Evaluation of Interactive Systems

Statistics

Caroline Appert - 2018/2019
Statistical analyses

Used to provide the mathematical characteristics of data

Used to describe how datasets are related to each other

*Used to estimate the probability that a hypothesis is correct*
Factors and measures

Factor = independent variable

interaction technique
expertise
...

Measure = dependent variable

completion time
number of errors
...

Data analysis

Type of data and experiment design determine applicable statistics

Data types

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>Categories (e.g., interaction technique)</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Ranking, natural order (e.g., Likert scales)</td>
</tr>
<tr>
<td>Interval</td>
<td>Ordinal with equal intervals (e.g., Temperature)</td>
</tr>
<tr>
<td>Ratio/Scalar</td>
<td>Intervals with a 0 point (e.g., duration)</td>
</tr>
</tbody>
</table>
Common variables in HCI

Factors

Often nominal (e.g., interaction technique, expertise...)

Sometimes ordinal or scalar (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

<table>
<thead>
<tr>
<th></th>
<th><strong>Descriptive</strong></th>
<th></th>
<th><strong>Inferential</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>describe, show or summarize a data sample</td>
<td></td>
<td>make generalization about the whole population based on one data sample</td>
<td></td>
</tr>
<tr>
<td><strong>Parametric</strong></td>
<td></td>
<td>Descriptive parametric statistics</td>
<td>Inferential parametric statistics</td>
<td></td>
</tr>
<tr>
<td>make assumption</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>about the data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>distribution for the</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>whole population</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Non parametric</strong></td>
<td></td>
<td>Descriptive non-parametric statistics</td>
<td>Inferential non-parametric statistics</td>
<td></td>
</tr>
<tr>
<td>no assumption</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>about the data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>distribution</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Descriptive and Inferential statistics

Descriptive statistics

describe, show or summarize a data sample

Inferential statistics

make generalizations about the population from which the samples were drawn.
Experiment & statistics

whole population

data sample

Experiment
Experiment & descriptive statistics

descriptive statistics describe only the sample collected in the experiment
Experiment & inferential statistics

Experiment data sample

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

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)
Distributions

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
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
\[
\text{mean} = \frac{1+1+2+5+7+1+5+6+12+5+2}{11} = 4.27
\]

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 equal groups of equal size

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

\[ \sigma^2 = \frac{\sum_{i}^{n} (x_i - \bar{x})^2}{n} \]

- \( n \) is the number of observations
- \( \bar{x} \) 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,
Inferential statistics and random

**Sampling error** - chance, random error

Inferential statistics take into account sampling error.

**Sample bias** - constant error, due to inadequate design

Inferential statistics do not correct for sample bias.

Your experiment design must ensure that the sample is representative!
Level of statistical significance

Inferential statistics uses the probability theory.

The $p$ value provides an estimate of how often we would get the obtained result by chance, if in fact the null hypothesis were true.

When a test outputs a given $p$ value, how can we decide to reject a hypothesis? The decision is based on the level of significance, i.e., the cutoff ($\alpha$) we fixed before running the test.

Usually, the cutoff is set to 0.05 (as proposed by Fisher).
Statistics and degrees of freedom

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

Example: mean is 20

\[
\frac{[\square + \square]}{2} = 20 \quad df = 1
\]

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

\[
\frac{[\square + \square + \square]}{3} = 20 \quad df = 2
\]

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.

The number of DFs determines the assumption that we can draw on the theoretical distribution of observations in our population (we will get back to it when introducing the t-test).
Inferential statistics

Testing and identifying a relation between a ratio factor and a ratio measure

correlation
linear regression
Correlation

The Pearson’s correlation coefficient, \( r \), measures how linear a relationship between two ratio variables is

\[
 r = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \times (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \quad (-1 \leq r \leq 1)
\]

Usually, \( X \) is a factor, \( Y \) is a measure

\( r^2 \) is interpreted as the proportion of the variability of \( Y \) that is associated with the variability of \( X \)

\( 1 - r^2 \) is the residual variance (not explained)
Correlation

Pearson’s correlation coefficient ($r$) tells how much one variable tends to change when the other one does.

- $r = 0$, there is no relationship
- $r > 0$, there is a trend that one variable goes up as the other one goes up
- $r < 0$, there is a trend that one variable goes up as the other one goes down

Correlation is a measure of dependence.

Correlation $\neq$ Causality
Linear regression

Computing linear regression means defining the *regression line* that best fits the bivariate distribution of data points (Makes the squared vertical distances between the data points and regression line as small as possible)

Linear regression can be used as a predictive model when your experiment design is sound enough and the result of statistical tests is significant to support a cause-effect relation
Example

Pointing performance of different types of magnifying lenses

3 factors (5 x 5 x 5 design) - 10 participants

Lens type:
- **ML** (Manhattan Lens)
- **FL** (Fisheye Lens)
- **BL** (Blending Lens)
- **SCF**
- **SCB**

Lens’ magnification: 2, 4, 6, 10, 14

Index of Difficulty (ID): 4.2, 5.3, 6, 7, 8

Task: target acquisition

Measure: Pointing time (in ms)
Example
Pointing performance of different types of magnifying lenses

3 factors (5 x 5 x 5 design) - 10 participants

Collected data
(log file lens_experiment.csv)

<table>
<thead>
<tr>
<th>Participant</th>
<th>Block</th>
<th>Trial</th>
<th>Lens</th>
<th>Magnification</th>
<th>ID</th>
<th>PointingTime</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>0</td>
<td>FL</td>
<td>6</td>
<td>6.0035549</td>
<td>2297</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0</td>
<td>FL</td>
<td>6</td>
<td>6.0035549</td>
<td>1485</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0</td>
<td>FL</td>
<td>6</td>
<td>6.0035549</td>
<td>2000</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0</td>
<td>FL</td>
<td>6</td>
<td>6.0035549</td>
<td>1843</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0</td>
<td>FL</td>
<td>6</td>
<td>6.0035549</td>
<td>1813</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>9</td>
<td>SCF</td>
<td>6</td>
<td>6.0035549</td>
<td>2375</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>9</td>
<td>SCF</td>
<td>6</td>
<td>6.0035549</td>
<td>2359</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>9</td>
<td>SCF</td>
<td>6</td>
<td>6.0035549</td>
<td>2313</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>9</td>
<td>SCF</td>
<td>6</td>
<td>6.0035549</td>
<td>2453</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>9</td>
<td>SCF</td>
<td>6</td>
<td>6.0035549</td>
<td>2187</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>9</td>
<td>SCF</td>
<td>6</td>
<td>6.0035549</td>
<td>2875</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>9</td>
<td>SCF</td>
<td>6</td>
<td>6.0035549</td>
<td>2688</td>
</tr>
</tbody>
</table>

...
Example

Pointing performance of different types of magnifying lenses

Hypothesis

Pointing with a lens of type ML follows Fitt’s law (i.e., if there exists a linear model that predicts Time according to ID)
Example
Pointing performance of different types of magnifying lenses

Hypothesis
Pointing with a lens of type FL follows Fitt's law (i.e., if there exists a linear model that predict Time according to ID)

R code for linear regression

```r
lens_experiment <- read_delim("lens_experiment.csv", ";", trim_ws = TRUE)

# file helper_functions.R contains methods (e.g., summarySE)
# that are useful for aggregating data frames
source("helper_functions.R")

# Optional: aggregate replications (note: this likely increases the r coefficient)
lens_experiment_summary <- summarySE(lens_experiment, measurevar="PointingTime",
groupvars=c("Participant", "Lens", "ID"))

# filter out data for keeping trials in Lens="FL" condition only
fisheye_data <- lens_experiment_summary[lens_experiment_summary$Lens == "FL", ]

linearmodel <- lm(fisheye_data$PointingTime ~ fisheye_data$ID)
summary(linearmodel)
```
Example
Pointing performance of different types of magnifying lenses

Hypothesis

Pointing with a lens of type FL follows Fitt’s law (i.e., if there exists a linear model that predict Time according to ID)

R output

```r
Call:
  lm(formula = fisheye_data$PointingTime ~ fisheye_data$ID)

Residuals:
     Min      1Q  Median      3Q     Max
-582.1 -311.9 -133.5   279.2   840.7

Coefficients:
            Estimate Std. Error  t value Pr(>|t|)
(Intercept) -2146.54     264.17  -8.126  1.42e-10 ***
  fisheye_data$ID    743.83      42.24  17.611  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 396.5 on 48 degrees of freedom
Multiple R-squared:  0.866,  Adjusted R-squared:  0.8632
F-statistic: 310.2 on 1 and 48 DF,  p-value: < 2.2e-16
```
Report a simple linear regression

Coefficient of determination $r^2$
(for simple linear regressions, $r$ is Pearson’s correlation coefficient)

F statistics and p value (more details later)

Regression line function (intercept and slope)

R output

Call:
`lm(formula = fisheye_data$PointingTime ~ fisheye_data$ID)`

Residuals:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>-582.1</td>
</tr>
<tr>
<td>1Q</td>
<td>-311.9</td>
</tr>
<tr>
<td>Median</td>
<td>-133.5</td>
</tr>
<tr>
<td>3Q</td>
<td>279.2</td>
</tr>
<tr>
<td>Max</td>
<td>840.7</td>
</tr>
</tbody>
</table>

Coefficients:

|                     | Estimate | Std. Error | t value | Pr(>|t|) |
|---------------------|----------|------------|---------|---------|
| (Intercept)         | -2146.54 | 264.17     | -8.126  | 1.42e-10*** |
| fisheye_data$ID    | 743.83   | 42.24      | 17.611  | < 2e-16 *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 396.5 on 48 degrees of freedom
Multiple R-squared: 0.866, Adjusted R-squared: 0.8632
F-statistic: 310.2 on 1 and 48 DF, p-value: < 2.2e-16

“A significant regression equation was found, with a $r^2$ of 0.86 ($F_{1,48} = 310, \ p<0.001$). Predicted Pointing Time (PT) in ms is equal to $-2146 + 744 \times \text{ID}$ (ID being the Index of Difficulty).”
Linear regression

Must be interpreted with caution

Needs to be visualised

Anscombe's quartet:
same linear regression line but very different datasets…
Example

Pointing performance of different types of magnifying lenses

Visualize Time as a function of ID

```r
library(ggplot2) # ggplot2 is a library for creating charts in R

lens_experiment_summary <- summarySE(lens_experiment, measurevar="PointingTime",
groupvars=c("Participant", "Lens", "ID"))

ggplot(lens_experiment_summary, aes(x=ID, y=PointingTime, color=Lens)) + # map chart attributes to data
  geom_point(shape=1) + # add a point layer (one circle per data point)
  geom_smooth(method=lm) + # add a regression line layer
  theme(text = element_text(size=30), legend.position = "top") # adjust label size and legend position
```
Example
Pointing performance of different types of magnifying lenses

Visualize Time as a function of ID

```r
library(ggplot2) # ggplot2 is a library for creating charts in R

lens_experiment_summary <- summarySE(lens_experiment, measurevar="PointingTime",
groupvars=c("Participant", "Lens", "ID"))

ggplot(lens_experiment_summary, aes(x=ID, y=PointingTime, color=Lens)) + # map chart attributes to data
  theme(text = element_text(size=30), legend.position = "top") + # adjust label size and legend position
  geom_point(shape=1) + # add a point layer (one circle per data point)
  geom_smooth(method=lm, se=FALSE) # add a regression line layer
```

Removes the confidence interval envelope
Example

Pointing performance of different types of magnifying lenses

Visualize Time as a function of ID

R code

```r
library(ggplot2) # ggplot2 is a library for creating charts in R

lens_experiment_summary <- summarySE(lens_experiment, measurevar="PointingTime", groupvars=c("Participant", "Lens", "ID"))

ggplot(lens_experiment_summary, aes(x=ID, y=PointingTime, color=Lens)) + # map chart attributes to data
  theme(text = element_text(size=30), legend.position = "top") + # adjust label size and legend position
  geom_point(shape=19, alpha=1/4) + # add a point layer (one circle per data point)
  geom_smooth(method=lm, se=FALSE) # add a regression line layer

ggsave("linearregression.pdf", plot = last_plot()) # save plot in a file named linearregression.pdf
```

Use solid circles, with transparency
Inferential statistics

Testing the effect of nominal factor(s) on a measure

- t-test
- ANOVA
- Chi-square
- etc.
Null Hypothesis Significance Test

State the null hypothesis based on your research hypothesis (H₀: the factor does not impact the measure value)

Decide the level of statistical significance α (usually 0.05), i.e., the probability of Type I error

A Null Hypothesis Significance Test proceeds as follows: it computes a statistic s (*) and tells what the probability of observing such a value is when the null hypothesis is true. If this probability (p value) is smaller than α, then we can reject the null hypothesis (based on the fact that there is very little chance to observe such a result if there was actually no difference).

(*) the specific test to be run depends on the type of your measure, assumption about the distribution, and the type of design
Type I and Type II 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 increases the chances of a type II error

A low level of statistical significance increases the chances of a type I error
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 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.
Confidence interval

A *Confidence Interval (CI)* is as a range of plausible values for the population mean (or another population parameter such as a correlation), calculated from our sample data.

e.g., a CI with a 95 percent confidence level has a 95 percent chance of capturing the population mean.

A confidence interval is a function of the *Standard Error of the mean (SE)*, i.e., the estimate of the standard deviation that would be obtained from the means of a large number of samples drawn from that population.
Effect size & Confidence interval

CIs can be computed in the units of measurement used by the experimenter (useful to report on graphs)

CIs can also be computed in the units of effect size

<table>
<thead>
<tr>
<th>CI includes the zero</th>
<th>The absolute value of the effect size</th>
<th>The range of CI</th>
<th>Implication</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>small</td>
<td>small</td>
<td>The effect apparently exists, but we are also sure that the effect is small. This is unlikely to happen.</td>
</tr>
<tr>
<td></td>
<td>small</td>
<td>large</td>
<td>The effect apparently exists, and we are sure that the effect is large.</td>
</tr>
<tr>
<td></td>
<td>large</td>
<td>small</td>
<td>The effect apparently exists and may be large, but we are not sure about the real size of the effect.</td>
</tr>
<tr>
<td></td>
<td>large</td>
<td>large</td>
<td>We are sure that the effect is small, but not sure whether the effect really exists.</td>
</tr>
<tr>
<td>Yes</td>
<td>small</td>
<td>small</td>
<td>We are not sure whether the effect exists. We thus need more data. This is unlikely to happen.</td>
</tr>
<tr>
<td></td>
<td>small</td>
<td>large</td>
<td>We are not sure whether the effect exists. We thus need more data.</td>
</tr>
<tr>
<td></td>
<td>large</td>
<td>small</td>
<td>We are not sure whether the effect exists. We thus need more data.</td>
</tr>
</tbody>
</table>

source: [http://www.theusrus.de/blog/happy-holidays/](http://www.theusrus.de/blog/happy-holidays/)
Power of a test

$\alpha$ 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

$\beta$ is the probability of Type II error

1-$\beta$ is the power of the test

<table>
<thead>
<tr>
<th></th>
<th>Reject the null hypothesis</th>
<th>Fail to reject the null hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>The null hypothesis is true.</td>
<td>Type I error (false positive)</td>
<td>True negative</td>
</tr>
<tr>
<td>The null hypothesis is false.</td>
<td>True positive</td>
<td>Type II error (false negative)</td>
</tr>
</tbody>
</table>
Power analysis

Power analysis means considering:

- The sample size,
- The effect size (the one of the underlying population and not the observed one),
- $\alpha$ and $\beta$

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)
Power analysis - Touchstone

Touchstone performs a power analysis of the current design with the typical values for $\alpha$ and $\beta$: $\alpha = 0.05$, $\beta = 0.80$
Parametric vs. Non-parametric

Non-parametric statistical tests

no assumption about the distribution of the measure value in the general population

Parametric statistical tests

measure values belong to a particular distribution that can be described with some commonly used parameters (e.g., the mean and standard deviation)
The Normal Distribution

The "bell curve"

Mean = Median = Mode

This graphics uses the standard deviation to scale the distribution $z = (x - \bar{x}) / \sigma$
Parametric and normal distribution

Many parametric statistical tests rely on the normality assumption.

Normality assumption: the population from which you take the sample forms the normal distribution, but your samples don't necessarily form the normal distribution.

Some statistical tests check the normality (e.g., one-sample Kolmogorov-Smirnov test).
Parametric tests that we study in this class

**t-test**
- compares the mean between two groups

**ANOVA test (Analysis of Variance)**
- test the difference in means among more than two groups

**Post-hoc tests**
- complements an ANOVA to find where the effect is
t-test (Student test)

Computes the t-ratio statistic (a function of the difference between the two means)

Watches where the computed t-ratio lies in the t-distribution when the null hypothesis is true

If, for our two experimental groups, we observe a t-value that is unlikely to happen if there was no difference between the two groups (the t-ratio belongs to the tails of the distribution), we reject the null hypothesis (with a given level of confidence)
t-distribution

A t-distribution shows the probability of observing a given t-value when the null hypothesis holds.

There are several t-distributions. The particular form of a t-distribution is determined by its degrees of freedom (df).
t-test explained again

https://www.youtube.com/watch?v=5Dnw46eC-0o
Paired vs. unpaired t-test

paired t-test (or "repeated measures" t-test)

if the two groups are correlated. For example, participants have been measured under the two technique conditions (within-subject design)

unpaired t-test

if the two groups are independent. For example, two groups of participants have been measured under the two different technique conditions (between-subject design)

additional assumption for unpaired t-test:
the variances of the population of the two groups are equal
Degrees of freedom

Paired t-test

size group 1 = size group 2 = size group

\[ df = \text{size group} - 1 \]

= number of participants - 1 (within-subject design)

Unpaired t-test

\[ df = \text{size group 1} + \text{size group 2} - 2 \]
t-test and effect size

The t-ratio allows you to tell if the difference in means is significant or not but does not give the size of the difference.

Two kinds of effect size metrics for a t-test

Cohen's $d$ (paired and unpaired)

Pearson's $r$ (unpaired only)

<table>
<thead>
<tr>
<th></th>
<th>small size</th>
<th>medium size</th>
<th>large size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohen’s $d$</td>
<td>0.2</td>
<td>0.5</td>
<td>0.8</td>
</tr>
<tr>
<td>Pearson’s $r$</td>
<td>0.1</td>
<td>0.3</td>
<td>0.5</td>
</tr>
</tbody>
</table>
Effect size for a paired t-test

Cohen's $d$ for a paired t test

$$d = \frac{|M|}{SD}$$

where $M$ is the mean of differences, and $SD$ is the standard deviation of differences

It represents the difference in terms of standard deviations (normalized)
Effect size for an unpaired t-test

Pearson's $r$ for an unpaired t test

$$r = \sqrt{\frac{t^2}{t^2 + df}}$$

where $t$ is the value of the test (t-ratio), and $df$ is the number of degrees of freedom.
[R] Unpaired and paired t-tests

When

use them to test if there is a difference between exactly two groups regarding a continuous (i.e., ratio) measure

Welch t-test for unpaired groups

Paired t-test for paired groups

How

compute two vectors:

mean measure value per participant for group 1
mean measure value per participant for group 2

Group1 has n participants p1_g1, ..., pn_g1
mean_measure(p1_g1), mean_measure(p2_g1), ..., mean_measure(pn_g1)
Group2 has m participants p1_g2, ..., pm_g2
mean_measure(p1_g2), mean_measure(p2_g2), ..., mean_measure(pm_g2)

use the t.test(…) function
Example

Pointing performance of two types of magnifying lenses

for illustration purpose, we assume being in a case where we evaluate only two types of magnifying lenses

10 participants

Hypothesis

As ML lenses suffer from occlusion issues, pointing with a ML lens is slower than pointing with a FL lens

R code

```r
ml_trials <- lens_experiment[lens_experiment$Lens == "ML", ]
ml_trials_summary <- summarySE(ml_trials, measurevar="PointingTime", groupvars=c("Participant"))
ml_time <- ml_trials_summary$PointingTime

fl_trials <- lens_experiment[lens_experiment$Lens == "FL", ]
fl_trials_summary <- summarySE(fl_trials, measurevar="PointingTime", groupvars=c("Participant"))
fl_time <- fl_trials_summary$PointingTime

# Option 1: if we were in a within-subject design (paired t-test)
ttest <- t.test(ml_time, fl_time, paired=TRUE)
# Option 2: if we were in a between-subject design (unpaired t-test)
ttest <- t.test(ml_time, fl_time, var.equal=FALSE)
# ttest is a data frame from which mean values for the two conditions, t-ratio, degrees of freedom and confidence interval can be extracted

ttest$estimate # mean values for the two conditions
ttest$statistic # t-ratio
ttest$p.value # p-value
ttest$parameter # degrees of freedom
ttest$conf.int # confidence interval for the difference between the conditions
```
Example
Pointing performance of two types of magnifying lenses

Hypothesis
As ML lenses suffer from occlusion issues, pointing with a ML lens is slower than pointing with a FL lens

R output

```r
> ttest$estimate # mean of the differences
  mean of the differences
    1098.775
> ttest$statistic # t-ratio
  t
    7.490894
> ttest$p.value # p-value
  [1] 3.728208e-05
> ttest$parameter # degrees of freedom
df
    9
> ttest$conf.int # confidence interval for the difference between the conditions
  [1]  766.9586 1430.5914
attr(,"conf.level")
  [1] 0.95
```

The effect of Lens is significant: p-value < 0.05
Hypothesis

As ML lenses suffer from occlusion issues, pointing with a ML lens is slower than pointing with a FL lens.

Effect size:
- **Cohen’s d** \( d = \frac{|M|}{SD} \)
- **Pearson’s r** \( r = \sqrt{\frac{t^2}{t^2 + df}} \)

R code

```r
# Option 1: compute effect size in a within-subject design (paired t-test)
cohen_effect_size <- abs(mean(ml_time - fl_time)) / sd(ml_time - fl_time)
cohen_effect_size

# Option 2: compute effect size in a between-subject design (unpaired t-test)
pearson_effect_size <- sqrt( ttest$statistic^2 / (ttest$statistic^2 + ttest$parameter) )
pearson_effect_size
```

Example

Pointing performance of two types of magnifying lenses
Report a t-test

t-ratio (e.g., 7.5)
degrees of freedom (e.g., 9)
p-value (e.g., 3.728208e-05) or confidence interval
effect size (e.g., 2.36)

For p-value, we typically report one of the three following orders of magnitude:
- < 0.001
- < 0.01
- < 0.05

e.g., we found a significant effect of factor Lens on pointing time (t(9) = 15.8, p < 0.001, Cohen's $d=2.36$), with lens FL outperforming lens ML
ANOVA test

Analysis of variance generalizes the t-test to an arbitrary number of groups (more than 2 dependent variable)

Making Student tests between pairs of categories does not give correct results

ANOVA separates the internal variability in each sample and the variability between samples

It computes the F-ratio (and watches where it lies in the F distribution when there is no difference between the groups - i.e., null hypothesis is true)
ANOVA test

The computation of the F-ratio depends on your design

If the groups are correlated (e.g., a within-subject design), use a repeated measures ANOVA test

One-way ANOVA

analyze one factor

n-way ANOVA

analyze two or more factors
ANOVA estimates how much of the total observed variance is due to the variance between groups.

$F = \frac{MS_{bg}}{MS_{wg}}$

$d_{bg} = k - 1$

$d_{wg} = \text{size group} - k$

also called $d_{error}$ or $d_{residuals}$
repeated measures ANOVA $df$

$F = \frac{MS_{bg}}{MS_{wg}}$

- $MS_{bg}$: a measure of the aggregate differences among the means of the $k$ groups
- $MS_{wg}$: a measure of the amount of random variability that exists inside the $k$ groups

$df_{bg} = k - 1$

$df_{wg} = (\text{number of observations} - 1) - (\text{number of participants} - 1) - (k-1)$

Example:
12 participants, 1 factor Technique with 3 levels (T1, T2, T3)

$df_{bg} = 3 - 1 = 2$

$df_{wg} = (3\times12-1) - (12-1) - (3-1) = 22$
ANOVA and effect size

Eta squared $\eta^2$

$$\eta^2 = \frac{SS_{effect}}{SS_{total}}$$

where $SS_{effect}$ is the sum of square for the factor and $SS_{total}$ is the total sum of square

<table>
<thead>
<tr>
<th></th>
<th>small size</th>
<th>medium size</th>
<th>large size</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\eta^2$</td>
<td>0.01</td>
<td>0.06</td>
<td>0.14</td>
</tr>
</tbody>
</table>
One-way ANOVA and One-way repeated measures ANOVA

When

use them to test if there is a difference between more than two groups (which differ by only one factor) regarding a continuous (i.e., ratio) measure

- One-way ANOVA for unpaired groups
- One-way repeated measures ANOVA for paired groups

How

One-way ANOVA

- create a summary table (mean measure value per participant for all groups)
- use the `aov(…)` function

One-way repeated measures ANOVA

- use the `ezANOVA(…)` function
Example

Pointing performance of five different types of magnifying lenses

Hypothesis

SCB and SCF will outperform other types of lenses for pointing tasks
Example
Pointing performance of five different types of magnifying lenses

Hypothesis
SCB and SCF will outperform other types of lenses for pointing tasks

R code

```r
# aggregate data per condition of interest (Participant x Lens)
lens_experiment_summary <- summarySE(lens_experiment, measurevar="PointingTime",
groupvars=c("Participant", "Lens"))

# for a between-subject design, use aov function
aov <- aov(PointingTime ~ Lens, lens_experiment_summary)
summary(aov)

library(lsr) # requires having installed package “lsr”
etaSquared(aov)
```

One-way anova in case of a between-subject design
Example
Pointing performance of five different types of magnifying lenses

Hypothesis
SCB and SCF will outperform other types of lenses for pointing tasks

R output

```
> lens_experiment_summary <- summarySE(lens_experiment, measurevar="PointingTime",
   groupvars=c("Participant", "Lens"))
> lens_experiment_summary

          Participant Lens N PointingTime        sd        se        ci
   1           1   BL 240     2426.696 1125.6468  72.66019 143.13616
   2           1   FL 240     2234.900 1019.6713  65.81950 129.66043
   ...          ...   ...     ...        ...        ...        ...
   49          10  SCB 240     1994.729  783.3901  50.56761  99.61513
   50          10  SCF 240     2505.529 1325.5238  85.56219 168.55233

> aov <- aov(PointingTime ~ Lens, lens_experiment_summary)
> summary(aov)

             Df  Sum Sq Mean Sq F value   Pr(>F)
    Lens         4 14189510 3547378   43.44 6.67e-15 ***
 Residuals   45  3674951   81666
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The effect of Lens is significant p-value < 0.05

```
> library(lsr)
> etaSquared(aov)

  eta.sq  eta.sq.part
Lens  0.794287     0.794287
```

One-way anova in case of a between-subject design

The effect size eta squared

\[ \eta^2 = \frac{SS_{effect}}{SS_{total}} \]
Example
Pointing performance of five different types of magnifying lenses

Hypothesis
SCB and SCF will outperform other types of lenses for pointing tasks

R code

```r
# for a within-subject design, use ezANOVA function
# (no need for aggregating data beforehand)
library(ez)  # requires having installed package "ez"
ezANOVA(data=lens_experiment, dv=.(PointingTime), wid=.Participant, within =.
          (Lens), detailed=TRUE)
```

One-way repeated measures anova in case of a **within-subject** design
Example

Pointing performance of five different types of magnifying lenses

Hypothesis

SCB and SCF will outperform other types of lenses for pointing tasks

R output

```r
> library(ez) # requires having installed package "ez"
> ezANOVA(data=lens_experiment,dv=.(PointingTime), wid=.(Participant), within =.(Lens), detailed=TRUE)
...
$ANOVA

   Effect DFn DFd SSn    SSd         F          p  p<.05     ges
1 (Intercept)  1   9 327972439 1628666 1812.3739 1.086312e-11  *  0.9889191
2    Lens      4  36  14189510 2046285   62.4085 1.077454e-15  *  0.7942870
```

The effect of Lens is significant
p-value < 0.05

effect size (generalized eta squared)

One-way anova in case of a **within-subject** design

The effect size (generalized eta squared) is already provided in the report.
Example
Pointing performance of five different types of magnifying lenses

Hypothesis

SCB and SCF will outperform other types of lenses for pointing tasks

Based on analyses in the previous slides, we only know that Lens has a significant effect on pointing. We do not know which types of lens are significantly different from each other.

We need to do post-hoc tests
Post-hoc tests

ANOVA says that there are significant effects

BUT does not say which group significantly differs from which other group

Post-hoc tests are used to find where the differences between groups are

The most common post hoc tests used in HCI are the Tukey’s test and the multiple pairwise t-test (with potentially some correction method like Holm or Bonferroni)
Tukey HSD post-hoc test

compares all possible pairs of means, and is based on a studentized range distribution (q)

Tukey's test differs from multiple t-tests as it prevents the probability of making a type I error increase

With multiple t-tests, probabilities get multiplied

\[
\text{Condition 1} \neq \text{Condition 2} \text{ (95\% confidence)} \\
\text{Condition 1} \neq \text{Condition 3} \text{ (95\% confidence)} \\
\Rightarrow \text{Condition 1} \neq \text{Conditions 2 and 3} \text{ (90\% confidence)}
\]
# post-hoc tests for a within-subject design (paired=TRUE).
# For a between-subject design, change to paired=FALSE
pairwise.t.test(lens_experiment$PointingTime, lens_experiment$Lens, paired=TRUE)

Pairwise comparisons using paired t tests

data: lens_experiment$PointingTime and lens_experiment$Lens

<table>
<thead>
<tr>
<th></th>
<th>BL</th>
<th>FL</th>
<th>ML</th>
<th>SCB</th>
</tr>
</thead>
<tbody>
<tr>
<td>FL</td>
<td>&lt;2e-16</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ML</td>
<td>&lt;2e-16</td>
<td>&lt;2e-16</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SCB</td>
<td>&lt;2e-16</td>
<td>&lt;2e-16</td>
<td>&lt;2e-16</td>
<td>&lt;2e-16</td>
</tr>
<tr>
<td>SCF</td>
<td>&lt;2e-16</td>
<td>0.049</td>
<td>&lt;2e-16</td>
<td>&lt;2e-16</td>
</tr>
</tbody>
</table>

P value adjustment method: holm

1. p-value for difference between SCB and ML
   It is < 0.001 => the difference is significant

2. p-value for difference between SCF and FL
   It is 0.049 => the difference is significant, but with a lower confidence level
Report an ANOVA test

1. F-ratio (e.g., 62)
2. number of degrees of freedom (e.g., 4 and 36)
3. p-value (p) (e.g., < 0.001)
4. effect size (e.g., 0.79)

Pairs of conditions that significantly differ

e.g., an ANOVA test revealed a significant effect of Lens on Pointing Time ($F(4,36) = 62, p < 0.001, \eta^2 = 0.79$). Post-hoc tests revealed that all pairs of lenses significantly differ.

```
> ezANOVA(data=lens_experiment, dv=(PointingTime), wid=(Participant), within=(Lens), detailed=TRUE)
...
$ANOVA
             Effect DFn DFd      SSn    SSD     F     p p<.05    ges
1 (Intercept)     1   9 327972439 1628666 1812.37 1.08e-11     * 0.9889191
2        Lens    4  36  14189510 2046285   62.41 1.08e-15     * 0.7942870
```
[R] Two-way ANOVA and Two-way repeated measures ANOVA

When

use them to test if there is a difference between more than two groups (which differ by more than one factor) regarding a continuous (i.e., ratio) measure

- Two-way ANOVA for unpaired groups
- Two-way repeated measures ANOVA for paired groups

How

Two-way ANOVA
not addressed in this class

Two-way repeated measures ANOVA
use the ezANOVA(…) function
Similar to one-way repeated measures: use ezANOVA function

```r
# Magnification is numeric (continuous), turn it into a categorical (nominal) factor
lens_experiment$Magnification <- as.factor(lens_experiment$Magnification)

ezANOVA(data=lens_experiment,dv=.(PointingTime), wid=.(Participant), within =.
(Lens,Magnification), detailed=TRUE, type=1)
```

### ANOVA

<table>
<thead>
<tr>
<th>Effect</th>
<th>DFn</th>
<th>DFd</th>
<th>SSn</th>
<th>SSd</th>
<th>F</th>
<th>p</th>
<th>p&lt;.05</th>
<th>ges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lens</td>
<td>4</td>
<td>36</td>
<td>70947550</td>
<td>10231427</td>
<td>62.4085</td>
<td>1.077454e-15</td>
<td>*</td>
<td>0.7438569</td>
</tr>
<tr>
<td>Magnification</td>
<td>4</td>
<td>36</td>
<td>330285802</td>
<td>4461948</td>
<td>666.2051</td>
<td>3.309458e-33</td>
<td>*</td>
<td>0.9311269</td>
</tr>
<tr>
<td>Lens:Magnification</td>
<td>16</td>
<td>144</td>
<td>95256469</td>
<td>9737027</td>
<td>88.0462</td>
<td>6.510501e-66</td>
<td>*</td>
<td>0.7958807</td>
</tr>
</tbody>
</table>

An ANOVA test revealed a significant effect of Lens on Pointing Time ($F(4,36) = 62, p < 0.001, \eta^2=0.74$), a significant effect of Magnification on Pointing Time ($F(4,36) = 666, p < 0.001, \eta^2=0.93$), as well as a significant Lens x Magnification interaction effect ($F(16,144) = 88, p < 0.001, \eta^2=0.79$).

[+ post-hoc tests to analyze pairwise comparisons]
Simple and Interaction effects

Two-way ANOVA with two independent variables (A and B)

Simple main effect (A or B): effect of a single independent variable on a given level of a second dependent variable

Interaction effect (AxB): a change in the simple main effect of one variable over levels of the second
A simple example

Satisfaction = Food x Condiment

Experiment: ask participants “How much do you appreciate condiment X on food Y?”

“It depends on the type of food!”

interaction effect between food and condiment

from http://statisticsbyjim.com/regression/interaction-effects/
Simple and Interaction effects

Main effect of A -- yes
Main effect of B -- no
Interaction -- yes

Main effect of A -- no
Main effect of B -- yes
Interaction -- no

Main effect of A -- yes
Main effect of B -- yes
Interaction -- no

Main effect of A -- no
Main effect of B -- yes
Interaction -- no

Main effect of A -- yes
Main effect of B -- no
Interaction -- yes

Main effect of A -- no
Main effect of B -- yes
Interaction -- yes

Main effect of A -- no
Main effect of B -- yes
Interaction -- yes

Main effect of A -- no
Main effect of B -- yes
Interaction -- yes
[R and ggplot] Convey differences between conditions visually

Create a graphics with the ggplot2 library

create a base plot (e.g., functions `ggplot` or `qplot`)

associate a dataset

associate some aesthetics

  how data are mapped on the plot (x-axis, y-axis, fill colors, etc.)

...

+ additional layers (e.g., `geom_bar`, `geom_line`, `geom_errorbar`, etc.)
# first, aggregate data to get means per Lens condition
lens_experiment_summary <- summarySE(lens_experiment, measurevar="PointingTime", groupvars=c("Lens"))
lens_experiment_summary

<table>
<thead>
<tr>
<th>Lens</th>
<th>N</th>
<th>PointingTime</th>
<th>sd</th>
<th>se</th>
<th>ci</th>
</tr>
</thead>
<tbody>
<tr>
<td>BL</td>
<td>2400</td>
<td>2666.405</td>
<td>1400.2672</td>
<td>28.58283</td>
<td>56.04961</td>
</tr>
<tr>
<td>FL</td>
<td>2400</td>
<td>2399.814</td>
<td>1266.7927</td>
<td>25.85830</td>
<td>50.70691</td>
</tr>
<tr>
<td>ML</td>
<td>2400</td>
<td>3498.589</td>
<td>3156.1004</td>
<td>64.42363</td>
<td>126.33173</td>
</tr>
<tr>
<td>SCB</td>
<td>2400</td>
<td>1881.055</td>
<td>658.1707</td>
<td>13.43485</td>
<td>26.34512</td>
</tr>
<tr>
<td>SCF</td>
<td>2400</td>
<td>2359.847</td>
<td>1162.8386</td>
<td>23.73634</td>
<td>46.54586</td>
</tr>
</tbody>
</table>

ggplot(lens_experiment_summary, aes(x=Lens, y=PointingTime, fill=Lens)) +
# plot data as is using a bar plot layer
  geom_bar(stat="identity") +
  # plot error bar layer using the confidence intervals
  geom_errorbar(aes(ymin=PointingTime-ci, ymax=PointingTime+ci), width=.2) +
  # make text smaller
  theme(text = element_text(size=20))

ggsave("mean_per_lens.pdf", plot = last_plot())
Means per condition
Two factors

Example: Visualize PointingTime per Lens x Magnification condition

# first, aggregate data to get means per Lens condition
lens_experiment_summary <- summarySE(lens_experiment, measurevar="PointingTime",
groupvars=c("Lens","Magnification"))
# Magnification should be a nominal factor with categories to make it suitable for a bar plot
lens_experiment$Magnification <- as.factor(lens_experiment$Magnification)
lens_experiment_summary

ggplot(lens_experiment_summary, aes(x=Lens, y=PointingTime, fill=Magnification)) +
   # plot data as is using a bar plot layer (use position_dodge to display conditions side-by-side)
   geom_bar(stat="identity", position=position_dodge()) +
   # plot error bar layer using the confidence intervals
   geom_errorbar(aes(ymin=PointingTime-ci, ymax=PointingTime+ci), width=.2, position=position_dodge(.9)) +
   # make text smaller
   theme(text = element_text(size=20))

ggsave("mean_per_lens_and_mag.pdf", plot = last_plot())
(Non-)parametric and HCI

In many cases, you can assume the normality for the population and use parametric tests.

If the type of measure is not ratio (continuous), you should use non-parametric tests.

Example: results from Likert-scale questions.

An assumption of a parametric test is violated when:

- largely unbalanced data (number of observations largely varies among groups),
- non-normal distribution,

Note, however, that many of the parametric tests are fairly robust against the non-normality.
Non-parametric tests in this class

Chi-test, McNemar

compares two groups with regard to a categorical (nominal measure)

Example: the measure, the answer to “which system do you use”?), is nominal

Results are presented as a contingency table (count of answers per category)

<table>
<thead>
<tr>
<th></th>
<th>windows</th>
<th>mac</th>
<th>linux</th>
</tr>
</thead>
<tbody>
<tr>
<td>young</td>
<td>16</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>old</td>
<td>21</td>
<td>8</td>
<td>1</td>
</tr>
</tbody>
</table>
Chi-square test

It computes the ratio $\chi^2$ (and watches where it lies in the Chi-squared distribution)

Effect size ($\phi$ or Cramer's $V$):

$$\phi = \sqrt{\frac{\chi^2}{N(k - 1)}}$$

where $N$ is the number of observation and $k$ is the smaller of the number of rows $r$ or columns $c$

If groups are paired, use a McNemar’s test
Which test when?

This class does not cover all possible statistical tests

In order to choose the right test, consider:

the experiment design:
  within-subject / between-subject

the type of your variables:
  number and types of independent variables (factors)
  and type of dependent variable (measure)
A simplified (and incomplete) decision tree for choosing the right statistical test

Is my measure of type ratio?

NO

How many factors do I have?

1

Is my factor tested with a within-subject design (i.e. are my groups paired)?

NO

Chi-square test

YES

McNemar’s test


NO

n >= 2

How many factors do I have?

1

How many levels does my factor have?

2

Is my factor tested with a within-subject design (i.e. are my groups paired)?

NO

t-test

YES

paired t-test

n >= 3

Is my factor tested with a within-subject design (i.e. are my groups paired)?

NO

One-way ANOVA

YES

One-way repeated measures ANOVA

Are my factors tested with a within-subject design (i.e. are my groups paired)?

NO

YES

n-way ANOVA

n-way repeated measures ANOVA