Introduction to HCI

Statistical Analysis

Prof. Ali Sarvghad
UMass Amherst

asarvr@cs.umass.edu
Courses, projects, papers, and more:
http://groups.cs.umass.edu/nmahyar/

Introduction

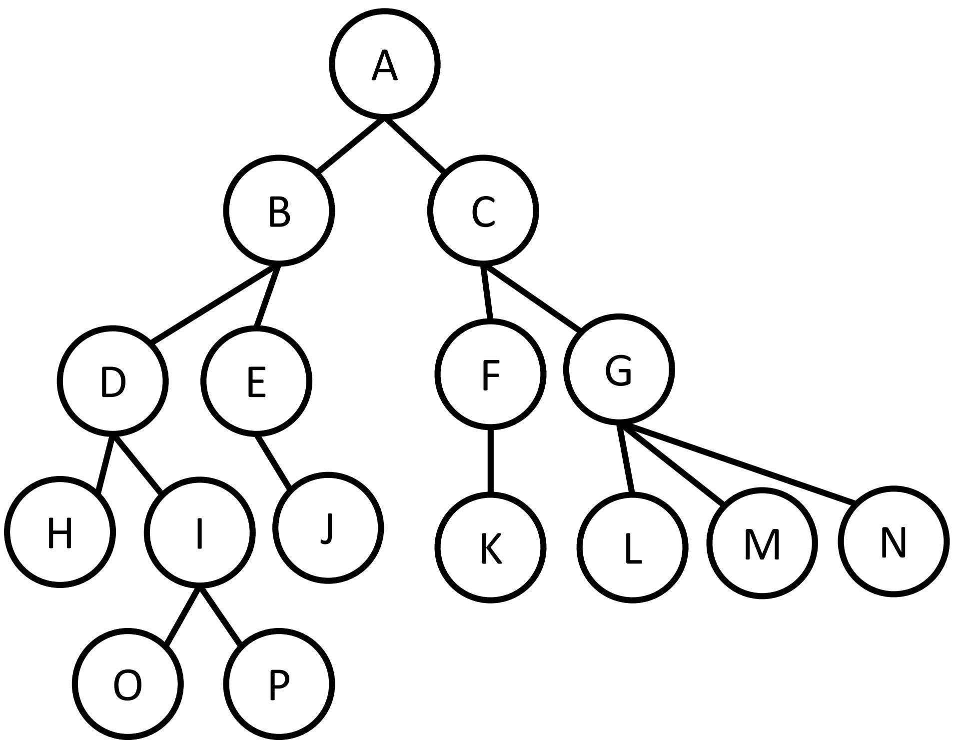
- **Inferential statistics** methods for hypothesis testing:
    - t-test
    - ANOVA
- You need to know:
    - Basics of descriptive statistics
        - Mean
        - variance
        - Standard deviation
    - Normal distribution
    - Basics of probability

Introduction

- We will use an example of a **designed experiment** to talk about t-test & ANOVA
- In a designed experiment, you can establish **causation**
- Control some factors (**independent variables**) to find their effects on some other factors (**dependent variables**)

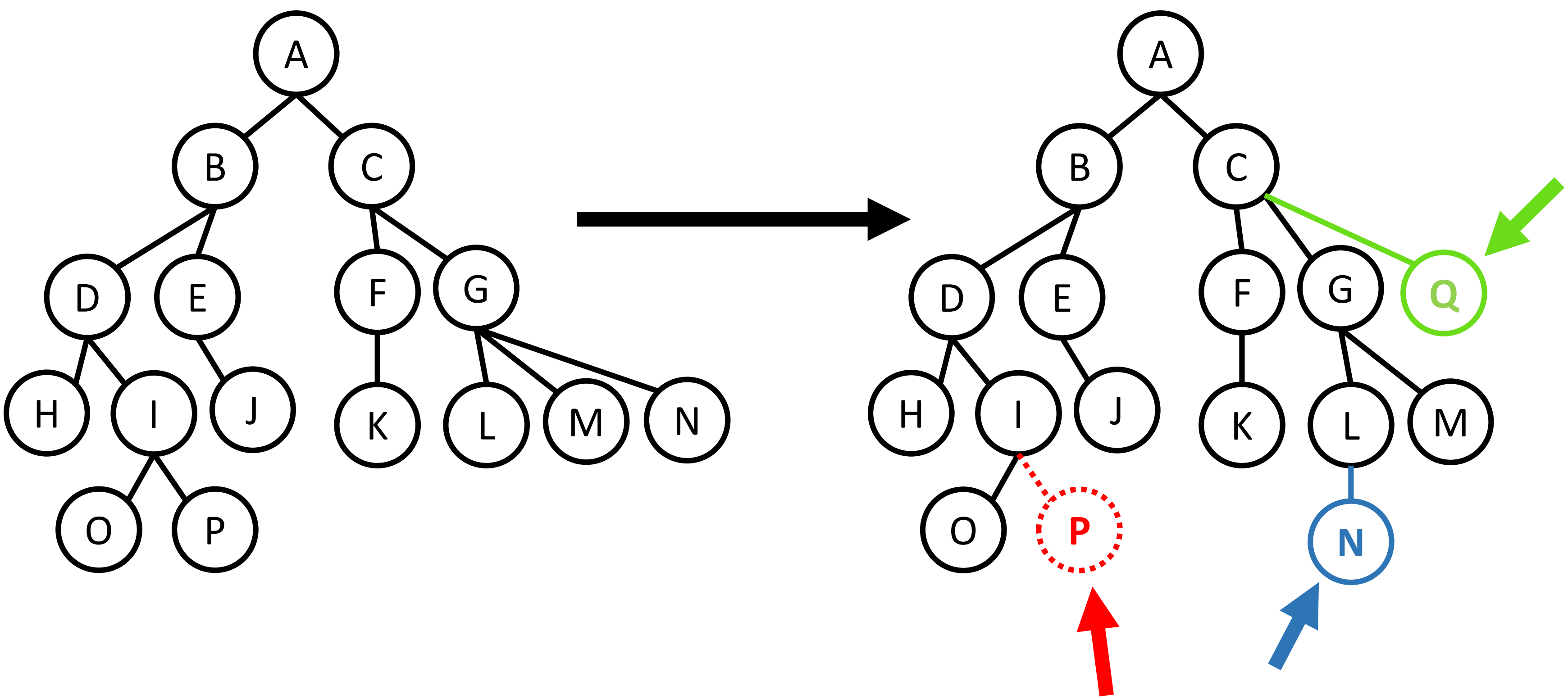
Example Designed Experiment

We designed a new technique for visually compare and understating changes in a hierarchical data



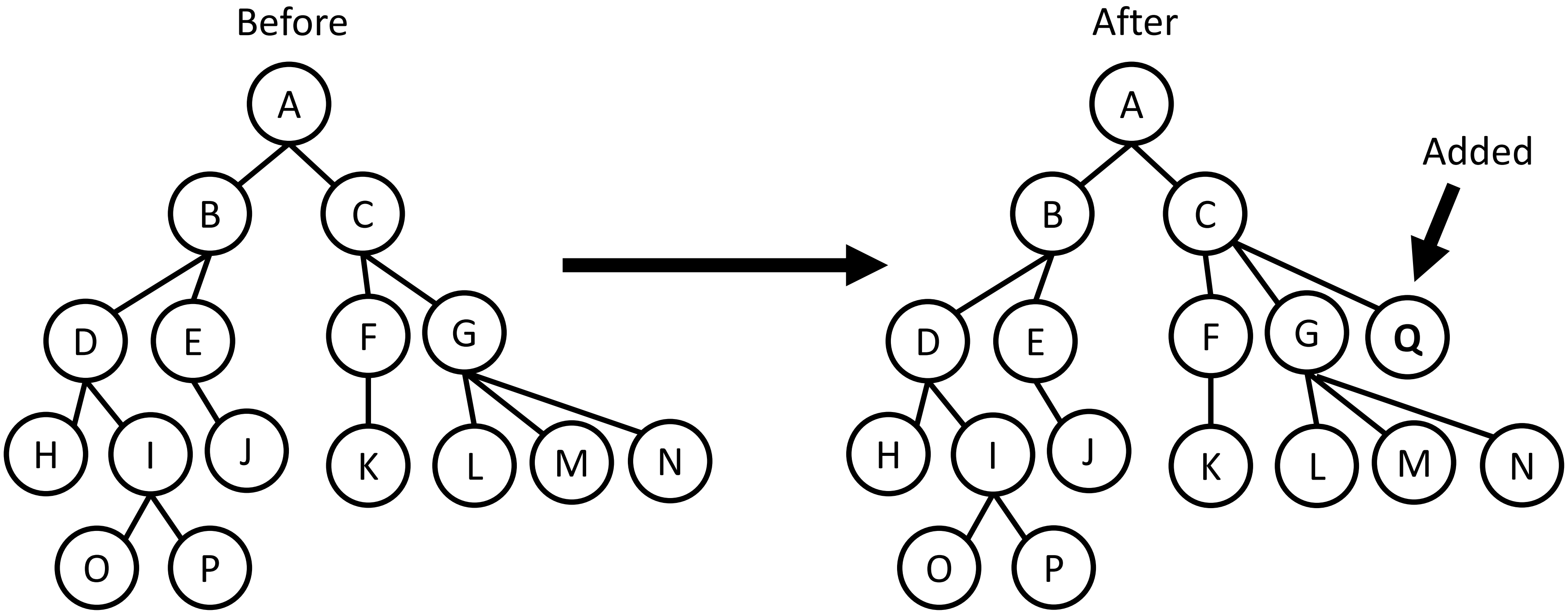
Example Designed Experiment

Change type**:** a node can be : **deleted**, **moved**, **added**



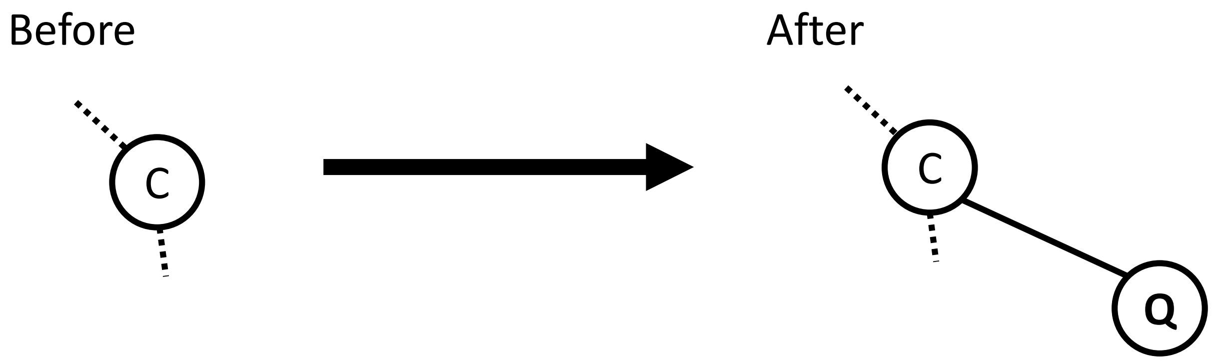
Example Designed Experiment

Comparison technique 1: **side by side**



Example Designed Experiment

Comparison technique2 : **reduced side by side (RSS)**

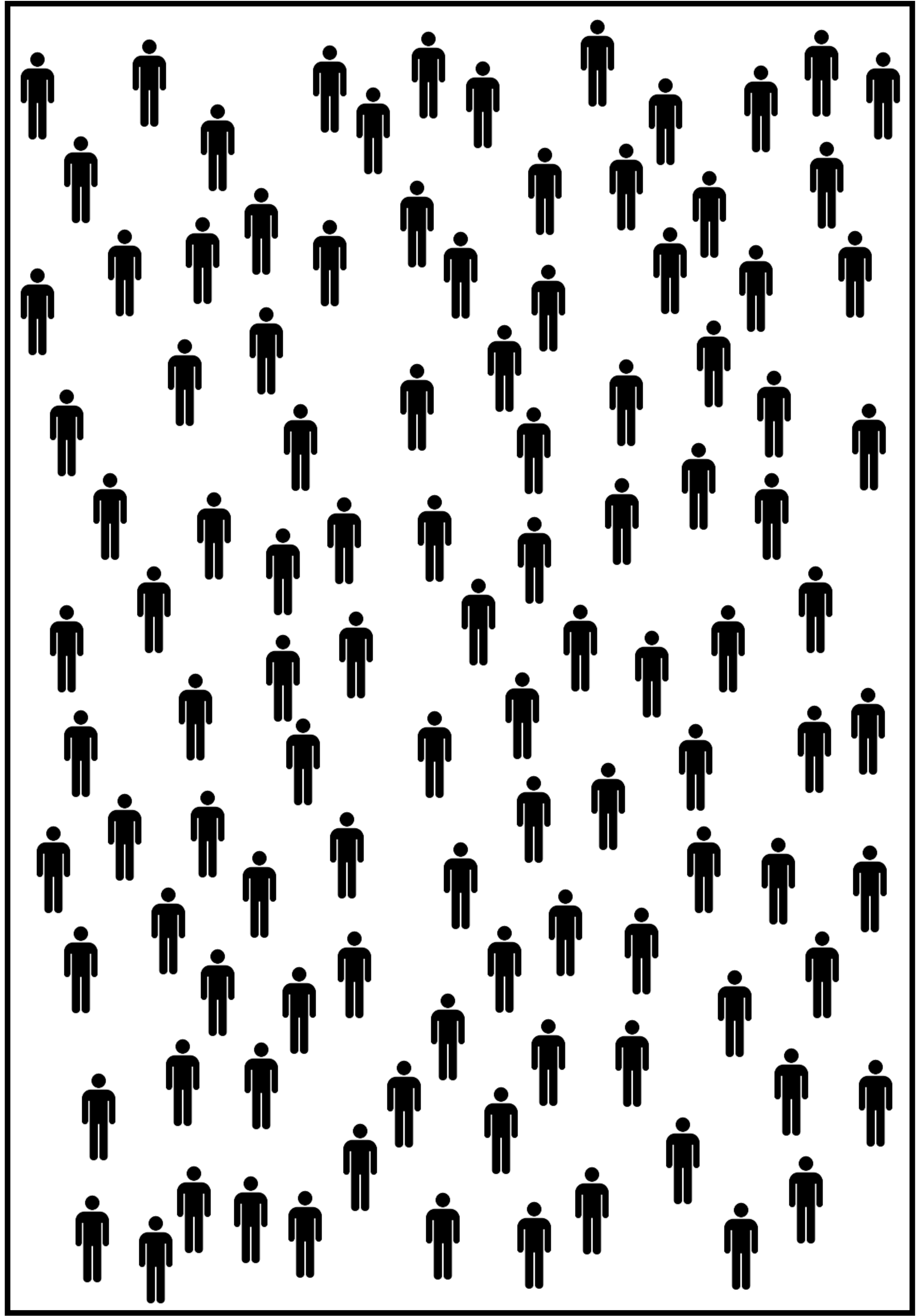


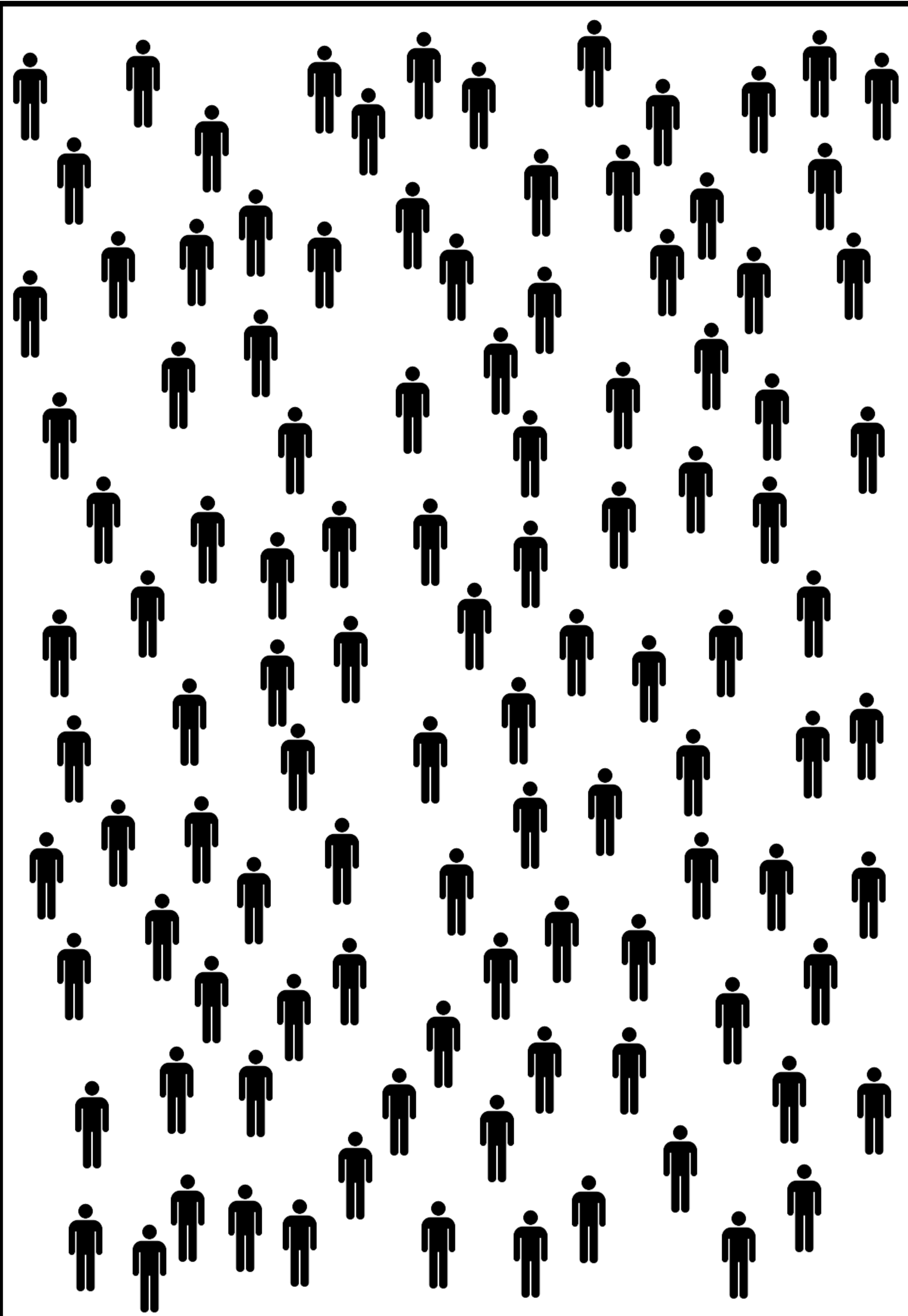
Controlled Lab Experiment

- **Goal:** Compare Side by Side , and Reduced Side by Side (RSS) techniques
- **H0:** There will be no difference between the two conditions
- **H1:** Users will be **faster** to identify the change using RSS
- Measured
  - Performance **time**

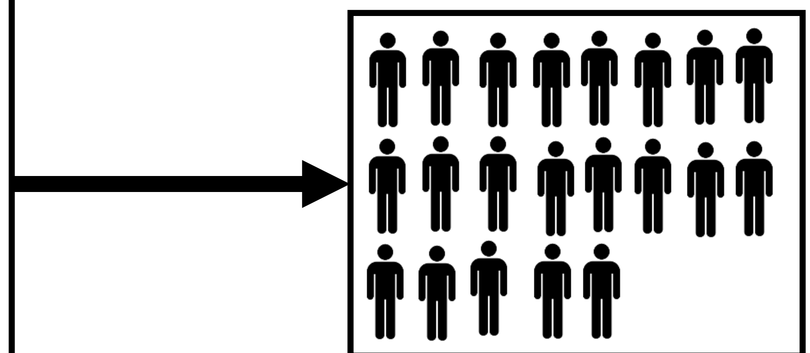
Target population.

It is often not possible to access and involve the entire target population in your study.

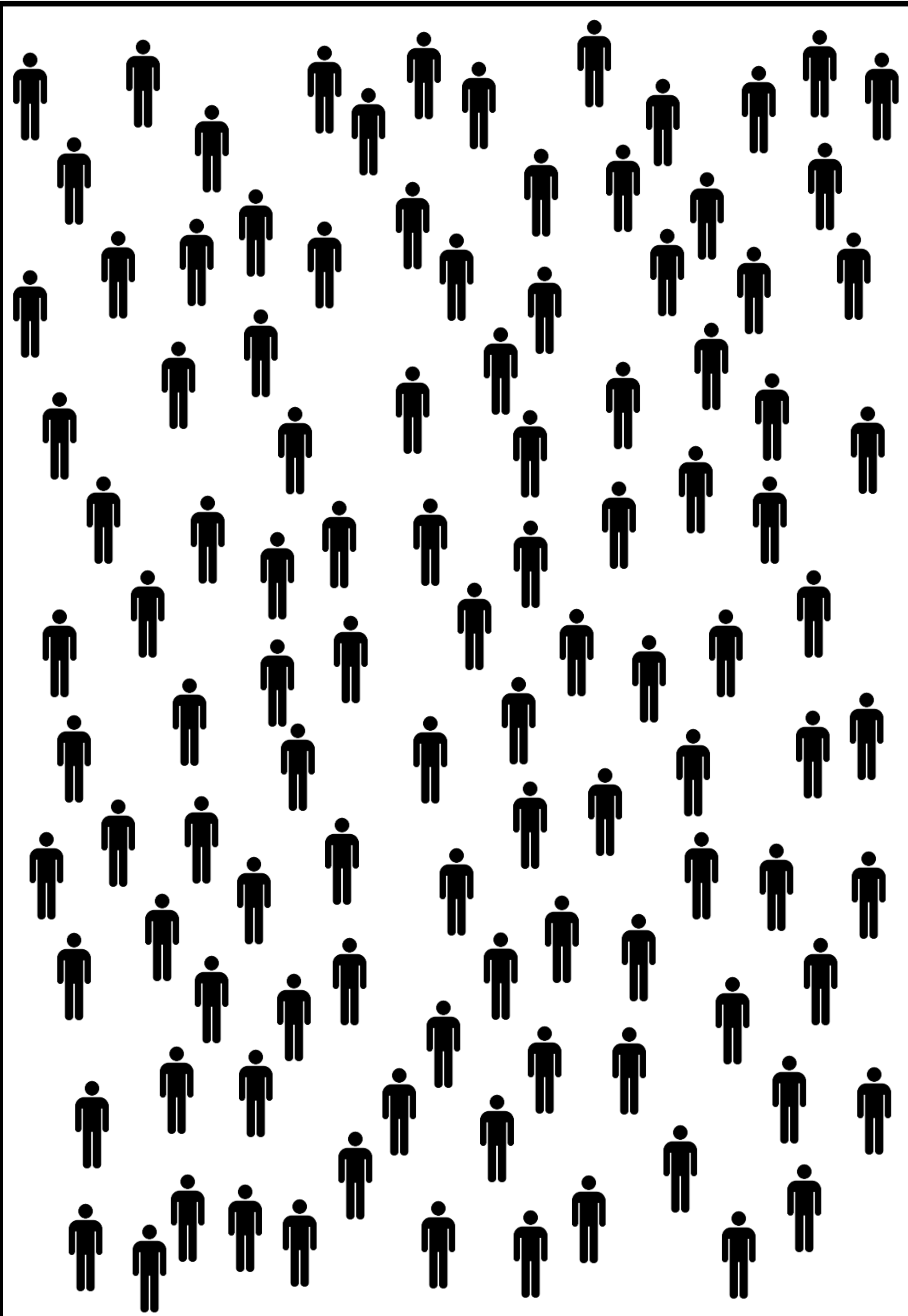




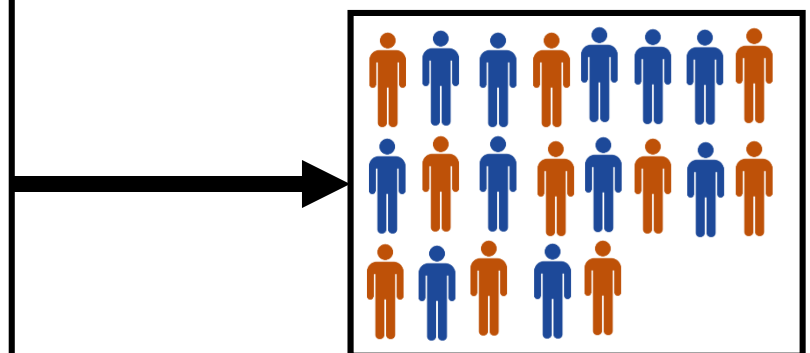
Sampling is the technique that we use when we can't access the entire target population.



Example, a sample of size = 20



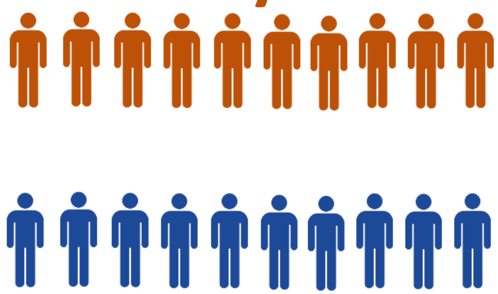
Randomly assign participants to your two experimental condition



Condition 1, Side by side, 10

Condition 2, RSS, 10

Condition 1,
Side by Side



Condition 2,
Reduce Side by Side

Study:

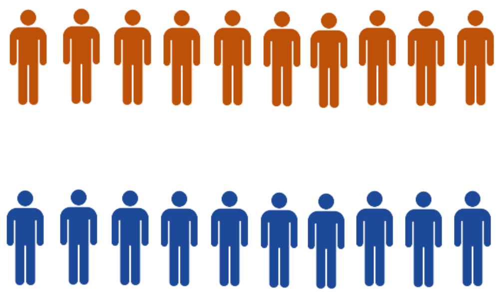
All the participant worked one the **same set of tasks** identifying changes in hierarchical data.

Participants in **condition 1** used **Side by Side** to identify the changes and participants in **condition 2** used **Reduced Side by Side**.

We collected **time** and **accuracy per task** for each participant.

Average time and accuracy for performing for the two conditions

Side by side



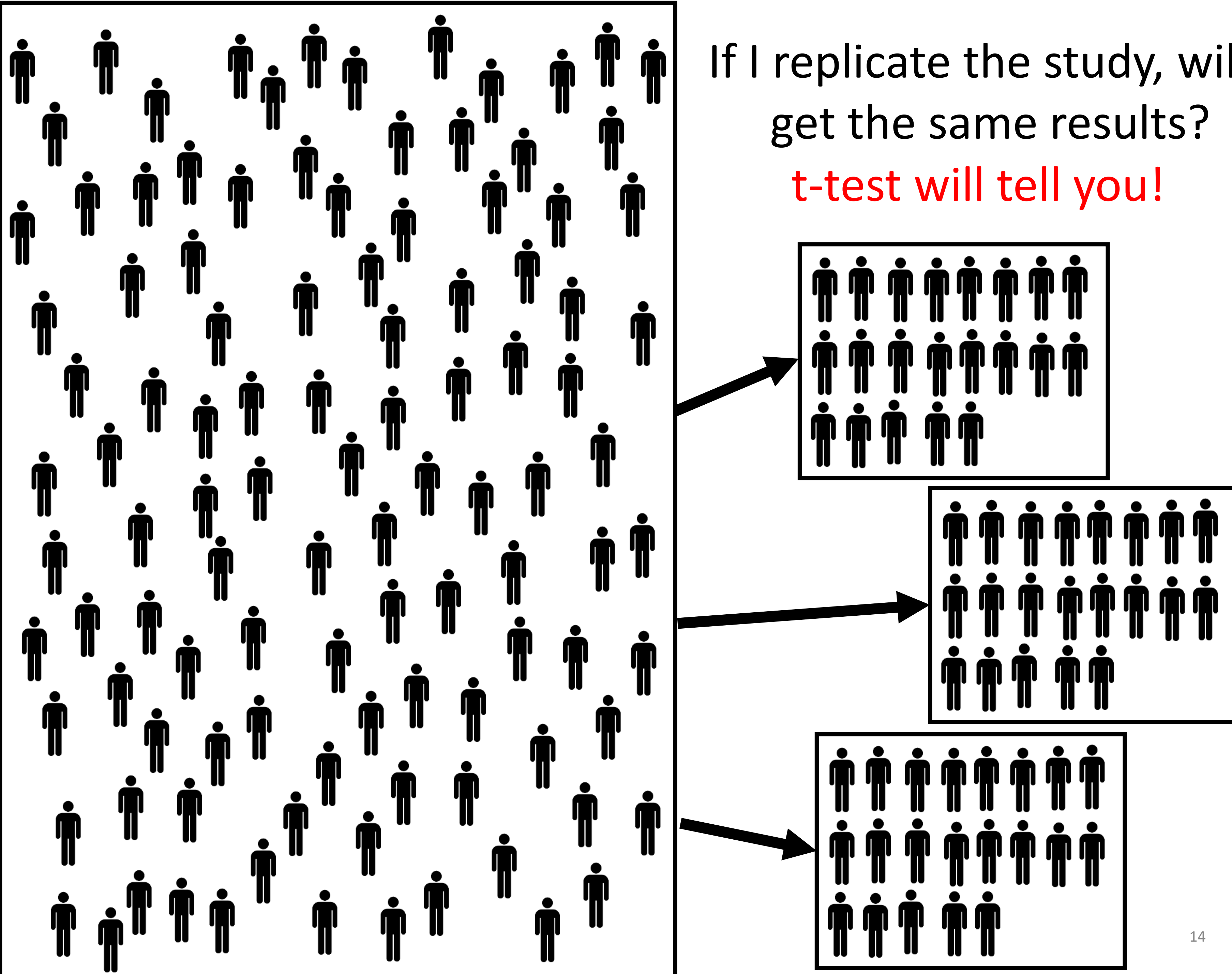
Reduced side by side

	Time
Avg.	**36.0 , 19.25**
StDev.	**21.12 , 18.12**

Is the difference between the time averages **significant**?

If I replicate the study, will I get the same results?

t-test will tell you!



T-Test

T-test tells you the **probability (p-value)** of getting the same outcomes if you replicate your experiments with a different sample from the target population

T-Test

Avg. Condition 1

Avg. Condition 2

$$t = \frac{\bar{X}1 - \bar{X}2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Variance Condition 1

Variance Condition 2

samples in each condition

Formula for independent samples, **df = n1 + n2 - 2**

Side by side

Reduced side by side

	Time
Avg.	**36.0 , 19.25**
StDev.	**21.12 , 18.12**
n	**10 , 10**

$$t = \frac{\overline{X}1 - \overline{X}2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

t value = 1.93

What does it mean?

How do we calculate the p value?

t-test

We use a number called **critical value** to decide whether we reject the null hypothesis based on our t value.

Our **t value < critical value**, we **don’t reject** the null hypothesis

Our **t value > critical value,** we r**eject** the null hypothesis

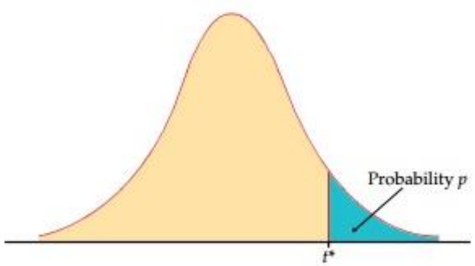
Degrees of freedom (df)

df = (n1 + n2) -2

df = 10+10–1 = 18

Significance threshold

One-tailed t-table



df	.25	.20	.15	.10	.05	.025	.02
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197

Degrees of freedom: https://statisticsbyjim.com/hypothesis-testing/degrees-freedom-statistics/

Side by side

Reduced side by side

	Time
Avg.	**36.0 , 19.25**
StDev.	**21.12 , 18.12**
n	**10 , 10**

t value **1.903 > 1.734** critical value

p-value = .03476 < .05

We **CAN** reject the null hypothesis!

In other words, the difference between **36.0** and **19.25** is significant and **not random**

t-test: Important points to note

There are fundamental questions you ask before doing a t-test:

1. Is your data is normally distributed?
2. Do you have enough samples? (Ideally between 20-30)
3. Are you doing a **two-tailed** or **one-tailed** t-test?
4. Is data **paired** or **unpaired** (independent)?

Unpaired & Paired Samples

Comparing two sets of unpaired observations

Usually different subjects in each group (number may differ as well)

Condition 1	condition 2
S1–s20	s21–s43

Which one is within-subject? Between-subject?

Paired observations

usually single group studied under separate experimental conditions

Data points of one subject are treated as a pair

Condition 1	condition 2
S1–s20	s1–s20

T-Test

The mean difference

$$t = \frac{\bar{d}}{\sqrt{\frac{s^2}{n}}}$$

Sample variance

Samples size

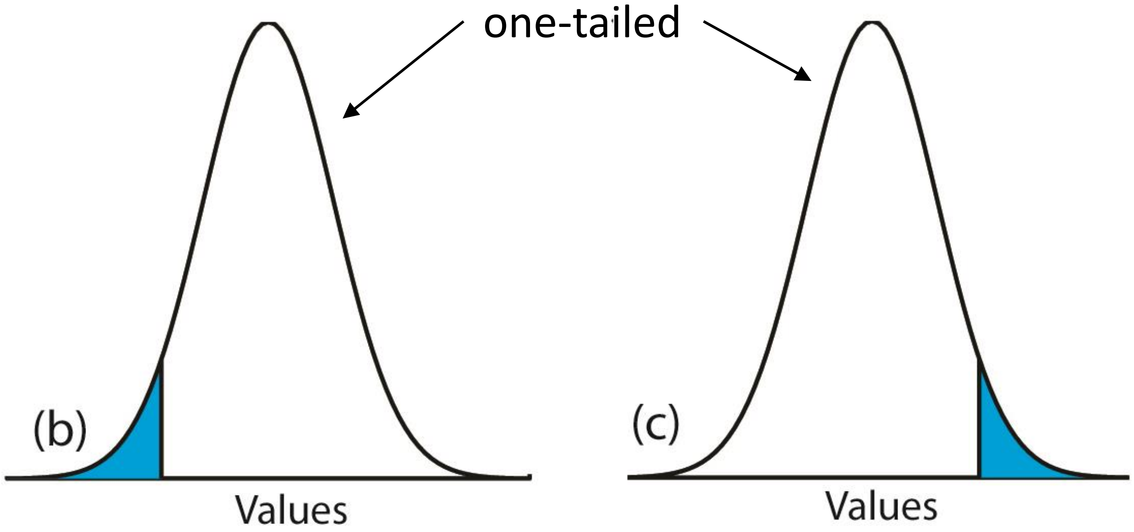
Formula for paired samples, **df = n -1**

t-test: One-tailed

If you have **two sample means, A & B:**

You do a one-tailed test when you restrict your null hypothesize is **A<B, A>B**,…

Example, the average height 8 year of boys is less than the average height of 8yeras old girls.



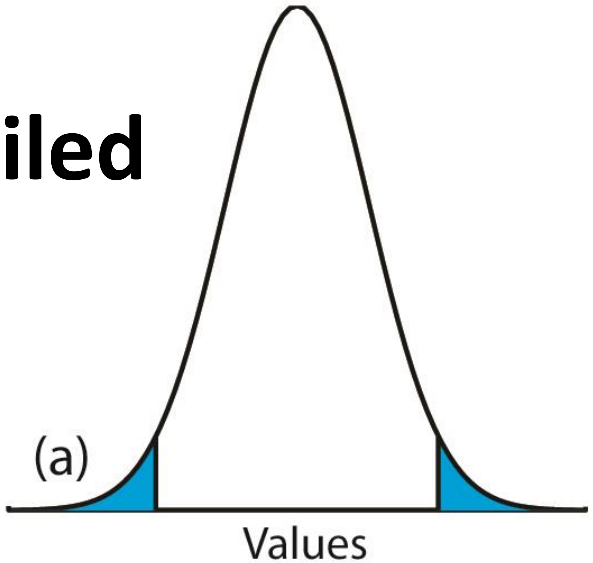
t-test: Tow-tailed

If you have **two sample means, A & B:**

You do a two-tailed test when your null hypothesize A=B,
so you combine the possibilities of A>B and A<B

Example, the average height 8 year of boys and girls are different.

two-tailed



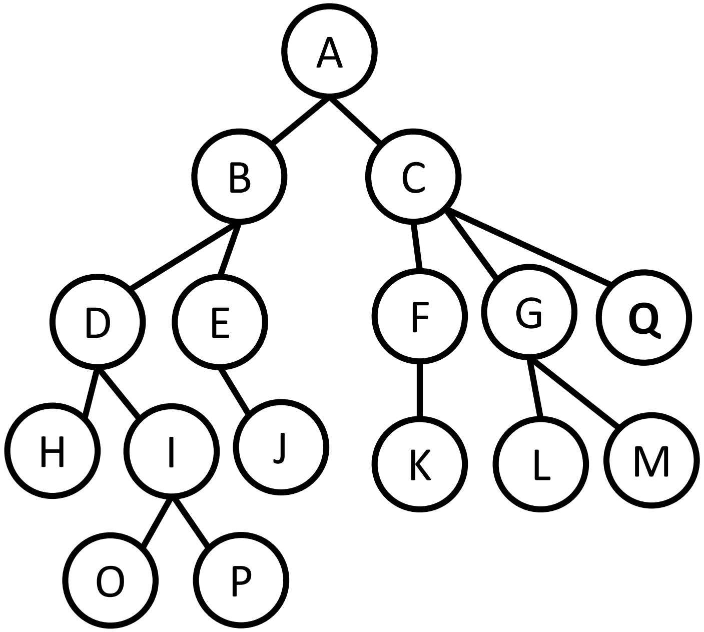
Two-tailed t table

Degrees of freedom	Significance level 20% (0.20)	10% (0.10)	5% (0.05)	2% (0.02)	1% (0.01)	0.1% (0.001)
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850

Example Designed Experiment

What happen if we add a third condition to our experiment?

Comparison technique 3: **animation**



H0 = the mean of performance time is the same between three conditions

ANOVA

How do we compare **three means** of the three experimental conditions?

ANOVA (Analysis of Variances) is a technique that we can use to do this

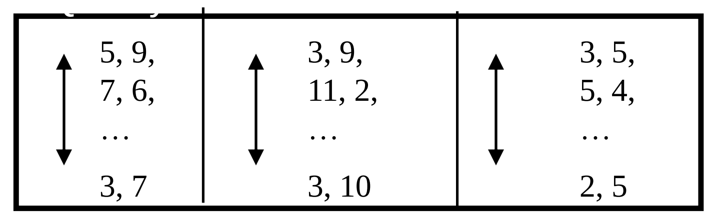
ANOVA is what we call an omnibus test

- tells us if ($\bar{x}1 = \bar{x}2 = \bar{x}3$) **IS NOT** true
- **doesn't tell** us **HOW** the means differ (i.e. $\bar{x}1 > \bar{x}2$)

ANOVA

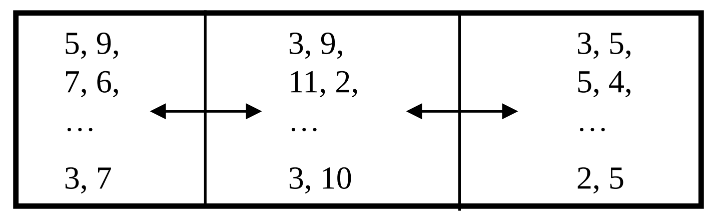
Within group variability (WG)

- Participants' differences
- Error (random + systematic)



Between group variability (BG)

- Conditions effects
- Individual differences
- Error (random + systematic)



These two variability's combine to **give total variability**

29

ANOVA

You want to make sure that the difference between conditions are because of the differences between the groups (BG), not the differences within the groups (WG)!

30

ANOVA

To do ANOVA, we calculate the **f statistic**

f = Between group variability (BG) / Within group variability (WG)

- f <= 1, if there are no treatment effects
- f > 1, if there are treatment effects

31

Cheers!

Analysis of variance (ANOVA)

- A workhorse
  - Allows moderately complex experimental designs (relative to t-test)

- Terminology
  - Factor
    - Independent variable
    - E.G., Keyboard, expertise, age
  - Factor level
    - Specific value of independent variable
    - E.G., Qwerty, novice, 10-12 year olds

ANOVA terminology

between subjects

- a subject is assigned to only one factor level of treatment
- problem: greater variability, requires more subjects

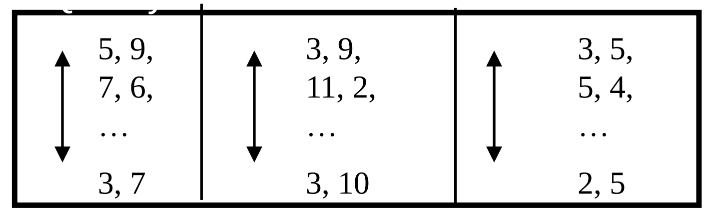
S1-20	*S21-40*	*S41-60*

within subjects

- subjects assigned to all factor levels of a treatment
- requires fewer subjects
- less variability as subject measures are paired
- problem: order effects (e.g., learning)
- partially solved by counter-balanced ordering

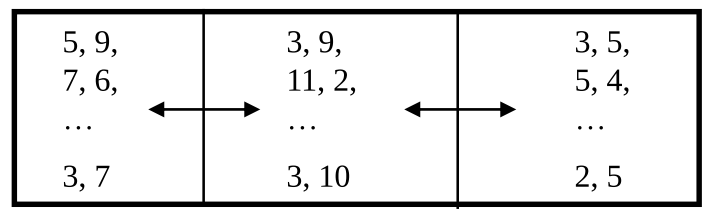
S1-20	*S1-20*	*S1-20*

f statistic



Within group variability (WG)

- Individual differences
- Error (random + systematic)



Between group variability (bg)

- Treatment effects
- Individual differences
- Error (random + systematic)

These two variability's combine to give total variability

- We are mostly interested in _______________ variability because we are trying to understand the effect of the treatment

f statistic

ANOVA is what we call an omnibus test

- tells us if ($\bar{X}_1 = \bar{X}_2 = \bar{X}_3$) IS NOT true
- doesn't tell us HOW the means differ (i.e. $\bar{x}_1 > \bar{x}_2$)

Intuition…

$$f = \frac{BG}{WG} = \frac{\text{treatment} + \text{id} + \text{error}}{\text{id} + \text{error}} = ?$$

= 1, if there are no treatment effects

> 1, if there are treatment effects

within-subjects design: the id component in numerator and denominator factored out, therefore a more powerful design

f statistic

- Similar to the t-test, we look up the f value in a table, for a given α and degrees of freedom to determine significance

- Thus, f statistic is sensitive to sample size
  - Big N big power easier to find significance
  - Small N small power difficult to find significance

- What we (should) want to know is the effect size
  - Does the treatment make a big difference (i.E., Large effect)?
  - Or does it only make a small difference (i.E., Small effect)?
  - Depending on what we are doing, small effects may be important findings

Statistical significance vs. Practical significance

- When N is large, even a trivial difference (small effect) may be large enough to produce a statistically significant result
  - E.G., Menu choice:
    mean selection time of menu A is 3 seconds;
    menu B is 3.05 seconds
- Statistical significance does not imply that the difference is important!
  - A matter of interpretation, i.E., Subjective opinion
  - Should always report means to help others make their opinion
- There are measures for effect size
  - Regrettably they are not widely used in HCI research

Single factor analysis of variance

- Compare means between two or more factor levels within a single factor
- E.G.:
  - Dependent variable: typing speed (time)
  - Independent variable (factor): keyboard
  - Between subject design

S1: 25 secs	S21: 40 secs	S51: 17 secs
S2: 29	S22: 55	S52: 45
…	…	…
S20: 33	S40: 33	S60: 23

ANOVA terminology

- Factorial design
  - Cross combination of levels of one factor with levels of another
  - E.G., Keyboard type (3) x expertise (2)

- Cell [or condition]
  - Unique treatment combination
  - E.G., Qwerty x non-typist

ANOVA terminology

- Mixed factor [split-plot]
  - Contains both between and within subject combinations

ANOVA

- Compares the relationships between many factors
- Provides more informed results
  - Considers the interactions between factors
  - E.G.,
    - Typists type faster on dvorak, than on alphabetic and qwerty
    - Non-typists are fastest on alphabetic

Other statistical tests commonly used in HCI

- Your reading does a very good job of covering these, and we won’t cover them further
  - Correlation
  - Regression
  - Non-parametric tests
    - Chi-squared
    - Mann-Whitney
    - Wilcoxon signed-rank
    - Kruskal-Wallis
    - Friedman’s