User testing

Experimental Design, Basic Statistics

Experimental Design

Example Experiment

How much do people like these crisps?

Which crisps do they like more?

Dependent variables

- What you want to measure
- For example, for the crisps could have 2:
  - *User satisfaction*:

    How much they say they like them, say on a scale from 1 to 7 (1=I really hate them, 7=I love them)
  - *Amount eaten*:

    How many (in grams) they eat if left alone with a bowl of crisps, without knowing this is what we are interested in

Independent variables

- Experimental factors the effect of which you want to measure
- For example, for the crisps:
  - The brand of crisp (Brand A, Brand B)

  Or

  - The crisp flavouring (neutral, cheese, chilli)

Aside: brand in this example has 2 so-called *levels* (Brand A and B), whilst crisp flavouring has 3 levels.

Another Example Experiment

How good are these running shoes?

Which running shoes are better?

Question: Variables

- What dependent variable(s) may we be interested in for running shoes?
- What independent variable(s)?

Often in User-Testing: Compare A to B

Compare

the performance (on the dependent variable) of a group of users who experienced a certain level of the independent variable (A)

with

the performance of a group of users who experienced another level of the independent variable (B)

Examples

- Compare two systems:
  - Is Virgin’s on-line shopping site more usable than Amazon’s?
  - Is the new system more usable than the old one?
- Compare two designs:
  - Is my system more usable with drop-down menus or with lists?

Aim of experimental design

Highlight effect of independent variable

while avoiding undesired effects

by strictly controlling the influence of irrelevant variables

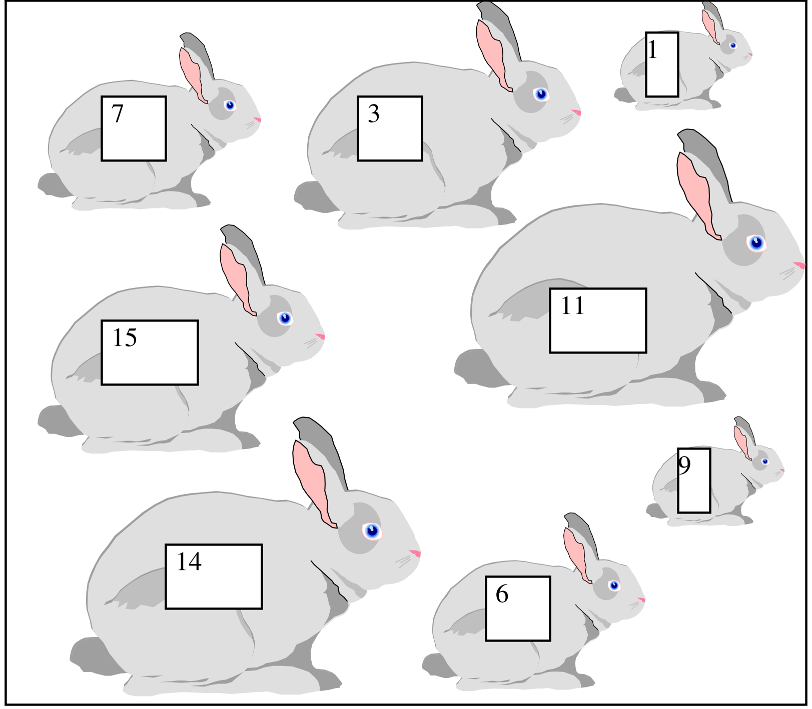
Irrelevant variables (1)

- User variables
  - Age
  - Sex
  - Education
  - Cultural background
  - Experience with computers
  - Etc. etc.

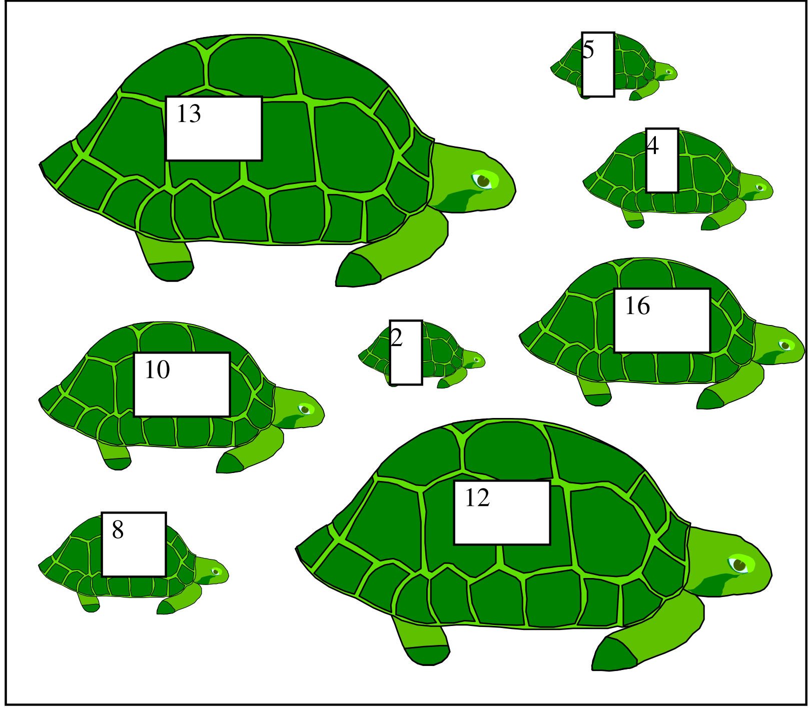
  (unless this is the purpose of your experiment)

Is this correct?

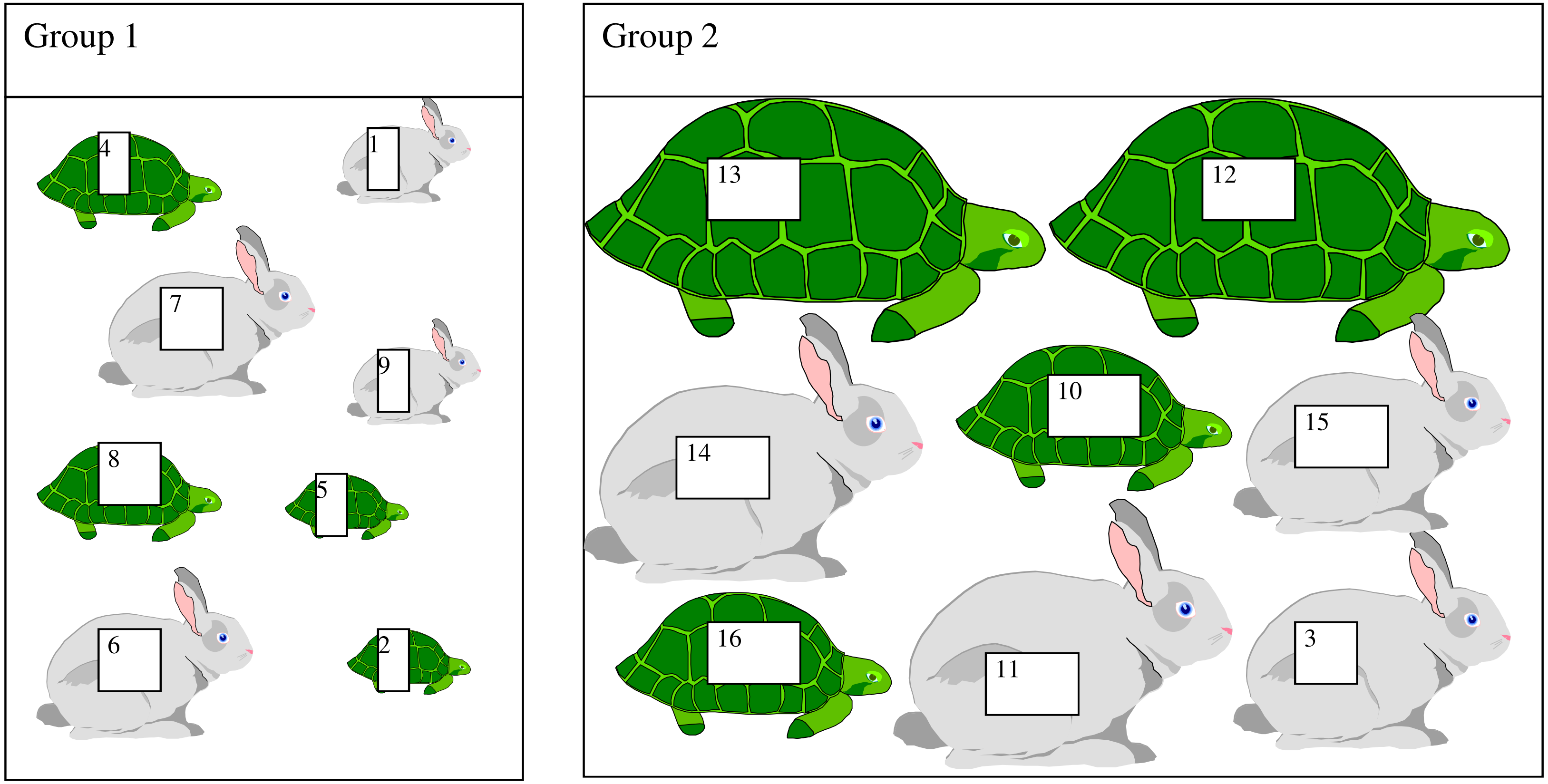
Group 1



Group 2



Is this correct?



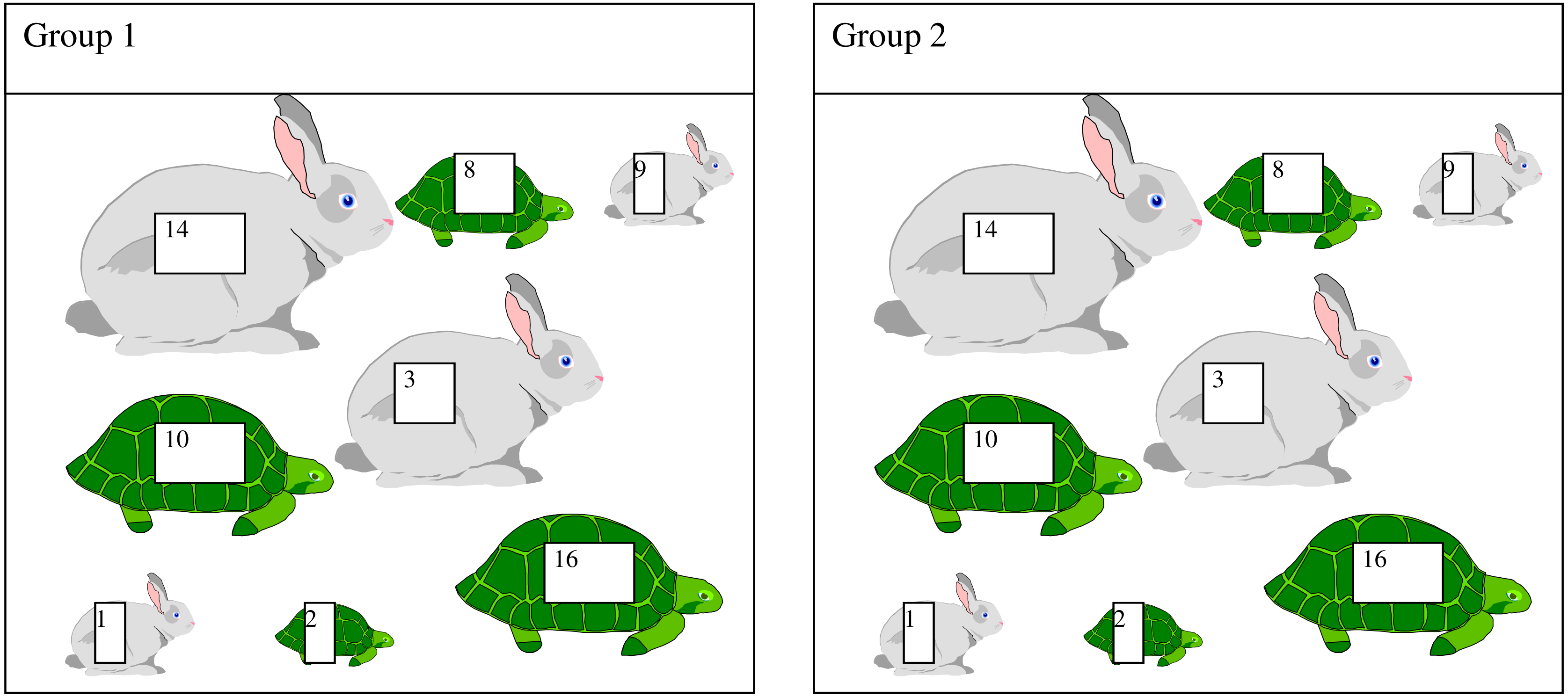
Let’s try making groups

- I need two volunteers
- I want you to select two groups of 4 people, so that both groups are as similar as possible

Within subjects design

Use the same subjects in each group

(also called Repeated Measures)



Within subjects design (2)

Advantage

- Subject variables (like age, sex, education, intelligence) are the same for both groups

Disadvantage

- More time needed per subject
- Order effect due to practice, fatigue, etc

Within subjects design (3)

Counterbalancing:

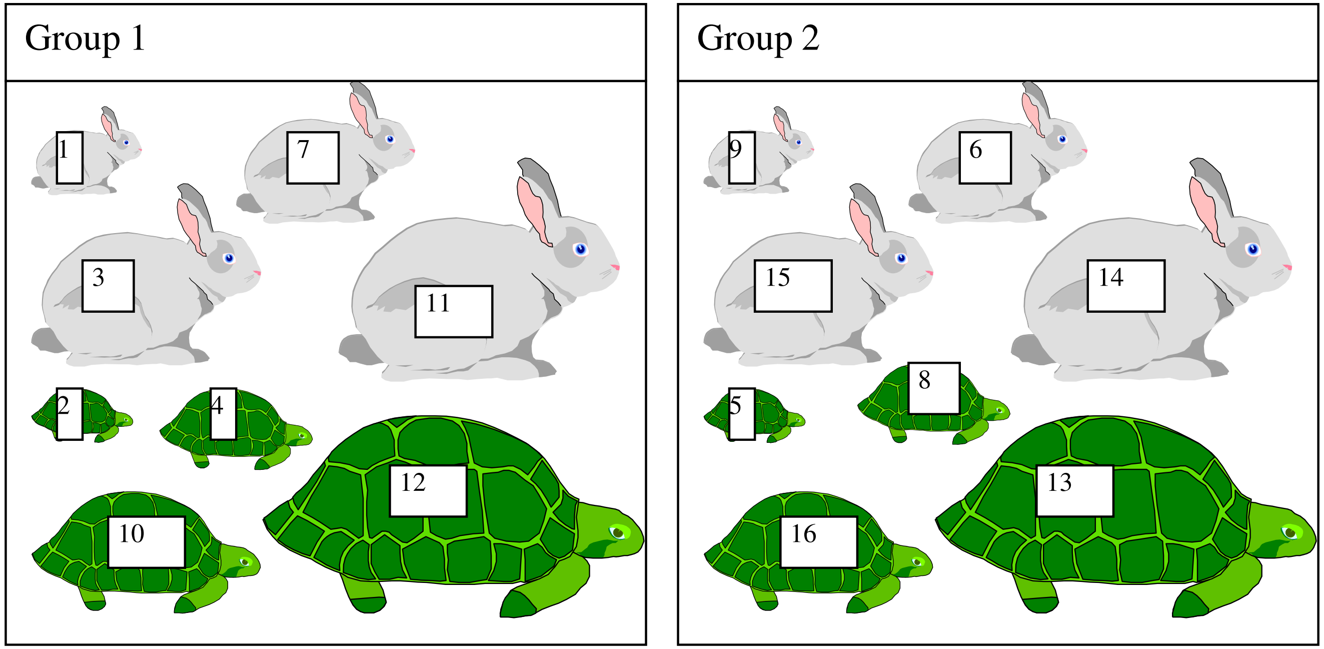
- half the subjects perform the tasks in one order (A-B)
- half the subjects perform the tasks in the reverse order (B-A)

Disadvantage

- Statistical analysis more difficult

Matched pairs design

- Divide subjects in pairs of similar subjects
- For each pair: randomly assign one subject to group 1 and the other to group 2



Matched pairs design (2)

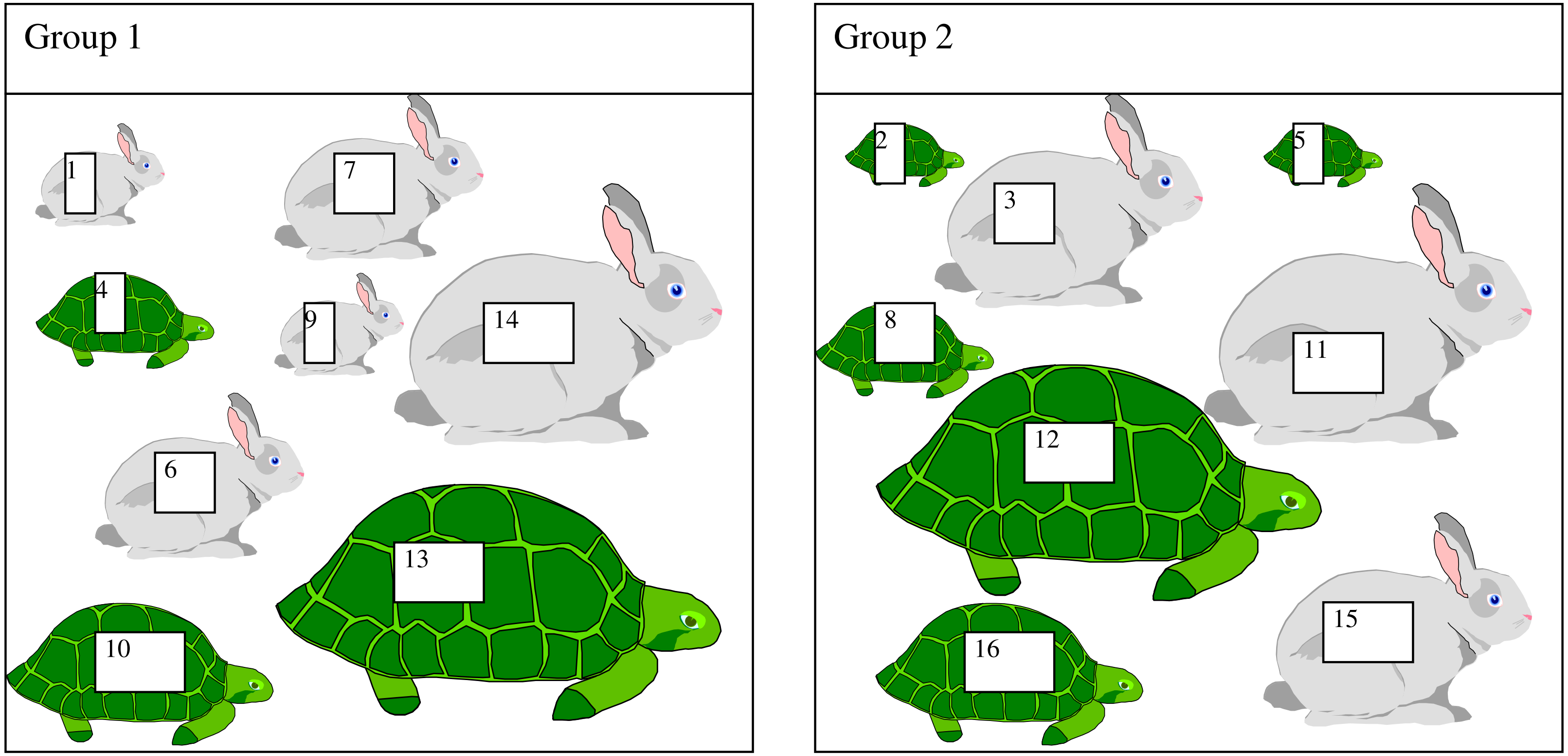
- When analyzing, compare pair-wise

Disadvantages:

- Difficult to match pairs
- Statistics more difficult

Between subjects design

Use two different groups of subjects, allocate subjects randomly



Between subjects design (2)

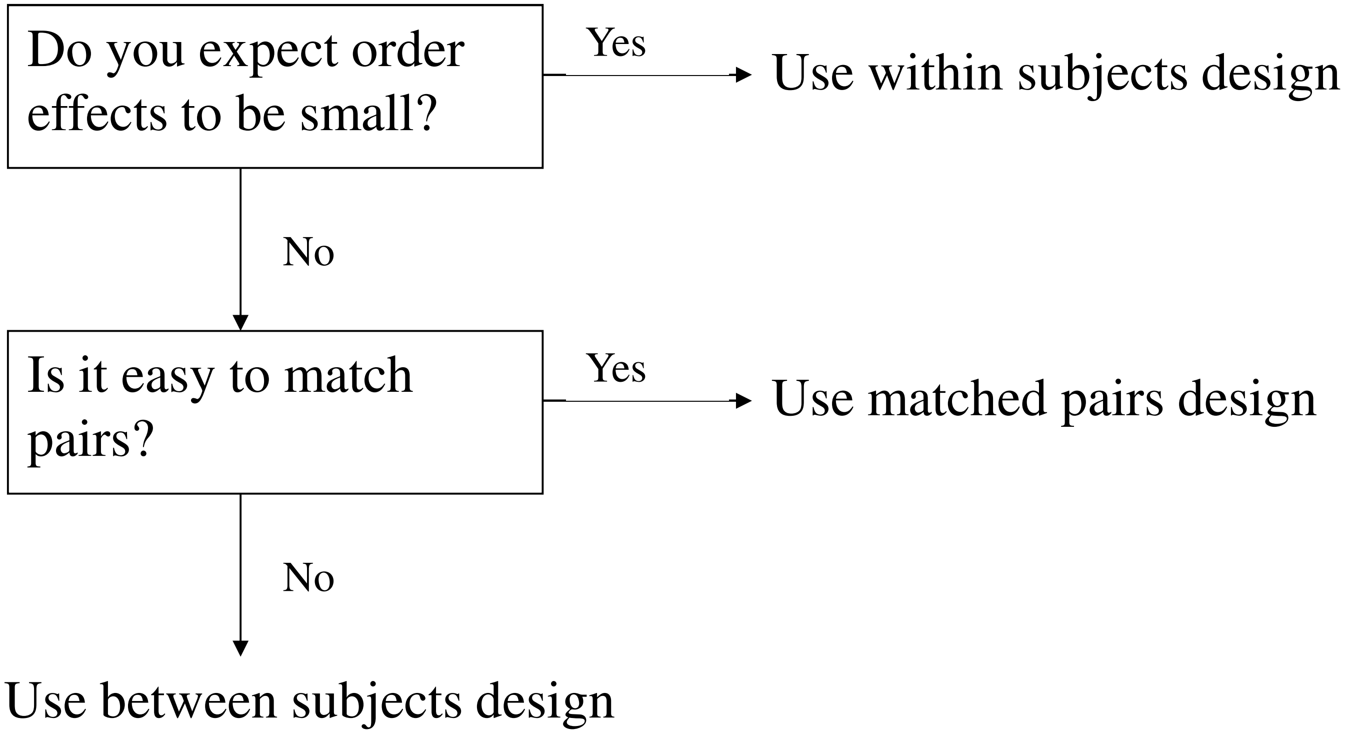
Advantages

- No order effects
- Less time needed per subject

Disadvantages

- More subjects needed
- Less powerful
- Need to carefully select subjects

Choosing the design



Do you expect order effects to be small?

Yes

Use within subjects design

No

Is it easy to match pairs?

Yes

Use matched pairs design

No

Use between subjects design

Irrelevant variables (2)

- Situational variables
  - Physical characteristics of the experimental room, like light, temperature, noise,..
  - Equipment

  - Experimenter

Real-life versus laboratory

Control vs Realism

- Control: Ensuring‘irrelevant’ variables do not influence the outcome
- Realism: Ensuring the results hold in the real world

This is a trade-off, cannot always do both…

Basic statistics

Misconception

- "With statistics you can prove anything…."

No, you cannot, but it is easy to confuse people with numbers.

I want you to be able to:

- criticize other people's misuse of numbers
- understand relation with experimental design

Critical thinking required!

- I will show you some ‘stories’ I have heard, which use numbers.
- Tell me what is wrong with them!

On the radio….

- “In *65%* of domestic violence cases, the offender is drunk. Alcoholism causes violence.”

In the train..

- “I work for the council, providing information to people in my area about the environment and recycling. The glass-recycling containers in my area are *much fuller* than in other areas. This shows how good I am at my job!”

In the newspaper

- The newspapers have published a *ranking list* for schools: a school is higher in the ranking if a higher percentage of its pupils do well at the national exams.

  A mother sees this ranking and decides that her daughter should go to a particular school because it is higher in the ranking than other schools in the area.

How does this relate to experimental design?

- The misconception in the stories came from someone overlooking some variables that may have influenced the measurements.
- These should have been controlled for through good experimental design.

Hypothesis testing

A hypothesis should be

- Stated in clear language
- A question with yes or no answer
- Answerable using a small set of independent and dependent variables.

Examples:

H_1 People will say they like crisps of Brand A more than crisps of Brand B.

H_2 The amounts people eat from A and B will differ

Statistical Hypothesis

Because of the way statistics works, statisticians want to *reject* a hypothesis!

Therefore, a statistician will state a hypothesis (called null hypothesis) in the opposite way.

H_0 There will be no signifant difference in how much

1. People say they like crisps of Brand A compared to B
2. Is eaten from A and B

Two types of errors stats can make

- Type I (also called false positive):
  Null Hypothesis rejected while actually it is true
  (so, original hypothesis ‘proven’ while actually it is false) This is very bad..

- Type II (also called false negative):
  Null Hypothesis not rejected while it is false.
  (so, original hypothesis not ‘proven’ though true)
  This is not so bad.. Found not enough evidence yet..
  May need more participants!

p-value

- p-value is the probability of a Type I error, so the probability that the effect you are seeing is due to chance
- Normally, you want p<0.05 to say that something is *statistically significant*
- You run a statistical test on your data, and it tells you the p-value.

What statistics to use? (1)

One independent variable, two treatments

(Example: brand of crisps, one group A, other B)

Use T-test*.

Computed based on the difference in means of the two groups and how spread out the data is (standard deviation).

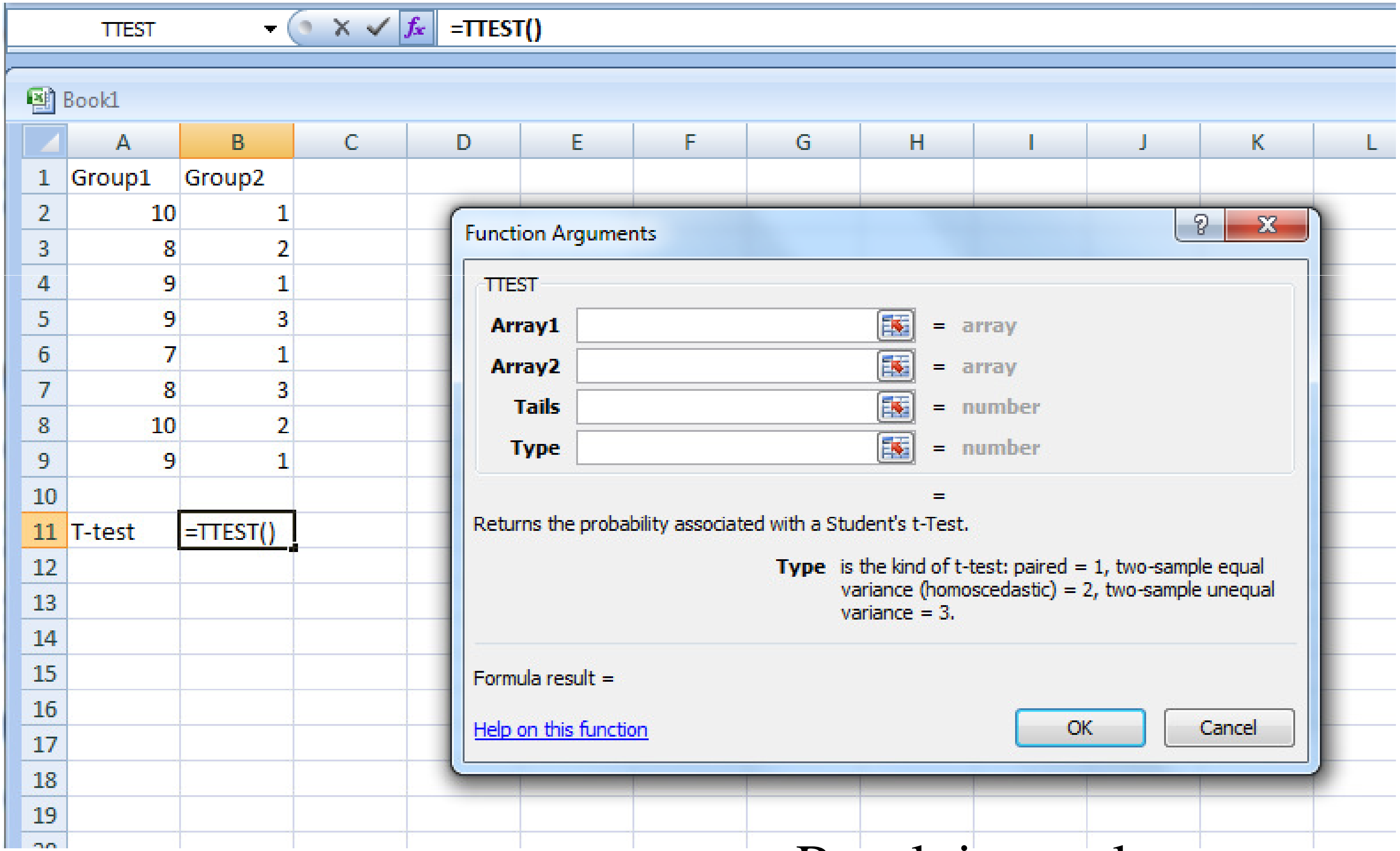
T-test is available as a function in Excel.

* If data is normally distributed.

Otherwise non-parametric test (Mann-Whitney U)

T-Test in Excel

Tails: 2 if test whether different
1 if test larger/smaller
Type: Paired: Within subject
Two-sample: Between subject



=TTEST()

	A	B
1	Group1	Group2
2	10	1
3	8	2
4	9	1
5	9	3
6	7	1
7	8	3
8	10	2
9	9	1
10		
11	T-test	=TTEST()

Function Arguments

TTEST

Array1 = array
Array2 = array
Tails = number
Type = number

=

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 =

Result is p-value

What statistics to use? (2)

One independent variable, more than two treatments.

Use one-way ANOVA*.

Two independent variables, any number of treatments for each one.

Use two-way ANOVA.

*If data is normally distributed. Otherwise Kruskal-Wallis.

Example User Test: Transport Project

Does our journey planner result in more sustainable use of transport?

Main issues

Transport use / experience depends on:

- people’s circumstances, such as home address, bus routes, having children
  => Used within-subject design
- people’s normal modes of transport used
  => Used stratified sampling
- weather, events on specific days
  => Used staggering & Added control group