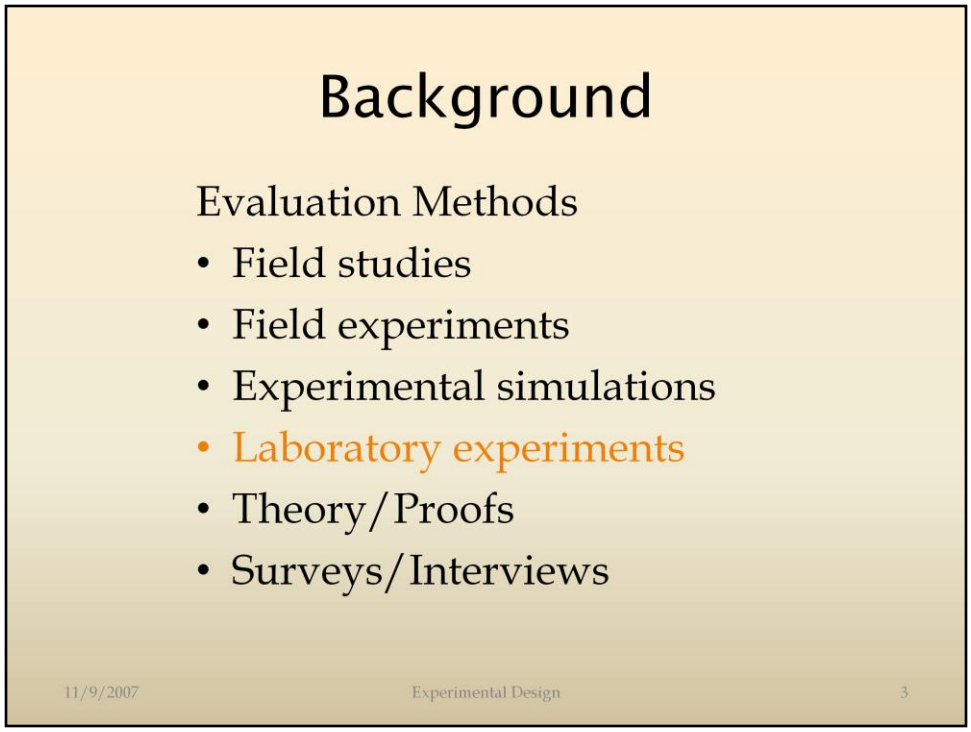
Experimental Design

November 7th, 2007

- Experiment
- Validity
- Reliability
- Statistics
- t-test
- ANOVA
- POWER
- Statistically significant
- Type I error
- Null hypothesis
- Negative result

Elements you might be familiar with



Laboratory experiments are only one of the multiple methods that we have to gather knowledge

Introduction

Experimental Design

Experiments:

- Quantitative
- Controlled
- Empirical

Learning Goals

- Decide when you need to design an experiment
- Recognize the main elements of experimental design
- Plan an experiment
- Avoid the main pitfalls of experimental design
- Interpret results from a experiment

Outline

- Part I: Why run experiments?
- Part II: From questions to hypotheses
- Part III: From hypotheses to design
- Part IV: Putting it all together

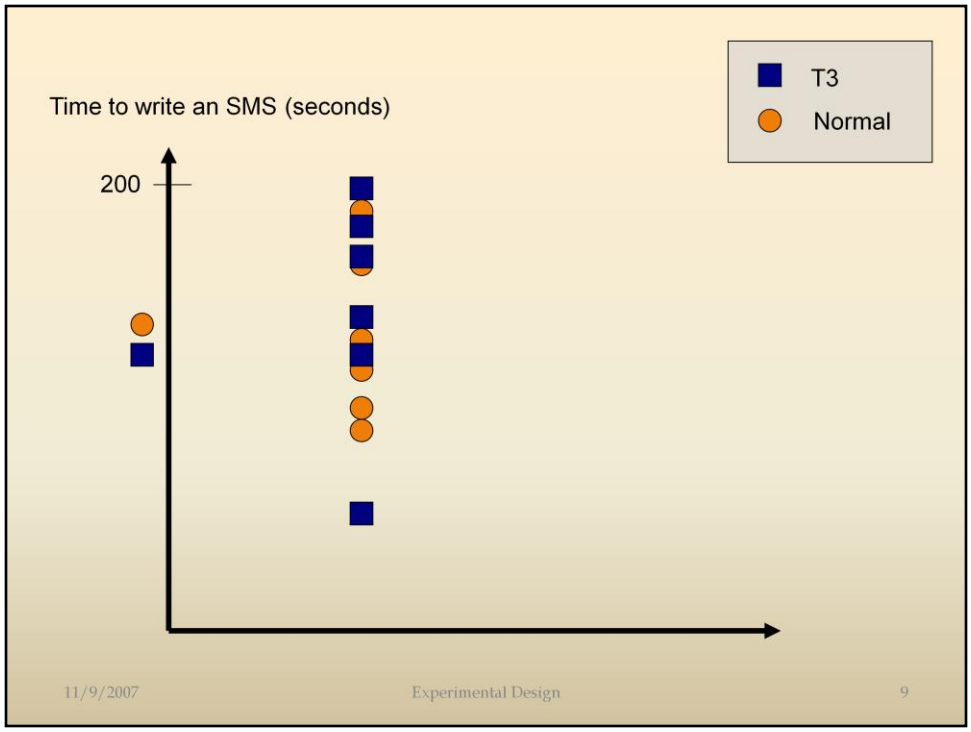
PART I

WHY RUN EXPERIMENTS?

Example: writing SMS

- Regular keyboard vs. T3 (automatic completion)

Normally you run an experiment because you want to know something… but why an experiment and not just observe?
Example: texting by using the normal “multi-click” technique, or a dictionary-based technique called T3.



Plain measures are difficult to interpret. Even averages don't tell us much.

Difficulties

- It is difficult to predict results with theory (the world is full of noise)
- It is difficult to know the real source of an observed difference
- It is difficult to interpret numeric data

The solution: systematic controlled experiments

- Repetition
- Control
- Statistics

What are our goals?

Reliability:

The extent to which a measurement is consistent, can be reproduced, and avoids error

Validity:

The extent to which a procedure measures what it is intended to measure

We want to maximize the reliability and validity of our tests.

What do we get from experiments?

- Evidence on a particular question
- A language to communicate research
- A measure of the certainty of our findings

What do we NOT get from experiments?

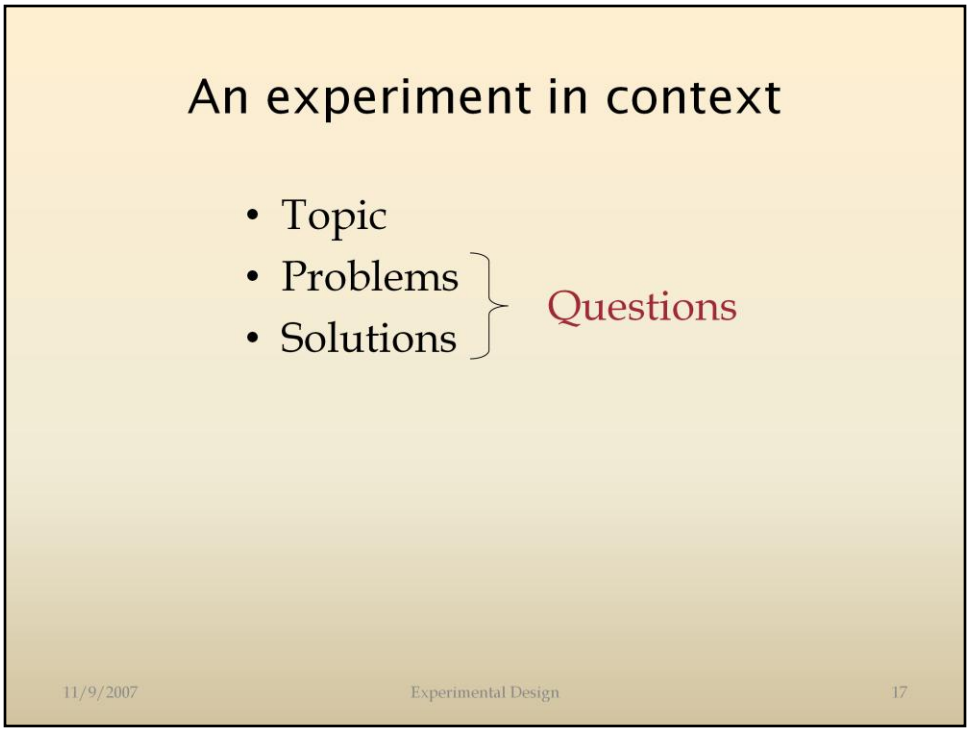
- A theory
- Interpretations of the data
- "Magic" results
- Absolute proof

Experiment as measuring tool

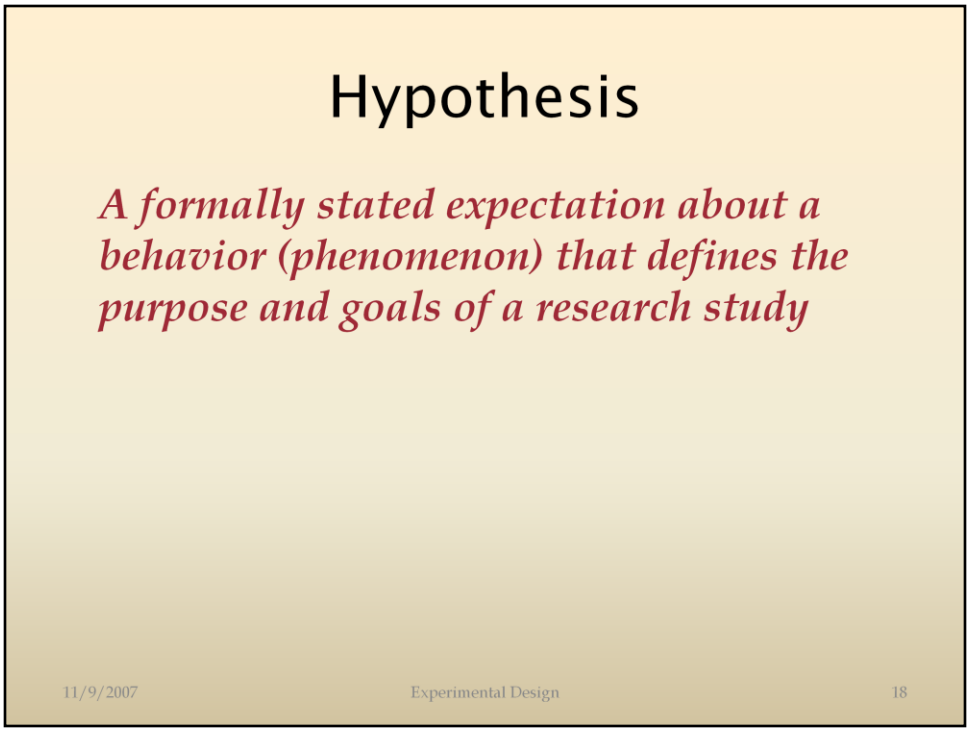
- Optimized every time
- Subject for interpretation
- With a measure of reliability

PART II

FROM QUESTIONS TO HYPOTHESES



Carl has taught you how to formulate your research in terms of problems and solutions etc.



With questions we make hypotheses. A hypothesis works as a possible answer for our question.

Examples of hypotheses

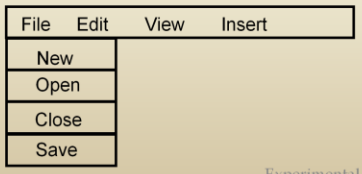
Example 1:

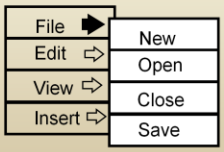
There is no difference in the number of cavities in children and teenagers using crest and no-teeth toothpaste when brushing daily over a one month period

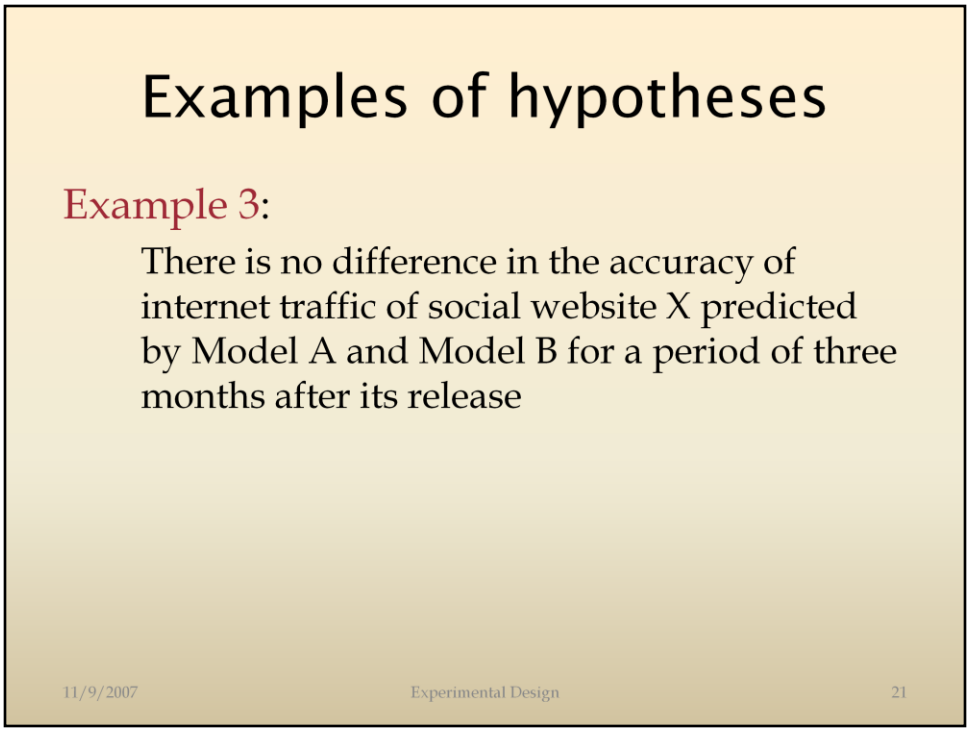
Examples of hypotheses

Example 2:

There is no difference in user performance (time and error rate) when selecting a single item from a pop-up or a pull down menu of 4 items, regardless of the subject's previous expertise in using a mouse or using the different menu types







Hypothesis have to be falsified. Notice also that they are normally stated in negative terms. Hypotheses formulated in these negative terms (there is no difference) are called NULL hypotheses.

Hypotheses and variables

A good hypothesis already includes:

- The main independent variable
- The main dependent variables

Dependent variable

In an experiment, the variable that is measured under each condition

Example Hypothesis (dep. var)

There is no difference in the number of cavities in children and teenagers using crest and no-teeth toothpaste when brushing daily over a one month period

Example Hypothesis (dep. var)

There is no difference in the number of cavities in children and teenagers using crest and no-teeth toothpaste when brushing daily over a one month period

Example Hypothesis (dep. var)

There is no difference in user performance (time and error rate) when selecting a single item from a pop-up or a pull down menu of 4 items, regardless of the subject's previous expertise in using a mouse or using the different menu types

Example Hypothesis (dep. var)

There is no difference in user performance (time and error rate) when selecting a single item from a pop-up or a pull down menu of 4 items, regardless of the subject's previous expertise in using a mouse or using the different menu types

Independent variable

In an experiment, the variable that is systematically changed or manipulated by the researcher; also called a factor

Examples of hypotheses (ind. var.)

There is no difference in user performance (time and error rate) when selecting a single item from a pop-up or a pull down menu of 4 items, regardless of the subject's previous expertise in using a mouse or using the different menu types

Examples of hypotheses (ind. var.)

There is no difference in user performance (time and error rate) when selecting a single item from a pop-up or a pull down menu of 4 items, regardless of the subject's previous expertise in using a mouse or using the different menu types

Examples of hypotheses (ind. var.)

There is no difference in the number of cavities in children and teenagers using crest and no-teeth toothpaste when brushing daily over a one month period

Examples of hypotheses (ind. var.)

There is no difference in the number of cavities in children and teenagers using crest and no-teeth toothpaste when brushing daily over a one month period

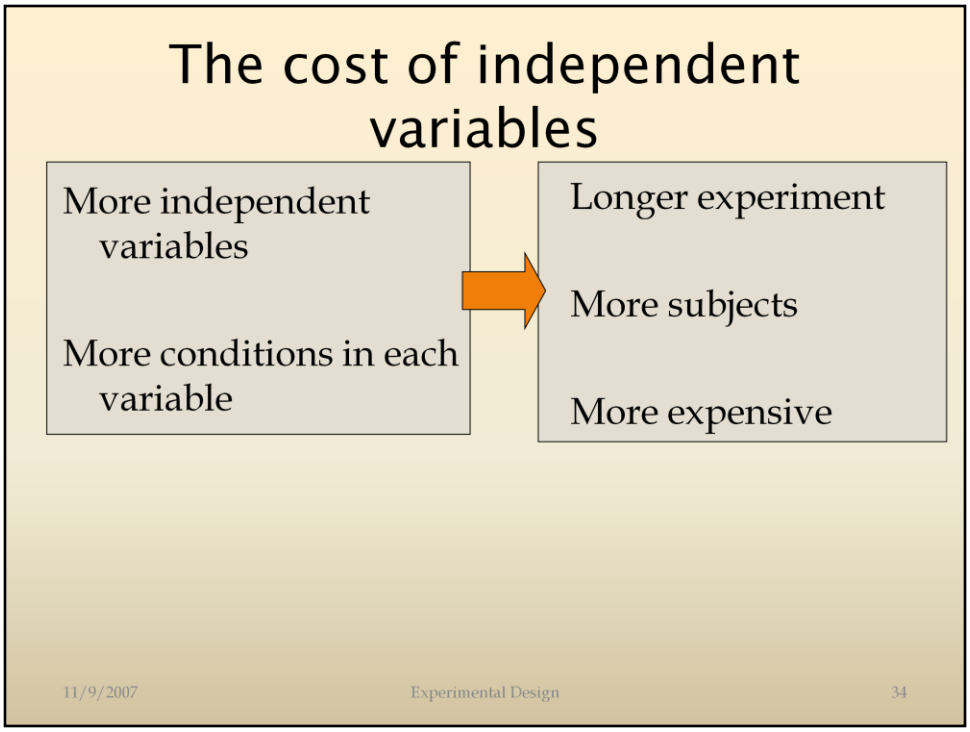
Independent variables for generality

Sometimes we use independent variables to show that our hypotheses are valid in more than one situation

Example (In menu experiment):

- menu type: pop-up or pull-down
- menu length: 3, 6, 9, 12, 15
- subject type (expert or novice)

The last two variables of our example, are independent variables chosen to give generality to our experiment



The more independent variables and conditions, the larger our experiment will have to be.

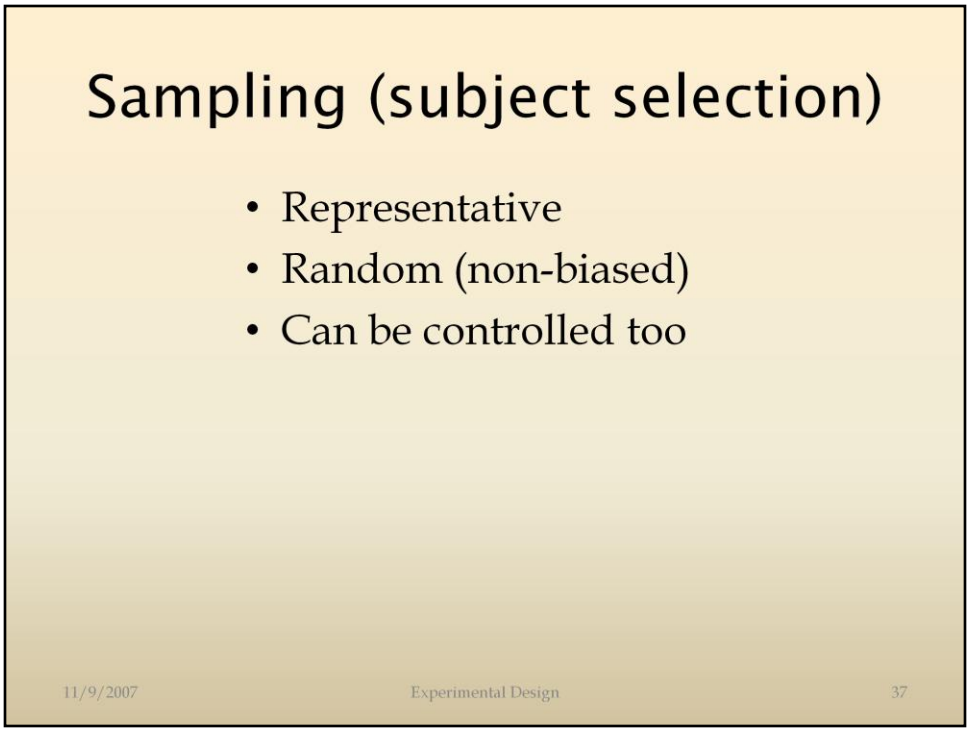
Pitfall 1

To try to measure too much in the same experiment

The cost of dependent variables

- No increased length of experiment
- Increased time of analysis
- Risk of "fishing for results"

More dependent variables do not increase the cost of running the experiment, but they do increase the cost of analyzing, and may lead to "fishing for results".

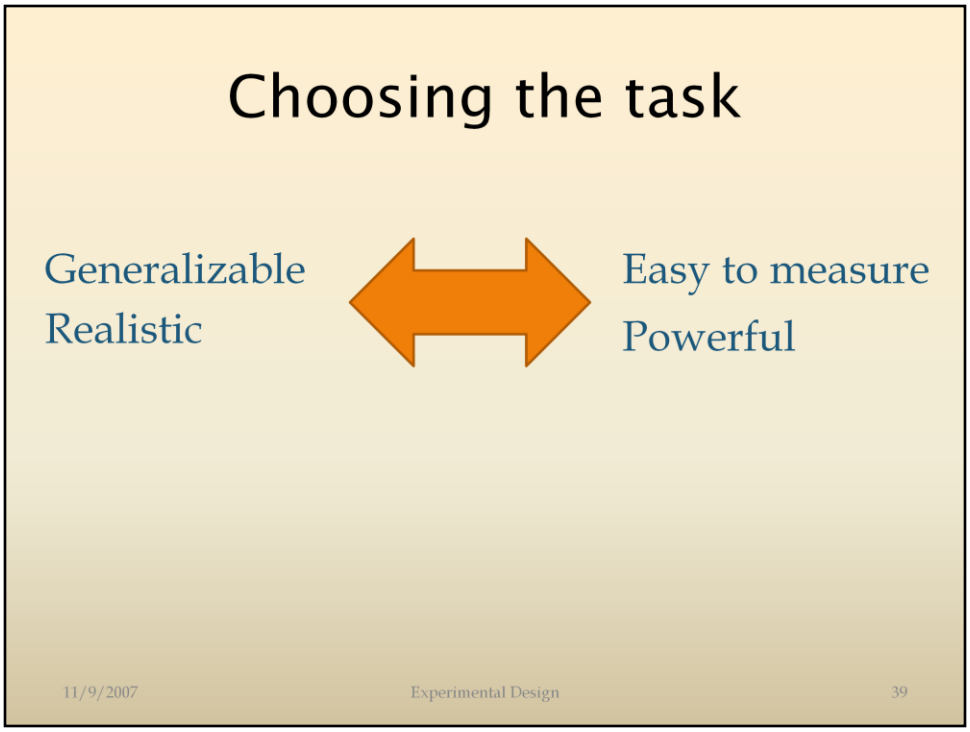


Representative means that you need enough number of different individuals from the target population

The target population refers to the group you are interested in generalizing to.

PART III

FROM HYPOTHESES TO DESIGN

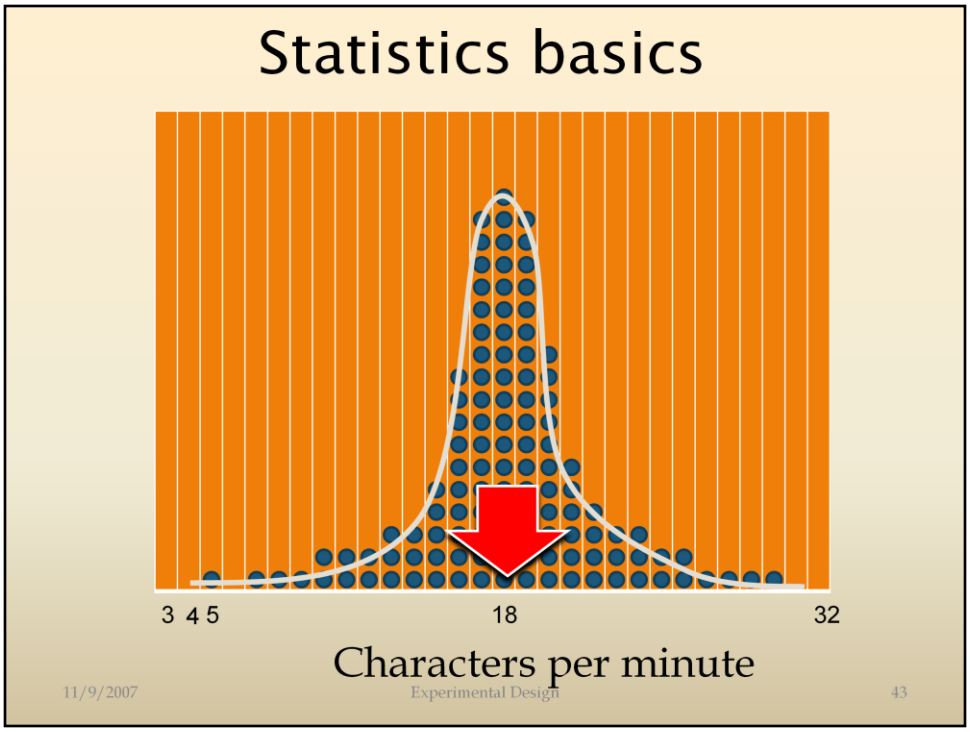


Try to strike a balance between a very realistic task that has a lot of noise and a very constrained task, which is easy to measure but might not be generalizable.

Take one of the texting interfaces.

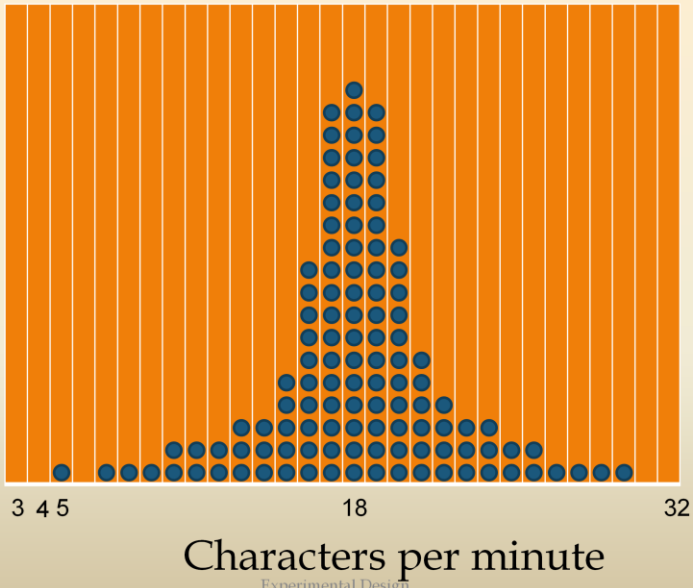
Box with different bins. We assume that we measure all the population (in this case, around 100 people) with one technique.

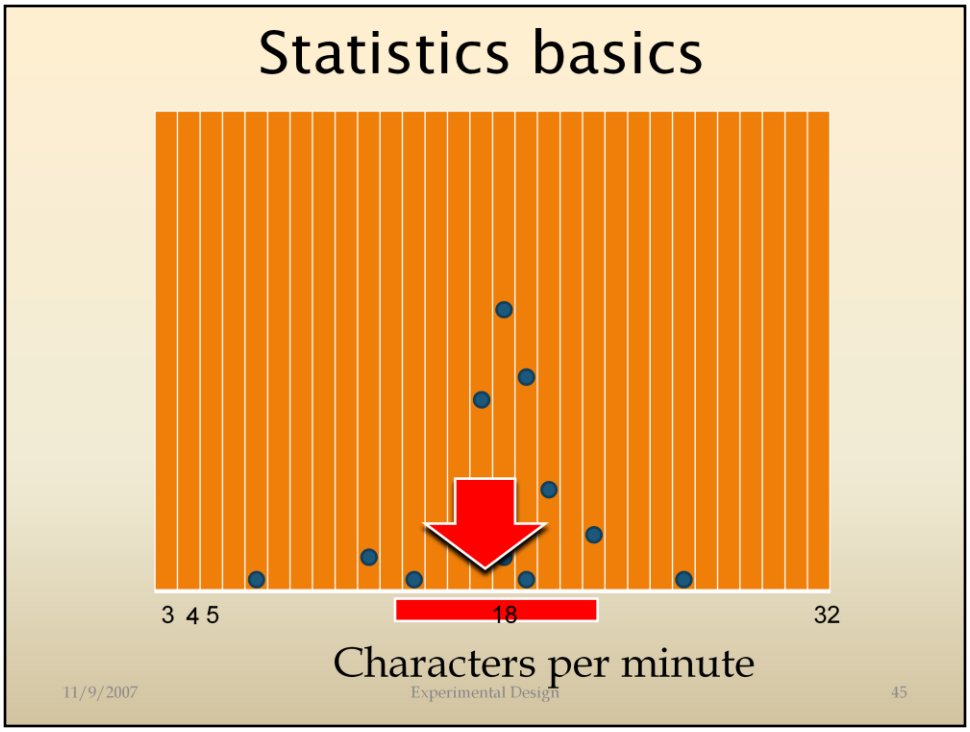
For each person that we measure, we put a marble in the bin that corresponds to their speed. The result is a histogram with the shape of a Normal curve (also called Gaussian)



Gaussian curves can be more or less pointy (depending on their variance), and be centered around different points.

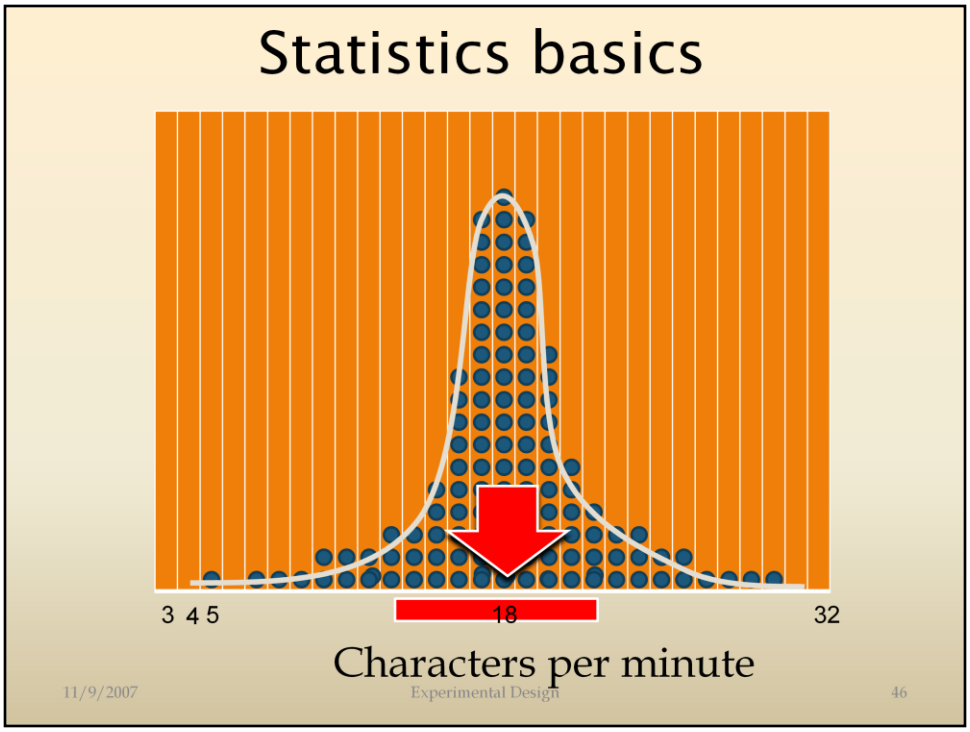
Statistics basics



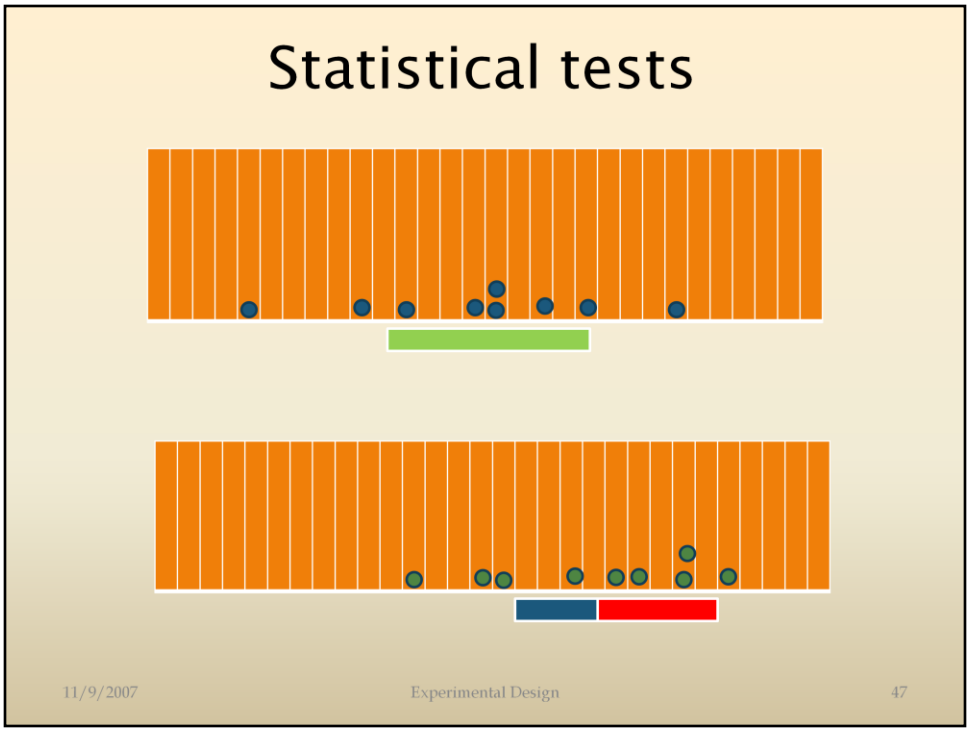


The problem is that we can never measure all the population. We only take a random sample of the total number of possible measures.

With just a few data points we can only estimate an interval in which the “real” average of the population will be.



Our job is to know where the “real” distribution is by using only the sample data. We get intervals.

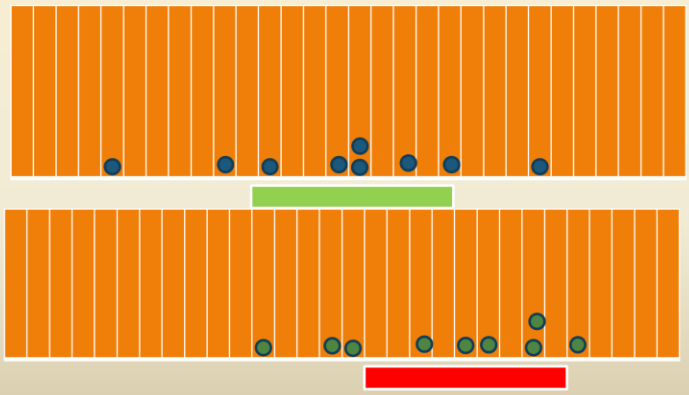


A statistical test is just a comparison between the intervals where we think the averages of two different populations are.

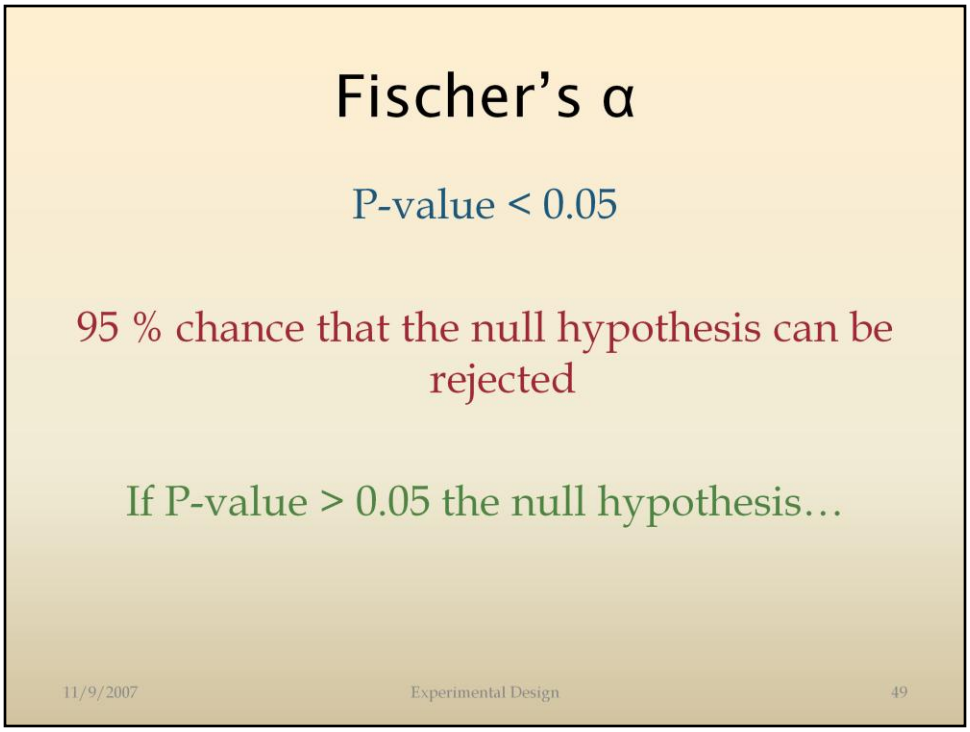
If the intervals overlap each other, we cannot know if the “real” distributions are one to the left of each other (they might as well be switched over). If the intervals don’t overlap we will have a high probability that one of the distributions is to the right of the other (that is, one technique is faster than the other).

Interpreting a result

There is no difference in the average number of words per second that adults between 20 and 30 can write using a normal text technique and the T3 text technique on a cell phone

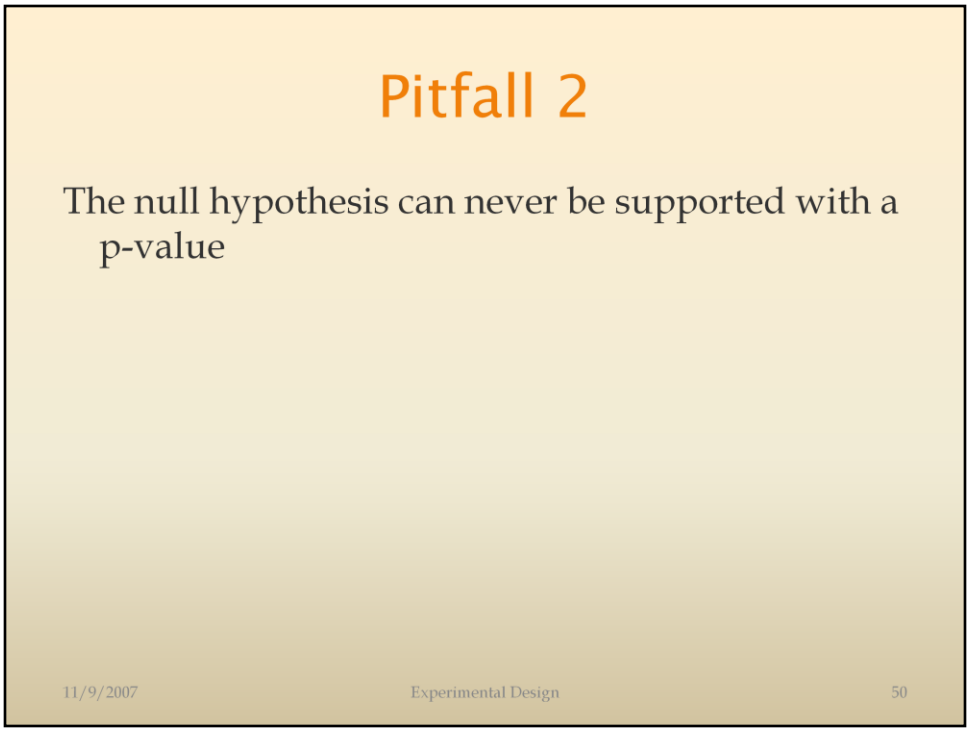


To interpret a result like this we have to take a look at the hypothesis. Overlapping intervals means that we cannot reject the null hypothesis.

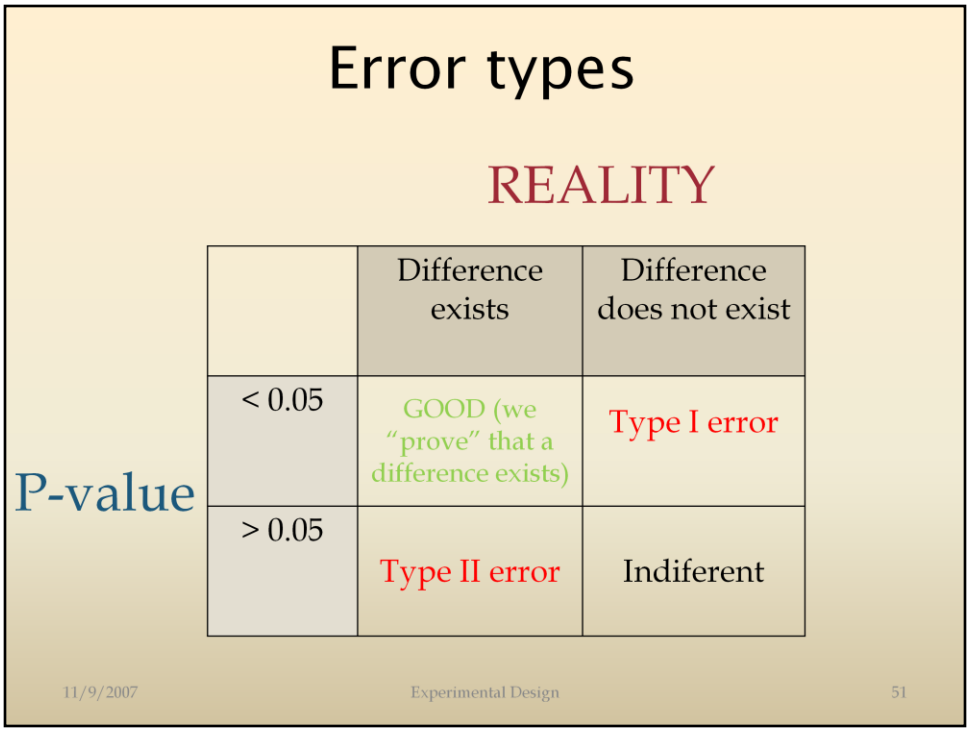


An statistical test gives us a p-value.

The probability that we have to make a mistake (of saying that the averages of the populations are separate, when the real ones are not) is set in advance, and is called the alpha value. An alpha value of 0.05 means that if we get a p-value smaller than the alpha, then we will reject the null hypothesis (e.g. we will accept that the two conditions – or techniques- are different).



A p value > 0.05 is a negative result, and shows nothing. When our p-value is larger than the chosen apha value (normally, 0.05 = 5% probability) it does NOT mean that the results support the null hypothesis. It could be that the averages are no different, but it could also be that there is too much noise in our experiment, or that we took few samples; therefore, a p value of more than 0.05 means nothing.



The standard apha is 0.05. This means a 5% probability of making a mistake in our judgement. It means that a 5% of the times the statistical test will find a difference that does not really exist. This is called a Type I error.

More often our statistical test will not find any difference, even though there exists a difference. This is called a Type II error. The job of a experimental designer is to make the experiment good enough (powerful enough) that this will not happen.

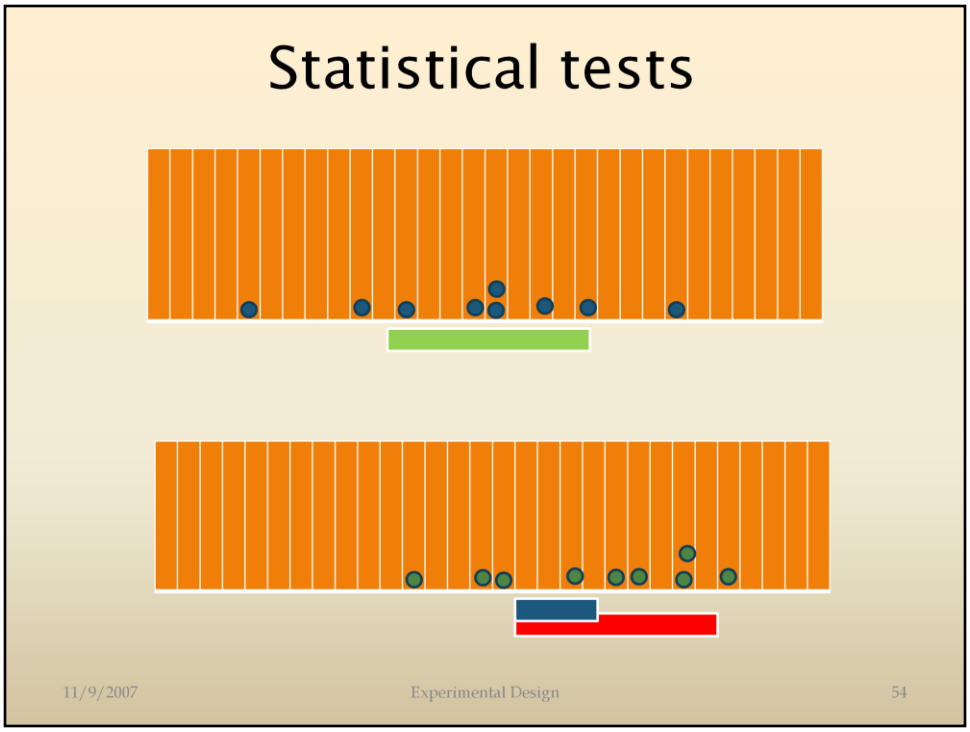
Pitfall 3

Fishing for results

Every statistical test has a probability of finding results that are just due to chance (5% chance with an alpha of 0.05). If we have many dependent variables and we perform many statistical tests, the overall chance of making a mistake goes up. Therefore, you should only test variables that are important, because otherwise you are bound to find something (something that is just due to chance).

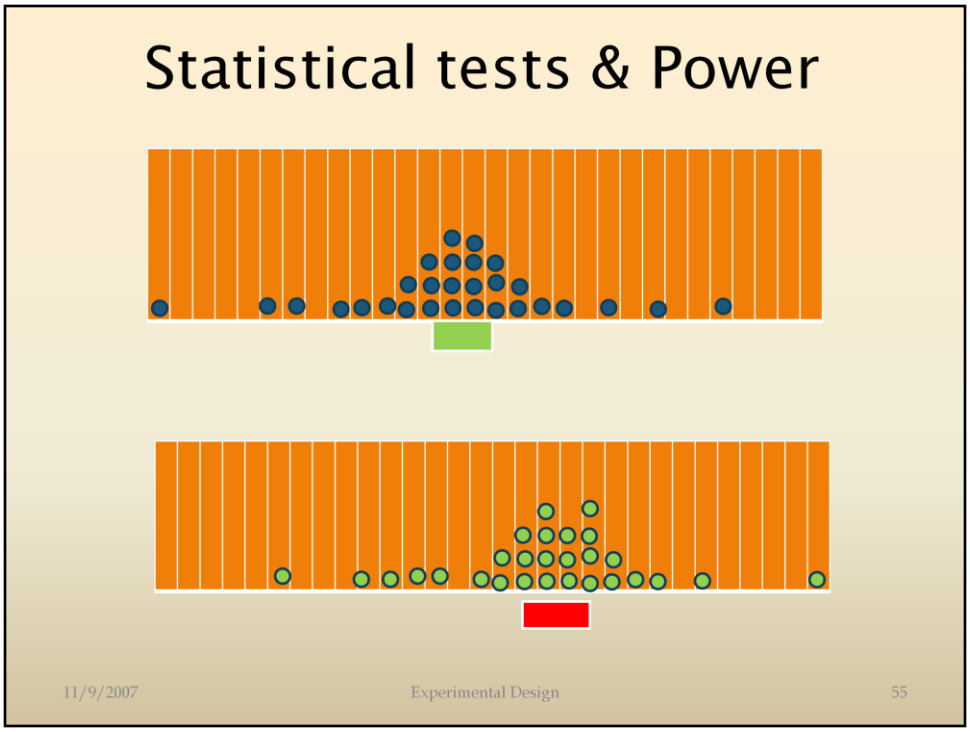
Power

The probability that a statistical test will detect a true relationship and allow the rejection of a false null hypothesis



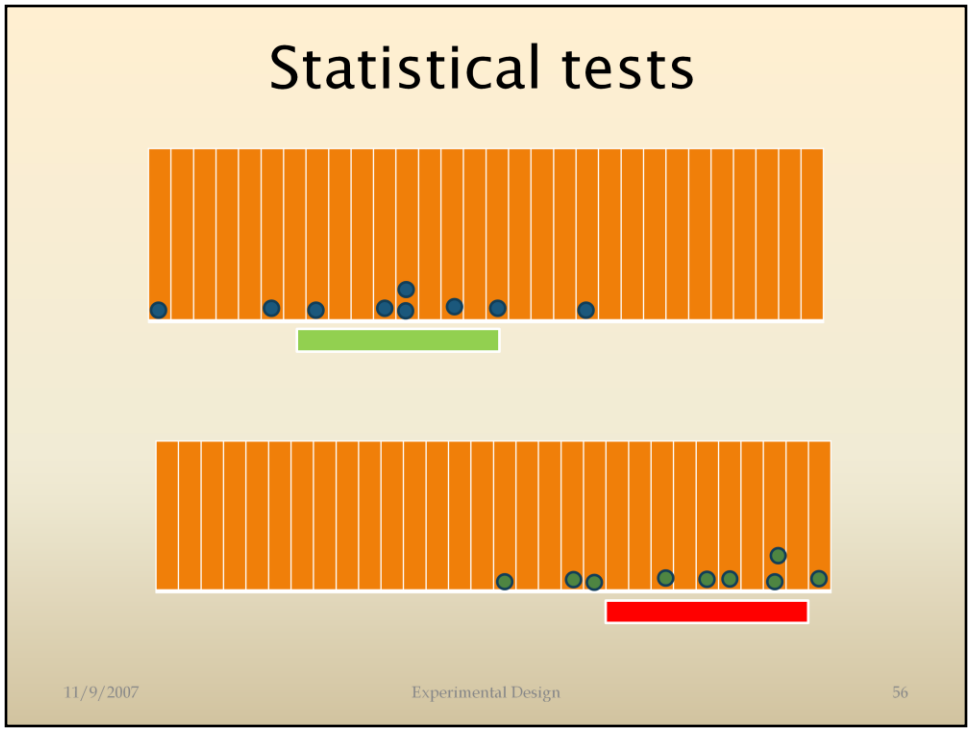
In this case there is overlap (blue). Our test had a $p > 0.05$.

There are several reasons why a test could find differences (see next slides)

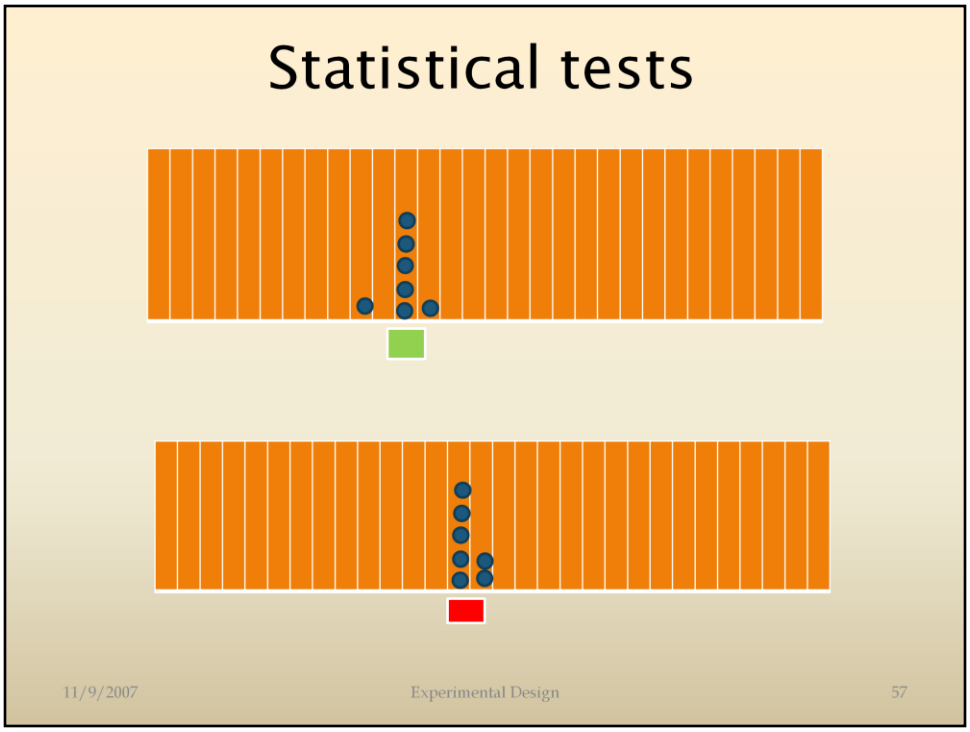


If we take more samples, we will have a better idea of where the actual populations are (the intervals are smaller).

If the intervals are smaller, we are more likely to have a p value < 0.05 (shorter intervals are more difficult to overlap)



The actual difference between two conditions might be so large that, even though we have just a few measures, the intervals will be far apart (showing a $p < 0.05$).



If the “real” distributions are very “pointy”, even with just a few samples we are going to be very sure of where the average is.

Noise (uncontrolled variability) is the enemy

There are three main ways to get more power:

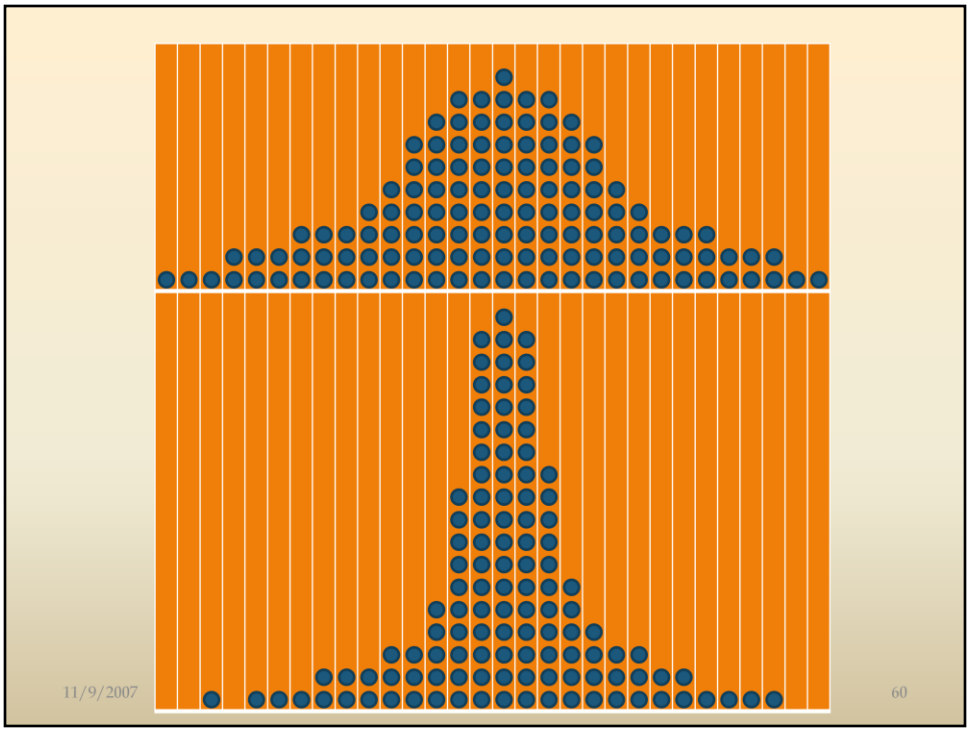
- Increase the number measures
- Eliminate noise from the experiment
- Test a task that will show big differences

Be a control freak

Pitfall 4

Let the noise take care of itself (intentionally or unintentionally)

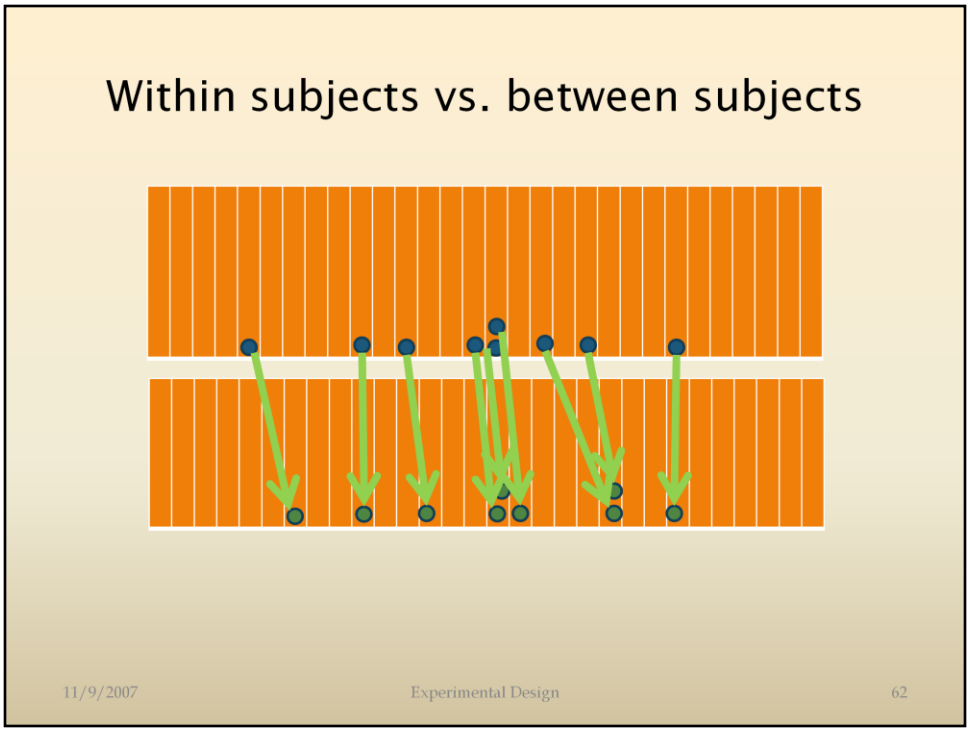
It is too easy to design a bad experiment that shows no difference… just let the noise creep in.



We want to have “real” distributions that look more like the one underneath, and less like the one up.

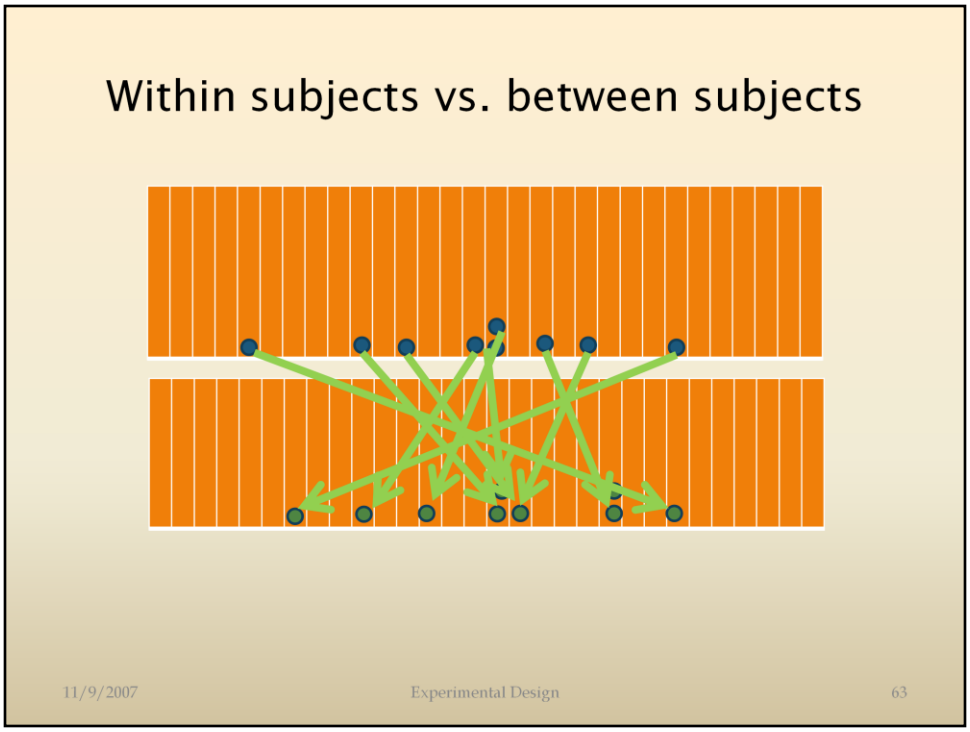
Strategies to reduce noise

- Be a control freak
- Use a within-subjects design if you can
- Reduce learning effects/tiredness
- Measure the same thing multiple-times and average

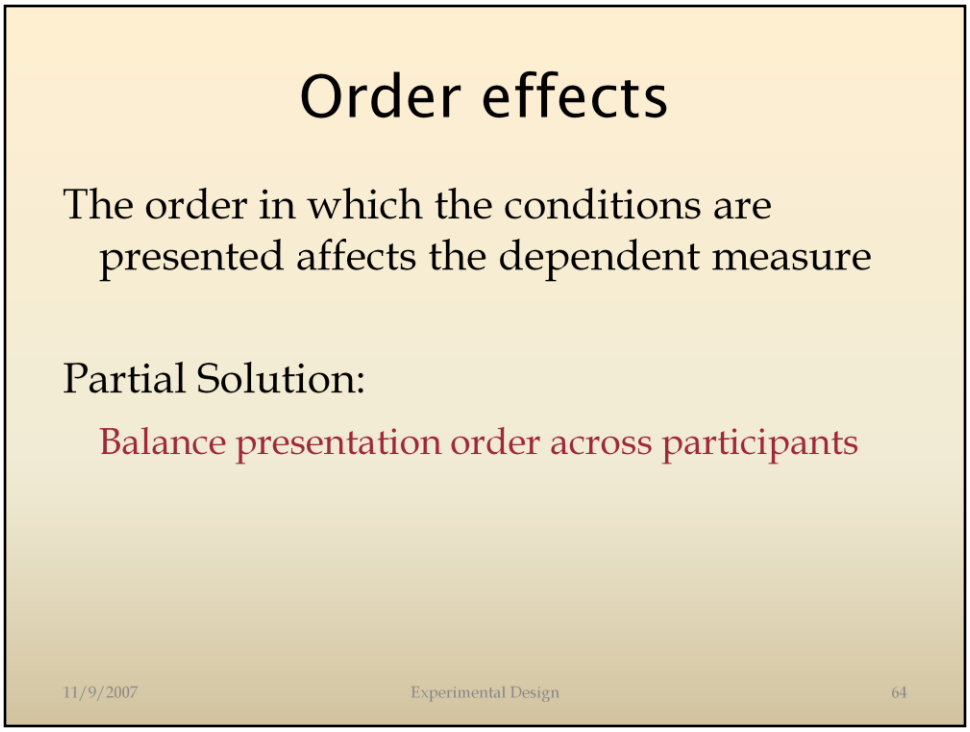


Within subjects uses the same subjects for several conditions. Between subjects uses different subjects for each condition.

Sometimes, within subjects help us eliminate noise that comes from the differences between users (some people are faster than others, but a particular technique might help everyone a little). Within subject designs require statistical tests that are designed for within subjects experiments.



In a case like this, a within subject design will not help us.



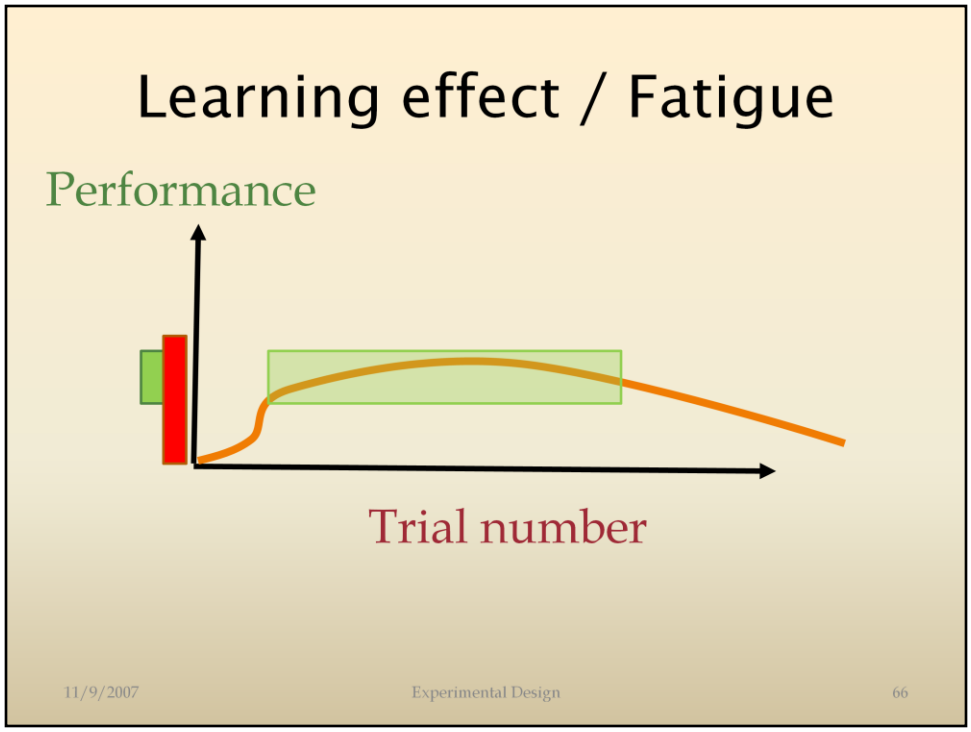
Sometimes, within subjects design are just not possible (you cannot use the same subject in two conditions). For example: testing two methods of learning a concept.

If you can use a within subject design, you have to take care of the order effect (think of the running with and without music experiment). Divide your subjects in different groups that perform the conditions in different orders.

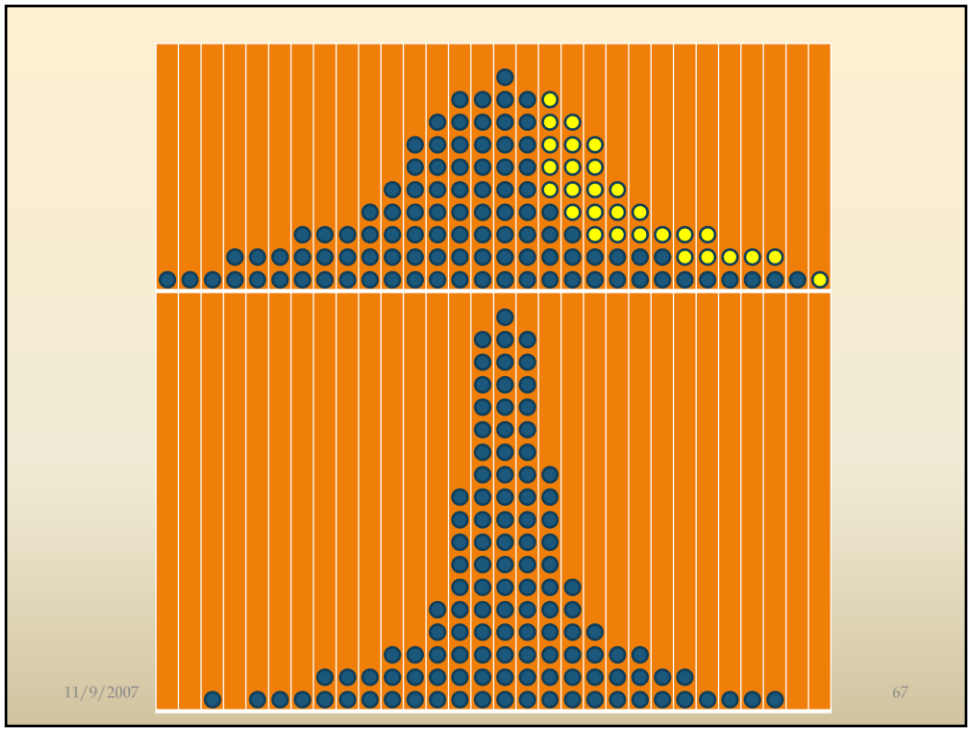
Pitfall 5

Balancing orders does not eliminate noise, it only distributes it evenly across conditions

You have to balance, but it is also important to avoid order effects as much as you can (e.g. get the people running in different weeks).



During learning time, participants are slower than normal, same thing when they are tired/bored. Reduce variability by measuring only the “flat” part of the curve: Give the users training that you will not consider real data, and don’t make an experiment too long.



Yellow marbles represent measures from fatigued/inexperienced users. If we remove them, our distribution is sharper.

Repeated measures

- Measure the same thing several times
- Then average

PART IV

PUTTING IT ALL TOGETHER

Significance and Relevance

A very low p-value means very high reliability (e.g. $p < 0.0001$)

But it does not say anything about the importance of your result

You could have a very high confidence that there is a difference between two techniques. However, if that difference is of a 0.0001% in efficiency, nobody will care.

Pitfall 6

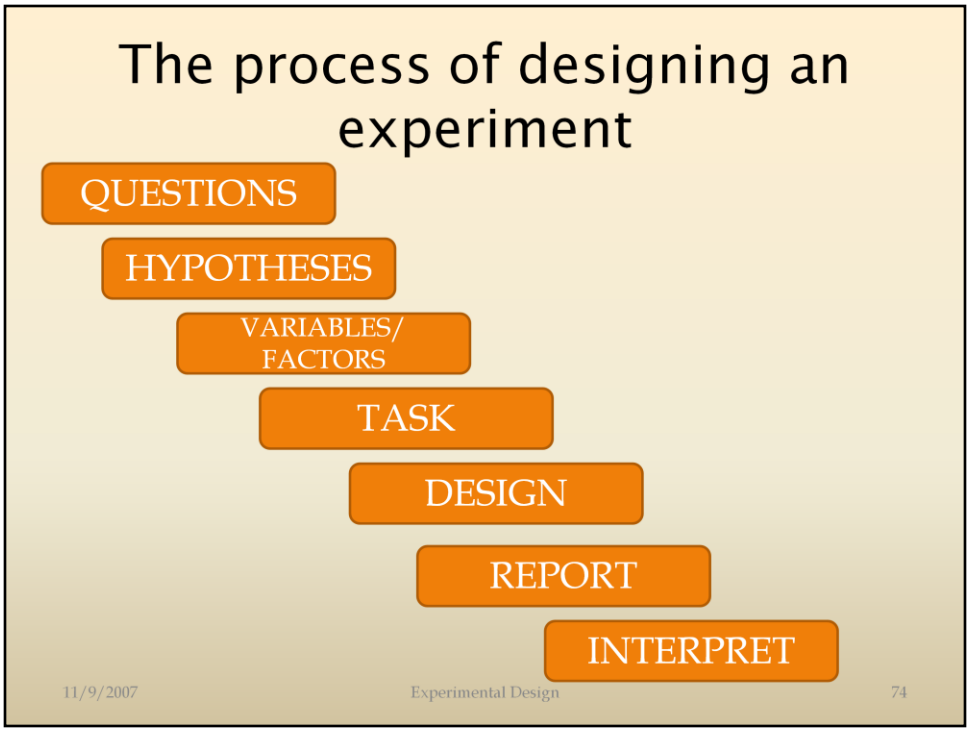
Not discussing the practical relevance of the differences

Not reporting the differences in mean (only report the p-values)

Interpreting the results

Pitfall 7

- Generalizing too much
- Constraining yourself to the strictly found



Summary of what we have seen.

It looks just as a waterfall model, but it doesn't really work like this. Try to imagine the last steps of the process before you start gathering data. Think about how you will interpret results, then work backwards.

Pilot studies can help you iterate several times and make less noisy experiments. It is normal to iterate 1 to 5 times with different configurations of pilot studies (studies in which you only test a few subjects, and where you do less strict statistical tests and eyeball trends etc.)

What you now know

- When and why to use controlled empirical evaluations
- The main elements of experimental design
- How to plan an experiment
- The main pitfalls of experimental design