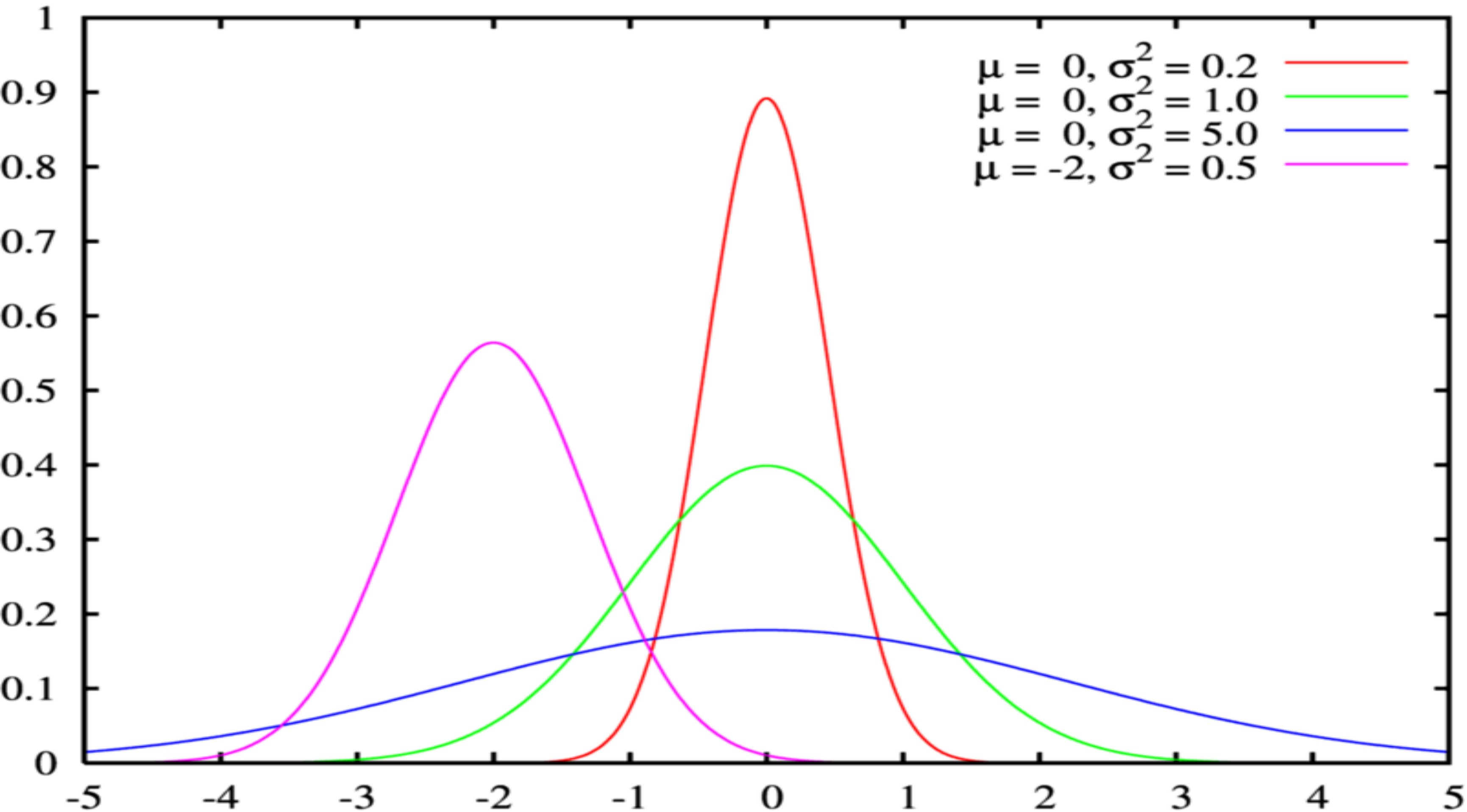
$$f(x)=ae^{-(x-b)^2/c^2}$$



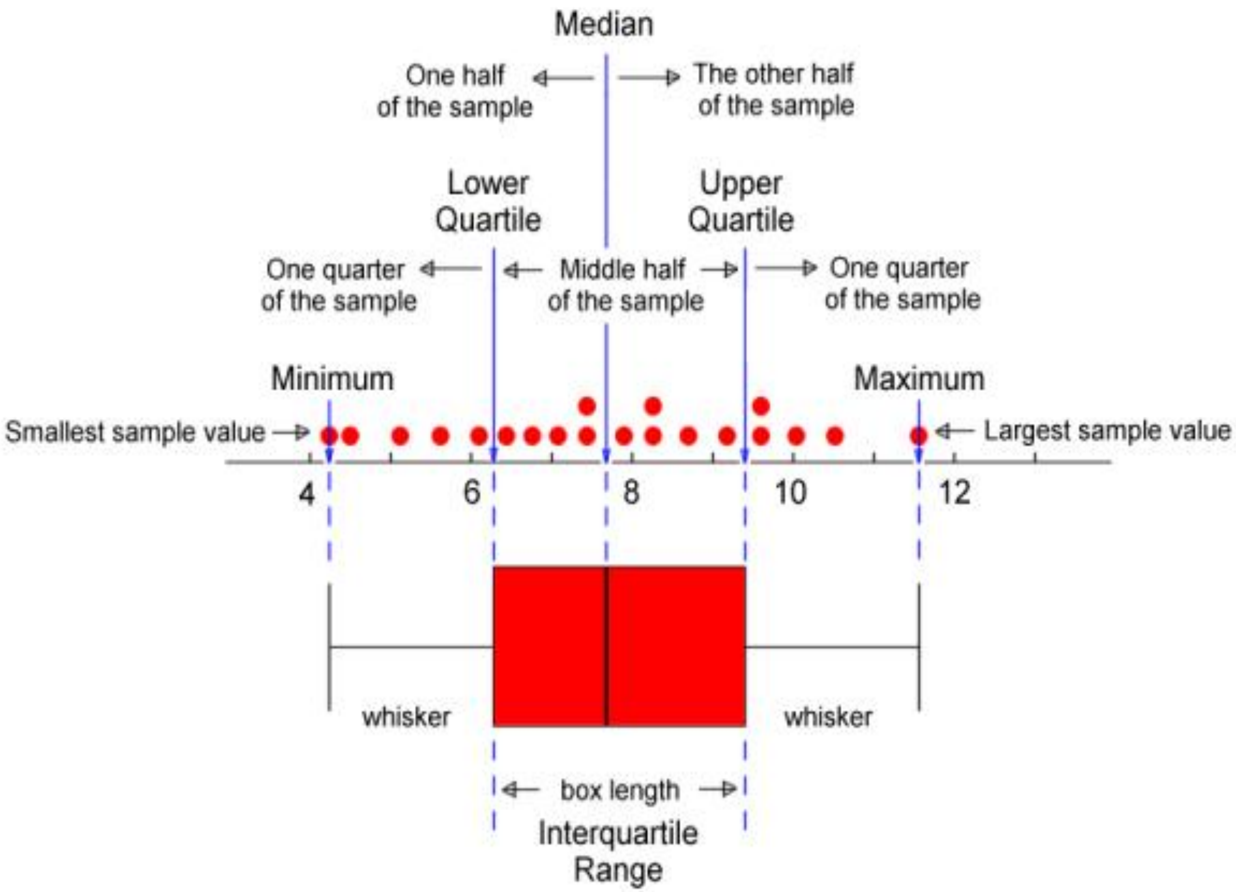


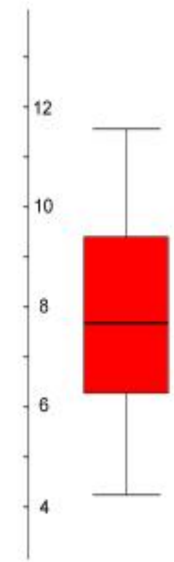
Usually the Gaussian is put at the center of the Y axis by putting X=X-μ

Some examples of normal distributions

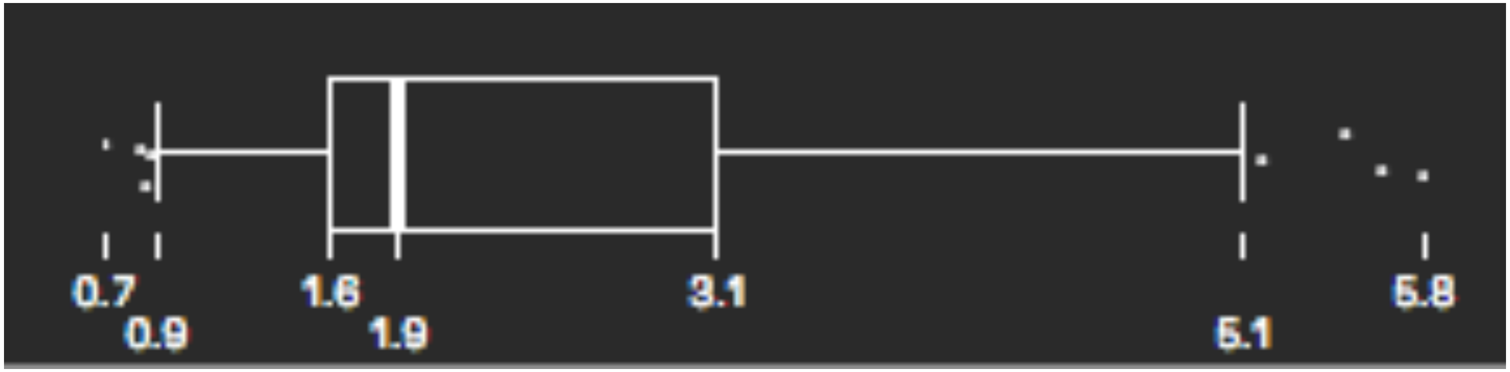


Graphical representation of parameters : Boxplot





Boxplot and outliers



Quality of a measure

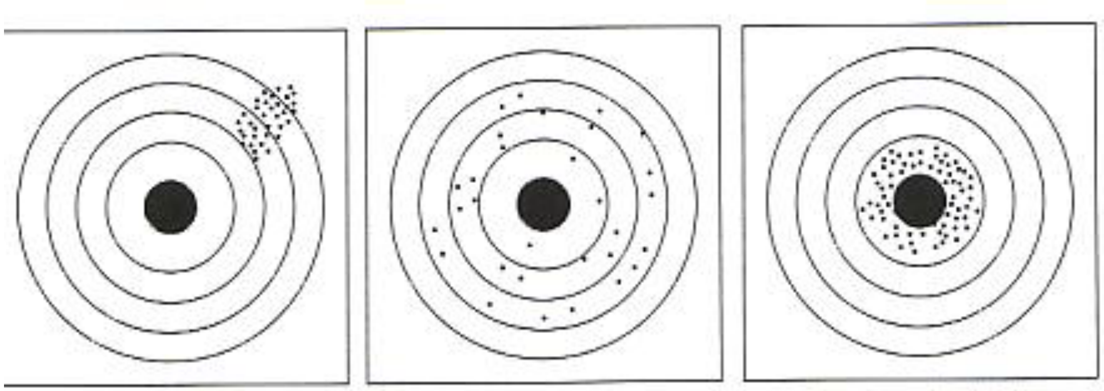
Reliability of a measure

- Reliability is the **consistency** of your measurement
  - The degree to which an instrument measures the same way each time it is used under the same condition with the same subjects
  - It is the **repeatability** of your measurement
  - A measure is considered reliable if a person's score on the same test given twice is similar
  - It is important to remember that reliability is not measured, it is **estimated**
  - Typically you can characterize this quality aspect by analyzing the variance σ^2 of repeated measures of the same value
  - The smaller σ^2 the more reliable the measure

Validity of a measure

- Validity is the strength of our conclusions, inferences or propositions
- Is the measure measuring what we actually are looking for?
- The best available approximation to the truth or falsity of a given inference, proposition or conclusion. Cook and Campbell (1979)
- In short, were we right?
- E.g., we want measure the comprehension of my classes
  - We can count the number of questions and use it as an indicator
  - No questions means full understanding?
- For more concrete measure it coincides with accuracy
  - E.g., weight, volume

Reliability and validity of a measure



Reliable and not valid

Valid and not reliable

Reliable and valid

Errors in measuring

- The result of a measure is a real number M which should capture the true value T of the phenomenon under analysis
- Experiences indicate that if we perform more measures of the same quantity rarely we obtain equal values
  - The measured values (M) are always different from the true value T
- The difference between the measured value and the true one is called total error (E_T)

$$M = T + E_T$$

Measure **True value** **Total error**

Errors in measuring

- By performing a measure we cannot determine with certainty the true value of the measured quantity, we produce an estimation
  - We have to consider the types of error in measuring

$$E_T = E_{\textbf{systematic}} + E_{\textbf{random}}$$

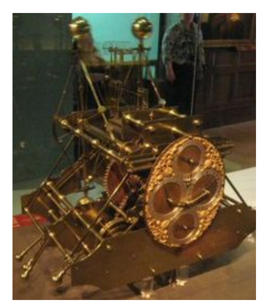
They are typically present in statistical methods

$E_{\textbf{systematic}}$ $\longrightarrow$ Influences validity

$E_{\textbf{random}}$ $\longrightarrow$ Influences reliability

Errors in measuring

- Systematic errors influence validity
- They occur constantly
- E.g. , a scale which configuration was wrong and adds always 1kg to the true weight
  - measure = T + 1kg + random variation : M= T + Es + Er
    - The measure is not valid
- If we assume that there can be only Er we have : M= T + Er
  - If Er is really due to a random event its contribution on the average can be ignored (expected value E(Er)=0)
  - The mean of an infinite number of measures (observations) is E(M)=T hence the measure is valid
  - A technique that exploits this principle is to repeat the measure N times and compute the mean
  - Longitude problem...
    - John Harrison clock



Errors in measuring

- What is this effect of a **random** error on the **reliability**?
- Intuitively, the smaller the error the lesser the influence

$$M = T + Er \rightarrow var(M) = var(T) + var(Er)$$

- The reliability of a measure is the ratio between the variance of the measured quantity and the variance of the metric

$$\rho_m = var(T)/var(M) = [var(M)-var(Er)]/var(M) = 1 - [var(Er)/var(M)]$$

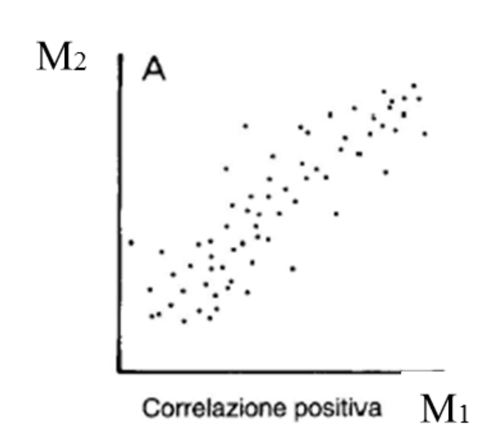
- Reliability value is between 1 and 0 (1 is the best value)

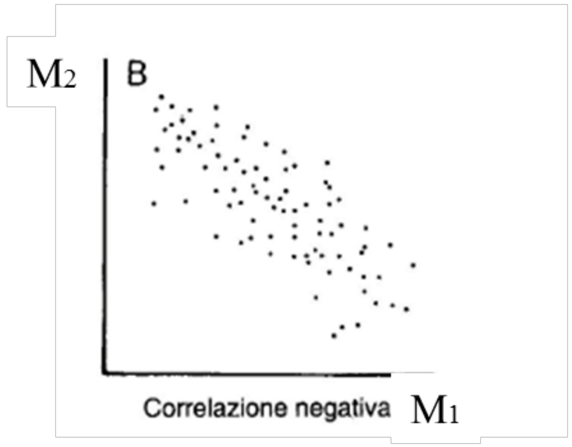
- In summary
  - Systematic errors influence validity
  - Random errors influence reliability that can be estimated through the variance

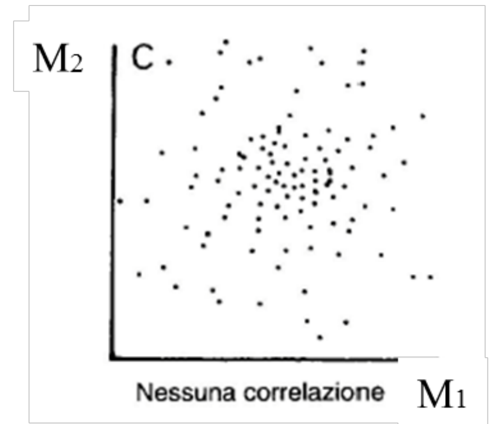
Correlation

- Indicates if a relationship between two variables holds
- The most popular correlation is Pearson's which can have values btw -1 (negative correlation) and +1 (positive correlation)
- Only for linear relations

Correlation







+1 -1 0

Inferential statistics

Reference parameters

- We want to analyze a population of M elements (M is unknown) through a sample of n elements $\{x_1, \dots, x_n\}$
- We identify the following parameters
  - Mean (sample) $\mu = (x_1 + x_2 + \dots + x_N)/n$
  - Variance (sample) $var = [(x_1-\mu)^2 + (x_2-\mu)^2 + \dots (x_N-\mu)^2]/(n-1)$
  - Standard deviation (sample) $\sigma = var^{1/2}$
  - Percentile /median / mode etc. (sample)
- These parameters are **random variables** which values depend on the causality of the sample
- Typically (hopefully) the elements of the sample have a normal distribution (Gaussian, bell-shaped) and we can perform an estimation

The problem

- We work with inferential statistics (vs. descriptive)
- We want to infer properties by using a sample of the data
  - we do not have all the data or
  - to save money or time
- The statistical characterization of our sample, e.g., the mean, is different from the actual data mean
  - The larger the sample size the smaller the error
- What is the trend of this error?
- Inferential statistics allows us to estimate it

Confidence interval

- Under the assumption of a normal distribution we can estimate the **probability** that the **mean** of a population **M** is **within** an **interval** centered on the mean of a sample of **n** elements of such population
- The size of the interval depends on the probability of error that we can bear
  - The error probability is proportional to the **standard deviation** of the sample and inversely proportional to the **size** of the sample

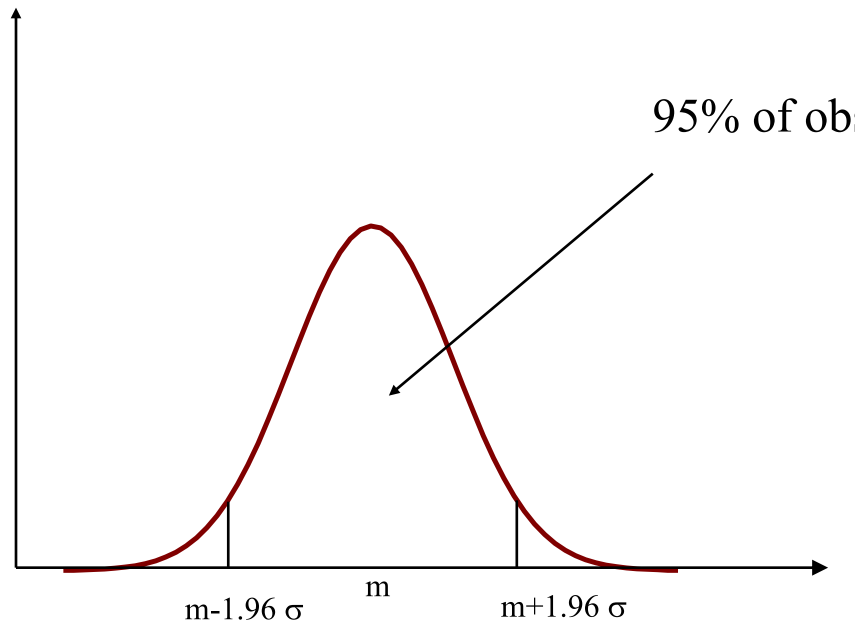
Confidence interval

- A **confidence interval** allows for estimating a population parameter
- Instead of estimating the parameter by a single value, we use an interval that is likely to include such a parameter
- Confidence interval size indicates the reliability of an estimate
- How likely the interval is to contain the parameter is determined by the **confidence level** or confidence coefficient
- A **confidence level** refers to the percentage of all possible samples that can be expected to include the true population parameter
- Increasing the desired confidence level will widen the confidence interval
- A confidence interval is always qualified by a particular confidence level, usually expressed as a percentage
  - E.g., 95% confidence interval

Confidence interval

- If the value of a **parameter** using a **sample** is x, with confidence interval [x-d, x+d] at confidence level P, then the actual population M parameter will be in [x-d, x+d] with a P probability
- 95% confidence intervals for the mean will be calculated by the following formulas
  - μ sample $\pm 2.77*\sigma/n^{1/2}$ if n=5
  - μ sample $\pm 2.26*\sigma/n^{1/2}$ if n=10
  - μ sample $\pm 2.09*\sigma/n^{1/2}$ if n=40
  - μ sample $\pm 1.96*\sigma/n^{1/2}$ if n ”large" (implemented in Excel)
- it is possible to use values greater than 95%
- the confidence intervals will be wider

Confidence interval



$1.96^{*}\sigma/n^{1/2}$ if n is large

Example

- We have collected 10 exam grades 26, 21, 29, 26, 21, 28, 27, 26, 29, 27
  $2.26*\sigma/n^{1/2}$ (n=10)

- μ= 26
- σ= 2.87
- $2.26*\sigma/10^{1/2}$ = 2.05
- The true mean (26.29), with a confidence level of 95%, falls within [26-2.05, 26+2.05] = [23.95, 28.05]

- If we collect 20 scores
- μ= 26.75
- σ= 3.02
- $2.26*\sigma/20^{1/2}$ =1.41
- The true mean (26.29), with a confidence level of 95%, falls within [26.75-1.41, 26.75+1.41] = [25.34, 28.16]

Exercise

- **Exercise 3 (7 points)** The quality manager of the ACME software house is going to assess the maintainability of the SW house software (50.000 Java classes). In order to do that s/he designs the metric

  USEFUL=number of useful comments/number of comments

- that is measured through manual inspection on seven classes, randomly chosen.
- The collected measures are [0.8, 0.92, 0.75, 0.95, 0.82, 0.88, 0.93].
- The candidate has to:
- a) indicate the scales of dividend and divisor;
- b) indicate the measure type (according to the theory of measure and ISO 25010)
- c) indicate the confidence interval for confidence level of 95%, using the following formulae
- d) explain what confidence value of 95% means
- e) compute how many samples are required to reduce the confidence interval to half (assuming the same value for σ)
- $2.77 * \sigma / N^{1/2}$ per N=5
- $2.26 * \sigma / N^{1/2}$ per N=10
- $2.09 * \sigma / N^{1/2}$ per N=40
- $1.96 * \sigma / N^{1/2}$ per N "big"

Exercise

- scale = absolute / absolute
- proportion
- mean=0.86   sigma=0.075
- $2.77*\sigma/N^{1/2}$ per N=5 = 0.079
- confidence interval [0.86-0.079, 0.86+0.079]=

[0.79, 0.94]

- $2.26*\sigma/x^{1/2}$ =0.079/2 → x=17

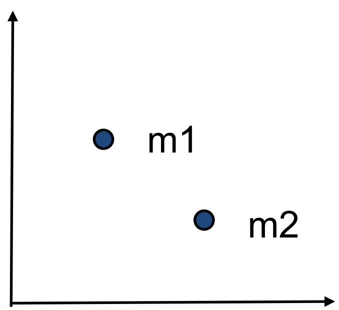
Hypotheses verification

Hypotheses verification

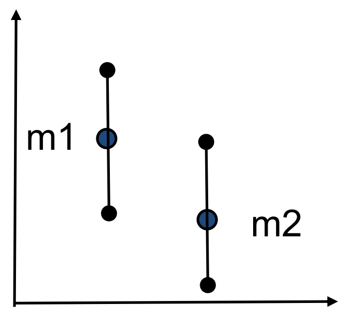
- Often we need to compare different repeated measures
  - e.g., results coming from different SE methods
- We can perform appropriate statistic tests
- A statistic test consist of challenging the hypothesis that the means of different samples are the same
- The hypothesis that all true means are equal indicates that we assume that all observed differences are random
  - This hypothesis is called **null hypothesis**
- The test is performed by fixing *a priori* the probability of having an error (α)

Two sample means intuitive discussion

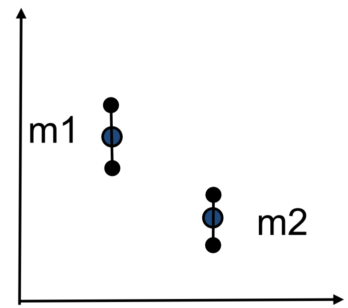
- Do the samples come from the same population?
- A rough answer using the confidence interval



Two means coming from two samples



Two confidence intervals that likely share the same mean



Two confidence intervals that likely DO NOT share the same mean

Hypotheses verification

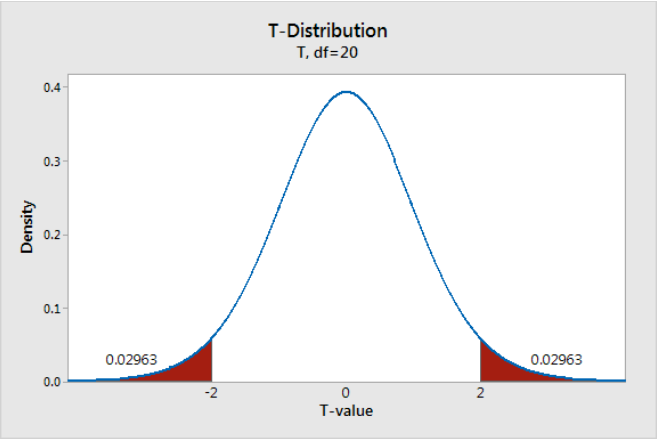
- A statistical hypothesis is a statement on the distribution of one or more **random variables** (in the following we will use the mean μ as an example)
- It is indicated by the letter H
- We pose the following question
  - **Given two samples (on which we compute μa and μb) what is the probability that they come from the same population?**
- We compare two opposite hypotheses
  - H_0 (null hypothesis) $\mu a = \mu b$ →the samples come from the same population
  - H_1 (alternative hypothesis) $\mu a \ != \mu b$ →the samples come from different populations
- We define **a formula** on the sample means capturing **data differences** (e.g., μa-μb, t-test, ANOVA). The **formula** generates a <u>random variable</u> and we compute the **p-value** with respect to a reference threshold, i.e., the probability that the random variable takes a value greater than the threshold IF the null hypothesis is true
- Little values of p indicate that the samples are likely not coming from the same population: <u>p is the probability of rejecting H_0 when it is true</u>
- We select for p an a priori **significance level α** (the maximum value for α is 0.05)
- In conclusion, if:
  - $p > \alpha$ -> we accept H_0
  - $p <= \alpha$ -> we accept H_1
- In making a critical decision you must use lower values for α (0.01, 0.001, etc.)

t-test

A toast to "Student"
William Sealy Gosset

- A technique that allows to compare the difference between the mean values of two samples (a and b, of n elements each)
- It exploits a comparison between means and standard deviation

$$t = \frac{\mu_a - \mu_b}{\sqrt{\frac{\sigma_a^2 + \sigma_b^2}{n}}}$$



- Assuming that a and b come from the **same** population we can compute the **probability density of t** and the probability of t being equal or greater than a value
  - It is expressed by a table that indicates the **probability** that t >= X, according to the degree of freedom (sample dimension)
  - Given n elements and the mean we have n-1 degree of freedom for a single sample:

$$\mu = (x_1 + x_2 + \ldots + x_N)/N$$

and the overall number of degree of freedom is 2(n-1)

t-critical values

5%

1 tail, double the values

Given **two** samples of size n=7
a=(a1,...,a7)
b=(b1,...,b7)
from the **same** distribution
what is the probability p(t>=x)?

$$t = \frac{|\mu_a - \mu_b|}{\sqrt{\frac{\sigma_a^2 + \sigma_b^2}{n}}} >= x$$

It depends on n=7
For ν=2*(7-1)=12
P(t>=1.356)= 20%
P(t>=1.782)=10%
P(t>=2.179)=5%
P(t>=2.681)=2%
P(t>=3.055)=1%
P(t>=3.929)=0.2%

ν	0.10	0.05	0.025	0.01	0.005	0.001
1.	3.078	6.314	12.706	31.821	63.657	318.313
2.	1.886	2.920	4.303	6.965	9.925	22.327
3.	1.638	2.353	3.182	4.541	5.841	10.215
4.	1.533	2.132	2.776	3.747	4.604	7.173
5.	1.476	2.015	2.571	3.365	4.032	5.893
6.	1.440	1.943	2.447	3.143	3.707	5.208
7.	1.415	1.895	2.365	2.998	3.499	4.782
8.	1.397	1.860	2.306	2.896	3.355	4.499
9.	1.383	1.833	2.262	2.821	3.250	4.296
10.	1.372	1.812	2.228	2.764	3.169	4.143
11.	1.363	1.796	2.201	2.718	3.106	4.024
12.	1.356	1.782	2.179	2.681	3.055	3.929
13.	1.350	1.771	2.160	2.650	3.012	3.852
14.	1.345	1.761	2.145	2.624	2.977	3.787
15.	1.341	1.753	2.131	2.602	2.947	3.733
16.	1.337	1.746	2.120	2.583	2.921	3.686

Big t values are associated with little probabilities

If the actual t value is >= than t-crit value for 5% probability it is possible to accept H_1

Example

- We have developed two interfaces A and B for the same application and we want to understand which one is perceived as the best one by the users
- We interview 7 users who have used interface A and 7 users that have used interface B
- We analyze the answers to the questions which are associated with a ratio scale from 1 to 6 (1 low degree of satisfaction, 6 high degree of satisfaction)
- We observe the following results

$$\mu_{a.} = (1+6+1+1+6+6+2)/7 = 3{,}286$$

$$\mu_{b.} = (5+3+1+6+2+4+1)/7 = 3{,}143$$

t-critical values

$$\mu_{a.} = (1+6+1+1+6+6+2)/7 = 3{,}286$$

$$\mu_{b.} = (5+3+1+5+2+4+2)/7 = 3{,}143$$

$$t = \frac{|\mu_a - \mu_b|}{\sqrt{\frac{\sigma_a^2 + \sigma_b^2}{n}}} = 0{,}126$$

5%

ν	0.10	0.05	0.025	0.01	0.005	0.001
1.	3.078	6.314	12.706	31.821	63.657	318.313
2.	1.886	2.920	4.303	6.965	9.925	22.327
3.	1.638	2.353	3.182	4.541	5.841	10.215
4.	1.533	2.132	2.776	3.747	4.604	7.173
5.	1.476	2.015	2.571	3.365	4.032	5.893
6.	1.440	1.943	2.447	3.143	3.707	5.208
7.	1.415	1.895	2.365	2.998	3.499	4.782
8.	1.397	1.860	2.306	2.896	3.355	4.499
9.	1.383	1.833	2.262	2.821	3.250	4.296
10.	1.372	1.812	2.228	2.764	3.169	4.143
11.	1.363	1.796	2.201	2.718	3.106	4.024
12.	1.356	1.782	2.179	2.681	3.055	3.929
13.	1.350	1.771	2.160	2.650	3.012	3.852
14.	1.345	1.761	2.145	2.624	2.977	3.787
15.	1.341	1.753	2.131	2.602	2.947	3.733
16.	1.337	1.746	2.120	2.583	2.921	3.686

p(t>=2.179)=5% Any value lower than 2.179 has p greater than 5%

- Compute the degrees of freedom: (7-1)+(7-1)=12
- Look at the corresponding row the value of the probability (p/2 in case of two-tails): 0.025-->2.179
- The value of the actual t, 0.126 is less or equal than 2.179 hence p is greater than 0.05 → we have to choose H_0

Another example

$$\mu_{a.} = (3+4+4+4+4+4+3)/7 = 3{,}714$$

$$\mu_{b.} = (3+3+3+3+3+4+3)/7 = 3{,}143$$

$$t = \frac{|\mu_a - \mu_b|}{\sqrt{\dfrac{\sigma_a{}^2 + \sigma_b{}^2}{n}}} = 2{,}449$$

t-critical values

$$\mu_{a.} = (3+4+4+4+4+4+3)/7 = 3{,}714$$

$$\mu_{b.} = (3+3+3+3+3+4+3)/7 = 3{,}143$$

$$t = \frac{|\mu_a - \mu_b|}{\sqrt{\frac{\sigma_a{}^2 + \sigma_b{}^2}{n}}} = 2{,}449$$

5%

ν	0.10	0.05	0.025	0.01	0.005	0.001
1.	3.078	6.314	12.706	31.821	63.657	318.313
2.	1.886	2.920	4.303	6.965	9.925	22.327
3.	1.638	2.353	3.182	4.541	5.841	10.215
4.	1.533	2.132	2.776	3.747	4.604	7.173
5.	1.476	2.015	2.571	3.365	4.032	5.893
6.	1.440	1.943	2.447	3.143	3.707	5.208
7.	1.415	1.895	2.365	2.998	3.499	4.782
8.	1.397	1.860	2.306	2.896	3.355	4.499
9.	1.383	1.833	2.262	2.821	3.250	4.296
10.	1.372	1.812	2.228	2.764	3.169	4.143
11.	1.363	1.796	2.201	2.718	3.106	4.024
12.	1.356	1.782	2.179	2.681	3.055	3.929
13.	1.350	1.771	2.160	2.650	3.012	3.852
14.	1.345	1.761	2.145	2.624	2.977	3.787
15.	1.341	1.753	2.131	2.602	2.947	3.733
16.	1.337	1.746	2.120	2.583	2.921	3.686

Any value greater than 2.179 has a probability less than 5%

- Compute the degrees of freedom: (7-1)+(7-1)=12
- Look at the corresponding row the value of the probability (p/2 in case of two-tail): 0.025-->2.179
- The value of t, 2.449 is greater than 2.179 hence p is smaller than 0.05 → we can reject H_0

p-value and **α**

- A value of $p >= \alpha$ indicates that the difference between the observed means is "random" and we have to select the null hypothesis
- The **α** reference level 0.05 is considered a boundary value:
  - A measurement is considered significant for values of $p <= 0.05$
- If:
  - $p <= 0.005$ the measurement is classified as statistically significant
  - $p <= 0.001$ highly significant

- These values are arbitrary, although they are widely used

Probability of error (α and β)

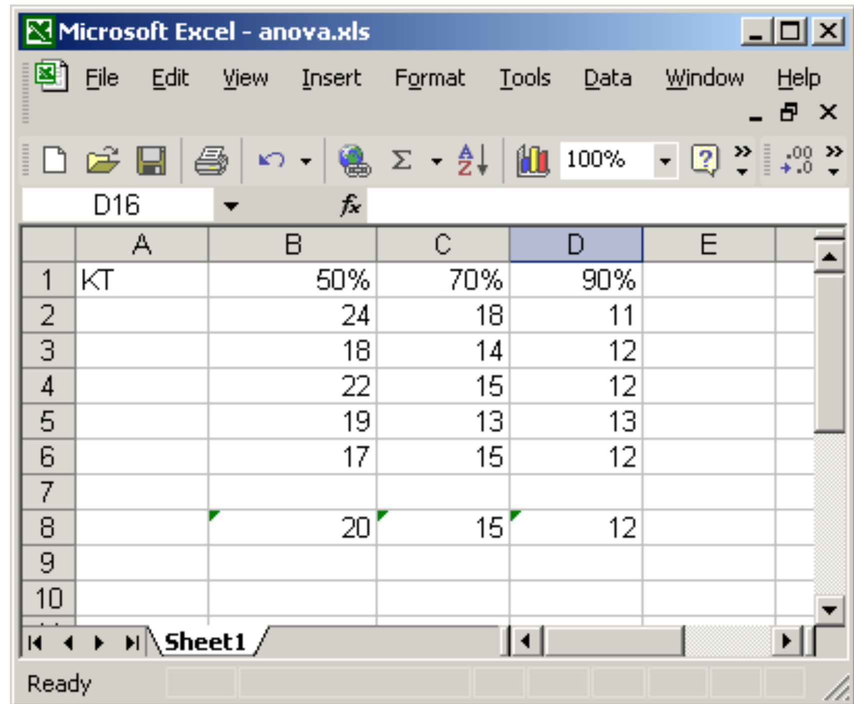
- H_0 is true
  - reject H_0 (α) Type **I** error (false positive)
  - accept H_0 ($1-\alpha$)
- H_0 is false
  - reject H_0 ($1-\beta$)
  - accept H_0 (β) Type **II** error (false negative)
- Type II errors arise frequently when the sample sizes are too small
- beta cannot generally be computed because it depends on the population mean which is unknown☹

More “short & dirty” readings

- https://machinelearningmastery.com/statistical-hypothesis-tests/
- https://online.stat.psu.edu/statprogram/reviews/statistical-concepts
- From Wikipedia (IT or EN depending on your preferred language):
  - https://it.wikipedia.org/wiki/Distribuzione_t_di_Student
  - https://it.wikipedia.org/wiki/William_Sealy_Gosset
  - https://it.wikipedia.org/wiki/Ipotesi_nulla
  - https://it.wikipedia.org/wiki/Test_t
  - https://it.wikipedia.org/wiki/Valore_p
  - https://it.wikipedia.org/wiki/Test_di_verifica_d%27ipotesi

What does it happen if we have more than two samples?

- If we perform [n*(n-1)/2]=3 t-test with α=0.05 the probability 0.95 of accepting true H_0 degrades 0.95*0.95*0.95= 0.86
- With n=5 we need 10 comparisons....
- For n>2 samples we use a technique based-on analysis of variance (ANOVA)
- ANOVA is conceptually similar to t-test, but it considers all means at once



	A	B	C	D	E
1	KT	50%	70%	90%	
2		24	18	11	
3		18	14	12	
4		22	15	12	
5		19	13	13	
6		17	15	12	
7					
8		20	15	12	
9					
10					

ANalysis **O**f **VA**riance

- ANOVA allows to analyze **two or more** samples comparing the internal variability **within** the groups (**Var_w**) with the variability **between** the groups (**Var_B**)
- The null hypothesis assumes that **all groups** have the same distribution, and that any observed difference in the samples is casual
- The idea is that if the **variability within** (**W**) the groups **is much higher** than the **variability between (B) the groups** than the observed difference is caused by the internal variability
- **The most popular** and known set of techniques is based on **comparing the variance**, and uses the random Snedecor variable F (similar to the t variable for the t test)
- Notice: t-test and ANOVA on **two** samples are **perfectly** equivalent, i.e., they produce the same p-value

ANOVA hypotheses

- ANOVA is a general technique that can be used to test the hypothesis that the means among two or more groups are equal, under the assumption that the sampled populations are normally distributed
- The hypotheses are the followings:
  - H_0: $\mu_1 = \mu_2 \ldots = \mu_I$
  - H_1: **at least** two among the means are different
- I samples of J items

- We have

I **> 2 different** *samples*: $\{C_{1, \dots,} C_I\}$

- Each sample is assumed to have the same number *J* of objects (although this is not mandatory)

Y_{ij} is the *j*-th observation on the *i*-th sample

Where:

- Mean of sample *i* :

$$\mu_i = (\sum_{j=1}^{J} Y_{ij}) / J$$

- General mean:

$$\mu = (\sum_{i=1}^{I} \mu_i) / I$$

I samples of J items

F-test (Fisher test)

- The **random** Snedecor variable F :

$$F = \frac{SS_B / (I-1)}{SS_W / [I(J-1)]}$$

$$SS_B = J \sum_{i=1}^{I} (\mu_i - \mu)^2$$

$$SS_W = \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{ij} - \mu_i)^2$$

- Once the degrees of freedom are known (for both numerator and denominator) it is possible to evaluate the probability (p-value) associated with the values of F
- This test tells us whether to:
- accept H_0: $p>\alpha$ (*F<F-crit*)
- reject H_0: $p<=\alpha$ (*F>=F-crit*)

Example: I=3 samples, J=5 elements each

$$F = \frac{SS_B/(I-1)}{SS_W/[I(J-1)]}$$

I-1=2
I(J-1)=12

The probability
that F>=3.89=0.05

numerator

denominator

nu \ nu	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,37
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78

F-critical values at α = **0.05**

Example (2 samples for the sake of simplicity)

I=2
J=7

$$\mu_a = (1+6+1+1+6+6+2)/7 = 3{,}28$$

$$\mu_b = (5+3+1+6+2+4+1)/7 = 3{,}14$$

$$\mu = (1+6+1+1+6+6+2+5+3+1+5+2+4+2)/14 = 3{,}21$$

$$SS_B = 7[(3{,}28-3{,}21)^2 + (3{,}14-3{,}21)^2] = 0{,}0714$$

$$SS_W = (1-3{,}28)^2 + (6-3{,}28)^2 + (1-3{,}28)^2 + (1-3{,}28)^2 + (6-3{,}28)^2 + (6-3{,}28)^2 + (2-3{,}28)^2 + \\ + (5-3{,}14)^2 + (3-3{,}14)^2 + (1-3{,}14)^2 + (6-3{,}14)^2 + (2-3{,}14)^2 + (4-3{,}14)^2 + (1-3{,}14)^2 = 62{,}29$$

$$F = \frac{0{,}0714/(2-1)}{62{,}29/[2(7-1)]} = 0{,}01376$$

<< F-crit $_{1,12}$ = 4.75

Accept H_o

Excel example

	A	B	C	D	E	F	G	H	I	J	K
1	A	B									
2	1	5									
3	6	3									
4	1	1			Anova: Single Factor						
5	1	6									
6	6	2			SUMMARY						
7	6	4			*Groups*	*Count*	*Sum*	*Average*	*Variance*		
8	2	1			A	7	23	3.285714	6.571429		
9					B	7	22	3.142857	3.809524		
10											
11											
12					ANOVA						
13					*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
14					Between Groups	0.071429	1	0.071429	0.013761	0.908556	4.747225
15					Within Groups	62.28571	12	5.190476			
16											
17					Total	62.35714	13				

$$SS_B/(I-1)$$

$$F = \frac{SS_B/(I-1)}{SS_W/[I(J-1)]}$$

SS= Sum of square
MS= Mean square
df= degree of freedom

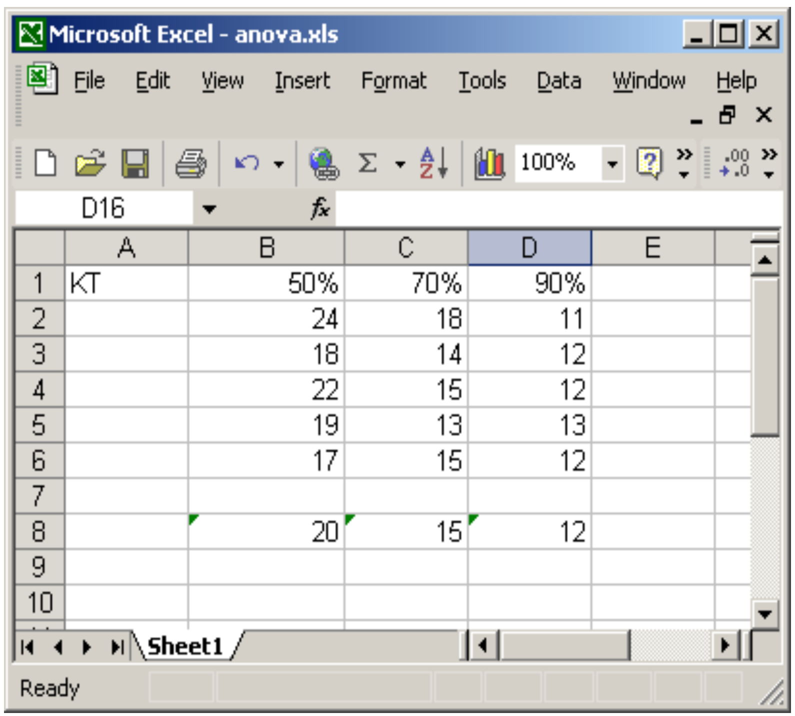
$$SS_W/[I(J-1)]$$

Analysis of variance (ANOVA) on tested KLOC and trend of defects

- Most often the attempt of demonstrating an hypothesis substantiates in searching a relation between two variables: if we change A then B changes (following a certain rule)
- For example: we try to demonstrate that if the proportion KT of tested KLOC increases then the rate of defect DR in the first year after the release decreases
- Proportion **KT**=(tested KLOC)/total KLOC
- Defect rate in the first year **DR**=(D/KLOC)*k (let k be 1)

Our sample set

- Let's assume that in the considered software house we consider three different fixed percentage values, i.e., we have 3 samples of five elements (e.g., java packages):
  - 50 %   70%   90%
- and we observe the 5 software packages for one year collecting their DR
- We compute the mean of DR and obtain the following table

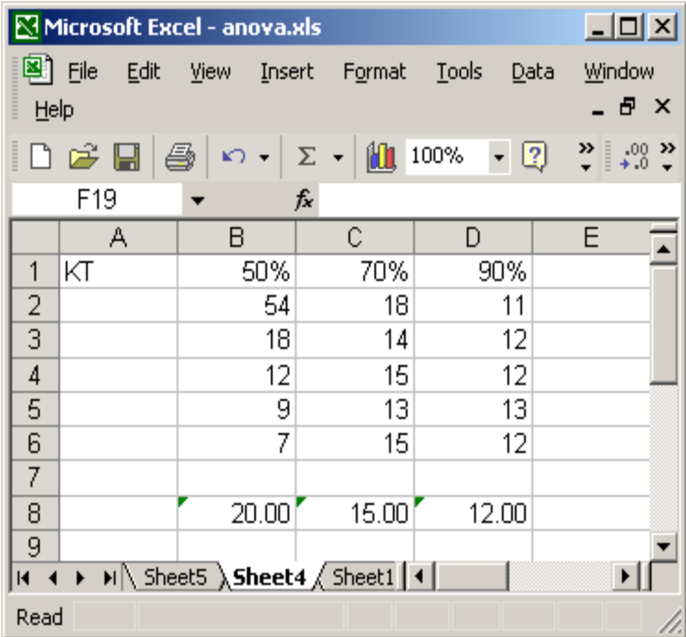


	A	B	C	D	E
1	KT	50%	70%	90%	
2		24	18	11	
3		18	14	12	
4		22	15	12	
5		19	13	13	
6		17	15	12	
7					
8		20	15	12	
9					
10					

ANOVA allows us to evaluate the probability that the differences among the means are random

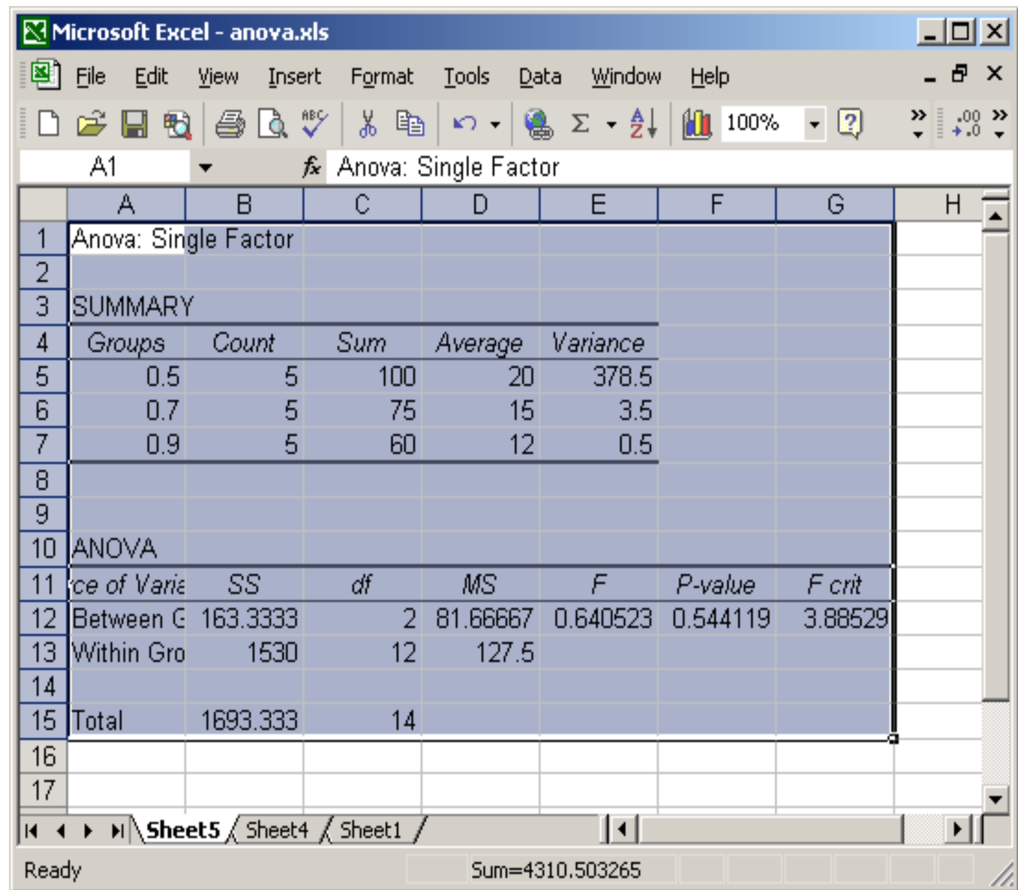
Dependent and independent variables

- In such analyses we call
  - **Independent** variables those ones that are manipulated to the aim of verifying a hypothesis
  - **Dependent** variables those ones that are observed and that depend (hopefully) on independent ones
- In our example
  - KT= independent variable with 3 values (50,70,90 %)
  - DR= dependent variable
- We have I=3 samples of J=5 elements



Microsoft Excel - anova.xls

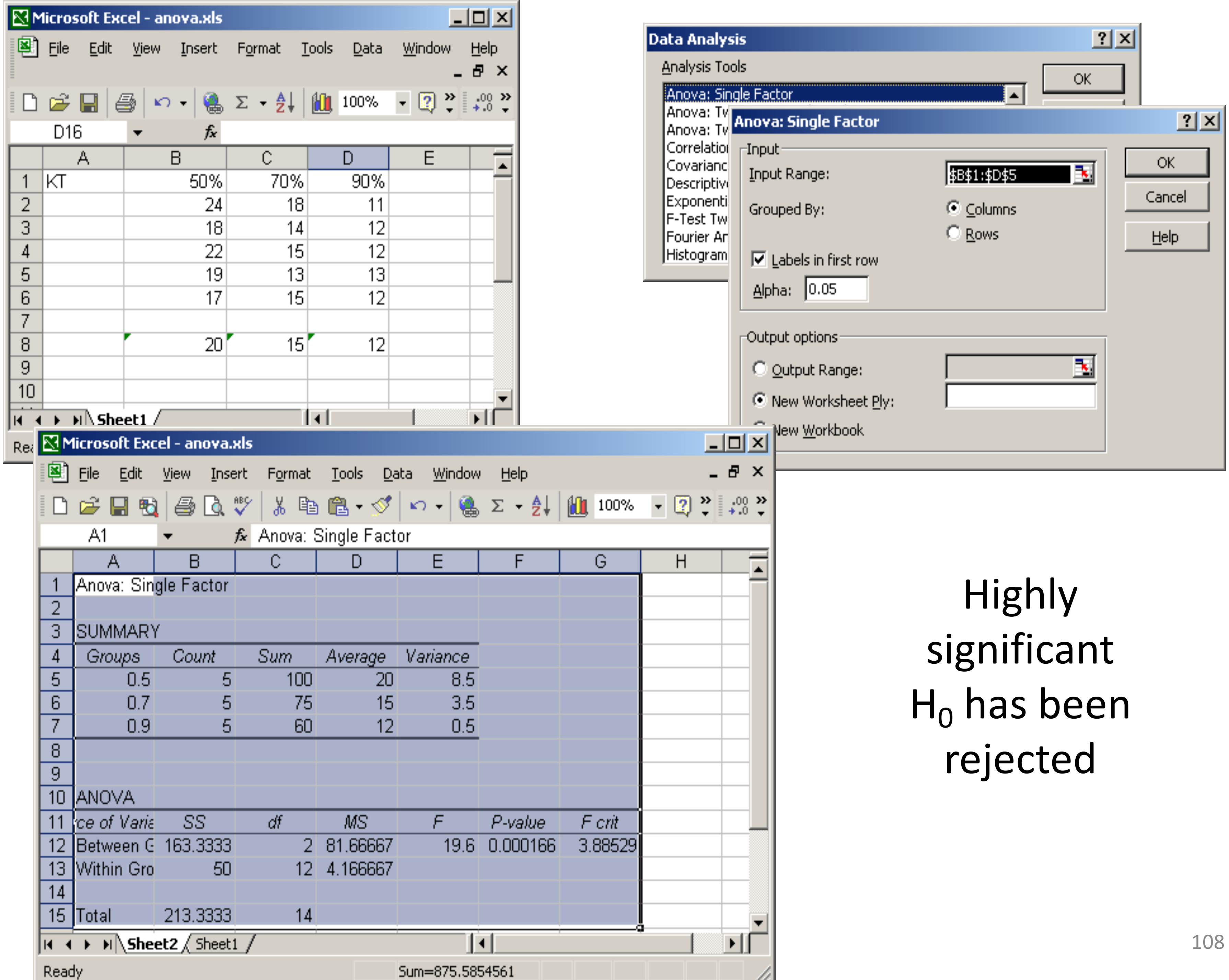
	A	B	C	D	E
1	KT	50%	70%	90%	
2		54	18	11	
3		18	14	12	
4		12	15	12	
5		9	13	13	
6		7	15	12	
7					
8		20.00	15.00	12.00	
9					



Microsoft Excel - anova.xls

	A	B	C	D	E	F	G
1	Anova: Single Factor						
2							
3	SUMMARY						
4	*Groups*	*Count*	*Sum*	*Average*	*Variance*		
5	0.5	5	100	20	378.5		
6	0.7	5	75	15	3.5		
7	0.9	5	60	12	0.5		
8							
9							
10	ANOVA						
11	*ce of Varia*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
12	Between G	163.3333	2	81.66667	0.640523	0.544119	3.88529
13	Within Gro	1530	12	127.5			
14							
15	Total	1693.333	14				

Not significant



Highly significant

H_0 has been rejected

Post hoc test (1)

- The experiment just performed tells us that the three samples **do not** belong to the same population
- This does not imply that they belong to **three** different populations! For example, the 70% group and the 90% group **might have** the same mean…
- It is necessary to compare the single pairs n*(n-1)/2=3
- The problem has been studied in the last years and yet there is not a definitive solution

KT	50%	70%	90%
	24	18	11
	18	14	12
	22	15	12
	19	13	13
	17	15	12
	20	15	12

Post hoc test : Fisher protected test

- This is the method most often applied
- Pairwise tests are used only **after** ANOVA has confirmed the significance f differences
- For example
  1. 50, 70 and 90 do not have the same mean (p=0.000166) ANOVA at 3
  2. 50 and 70 have different means (P1=0.012) ANOVA at 2 or t-test
  3. 70 and 90 have different means (P2=0.010) ANOVA at 2 or t-test
  4. 50 and 90 have different means (P3=0.00037) ANOVA at 2 or t-test
- In this case we can say that 50, 70, and 90 have all different means

KT	50%	70%	90%
	24	18	11
	18	14	12
	22	15	12
	19	13	13
	17	15	12
	20	15	12

50-70

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
0.5	5	100	20	8.5
0.7	5	75	15	3.5

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	62.5	1	62.5	10.41667	0.012103	5.317655
Within Groups	48	8	6			
Total	110.5	9				

70-90

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
0.7	5	75	15	3.5
0.9	5	60	12	0.5

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	22.5	1	22.5	11.25	0.010019	5.317655
Within Groups	16	8	2			
Total	38.5	9				

50-90

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
0.9	5	60	12	0.5
0.5	5	100	20	8.5

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	160	1	160	35.55556	0.000337	5.317655
Within Groups	36	8	4.5			
Total	196	9				

Exercise with ANOVA

Experiment on naming convention

- The quality management system of Acme software house is studying the influence of naming convention on the readability of code, with special attention to the integration phase
- More specifically, the quality management system set up 3 different techniques of naming convention NC1, NC2, NC3 and wants to investigate if there is a relationship between the applied <u>technique</u> and the <u>number of errors</u> identified during the integration test phase
- Design an experiment for validating this hypothesis

Dependent and independent variables Structure of the experiment

- Independent variable: type of adopted naming convention technique with three values (NC1, NC2, NC3)
- Dependent variable: number of errors found during the integration test /KLOC
- The different techniques for naming convention are used in three similar projects A, B, C
  - Similar? An attempt to minimize the influence of other factors, like project type
- The developers involved in and the followed methodology by the three projects have similar skills
  - Similar? An attempt to minimize the influence of other factors, like developers' skill

- During the integration phase 18 classes have been analyzed: 6 classes of project A, 6 classes of project B, and 6 classes of project C, collecting errors:
  - A (14, 15, 19, 20, 20, 22) mean 18.3 var 9.9
  - B (19, 20, 22, 23, 25, 28) mean 22.8 var 11.0
  - C (15, 15, 16, 16, 18, 21) mean 16.8 var 5.3

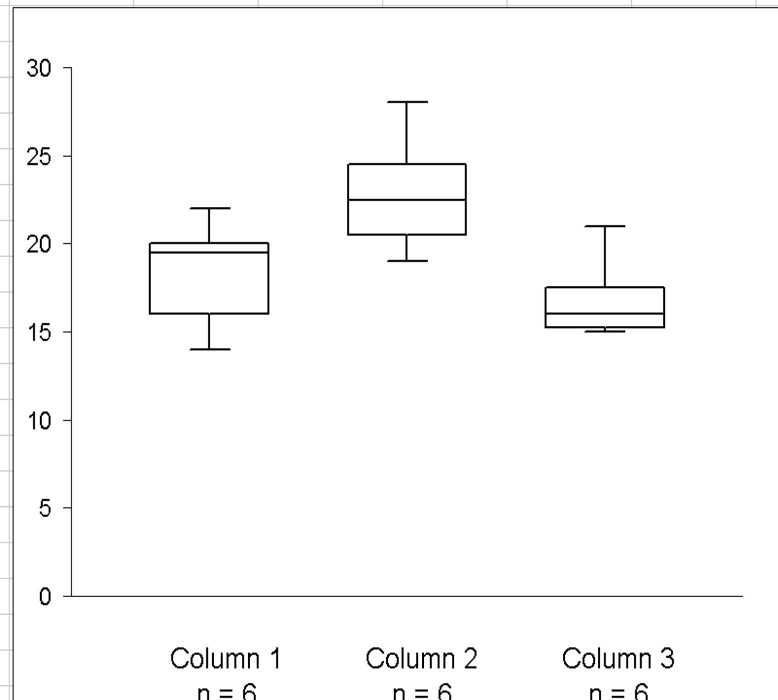
3 hypotheses for applying ANOVA

- 1) Independence
  - I samples taken **randomly** from k populations
- 2) Normal distribution
  - Samples must have with normal distribution
  - or symmetry between distributions of small samples (boxplot)
  - **or big large sample (n>30) : central limit theorem**
  - **alternative: use non-parametric tests**
- 3) Same variance
  - Samples must with same variance
  - Levene test (not supported by Excel!)
  - or empirical test (Excel)

Normal distribution?

Microsoft Excel - 05_EsempioVoti.xls

	A	B	C	D
1		Prog. A	Prog. B	Prog. C
2				
3		14	19	15
4		15	20	15
5		19	22	16
6		20	23	16
7		20	25	18
8		22	28	21
9				
10	Media	18.33	22.83	16.83
11	Varianza	9.87	10.97	5.37
12	Mediana	19.5	22.5	16



Reasonably symmetric wrt the mean

The central limit theorem

- Let $X_1, X_2, X_3, \ldots, X_n$ be a sequence of n independent and identically distributed random variables each having finite values of expected μ and variance $\sigma^2 > 0$

- The central limit theorem states that as the sample size n increases the distribution of the **sample average** of these random variables approaches the normal distribution with a mean μ and variance σ^2/n irrespectively of the distribution of the individual terms X_i

Let’s assume the hypotheses are satisfied
ANOVA single factor

	Prog. A	Prog. B	Prog. C
	14	19	15
	15	20	15
	19	22	16
	20	23	16
	20	25	18
	22	28	21
Media	18.33	22.83	16.83
Varianza	9.87	10.97	5.37
Mediana	19.5	22.5	16

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	6	110	18.3333	9.86667
Column 2	6	137	22.8333	10.9667
Column 3	6	101	16.8333	5.36667

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	117	2	58.5	6.69847	0.00834	3.68232
Within Groups	131	15	8.73333			
Total	248	17				

Null hypothesis H_0-> $\mu_a=\mu_b=\mu_c$

We can reject the null hypothesis (P=0.00834)

Hence at least one mean is different

Post hoc test : Fisher protected test

- After a ANOVA test that reject the null hypothesis we can perform n(n-1)/2 t-test

	Prog. A	Prog. B	Prog. C			Fisher protected test		
	14	19	15		A-B	0.03637		ok: A and B have different means
	15	20	15		A-C	0.36869		I have to accept the null hypothesis
	19	22	16		B-C	0.00456		ok: B and C have different means
	20	23	16					
	20	25	18					
	22	28	21					
Media	18.33	22.83	16.83					
Varianza	9.87	10.97	5.37					
Mediana	19.5	22.5	16					

B>A

B>C

$\mu_a = \mu_c$

Hence we can discard B

We don’t know what to chose between A and C