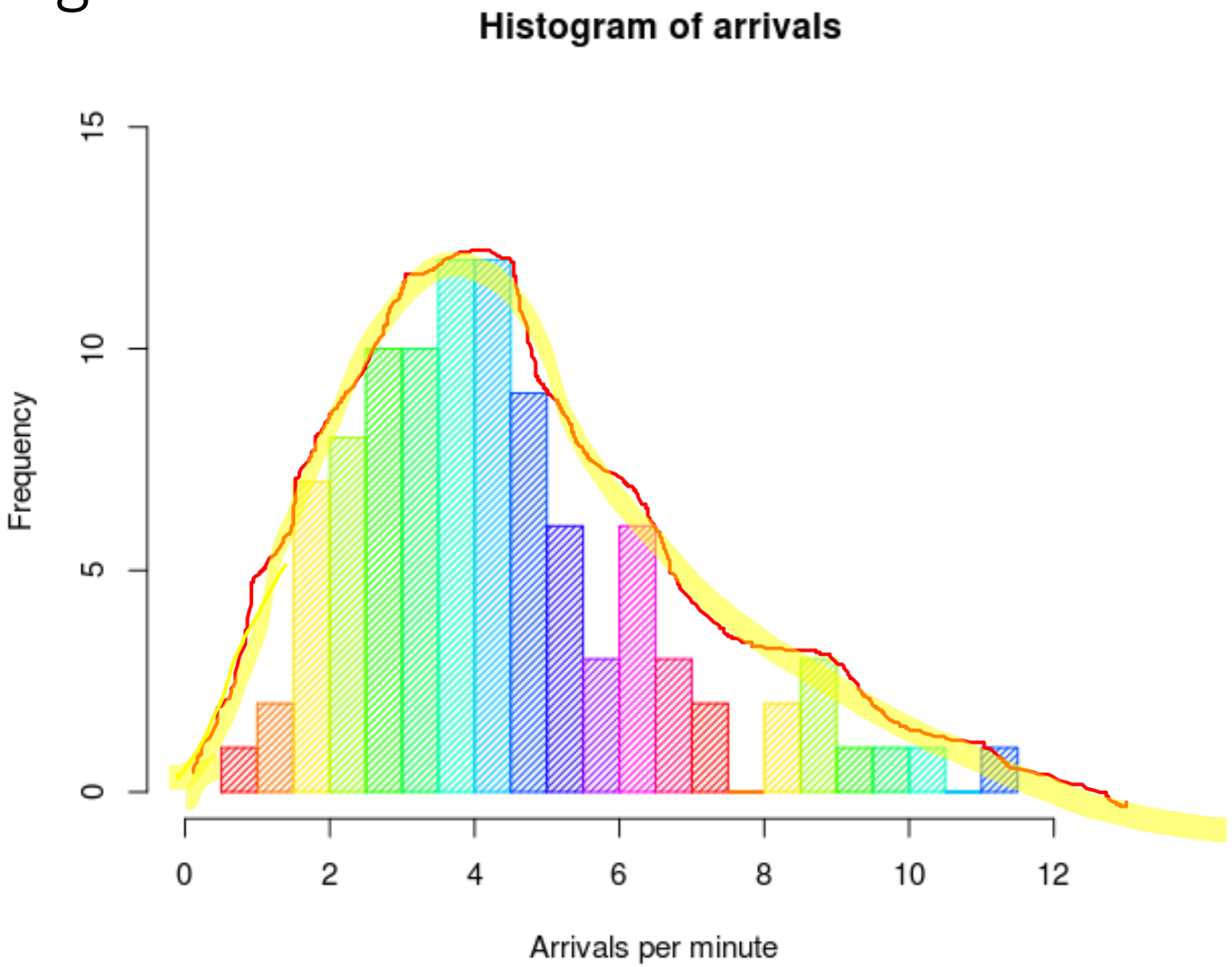
Human Computer Interaction

Conducting User Trial

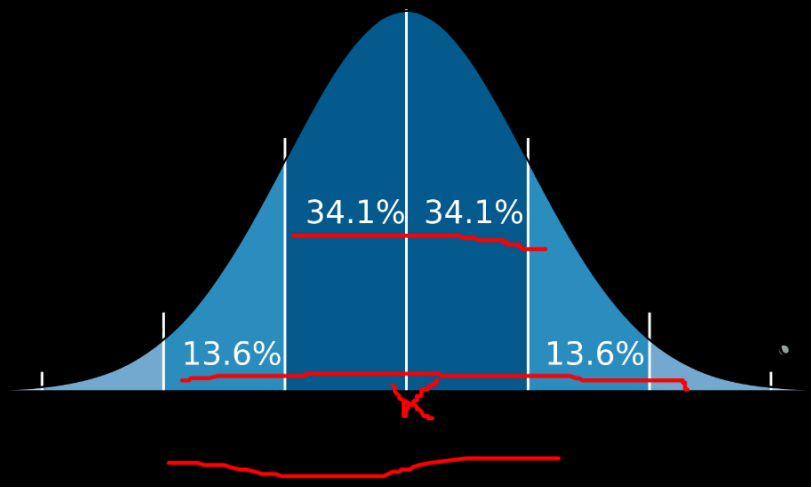
Dr Pradipta Biswas, PhD (Cantab)
Associate Professor
Indian Institute of Science
https://cambum.net/

Prerequisites

Histogram


Histogram of arrivals
Frequency
0
5
10
15
0
2
4
6
8
10
12
Arrivals per minute

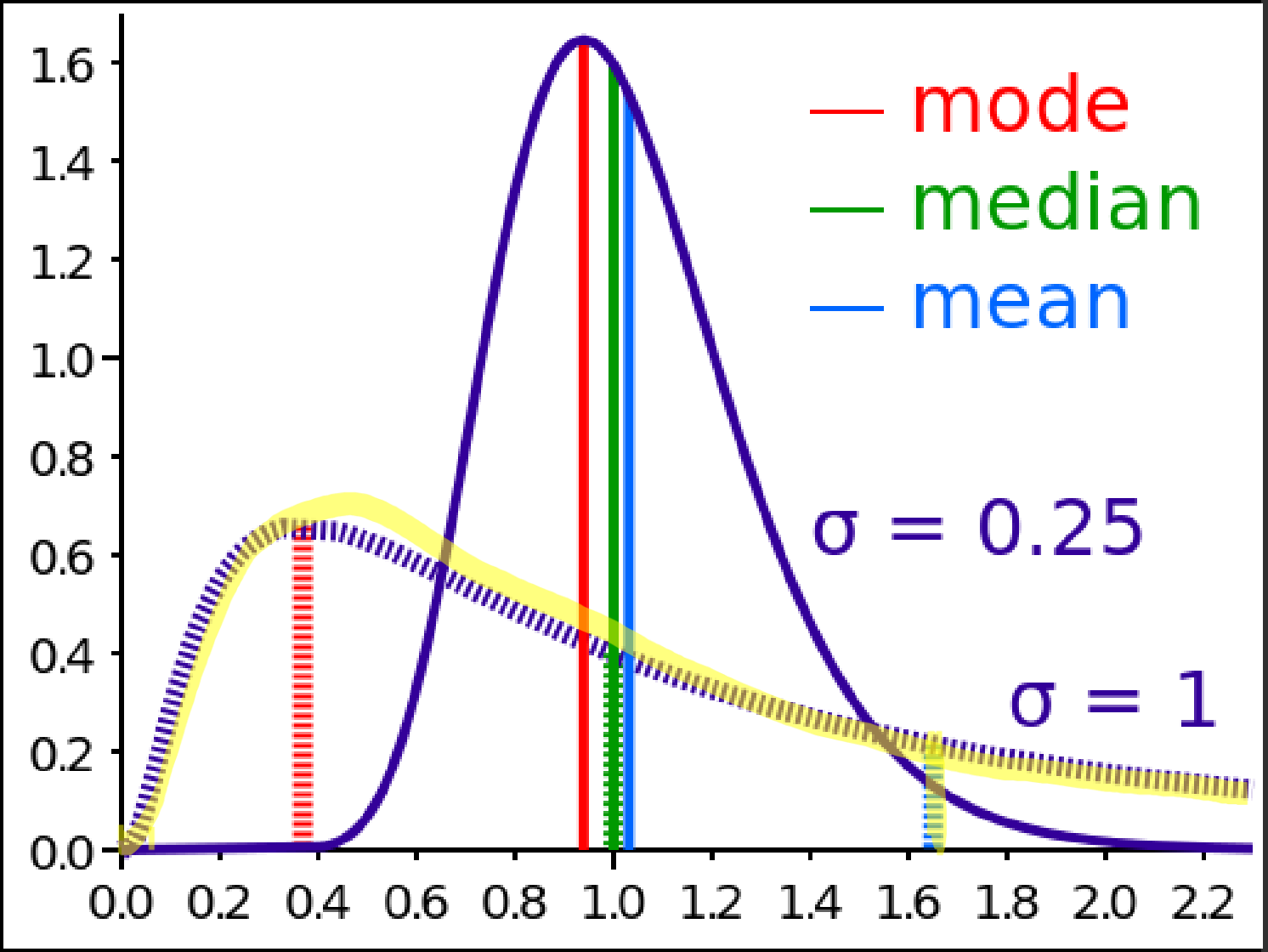
Central Limit Theorem



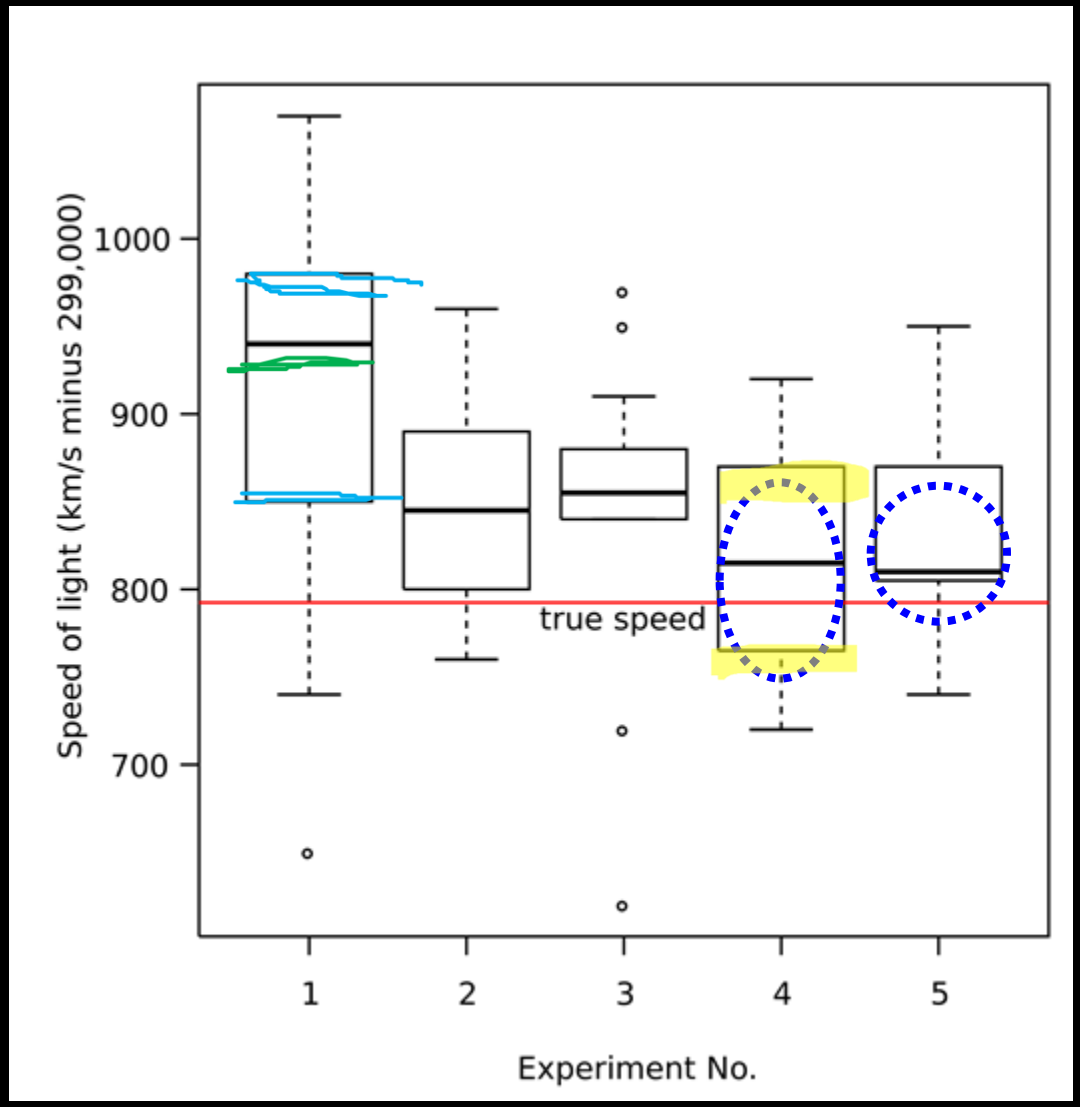
- A sample is obtained containing many observations, each observation being randomly generated in a way that does not depend on the values of the other observations, and that the arithmetic mean of the observed values is computed.
- If this procedure is performed many times, the central limit theorem says that the probability distribution of the average will closely approximate a normal distribution.

If $X_1, X_2, \ldots, X_n$ are n random samples drawn from a population with overall mean μ and finite variance σ^2, and if $\bar{X}_n$ is the sample mean, then the limiting form of the distribution, $Z = \lim_{n\to\infty} \sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right)$, is a standard normal distribution.

Central Tendency



Box plot


Speed of light (km/s minus 299,000)
1000
900
800
700
true speed
1
2
3
4
5
Experiment No.

User Trial

Remember: This presentation should not be taken as a replacement of a standard statistics course.

- Null hypothesis: No difference
- Alternative hypothesis: There is difference

Variables

- Variables are things that change

- The **independent variable** is the variable that is purposely changed. It is the manipulated variable.

- The **dependent variable** changes in response to the independent variable. It is the responding variable.

Variables

Constant Variables

- Factors that are kept the same and not allowed to change.

- It is important to control all but one variable at a time to be able to interpret data

Hypothesis

- Your best thinking about how the change you make might affect another factor.
- Tentative or trial solution to the question.
- An if ..... then ..... statement.
- Should be expressed in **measurable** terms

How to test

- Sampling: Select a set of participants
- Method: Design a study to collect data
- Material: Get instruments
- Procedure: Collect data from participants
- Result: Analyze result

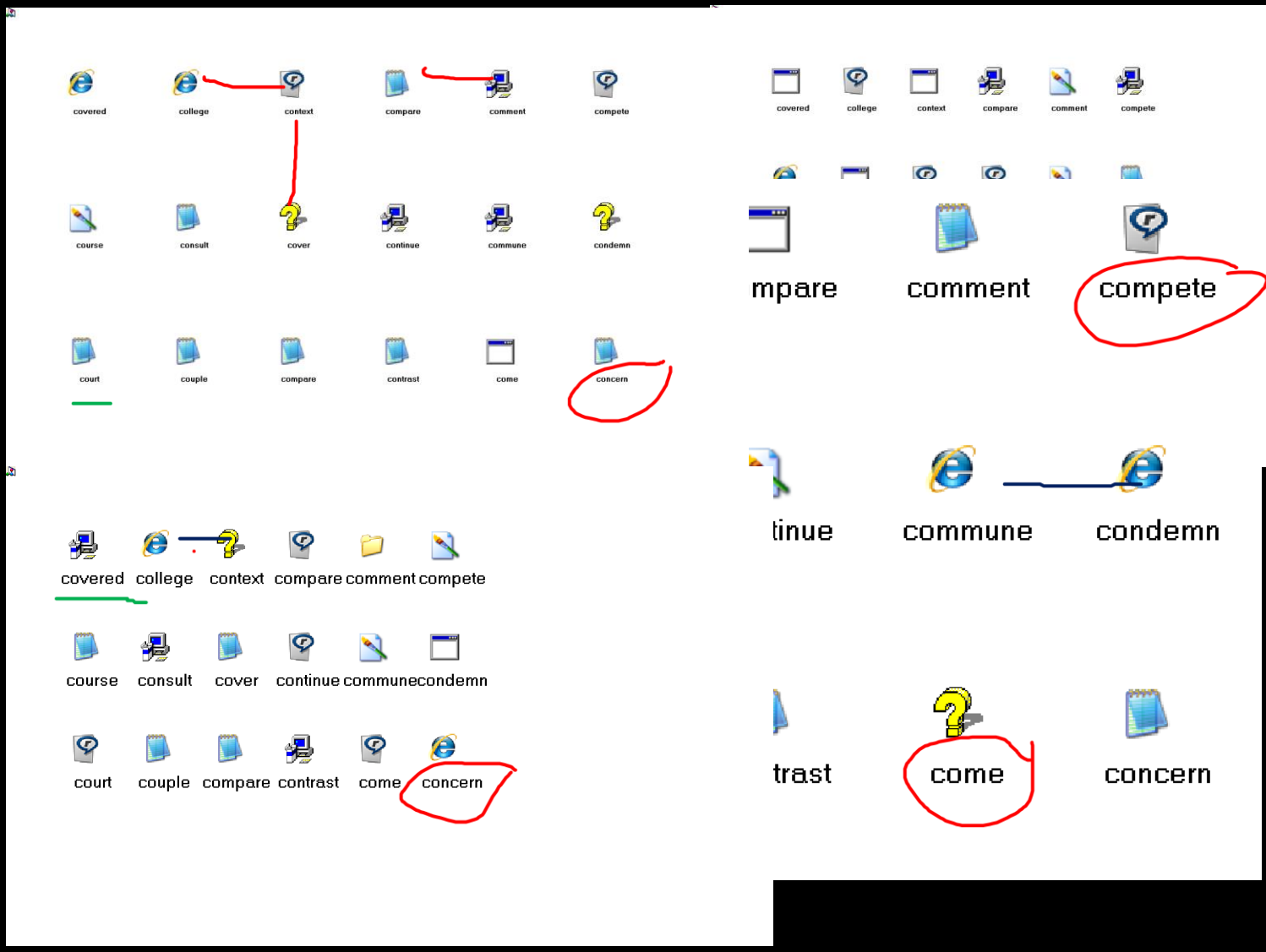
Result

$$\textbf{Test Statistics} = \frac{\text{Variance explained by model}}{\text{Variance not explained by model}} = \frac{\textbf{Effect}}{\textbf{Error}}$$

Significant → The probability that the model is explaining variance by chance < 0.05

Case Study

Which interface reduces interaction time



Independent variables

- Spacing × FontSize × Group

- Spacing
  - Sparse
  - Medium
  - Dense

- FontSize
  - 10 pt
  - 14 pt
  - 20pt

- Group
  - Able-bodied
  - Visually-impaired
  - Motor-impaired

Dependent variables

- Pointing time
  - Time needed to click on an icon
- Number of wrong selection

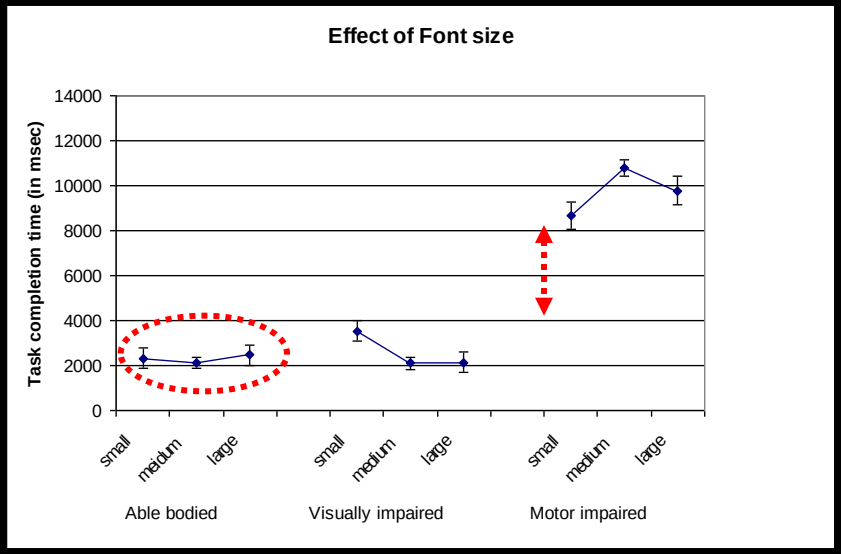
Hypothesis

Increasing *font size* and *button spacing* will reduce p*ointing time* and *number of errors*

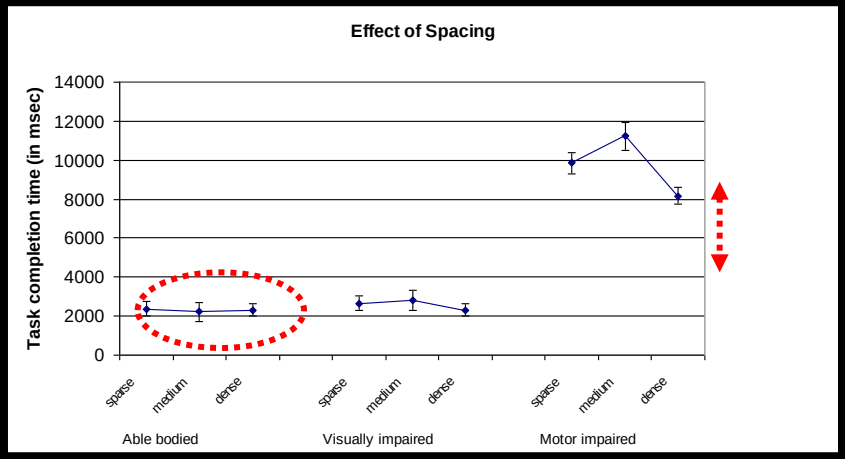
Experiment

- Show an icon to user
- Instruct him to remember it
- Show a list of icons
- Ask him to find and click on the icon
- Measure the time interval between presentation of the set of icons and click instance

Result

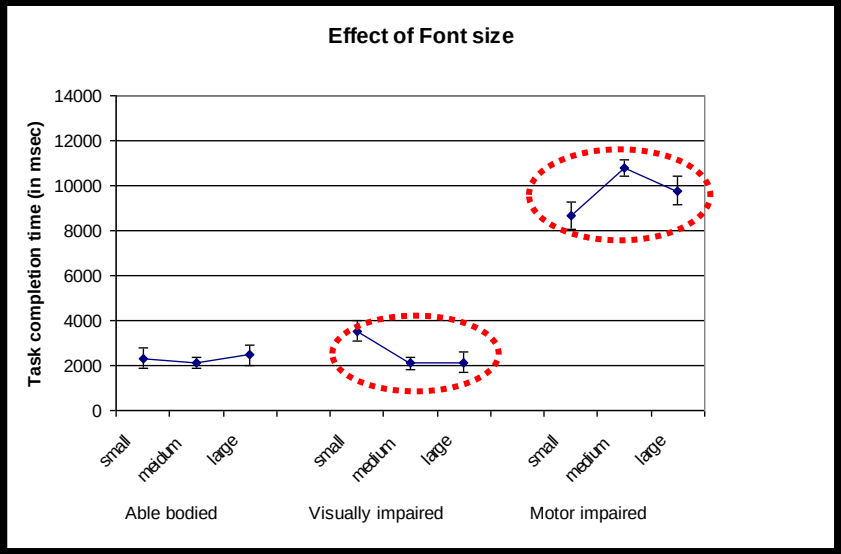


- Motor impaired users took more time to point

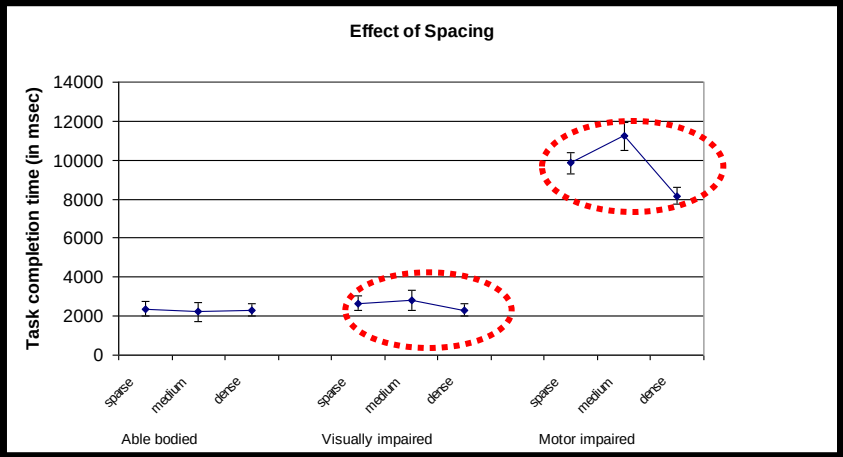


- Able bodied users did not show any effect for change in font size and spacing

Result



- Visually impaired users took less time to point for bigger (medium or large) font size and dense spacing



- Motor impaired users took less time in dense spacing and small font size

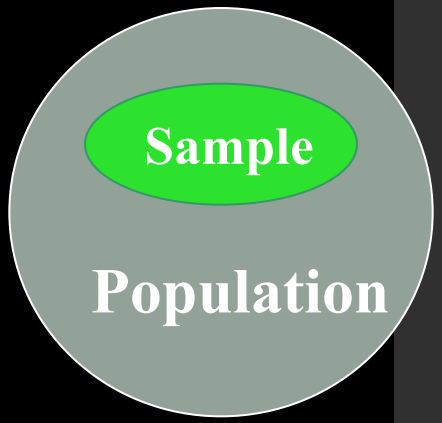
Theoretical detail

How to proceed

- Generate research question
  - Unstructured or semi-structured interview
  - Observational methods
  - Coding and theoritizing (discussed later)
- Postulate hypothesis
  - Define variables
  - Design experiment
- Measure
  - Conduct experiment
  - Analyse result

Sampling

- Size
  - No straight forward answer !!
  - Can be estimated statistically
  - Bigger the better
    - more representative of population
  - Often limited by availability



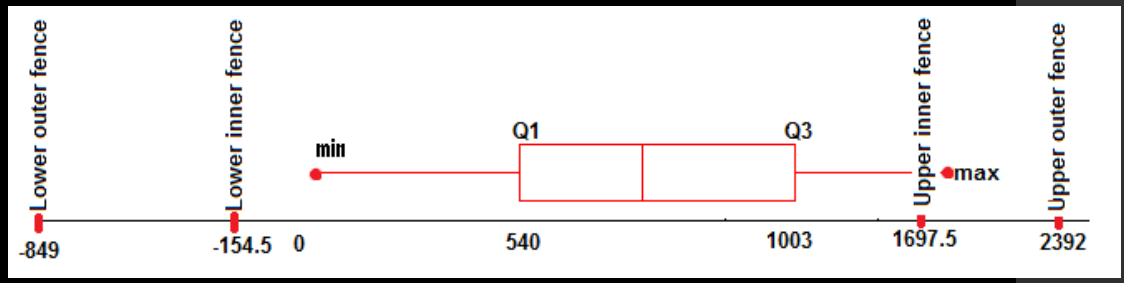
- Quality
  - Random sampling
  - Group based sampling
  - Purpose based sampling

Experiment

- Matched pair
- Repeated measure
- Latin square

Data Screening

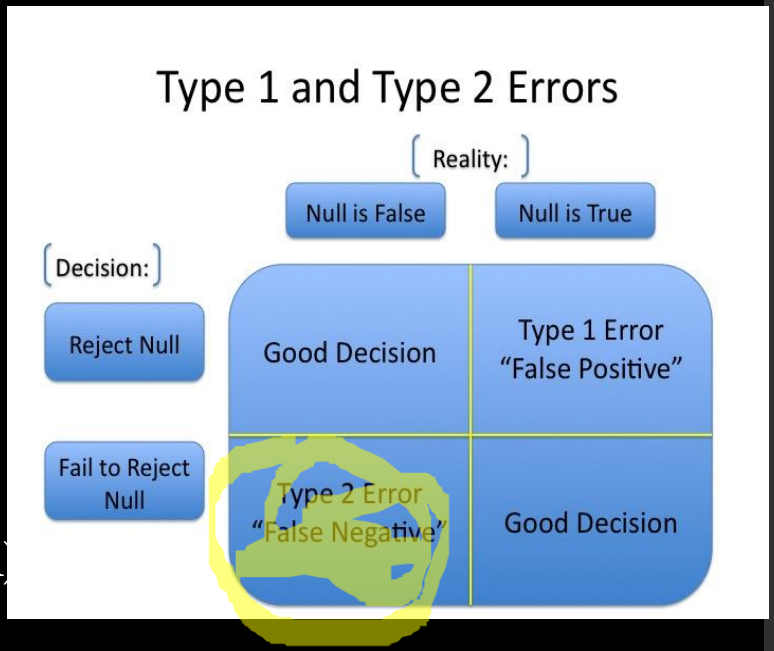
- Outliers
  - Inner Fence
  - Outer Fence
- Skewing
  - In opposite direction
- Unequal Variance
- Missing Values
- Data Transformation



Data Analysis

Important terms

- Degrees of freedom (df)
- One tail and two tail tests
  - Better/Worse or just different
- Type I (α) and Type II (β) error
- Sphericity assumption (for ANOVA)



Just google the terms in case you forget later

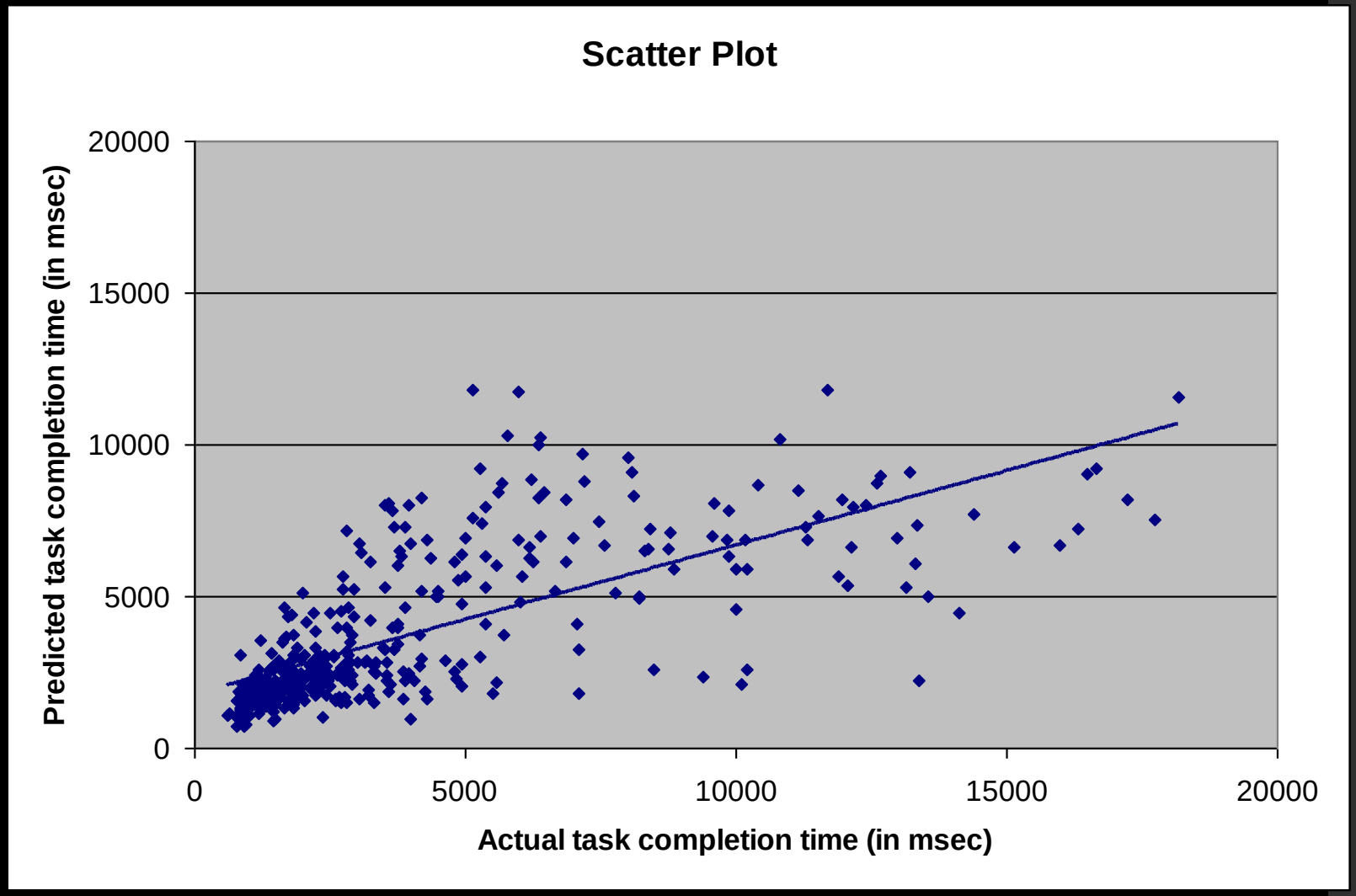
Test Selection

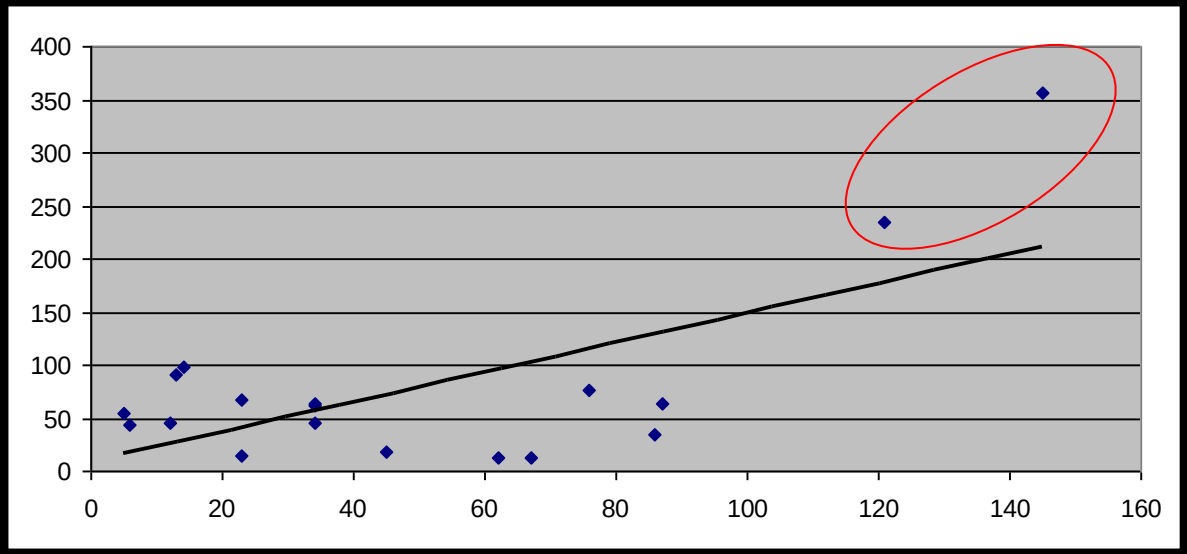
- Data normally distributed - Kolmogorov-Smirnov (K-S) test, Shapiro-Wilk test
  - Parametric / Non-parametric
- Relationship between two columns of data
  - Correlation (Pearson's *r* / Spearman's ρ)
- Comparing means between two columns of data
  - T-test / Mann-Whitney U-Test / Wilcoxon signed rank test
- More than two columns
  - ANOVA / Kruskal-Wallis H test /Scheirer Ray Hare Test
- Categorical Data – Chi Square

Cont...

Independent

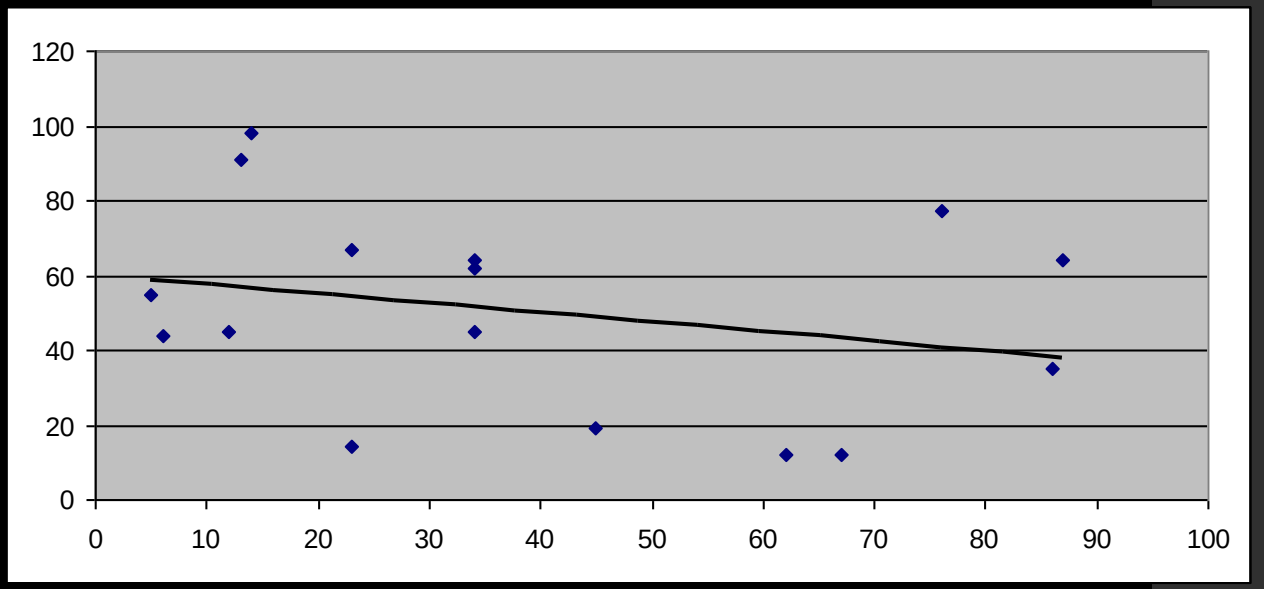
Dependent

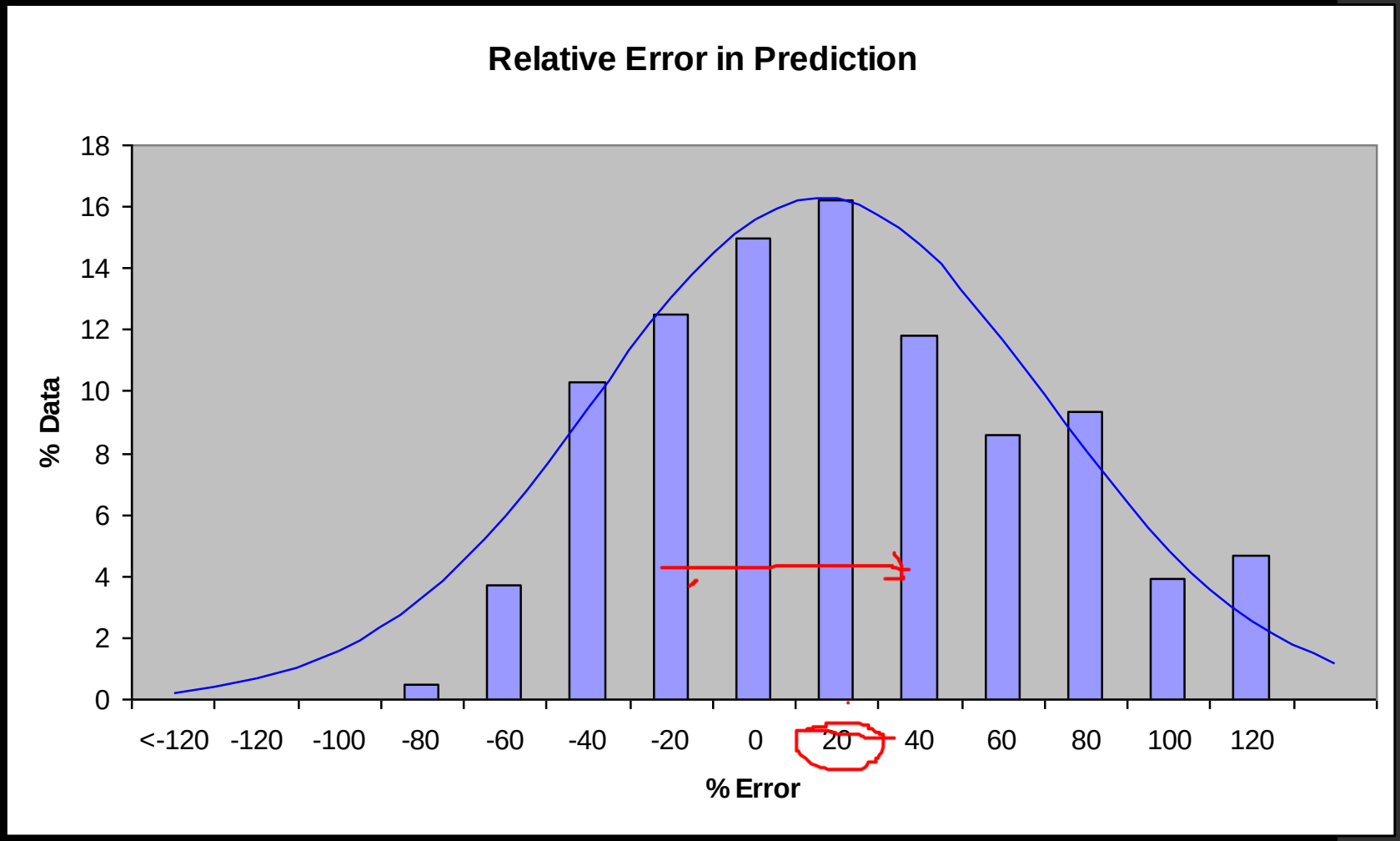

Scatter Plot
Predicted task completion time (in msec)
20000
15000
10000
5000
0
0
5000
10000
15000
20000
Actual task completion time (in msec)


400
350
300
250
200
150
100
50
0
0
20
40
60
80
100
120
140
160

$r = 0.66$

$r = -0.27$


120
100
80
60
40
20
0
0
10
20
30
40
50
60
70
80
90
100


Relative Error in Prediction
% Data
18
16
14
12
10
8
6
4
2
0
<-120
-120
-100
-80
-60
-40
-20
0
20
40
60
80
100
120
% Error

Test Selection

- Data normally distributed - Kolmogorov-Smirnov (K-S) test, Shapiro-Wilk test
  - Parametric / Non-parametric
- Relationship between two columns of data
  - Correlation (Pearson's r / Spearman's ρ)
- Comparing means between two columns of data
  - T-test / Mann-Whitney U-Test / Wilcoxon signed rank test
- More than two columns
  - ANOVA / Kruskal-Wallis H test /Scheirer Ray Hare Test
- Categorical Data – Chi Square

Comparing means – t-test

	CTRL	EXP
P1	6	10
P2	1	6
P3	2	8
P4	9	10
P5	6	8
P6	10	10
P7	6	8
P8	10	10
P9	10	10
P10	10	10
P11	10	10
P12	9	8
AVG	**7.42**	**9.00**
Stdev	**3.23**	**1.35**
Ttest		120,2,1)

Function Arguments

TTEST

Array1 G9:G20 = {6;1;2;9;6;10;6;10;

Array2 H9:H20 = {10;6;8;10;8;10;8;1

Tails 2 = 2

Type 1 = 1

= 0.034538833

Returns the probability associated with a Student's t-Test.

Type is the kind of t-test: paired = 1, two-sample equal variance (homoscedastic) = 2, two-sample unequal variance = 3.

Formula result = 0.034538833

Help on this function

OK Cancel

The basic ANOVA situation

Two variables: 1 Categorical, 1 Quantitative

Main Question: Do the (means of) the quantitative variables depend on which group (given by categorical variable) the individual is in?

If categorical variable has only 2 values:

- 2-sample t-test

ANOVA allows for 3 or more groups

An example ANOVA situation

Subjects: 25 patients with blisters
Treatments: Treatment A, Treatment B, Placebo
Measurement: # of days until blisters heal

Data [and means]:

- A: 5,6,6,7,7,8,9,10 [7.25]
- B: 7,7,8,9,9,10,10,11 [8.875]
- P: 7,9,9,10,10,10,11,12,13 [10.11]

Are these differences significant?

Informal Investigation

Graphical investigation:

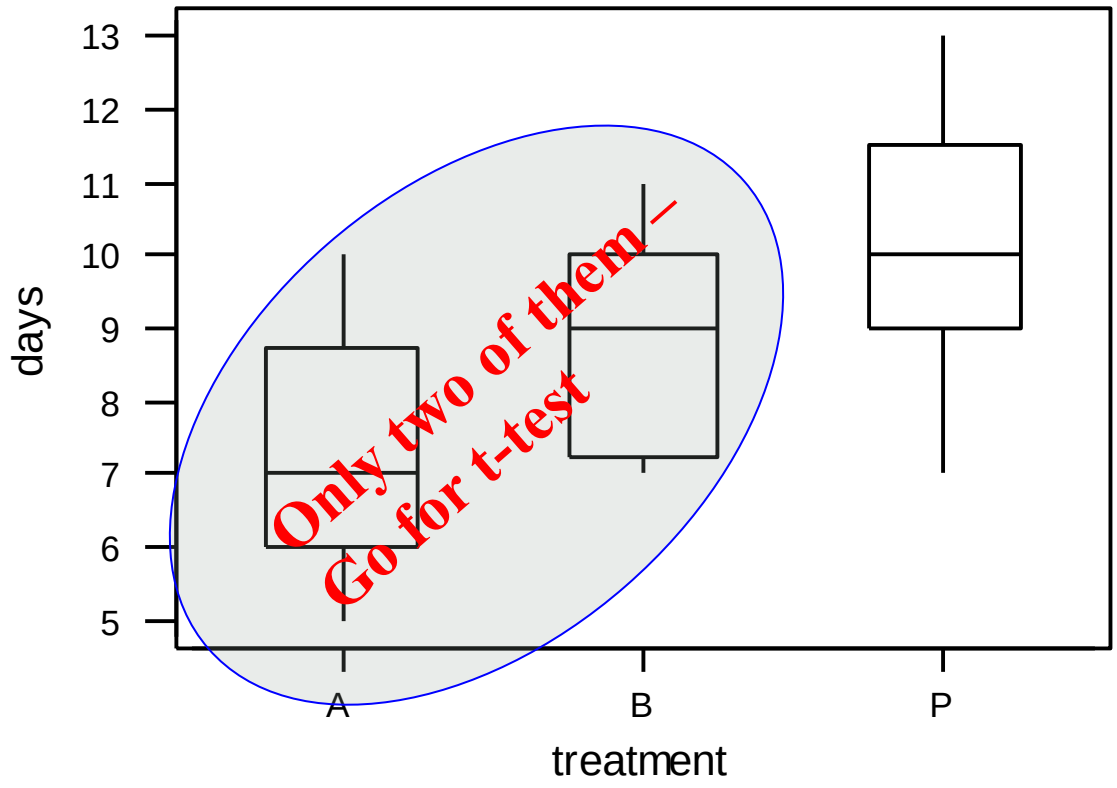
- side-by-side box plots
- multiple histograms

Whether the differences between the groups are significant depends on

- the difference in the means
- the standard deviations of each group
- the sample sizes

ANOVA determines P-value from the F statistic

Side by Side Boxplots



What does ANOVA do?

At its simplest (there are extensions) ANOVA tests the following hypotheses:

H_0: The means of all the groups are equal.

H_a: Not all the means are equal

- doesn't say how or which ones differ.
- Can follow up with "multiple comparisons"

Note: we usually refer to the sub-populations as "groups" when doing ANOVA.

Assumptions of ANOVA

- each group is approximately normal
  - check this by looking at histograms and/or normal quantile plots, or use assumptions
  - can handle some nonnormality, but not severe outliers

- standard deviations of each group are approximately equal
  - rule of thumb: ratio of largest to smallest sample st. dev. must be less than 2:1

Normality Check

We should check for normality using:

- assumptions about population
- histograms for each group
- normal quantile plot for each group

With such small data sets, there really isn't a really good way to check normality from data, but we make the common assumption that physical measurements of people tend to be normally distributed.

Standard Deviation Check

Variable	treatment	N	Mean	Median	StDev
days	A	8	7.250	7.000	1.669
	B	8	8.875	9.000	1.458
	P	9	10.111	10.000	1.764

Compare largest and smallest standard deviations:

- largest: 1.764
- smallest: 1.458
- 1.458 x 2 = 2.916 > 1.764

Note: variance ratio of 4:1 is equivalent.

How ANOVA works (outline)

ANOVA measures two sources of variation in the data and compares their relative sizes

- variation BETWEEN groups
  - for each data value look at the difference between its group mean and the overall mean

$$\left(\bar{x}_i - \bar{x}\right)^2$$

- variation WITHIN groups
  - for each data value we look at the difference between that value and the mean of its group

$$\left(x_{ij} - \bar{x}_i\right)^2$$

F- statistics

The ANOVA F-statistic is a ratio of the Between Group Variation divided by the Within Group Variation:

$$F = \frac{Between}{Within} = \frac{MSG}{MSE}$$

A large F is evidence *against* H_0, since it indicates that there is more difference between groups than within groups.

ANOVA Output

$F(2,22) = 6.45,\ p < 0.05$

Analysis of Variance for days

Source	DF	SS	MS	F	P
treatment	2	34.74	17.37	6.45	0.006
Error	22	59.26	2.69		
Total	24	94.00			

MS = SS /df

F = MSG /MSE

ANOVA result – Previous example

Source	Sum of Squares	df	Mean Square	F	Sig.
LAYOUT	**72183079.23**	**1.745**	**41367738**	**5.435**	**0.009**
LAYOUT * IMP	**83606114.83**	**3.49**	**23957109**	**3.148**	**0.025**
Error(LAYOUT)	491364031.8	64.562	7610760		
FONT	2534752.503	1.702	1489441	0.217	0.77
FONT * IMP	**117226676.6**	**3.404**	**34441665**	**5.017**	**0.002**
Error(FONT)	432286058.5	62.967	6865263		
LAYOUT * FONT	47748673.96	2.331	20482440	1.03	0.37
LAYOUT * FONT * IMP	76535737.42	4.662	16415520	0.825	0.528
Error(LAYOUT*FONT)	1715764114	86.254	19891898		

Effect size and Power

- Effect size
  - Percent of variance explained
  - Standardized measure of magnitude of effect
  - Cohen's *d*, correlation coefficient, η^2

- Power
  - Power of a test to detect significant effect
  - (1 – Type II error)
    - Type II error (β) → probability of not detecting an effect
  - Can be used to estimate sample size

Other tests

Other important tests won't be discussed in detail but relevant to HCI trials

- Non Gaussian distribution → Non parametric tests
- Comparing ranks → Sign test
- ANCOVA
- MANOVA and so on

Qualitative data

- Supplement and illustrate quantitative data
- No clear or single convention of data analysis
- Can be collected from different sources
  - Observational methods
  - Interview
  - Transcript
  - Written documents, reports

Steps to analyse

- Coding
- Adding memo
- Content analysis to find
  - Similar phrases
  - Patterns
  - Themes
  - Relationship
- Elaborating initial set of generalizations
- Linking generalizations to theory

Content analysis

- Key word in context
- Word frequency list
- Category counts
- Collocations

Non-numerical Unstructured Data Indexing, Searching and Theorizing (NUD*IST software)

Reporting

- Title
- Abstract
- Introduction
- Method
  - Participants
  - Materials
  - Design
  - Procedure
- Results
- Discussion
- References
- Appendix (Optional)

Take away points

- Introduction to the process of conducting a user trial
- Basic quantitative and qualitative data analysis techniques
- Basic statistical methods and terms associated with conducting controlled experiment
- Reporting a study following standard format