Experimental design and analysis

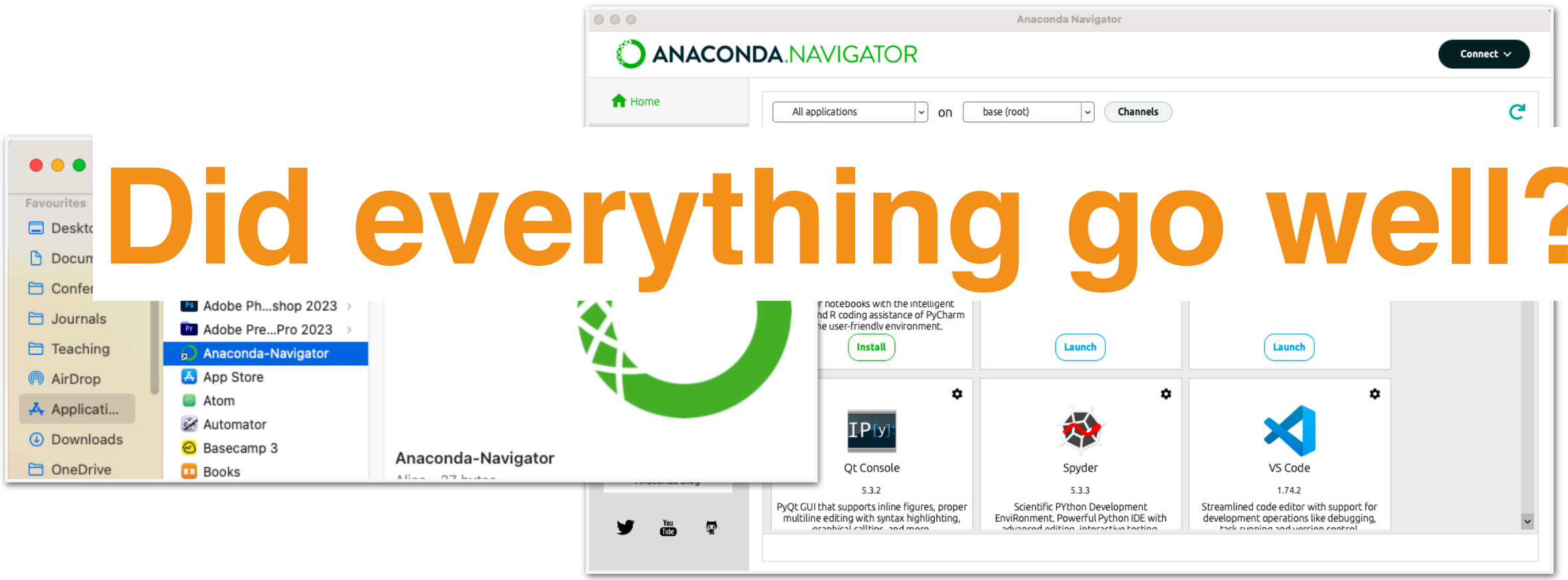
Statistical analyses for laboratory experiments

https://www.lri.fr/~appert/eval/

Caroline Appert

TODO

We will use Jupyter notebook for statistical analyses.
You have to install ANACONDA.NAVIGATOR
https://docs.anaconda.com/anaconda/install/

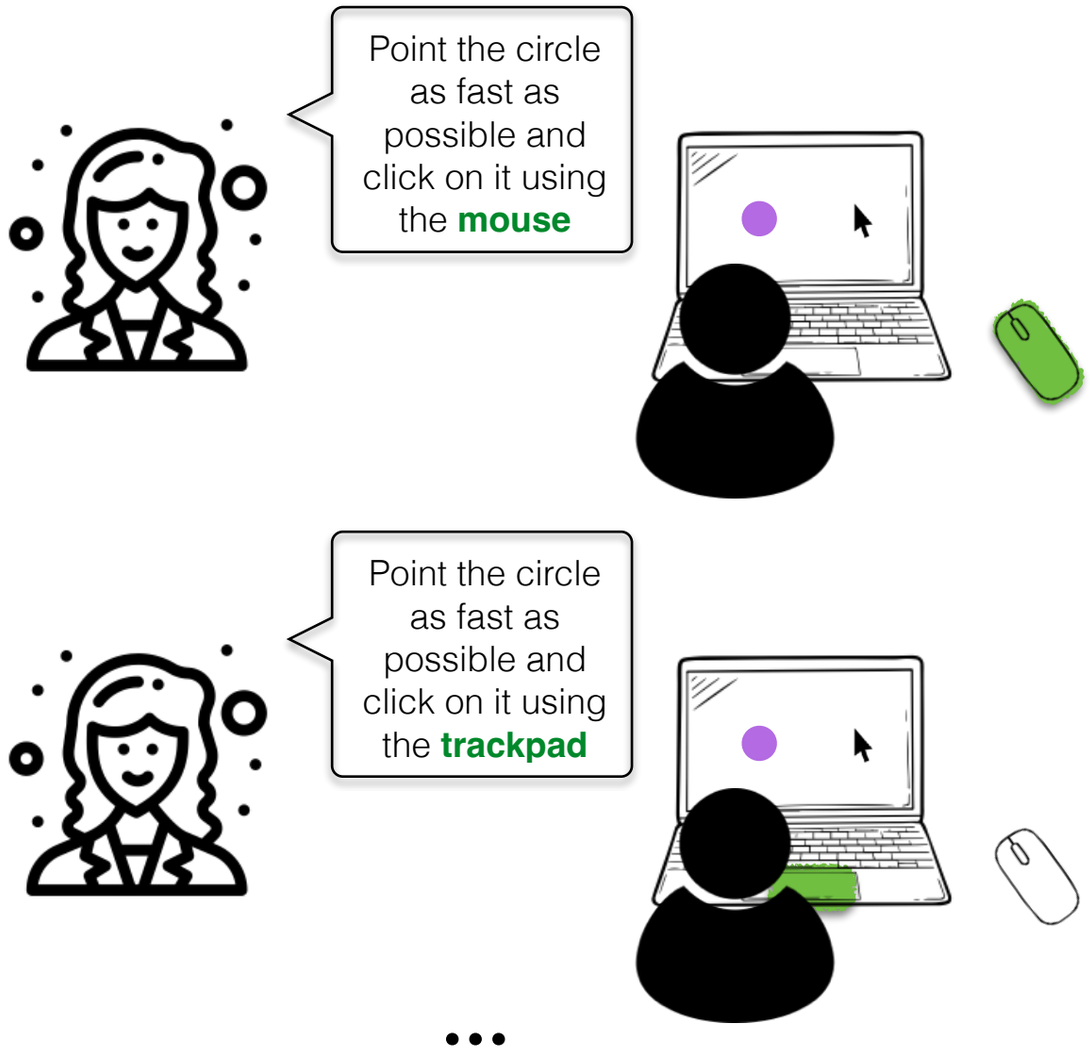


Did everything go well?

Analyzing experiment data

Laboratory experiment - overview

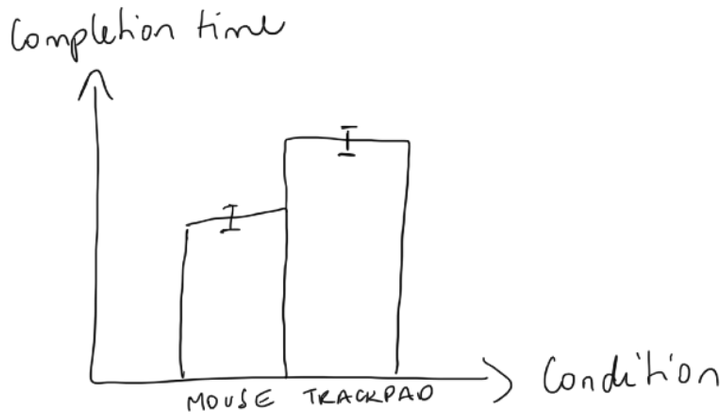
Hypothesis: Users point faster with a mouse than with a trackpad



1. The operator asks participants to complete tasks under **specific conditions**

Participant	Condition	Completion time
P1	mouse	403
P1	trackpad	527
⋮	⋮	⋮
P2	mouse	522
P2	trackpad	608
⋮	⋮	⋮

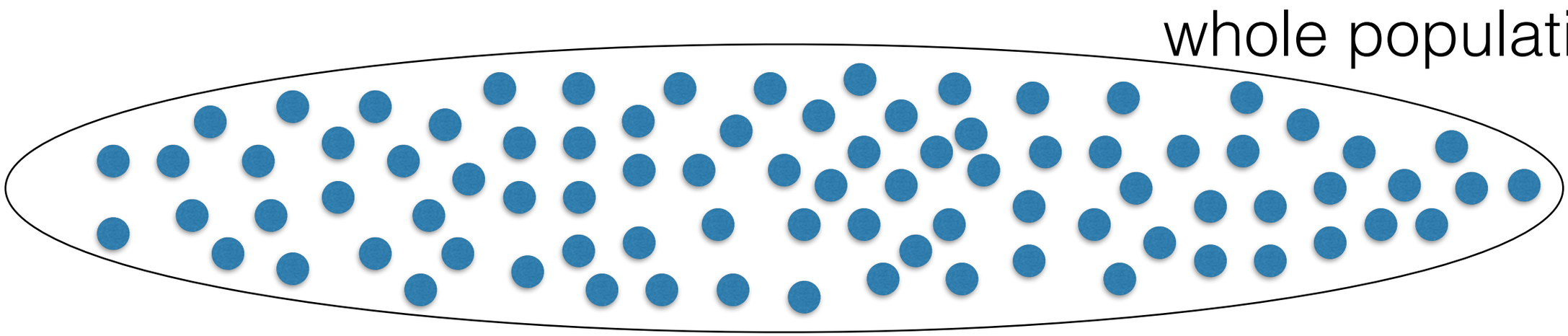
2. Participants' performance is recorded in log files (observations)

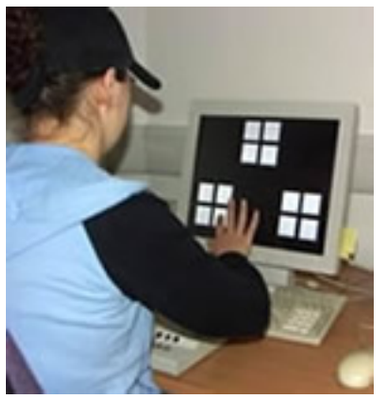


3. Observations are analyzed with statistical procedures to test the research hypothesis

image source: https://thenounproject.com/

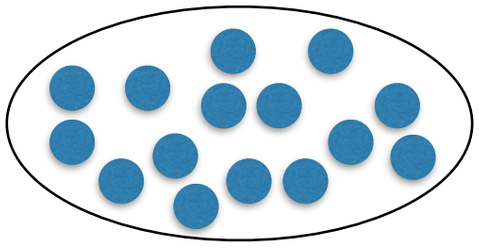
Experiment & statistics





Experiment

data sample



in tabular format

```
Participant,Practice,Block,Trial,Device,Difficulty,PointingTime
0,true,0,0,Trackpad,Easy,1632
0,true,0,1,Trackpad,Medium,1552
0,true,0,2,Trackpad,Hard,2030
0,false,1,0,Trackpad,Hard,1582
0,false,1,1,Trackpad,Medium,1639
...
11,false,3,19,Mouse,Easy,1582
11,false,3,20,Mouse,Hard,1639
```

● one observation (e.g., time for completing a pointing task)

What do we do now...

...that we have observations?

```
Participant,Practice,Block,Trial,Device,Difficulty,PointingTime
0,true,0,0,Trackpad,Easy,1632
0,true,0,1,Trackpad,Medium,1552
0,true,0,2,Trackpad,Hard,2030
0,false,1,0,Trackpad,Hard,1582
0,false,1,1,Trackpad,Medium,1639
...
11,false,3,19,Mouse,Easy,1582
11,false,3,20,Mouse,Hard,1639
```

General info

Factors

Measures

How do we conclude anything with respect to our initial hypothesis?

We use statistical analyses

...to provide the mathematical characteristics of data (descriptive statistics)

...to estimate the probability that a hypothesis is correct (inferential statistics)

e.g., our hypothesis was about a difference in pointing time between a mouse and a trackpad. We use statistics to describe observations in the mouse condition and observations in the trackpad condition in order to observe user performance. We also use statistics to estimate the probability that the difference observed in our sample is true for the whole population.

...to describe how datasets are related to each other (correlation -- both descriptive and inferential)

e.g., our hypothesis was that pointing time is a linear function of pointing difficulty. We use statistics to compute the linear correlation between pointing time and pointing difficulty. We can also estimate the probability that this linear relationship we observe in our sample is true for the population as a whole.

Data analysis and type of data

Applicable statistics depend on:

- Experiment design (between- vs. within-subject and the number of factors)
- Type of factors and measures (nominal, ordinal, etc.)

Possible data types:

Nominal	Categories (e.g., interaction technique)
Ordinal	Ranking, natural order (e.g., Likert scales)
Interval	Ordinal with equal intervals (e.g., Temperature in celsius degrees)
Ratio/Scalar	Intervals with a 0 point (e.g., duration)

Common types in HCI

Factors

- Often nominal (e.g., interaction technique, expertise...)
- Sometimes ordinal or ratio (e.g., Number of items in a menu, ID (index of difficulty) of a pointing task...)

Measures

- Mostly Ratio (e.g., completion time, number of errors...)
- and Ordinal (e.g., Level of preference for a technique...)

Families of statistics

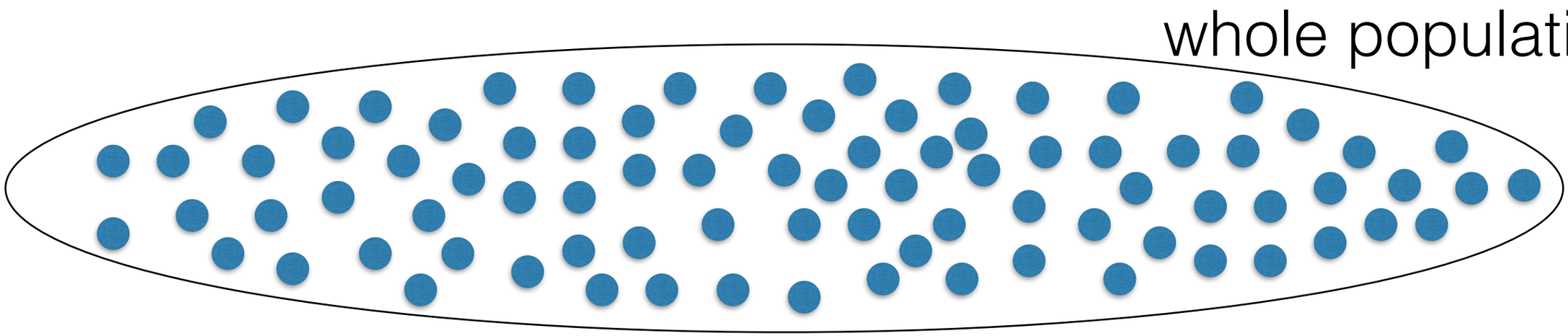
	Descriptive describe, show or summarize a data sample	**Inferential** make generalization about the whole population based on one data sample
Parametric make assumption about the data distribution for the whole population	Descriptive parametric statistics	Inferential parametric statistics
Non parametric no assumption about the data distribution	Descriptive non-parametric statistics	Inferential non-parametric statistics

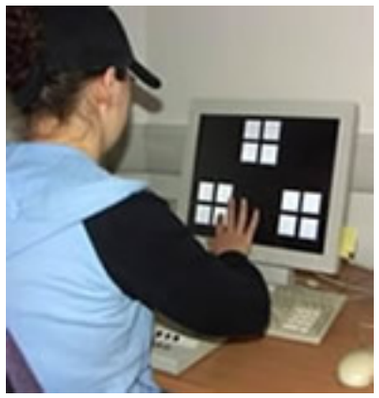
Descriptive and Inferential statistics

	Descriptive describe, show or summarize a data sample	**Inferential** make generalization about the whole population based on one data sample
Parametric make assumption about the data distribution for the whole population	Descriptive parametric statistics	Inferential parametric statistics
Non parametric no assumption about the data distribution	Descriptive non-parametric statistics	Inferential non-parametric statistics

describe, show or summarize a data sample

Experiment & descriptive statistics

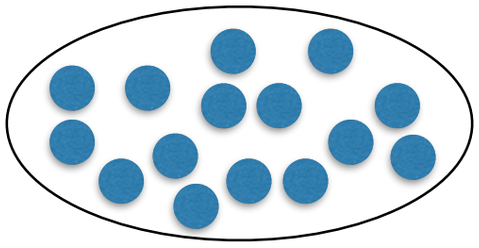




Experiment

data sample

(e.g., times for completing a pointing task)



descriptive statistics describe only the sample collected in the experiment

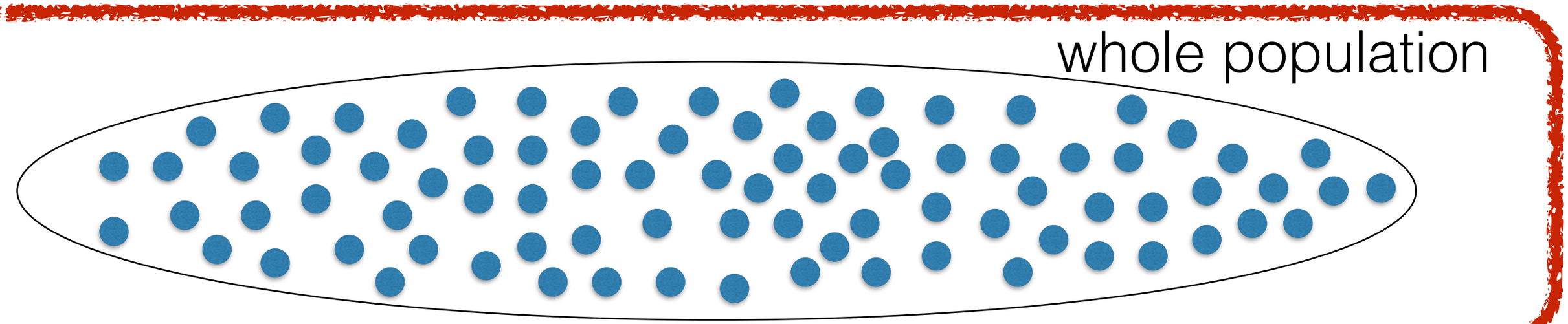
one observation (e.g., time for completing a pointing task)

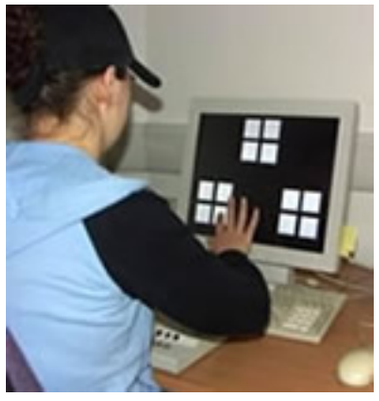
Descriptive and Inferential statistics

	Descriptive describe, show or summarize a data sample	**Inferential** make generalization about the whole population based on one data sample
Parametric make assumption about the data distribution for the whole population	Descriptive parametric statistics	Inferential parametric statistics
Non parametric no assumption about the data distribution	Descriptive non-parametric statistics	Inferential non-parametric statistics

make generalizations about the population from which the sample was drawn

Experiment & inferential statistics

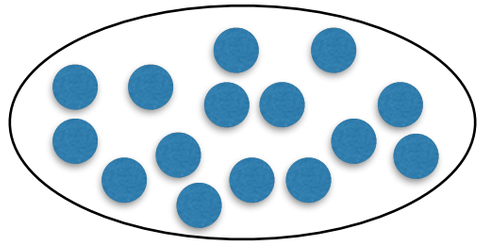




Experiment

data sample

(e.g., times for completing a pointing task)



inferential statistics use the sample to provide a description of the whole population that can be trusted only with a given probability

one observation (e.g., time for completing a pointing task)

Parametric and non-parametric statistics

	Descriptive describe, show or summarize a data sample	**Inferential** make generalization about the whole population based on one data sample
Parametric make assumption about the data distribution for the whole population	Descriptive parametric statistics	Inferential parametric statistics
Non parametric no assumption about the data distribution	Descriptive non-parametric statistics	Inferential non-parametric statistics

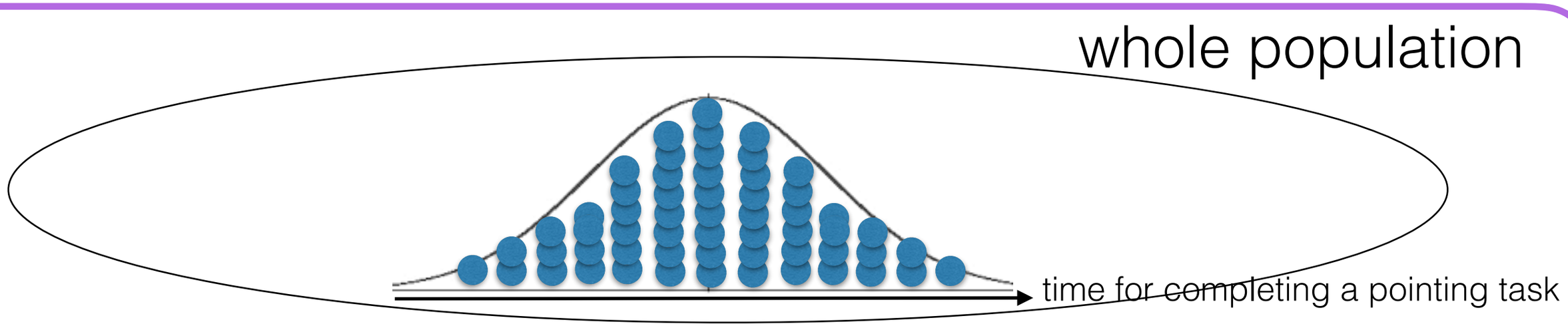
assumption about the data distribution for the population

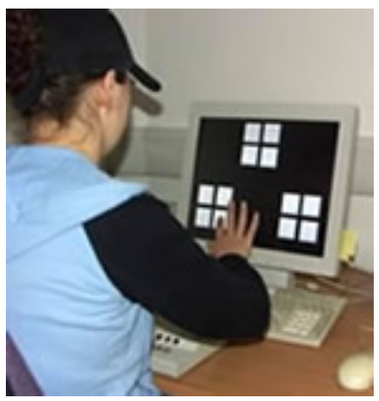
Parametric and non-parametric statistics

	Descriptive describe, show or summarize a data sample	**Inferential** make generalization about the whole population based on one data sample
Parametric make assumption about the data distribution for the whole population	Descriptive parametric statistics	Inferential parametric statistics
Non parametric no assumption about the data distribution	Descriptive non-parametric statistics	Inferential non-parametric statistics

no assumption about the data distribution for the population

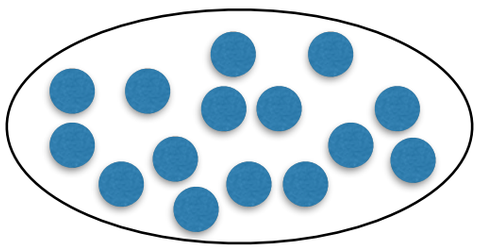
Experiment & parametric statistics





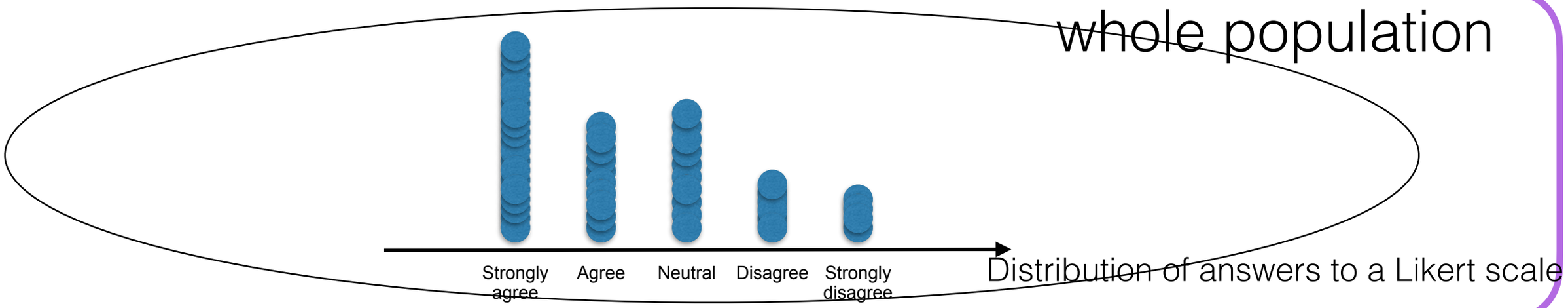
Experiment

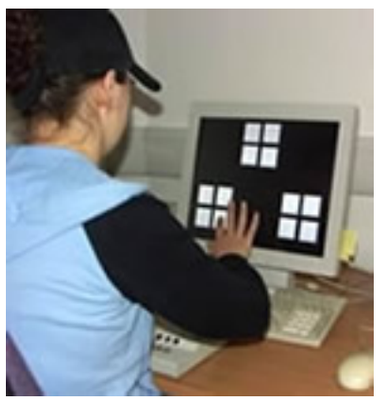
data sample



parametric statistics assumes that the data sample comes from **a population that follows a probability distribution** based on a fixed set of parameters (for example, a normal distribution)

Experiment & non-parametric statistics





Experiment

data sample

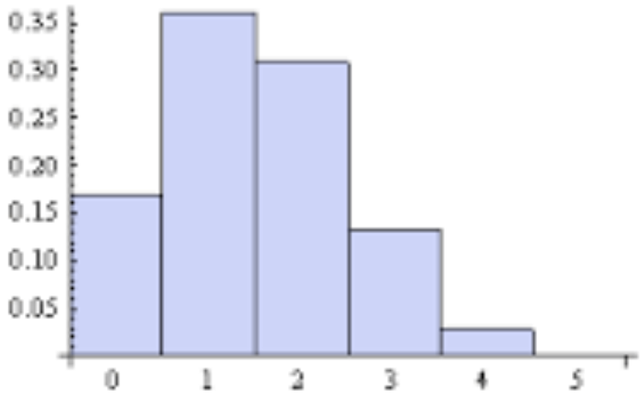
we assume nothing about the population

What is a distribution?

A frequency distribution is a table/graph that displays the frequency of various outcomes in a dataset (sample or whole population)

discrete outcome

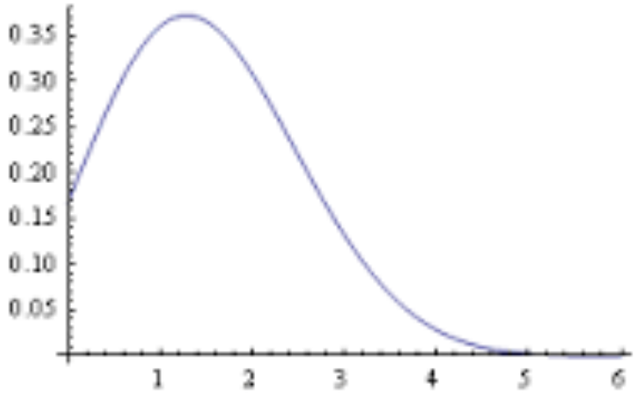
y: proportion among all observations



x: outcomes

continuous outcome (density)

y: proportion among all observations



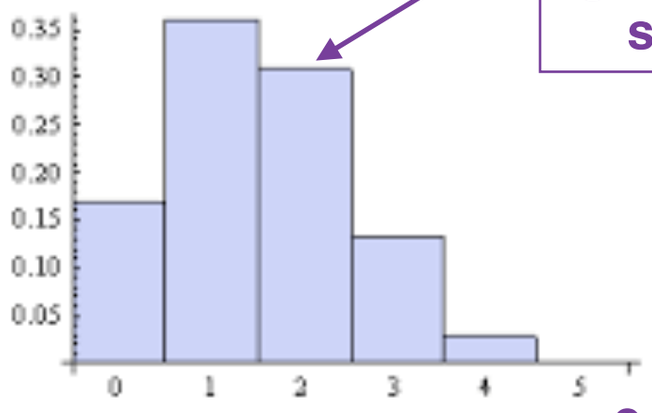
x: outcomes

What is a distribution?

A frequency distribution is a table/graph that displays the frequency of various outcomes in a dataset (sample or whole population)

discrete outcome

y: proportion among all observations



30% of participants gave a fatigue score of 2

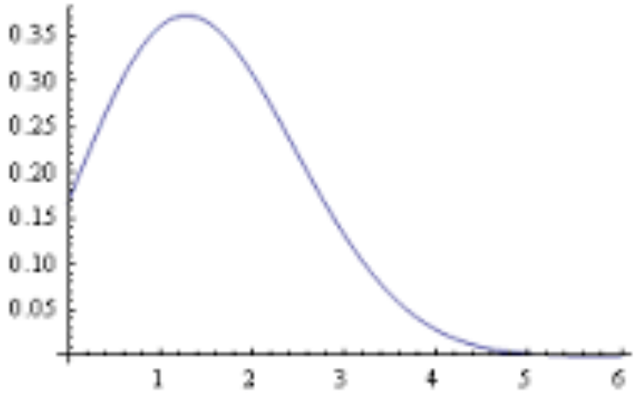
x: outcomes

e.g., fatigue score

(estimation on a Likert scale of the perceived fatigue)

continuous outcome (density)

y: proportion among all observations



x: outcomes

e.g., completion time in s

Descriptive statistics

describe a sample

Descriptive statistics

Describe the distribution of observed values

For ratio variables,

- Central tendency (mean, median, mode)
- Spread (variance, standard deviation)
- Correlation

For any type of variable,

- Range ([min, max]) *(except nominal)*
- frequency distribution (number of observations per value)

Central tendency

Ex: variable values {1, 1, 2, 5, 7, 1, 5, 6, 12, 5, 2}

Mean: sum of values divided by their number

mean = (1+1+2+5+7+1+5+6+12+5+2)/11 = 4.27

Median: “middle” value of the N sorted values

N is odd: {1,1,1,2,2,5,5,5,6,7,12}, median = 5

N is even: {1,1,1,1,2,2,5,5,5,6,7,12}, median = (2+5)/2 = 3.5

Mode: the most frequent value(s)

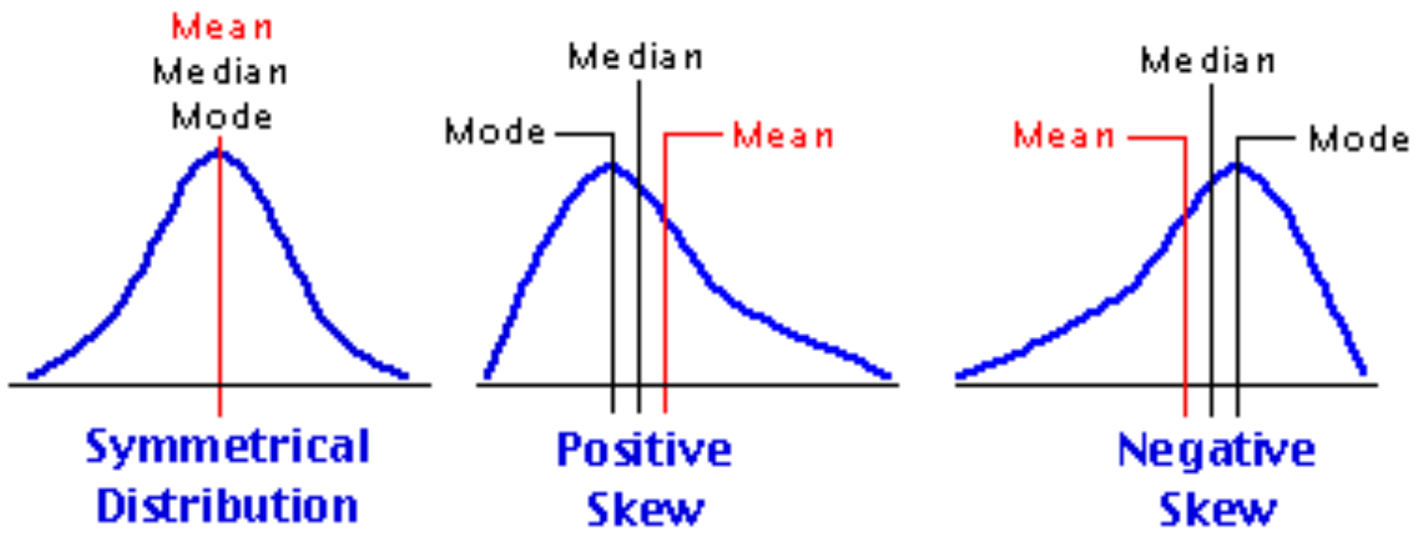
{1,1,1,2,2,5,5,5,6,7,12}, modes = 1 and 5

{1,1,1,1,2,2,5,5,5,6,7,12}, mode = 1

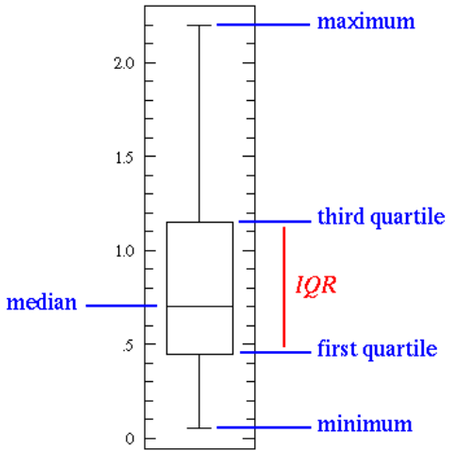
Mean or Median?

Mean is best for symmetric distributions without outliers

Median is useful for skewed distributions or data with outliers



Quartiles



Quartiles of a set of values divide the data set into four groups of equal size (i.e., each group has the same number of values)

first quartile = lower quartile = splits lowest 25% of data = 25th percentile

second quartile = median = cuts data set in half = 50th percentile

third quartile = upper quartile = splits highest 25% of data = 75th percentile

Spread

Variance gives the tendency for the individual measures to spread out away from each other

$$\sigma^2 = \frac{\sum_{i}^{n}(x_i - \overline{x})^2}{n}$$

n is the number of observations
$\overline{x}$ is the mean

Squares eliminate the negatives

Inferential statistics

use a sample to describe the whole population (with some uncertainty)

Inferential statistics

Make generalizations from a sample to a population

Use the probability theory to make inferences from the sample to the population

Many tests in inferential statistics according to the types of your variables and your design

t-test, Wilcoxon test, Chi-square test, ANOVA, ...

Experiment & inferential statistics

Informal

Inferential statistics compute a statistic (e.g., t, F, etc.) that assesses the difference between experimental conditions and estimates the probability of observing such a value by chance.

The specific statistic (e.g., t, F, etc.) depends on the experiment design, number of experimental conditions, the type of the measure, etc.

BUT, whatever the specific test you run, the reasoning behind an inferential statistical test is always the same. The test:

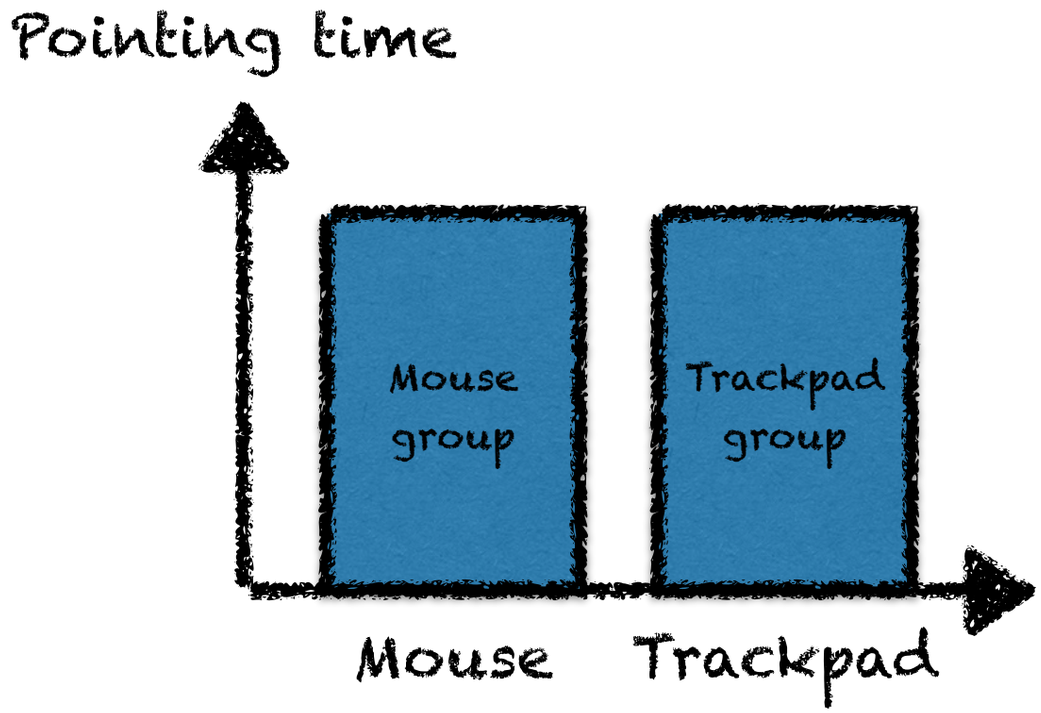
- calculates a statistic (i.e., a number) for your observations that is a measure of the difference between experimental conditions,
- assesses how likely it is for such a value to occur by chance if there is no difference between the experimental conditions being compared (Null hypothesis reasoning), and
- concludes that if it is unlikely to observe such a value by chance (low p-value), there is very likely a difference between the experimental conditions.

Experiment & inferential statistics

Informal

$H_{research}$: Users point faster with a Mouse than with a Trackpad

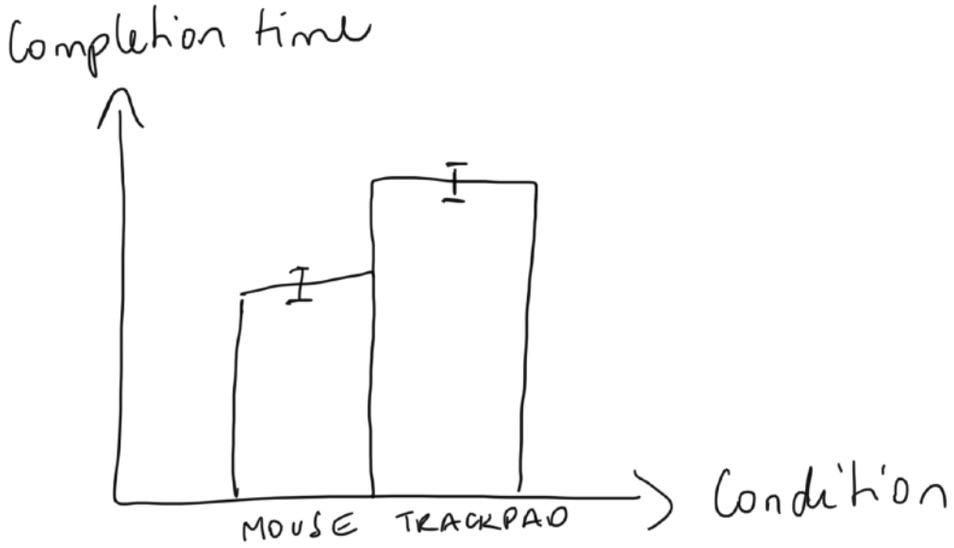
H_{null}: Users point as fast with a Mouse as with a Trackpad
(i.e., there is no difference between our two experimental groups)



Experiment & inferential statistics

Informal

Actual observations after having run the experiment:



In this example, because we compare a continuous measure between two experimental conditions, we run a t test (compute a t statistic)

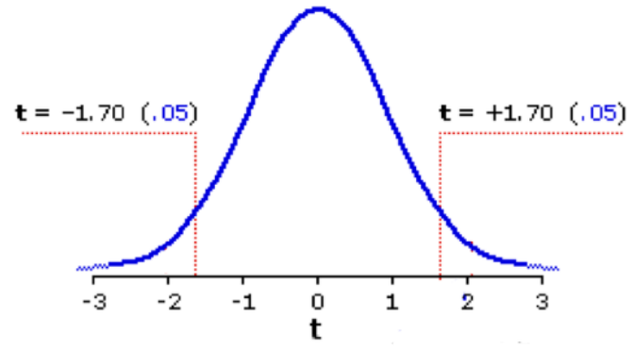
$$t_{obs} = 1.9$$

What does that mean?

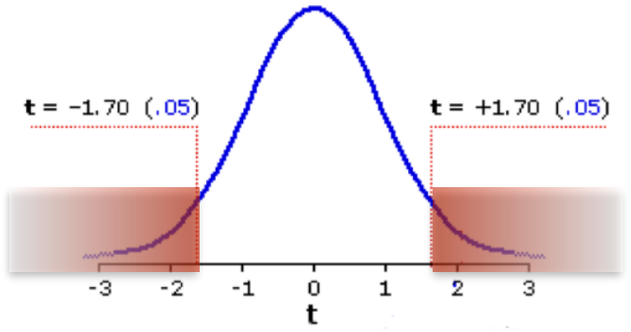
Null Hypothesis Significance Test

Informal

The statistical test knows what the distribution of the possible values for t is <u>when the null hypothesis is true</u>

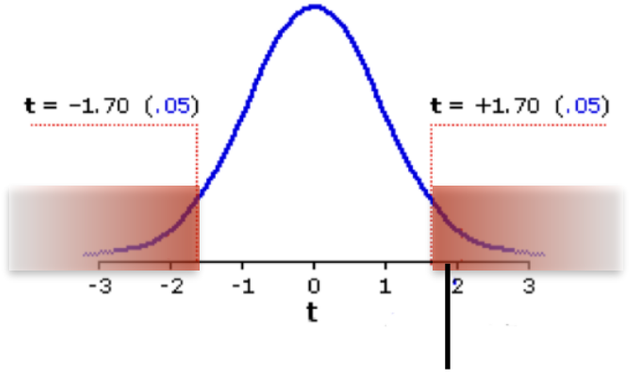


Before running the experiment, we decide on the level of statistical significance α (usually 0.05, i.e. 5%). We decide on the "red zones", *i.e.*, the values that can happen by chance with a probability lower than α.



In red: values for the statistic t that we are unlikely to observe if the null hypothesis is true (less than 5% of the cases if $\alpha = 0.05$).

The test computes a statistic value for our observations (say t_{obs}). If this value lies within the "red zones", we reject the null hypothesis. We observe a difference that is statistically significant.



$t_{obs} = 1.9$

$\Rightarrow p < 0.05$

Inferential statistics and random

Sampling error - chance, random error

Sample bias - constant error, due to inadequate design

loaded dice

Inferential statistics take into account sampling error.

Inferential statistics do not correct for sample bias.

⇦ **typically happens if the experiment design is not internally valid**

Your experiment design must ensure that the sample is representative!

From http://psychology.ucdavis.edu/sommerb/sommerdemo/stat_inf/intro.htm

Level of statistical significance

Inferential statistics use the probability theory

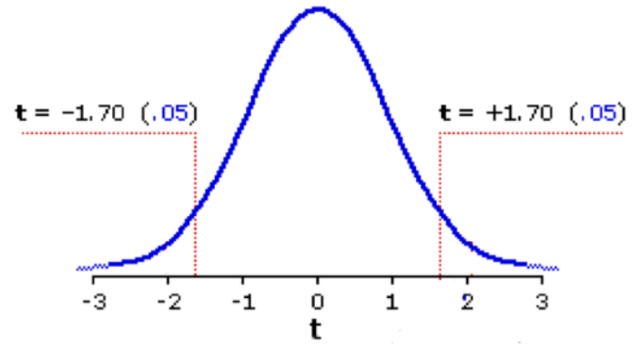
A test based on inferential statistics outputs a p value. This p value provides an estimate of how often we would get the obtained result by chance, if in fact the null hypothesis were true.

How can we decide to reject a hypothesis? We decide on a level of significance, *i.e.*, the cutoff (α) before running the test. We reject the null hypothesis when $p < \alpha$.

Usually, the cutoff is set to 0.05 (as proposed by statistician Fisher)

Statistics and degrees of freedom *df*

The distribution of the possible values for a statistic depends on its degrees of freedom.



The degrees of freedom of an inferential statistic is the number of values in the final calculation of that statistic that are free to vary.

Degrees of freedom typically relate to the size of the sample.

Intuitive explanation of what degrees of freedom are in a calculation.
For example: if we know that mean = 20

(□ + □) / 2 = 20 *df* = 1

the value of one cell is set once the other one is set

(□ + □ + □) / 3 = 20 *df* = 2

the value of one cell is set once the other two are set

Degrees of freedom (df)

Why are dfs important?

dfs are related to the sample size, then informing whether the statistic has been computed based on a large number of observations or not.

Higher degrees of freedom generally mean larger sample sizes, and thus more "trust" (*power*) in the statistical result.

Type I and Type II errors

When using inferential statistics, we draw conclusions with some uncertainty. We can thus make errors.

Type I error

> Reject the null hypothesis when it is true

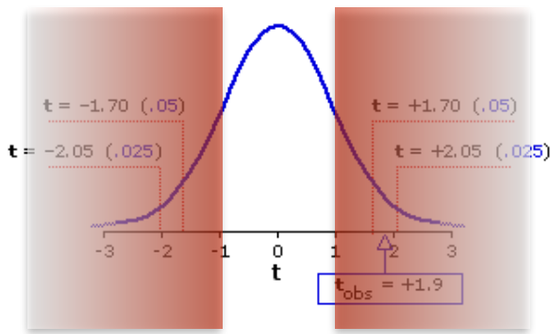
Type II error

> Accept the null hypothesis when it is false

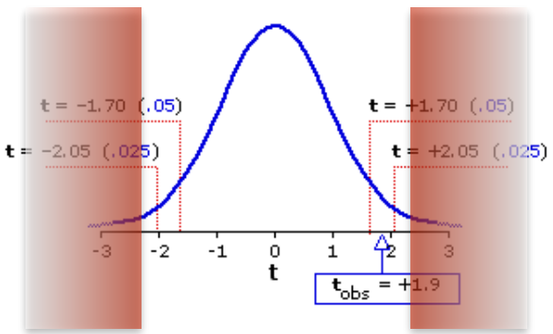
A high statistical significance level (α) increases the chances of a type I error

A low level of statistical significance (α) increases the chances of a type II error

High level of statistical significance



Low level of statistical significance



Power of a test

	Reject the null hypothesis	Fail to reject the null hypothesis
The null hypothesis is true.	Type I error (false positive)	True negative
The null hypothesis is false.	True positive	Type II error (false negative)

α is the level of significance. The statistical test outputs a p value ($p < \alpha$ means significant)

p value is the probability of Type I error

β is the probability of Type II error

1-β is the power of the test

Significance and Effect size

Rejecting the null hypothesis means that we observe a **significant effect**

Significant does not mean large

With a large sample size, very small differences will be detected as significant (the larger the sample, the smaller the standard error is)

You must use a complementary test to measure the size of the effect

Effect size

An effect size is a measure of the strength of a phenomenon

An effect size is a descriptive statistic that does not make any statement about whether the apparent relationship in the sample reflects a true relationship in the whole population.

Effect sizes complement inferential statistics

Effect size

Whether an effect size should be interpreted as small, medium, or large depends on its substantive context and its operational definition

It is usually a normalized value, with recommendations to interpret it

Power analysis

Power analysis means considering:

The sample size,

The effect size (the one of the underlying population and not the observed one),

α (usually 0.05) and β (usually 0.80)

Running a power analysis means estimating one of these values given the three other ones

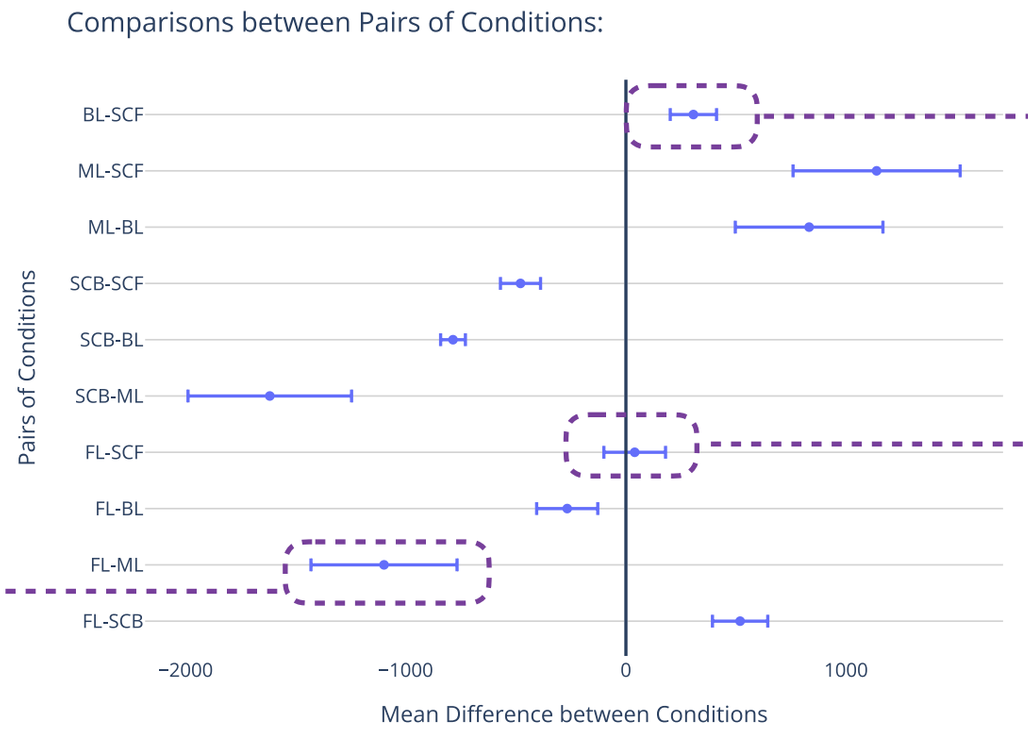
usually used to compute the minimum sample size
(we won't go into the computation details in this class but we will see how to run such a power analysis with TouchStone 2)

The null hypothesis reasoning: critics

Dichotomy regarding how to accept or reject the null hypothesis on an arbitrary cut-off can be criticized (and actually is criticized)

p = 0.051 => reject, p = 0.049 => accept, mmm...

Some analysts prefer to focus on confidence intervals regarding the difference in means to look at their data and make more nuanced conclusions.



The confidence interval for the difference between these two conditions is *close to* 0
=> some (weak) evidence

The confidence interval for the difference between these two conditions includes 0
=> inconclusive

The confidence interval for the difference between these two conditions is *far from* 0
=> strong evidence