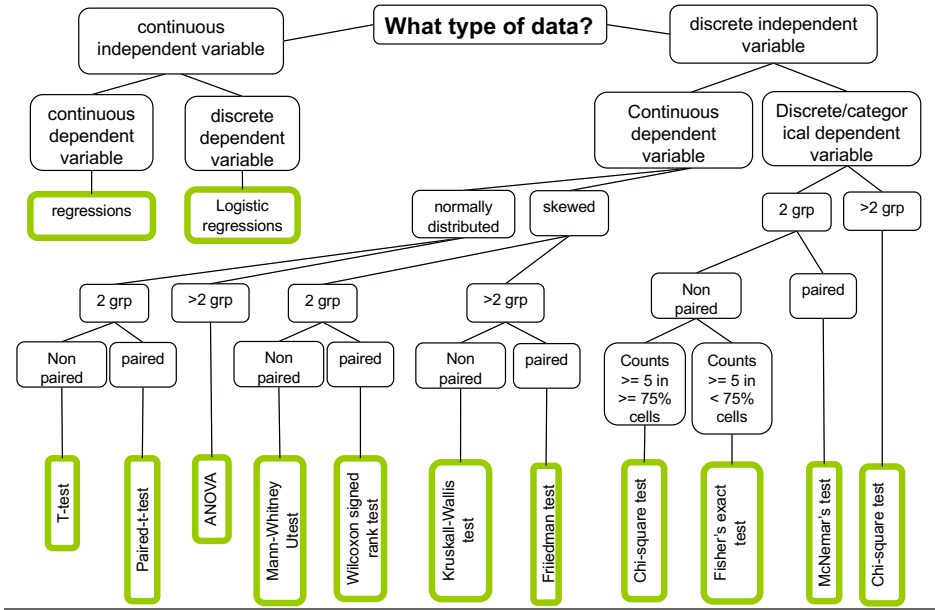
COMS10011
Dr. Anne Roudaut
csxar@bristol.ac.uk
https://github.com/coms10011

Probability and Statistics

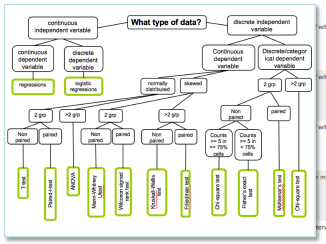


by the end of this unit you will know **what statistics tests** to perform depending on your data and **how**

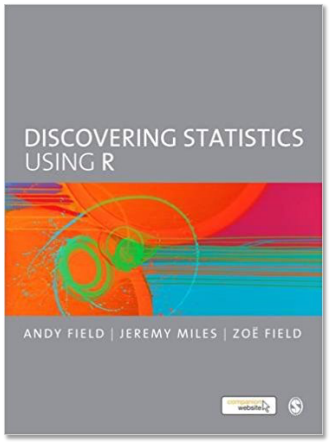
1. Regression
2. Hypothesis testing, comparing things
3. Experimental design
4. Parametric tests
5. Normality tests
6. Non-parametric tests
7. Categorical data: Chi-square
8. Sample size, power and effect size
9. P-hacking

unit menu

resources

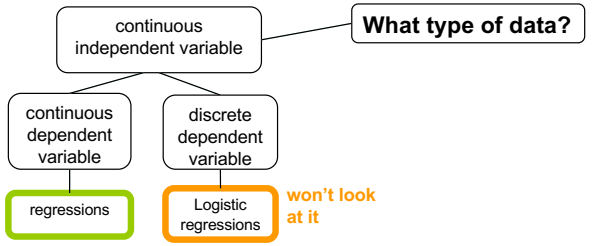


https://stats.idre.ucla.edu/other/mult-pkg/whatstat/



the **text book** I am using and a suggestion of **YouTube video channel**

1 regression



we are looking at this part of the graph

let's start with an example

imagine you are designing a graphical interface for a new application on a laptop

how big should the buttons/icons be?

Fitts' law ::

the time required to **acquire a target** of size *w* at distance *d* can be described as T = a + b log (1 + d/w)

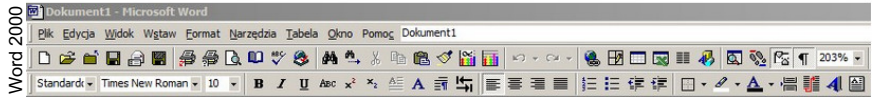
Fitts' law ::

the time required to **acquire a target** of size *w* at distance *d* can be described as T = a + b log (1 + d/w)

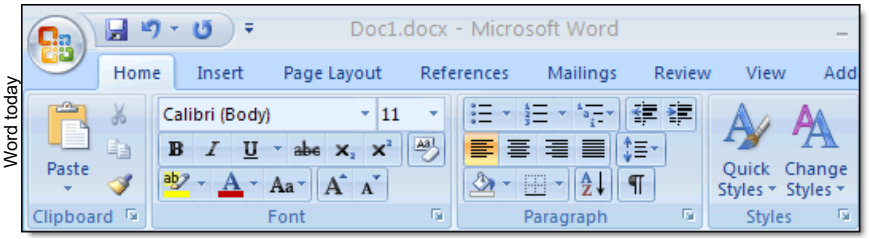
T = a + b $I_{ndex}D_{ifficulty}$

(depends on input device)

Word 2000



Word today

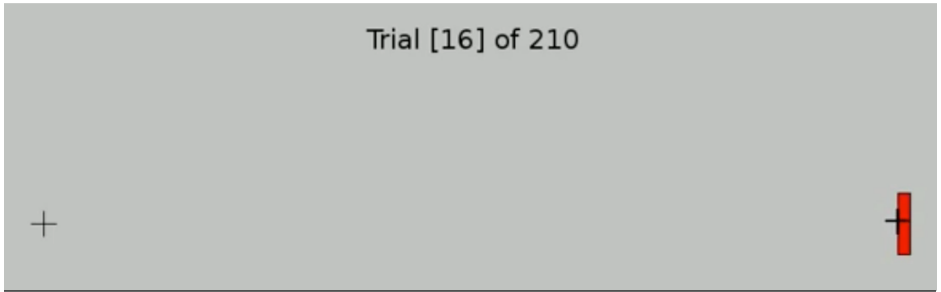


e.g. one reason why we have ribbons in Word now

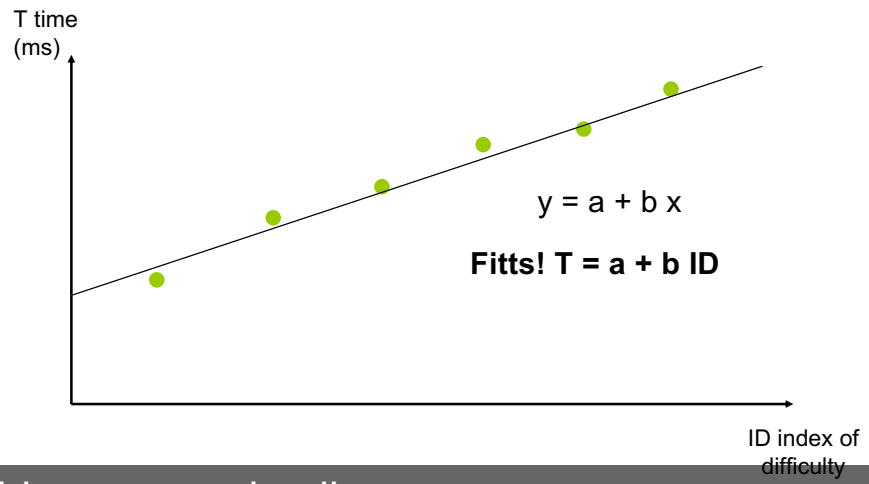
Fitts' law :: T = a + b ID

time required to **acquire a target**

but where does this equation come from?



let's run an experiment and ask one participant to click on targets of different IDs

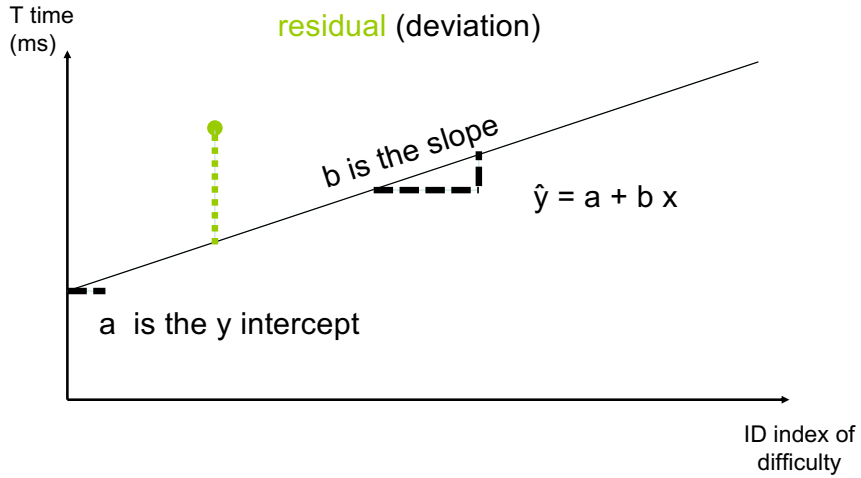


this a regression line

regression ::

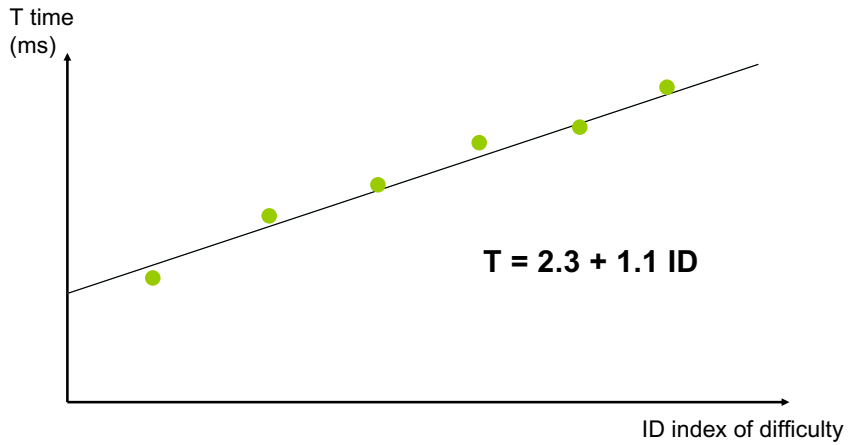
a technique for determining the statistical relationship between two or more variables where a change in a **dependent variable** is associated with, and depends on, a change in one or more **independent variables**

arguably the most basic technique for **machine learning**

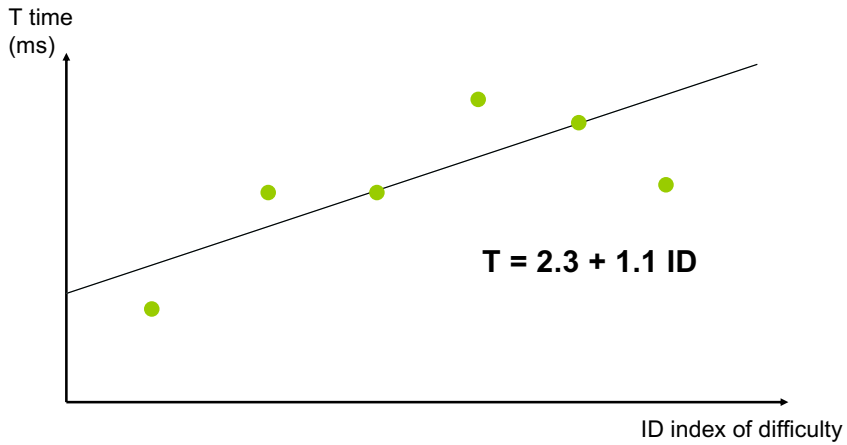


terminology

goodness of fit



how can you be sure this line is a good fit to our data?



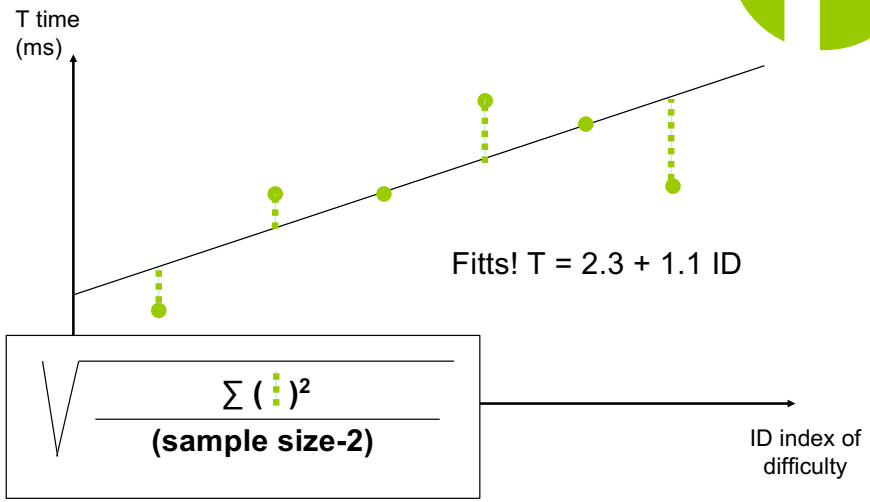
how can you be sure this line is a good fit to our data? what about now?

<brainstorming with your neighbor>

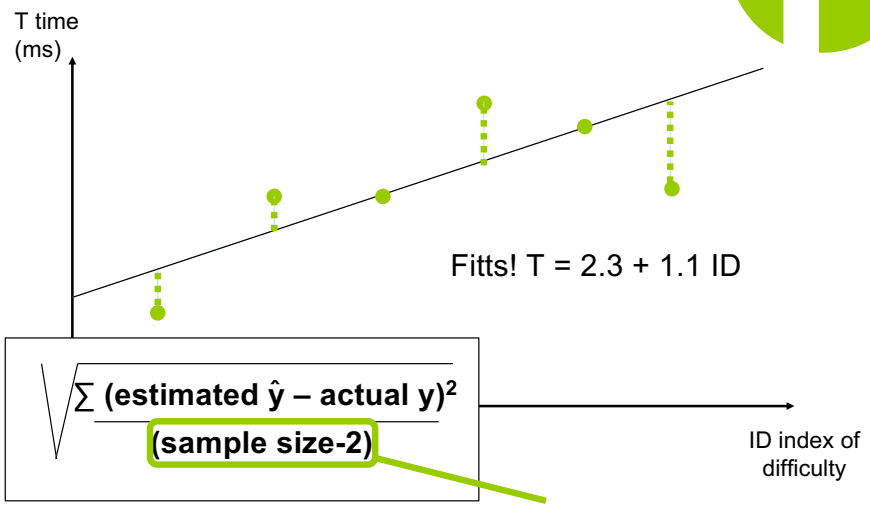
we can compute the **goodness of fit** with several methods

e.g. standard error of the estimate
or R squared

standard error of the estimate



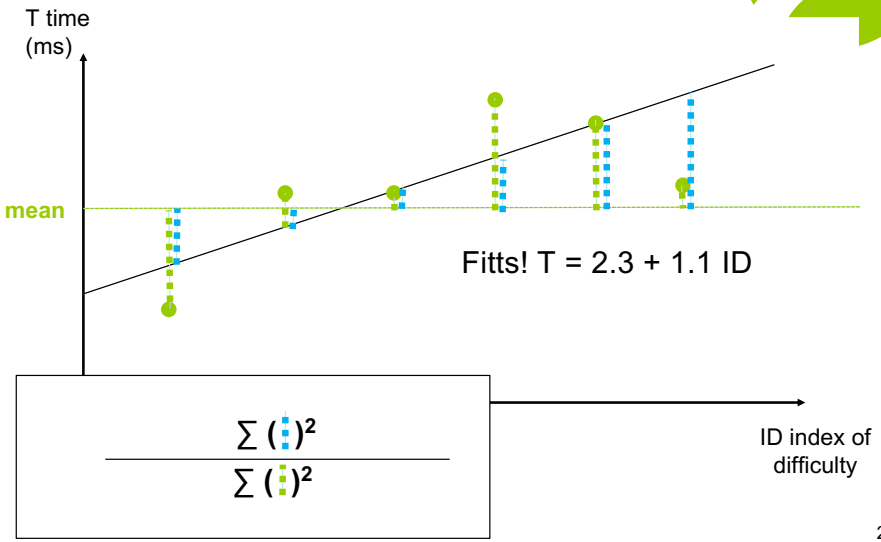
standard error of the estimate



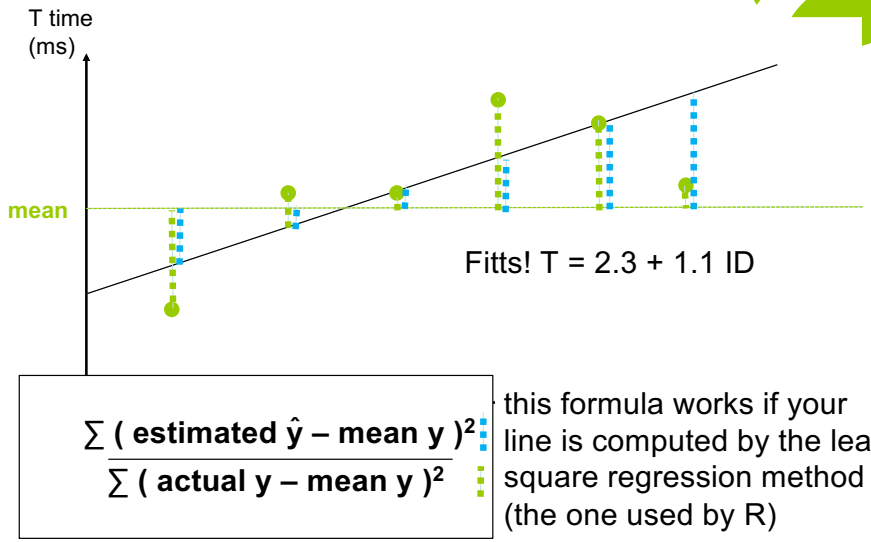
also called degree of freedom

S gives a standard error in the metric of the data (the less the better)

R squared

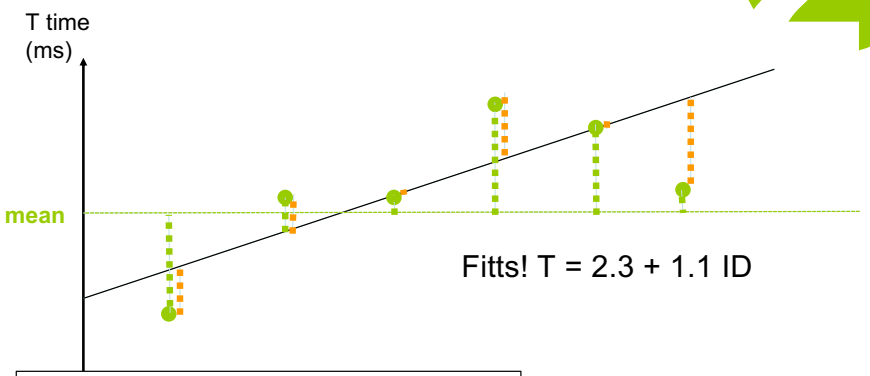


R squared



this formula works if your line is computed by the least square regression method (the one used by R)

R squared

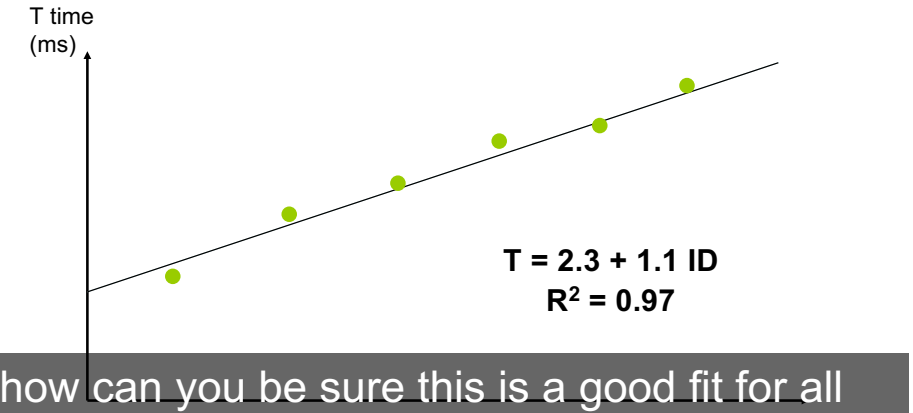


Fitts! T = 2.3 + 1.1 ID

$$1 - \frac{\sum (\text{actual y} - \text{estimated } \hat{y})^2}{\sum (\text{actual y} - \text{mean y})^2}$$

you could also find it in a this format which is **more genetic**

R^2 gives a percentage and 100% means perfect fit (>70% is better)



T = 2.3 + 1.1 ID
R^2 = 0.97

how can you be sure this is a good fit for all human?

ID index of difficulty

to **gain additional confidence** we **repeat**

we gain trust in a model if it fits the data with **little error** when
1.it is verified with a **lot of data**
2.it holds across very **different people**
3.it is verified in **independent studies**

The information capacity of the human motor system in controlling the amplitude of movement.

PM Fitts - Journal of experimental psychology, 1954 - psycnet.apa.org

Reports of 3 experiments testing the hypothesis that the average duration of responses is directly proportional to the minimum average amount of information per response. The results show that the rate of performance is approximately constant over a wide range of movement amplitude and tolerance limits. This supports the thesis that" the performance capacity of the human motor system plus its associated visual and proprioceptive feedback mechanisms, when measured in information units, is relatively constant over a considerable …

☆ 99 Cited by 7707 Related articles All 18 versions Web of Science: 3367

Fitts's paper probably most cited in HCI, studies done and redone many times

practically


```
vpn-user-244-044:~ neniseas$ R
```

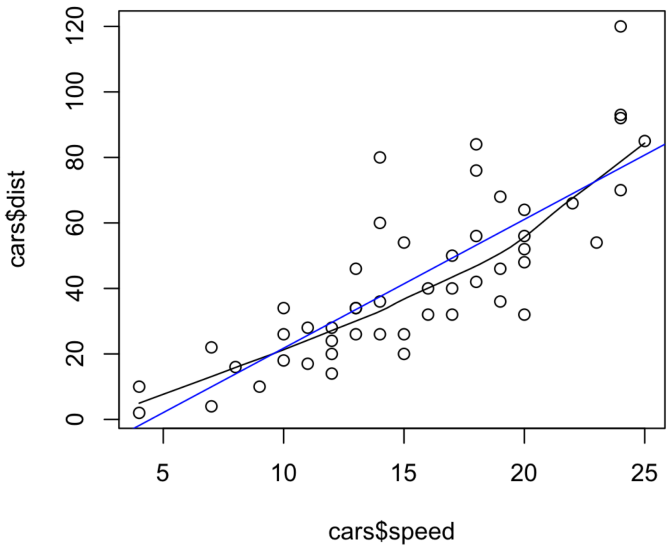
Install at https://www.r-project.org/

```
> print ("hello world!")
[1] "hello world!"
>
```

we will be using **R** and I will try to give you as much as possible of examples

in your terminal

```
head(cars) # cars is a table that already comes with R and
contain 50 observations of speed and distance in two rows

scatter.smooth(x=cars$speed, y=cars$dist, main="Dist ~
Speed")

linearMod <- lm(dist ~ speed, data=cars) # build linear
regression model

abline(linearMod, col="blue") # draw the regression line

summary(linearMod) # goodness of fit
```




```
Call:
lm(formula = dist ~ speed, data = cars)


Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

dist = -17.5791 + 3.9324 * speed


```
Call:
lm(formula = dist ~ speed, data = cars)


Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

also note this
pvalue<0.01
pvalue <0

(will explain next week)

usages of regressions

predicting ebay's online auction prices using functional data analysis

41

predicting the number of passerby who will pass in front of a public display and use the data for choosing advertisement prices

42

predicting the number of claim in a given time period (used by insurance companies)

43

understand the impact of a machine (on a production line) on the quality of the products created

44

45

46

you can also fit a curve =
polynomial fitting

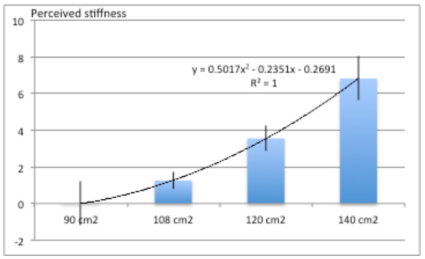


Figure 11. Bradley-Terry-Luce model output as well as a polynomial regression.

Our results are illustrated in Figure 11. We observed a clear distinction between the perceived stiffness of the 4 patches, the size of the patch increasing the perceived stiffness. In particular A is the least restrictive, followed by B, then C and then D is the most restrictive. We found that each paired comparison was significant ($p<0.0125$). This thus allows us to compare the different patches and conclude that D is the most efficient patch. We also performed a polynomial regression on our data and found a very accurate fit: $y=0.5017x2-0.2351x-0.2691$ ($R^2=1$). This suggests a quadratic correlation between the area of the patch and the perceived stiffness, which allows us to imagine bigger patches in order to restrict movements of the knee, which would require more stiffness. Of course further investigations need to be done to confirm this.

betwe
the ai
jam th

Prelin

The g
lab. A
the co
potent

First
One p
hands
sugge
sugge
player
the u
"defro
where
player
for ter
mover
our id

We al
becon
that th
in two
impro
particl
imple

Patch

47

polynomial regression model

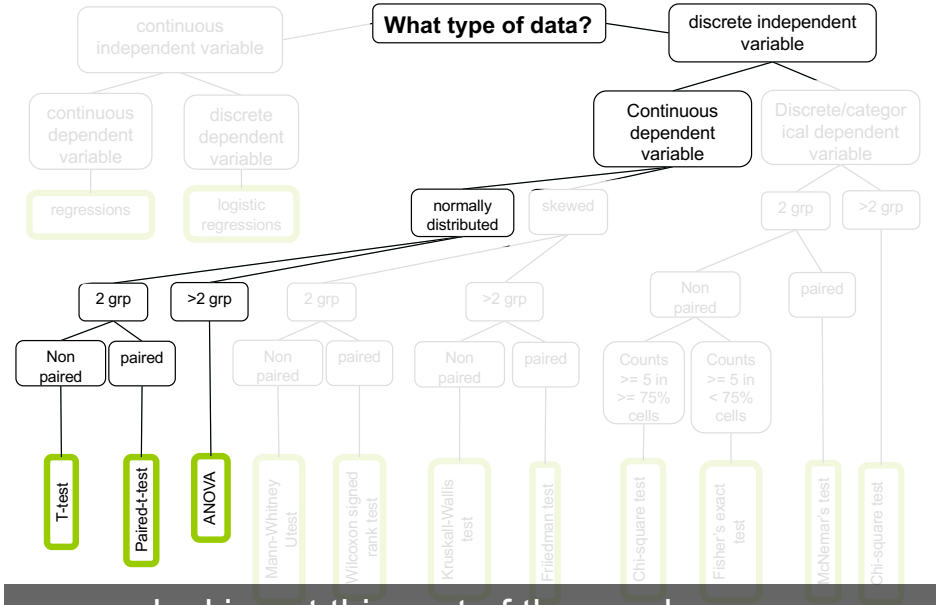
```
Mod2 <- lm(dist~poly(speed,2,raw=TRUE), data=cars)

Mod3 <- lm(dist~poly(speed,3,raw=TRUE), data=cars)

Mod4 <- lm(dist~poly(speed,4,raw=TRUE), data=cars)
```

48

1. Explain what is a linear regression
2. Give the terminology of a regression line
3. Give the two formulas of goodness of fit
4. Be able to compute the two formulas given a few observations

take away

2 hypothesis testing



we are looking at this part of the graph

let's start with a contrived example…

you own a magic pair of shoes. You have run the 10 meters with it a lot and you **have a long log of the time you have run with it.**

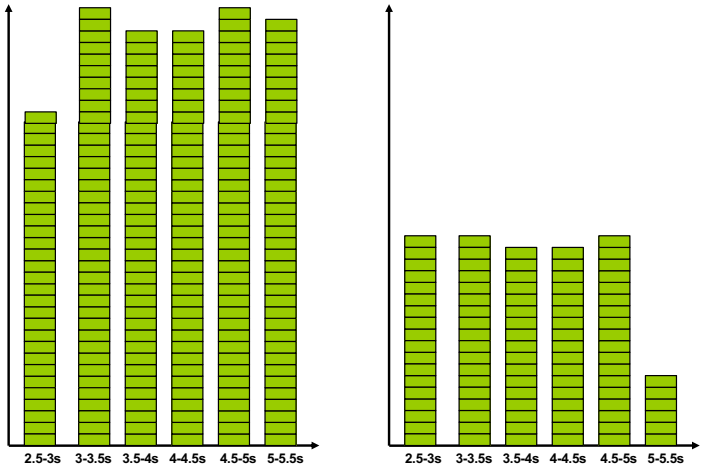
one day, you come home and it seems like someone has moved your shoes. You get concerned that someone might have taken your (beloved) shoes and instead **replaced it with identical looking shoes.**

you inspect the shoes long and hard and they *look* the same. But still worried. **How do you verify that they are the same?**

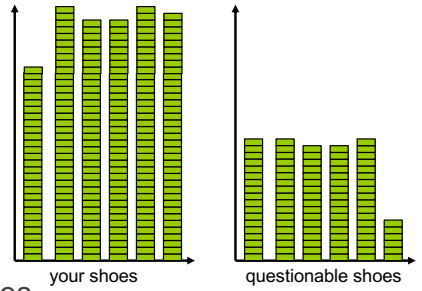
<30 sec brainstorming>

ok, so **you run the 10 meters with the "questionable" shoes** a bunch of times.

here is what you see…

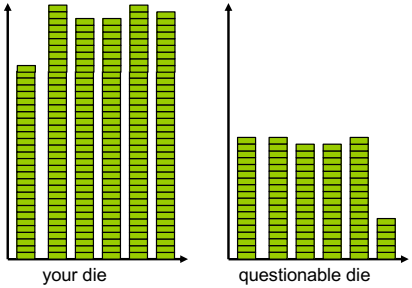


your shoes questionable shoes



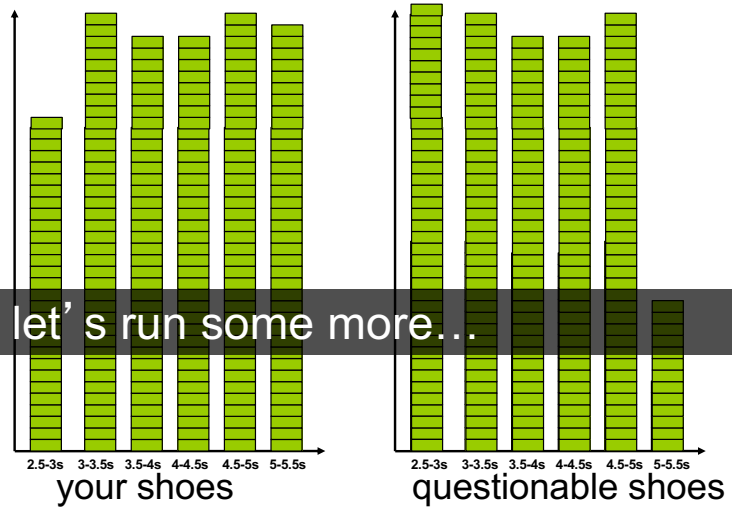
[ ] this is still the original shoes
[ ] someone has replaced my shoes
[x] could be the original or replacement shoes

<let's vote>

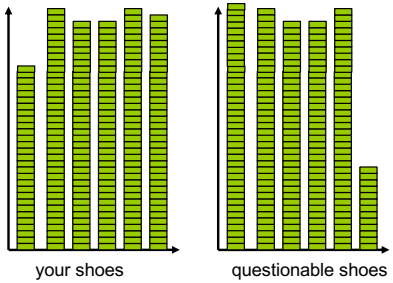


the distribution looks different from the shoes you know

while it is **possible** that it is the same shoes, it seems somewhat **unlikely**



ok, let's run some more…

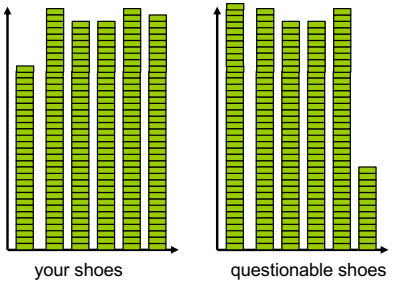


[ ] this is probably still the original shoes

[x] someone probably has replaced my shoes

[x] could be the original or replacement shoes

<let's vote>



again, the distribution could have happened by chance, but it seems **even more unlikely.** This is **probably not** your shoes

are you **sure** this it is not your shoes?

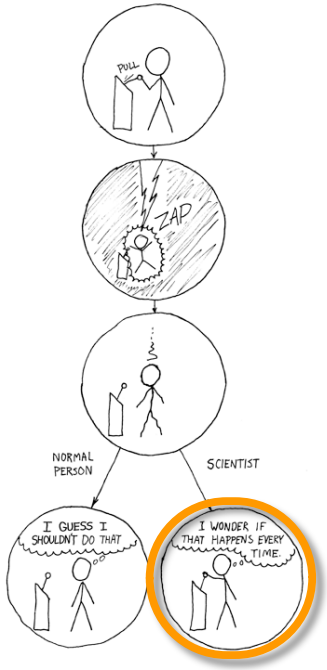
you are **not sure.**

what can you do **to be sure?**

there is **nothing** you can do,
you can **never be sure**

it is a limitation of science:

no matter how often you pull the lever, it could **always** be chance



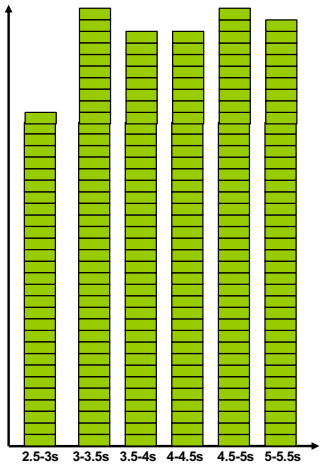
the good news:

the more sample (# of runs), the more your confidence increases
→ you can be **arbitrarily sure**

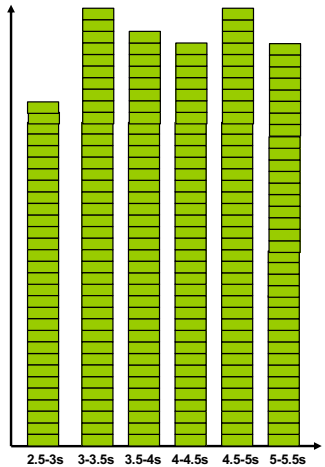
another round

ok, you get your original shoes back

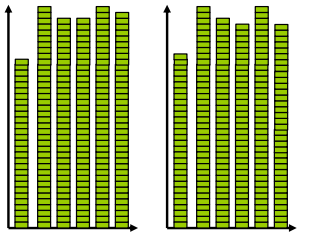
a week later the same thing **happens again.**
again, **you run 10 meters a few times** with the questionable shoes …



your shoes

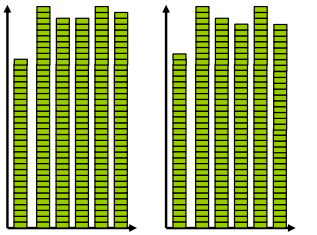


questionable shoes

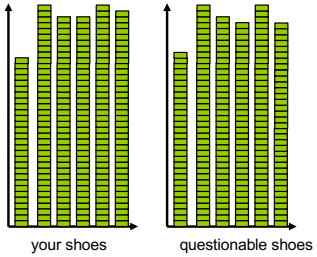


[ ] this is still the original shoes
[ ] someone has replaced my shoes
[ ] could be the original or replacement shoes

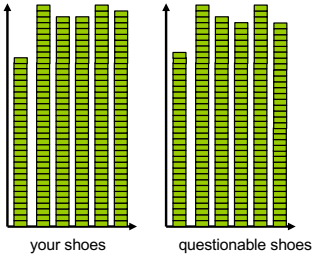
<let's vote>



[ ] this is still the original shoes
[ ] someone has replaced my shoes
[x] could be the original or replacement shoes



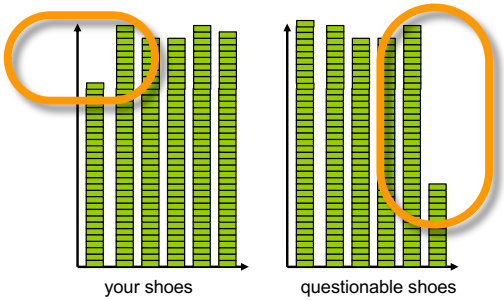
it could be your original shoes, or one that just happens to **behave the same.** a very, very, good copy maybe.



what are **the odds** of this being a different shoes?

you **cannot compute** the odds.

that's strange! why not?



this seems **unlikely…**

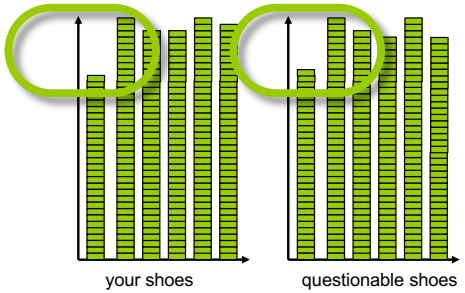
in both cases, there are two explanations

1.same shoes

2.different shoes

…thus this **must be true**

…now, in the other case



this **does seem not unlikely…**

in both cases, there are two explanations

1.same shoes

2.different shoes

we still have **two possible explanations**
→ we **cannot conclude** anything

…thus… **no thus**

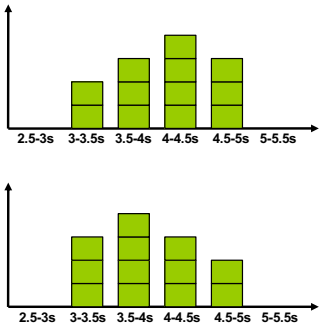
let's use stats

statistical significance ::

a result is called statistically significant if it is **unlikely to have occurred by chance**

before I show you how to compute, let's test our **intuition**

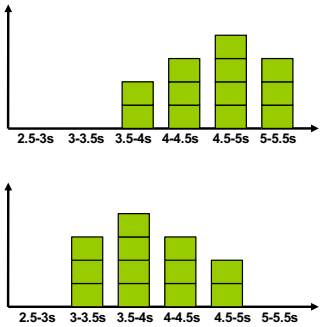
I show you pairs of distributions,
you tell me if the differences are **"statistically different"**



could have happened **by chance**
(45% dissimilar)

<30sec brainstorming>

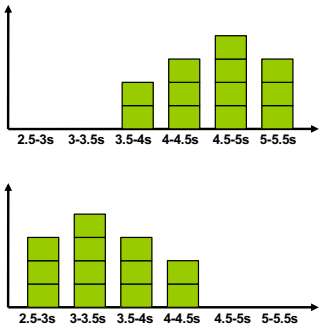
TTEST pvalue = 0.4548



still could have happened **by chance**
(14% dissimilar)

<30sec brainstorming>

TTEST pvalue = 0.1423



unlikely to have happened **by chance**
(0.1% dissimilar)

<30sec brainstorming>

TTEST pvalue = 0.00097

(student's) t-test
return a **p-value**

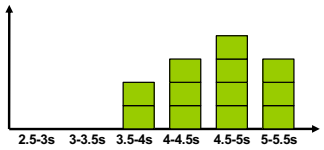
want to verify this: **run a t-test**

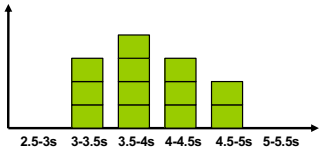
significance level ::

If a test of significance gives a **p-value lower** than the significance level, such results are informally referred to as 'statistically significant' .

Popular levels of significance are 10% (0.1), 5% (0.05), 1% (0.01), 0.5% (0.005), and 0.1% (0.001).

i.e., oddly, when we want to prove that they are different, we ask **whether they are the same…**





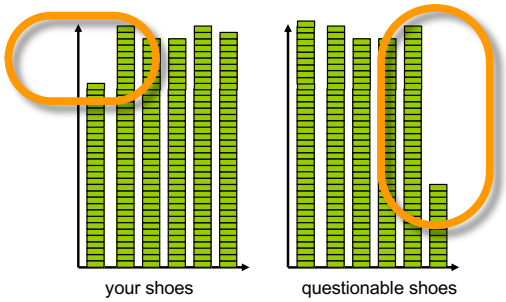
VS

null hypothesis: both data sets are from same mechanism

we are running stats in the hope that we will be able to **reject** the null hypothesis

→ if comparison of two groups reveals no statistically significant difference between the two, it does not mean **that there is no difference in reality.**

It only means that there is not enough evidence to reject the null hypothesis (it **fails to reject the null hypothesis).**

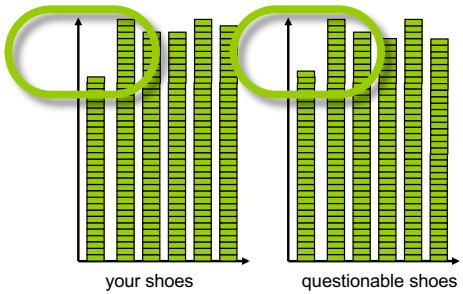


this seems **unlikely…**

in both cases, there are two explanations
1.same shoes
2.different shoes

…thus this **must be true**

…now, in the other case



this **does seem not unlikely…**

in both cases, there are two explanations
1.same shoes
2.different shoes

…thus… **no thus**

we still have **two possible explanations**
→ we **cannot conclude** anything

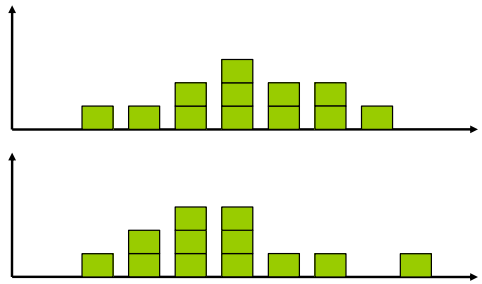
a classic screw-ups

you are making a new input device. You know that it cannot be better than a mouse, but you want to show that it is **as good as the mouse.**

how do you proceed?

<30sec brainstorming

how about you run a test and if stats come out non-significant you write "our tests showed that there was **no difference**"?

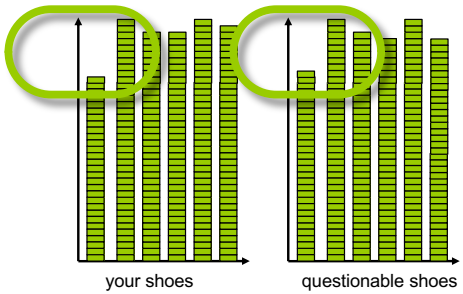


nope!

significant difference → mechanisms different

no significant difference → **nothing**

so how do you prove that two mechanisms **are the same?**



in both cases, there are two explanations

1.same shoes

2.different shoes

this **does seem not unlikely…**

…thus… **no thus**

you cannot

how to report non-significant results:
“our test did not **find** a significant difference”

practically

you want to test the effect of **two soporific drugs (independent variable)** on **amount of sleep(dependent variable)**. You take 10 participants and make then sleep to get their basic (control) sleep time. Then you give them drug 1 and note the difference of sleep time. You do the same for drug 2.

sleep extra drug 1	sleep extra drug 2
1 0.7	1 1.9
2 -1.6	2 0.8
3 -0.2	3 1.1
4 -1.2	4 0.1
5 -0.1	5 -0.1
6 3.4	6 4.4
7 3.7	7 5.5
8 0.8	8 1.6
9 0.0	9 4.6
10 2.0	10 3.4

in your terminal

```
head(sleep) # sleep is the table that already comes with R
and contain 20 observations on 10 patients to show the
effect of two soporific drugs on the increase in hours of
sleep

plot(extra ~ group, data = sleep)
```

you then have two options for t-test: paired or unpaired

```
with(sleep, t.test(extra[group == 1], extra[group == 2]))#
unpaired

with(sleep, t.test(extra[group == 1], extra[group == 2],
paired = TRUE))# paired
```

you want to test the effect of **two soporific drugs (independent variable)** on **amount of sleep(dependent variable)**. You take 10 participants and make then sleep to get their basic (control) sleep time. Then you give them drug 1 and note the difference of sleep time. You do the same for drug 2.

all participants did both conditions, i.e. had both drugs
= within subject experiment so the data is **paired**

otherwise (e.g. take 10 new participants for drug 2)
= between subject experiment so the data is **unpaired**

between subject experiment

```
with(sleep, t.test(extra[group == 1], extra[group == 2]))#
unpaired
```

```
data:  extra[group == 1] and extra[group == 2]
t = -1.8608, df = 17.776, p-value = 0.07939
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval: -3.3654832  0.2054832
sample estimates: mean of x mean of y 0.75      2.33
```

what you would write

"An unpaired student t-test showed no significant difference between the two drugs."

within subject experiment

```
with(sleep, t.test(extra[group == 1], extra[group == 2],
paired = TRUE))# paired
```

```
data:  extra[group == 1] and extra[group == 2]
t = -4.0621, df = 9  p-value = 0.002833
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval: -2.4598858 -0.7001142
sample estimates: mean of the differences -1.58
```

what you would write

"A paired student t-test showed significant difference between the two drugs (two-tailed t(9)=-4.0621, $p < 0.05$)"

(note: try to design your studies within-subject as it will increase the chance to reach a smaller p-value
… otherwise need twice more participants!)

one vs. two tail?

```
with(sleep, t.test(extra[group == 1], extra[group == 2],
paired = TRUE, alternative="less")) # or "greater"
```

```
Paired t-test

data:  extra[group == 1] and extra[group == 2]
t = -4.0621, df = 9, p-value = 0.001416
alternative hypothesis: true difference in means is less
than 0
95 percent confidence interval: -Inf -0.8669947
sample estimates: mean of the differences      -1.58
```

what you would write

"A paired student t-test showed significant difference between the two drugs (one-tailed t(9)=-4.0621, p < 0.001)."

two-tails: effect of drug 1 is > and/or < Drug 2

one-tail: only one side of the effect, i.e.

effect of Drug 1 > Drug 2 (less)

or

effect of Drug 1 < Drug 2 (greater)

e.g. created a shampoo for hair loss and want to know if better than concurrent one

(note: you will mostly use two-tails but if you can use a one-tail do it, it will increase the chance to reach a smaller p-value!)

multiple variables

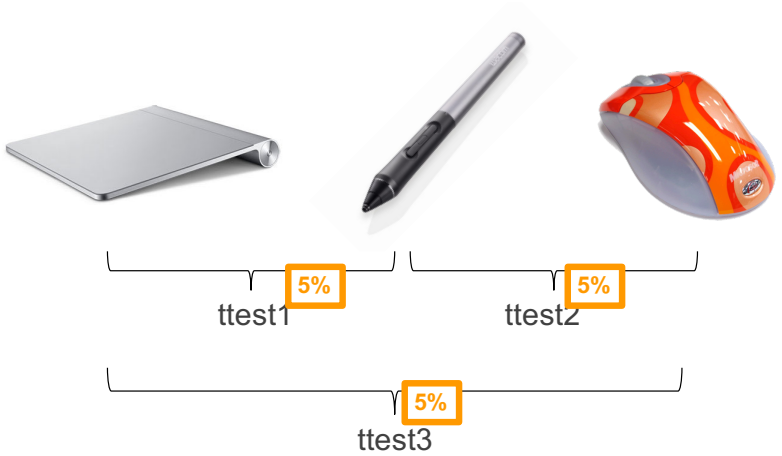
what if we have more than two variables?

you are making two new input devices, a track pad and a stylus. You want to know which one is better and if they are also better than a mouse.

how do we proceed?

<30sec brainstorming>

a simple solution would be to do this …



a simple solution would be to do this …

problem: any given test has a 5% chance of lying to you so when you use them multiple time you increase your risk of having errors (statisticians call this a “type I error”)

so there are two solutions to that:

bonferroni correction ::

when testing *n* hypotheses, test each one **against 0.05/n**

bonferroni correction ::

when testing *n* hypotheses, test each one **against 0.05/n**

in our example we would need to use **0.05/3** as a significant threshold instead of 0.05

or you could also use an

anova::

analyze of variance to compare multiple variables

one-way anova = one variable with multiple levels

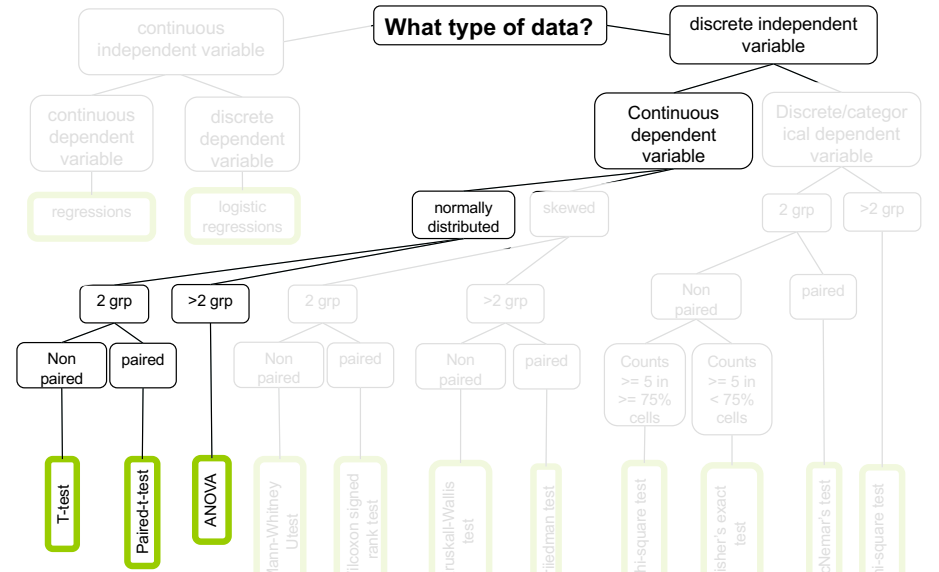
two-way anova = two variables with multiple levels

… we will look at ANOVA in the next lectures

1. Explain what is hypothesis testing
2. Identify the limit of hypothesis testing (we cannot prove that things are similar)
3. Explain what is a p value and a significance value
4. Explain what is a t-test and when to use it
5. Explain the difference between within and between subject studies
6. Explain what is a Bonferroni correction and find the new significance level given an experimental design

take away

3 experimental design



we are still at this part of the graph

let’s do an experiment!

memorization game

group 1

memorize as much as you can

group 2

if you beat group 1 = chocolate!

take a piece of paper and a pen

I will tell a list of numbers
“1,2,3,6,write”

only when “write” -> write the list on paper

I will show the list
1, 2, 3, 6

if you are **correct** continue the game

if you wrong stop the game, remember *best score*

practice trials

1, 4, 9 (size=3)

practice trials

8, 7, 3, 5, 6, 1 ,2 (size=7)

let’s start the real experiment!

trial

3, 2, 8 (size=3)

trial

4, 2, 5, 1 (size=4)

trial

7, 2, 5, 3, 1 (size=5)

trial

6, 2, 9, 8, 5, 1 (size=6)

trial

7, 4, 1, 8, 6, 3, 2 (size=7)

trial

2, 7, 4, 9, 3, 1, 5, 9 (size=8)

trial

1, 6, 7, 8, 5, 3, 1, 4, 6 (size=9)

trial

6, 4, 1, 9, 3, 8, 2, 1, 7, 9 (size=10)

trial

2, 7, 4, 1, 5, 7, 3, 8, 6, 4, 7 (size=11)

what is your best score (size of the list)?

enter it at

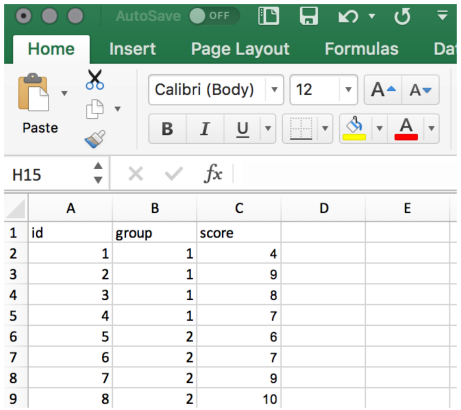
https://tinyurl.com/COMS10011

let's first look at the results

1. research question / hypothesis?
2. in(dependant) variables?
3. a. within or between subjects?
   b. counterbalancing?
4. how many repetitions/trials?
5. look at raw data
6. look at distributions
7. a. check for normality
   b. run some stats
8. conclude

look at raw data

let's put everything in a table (excel is great for that)



	A	B	C	D	E
1	id	group	score		
2	1	1	4		
3	2	1	9		
4	3	1	8		
5	4	1	7		
6	5	2	6		
7	6	2	7		
8	7	2	9		
9	8	2	10		

save your file as a .csv (comma separated virgule is a format to store tables as text files)

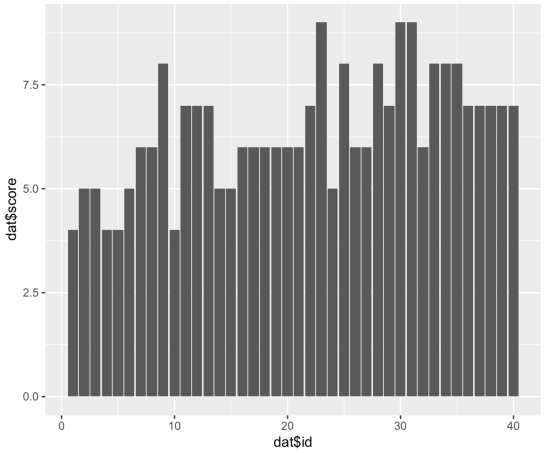
you can open csv with excel, text file an many other software

HCIXP.csv

```
id,group,score
1,1,4
2,1,9
3,1,8
4,1,7
5,2,6
6,2,7
7,2,9
8,2,10
```


```
dat = read.csv("HCIXP.csv", header = TRUE)
print(dat) # look at the file in R
```

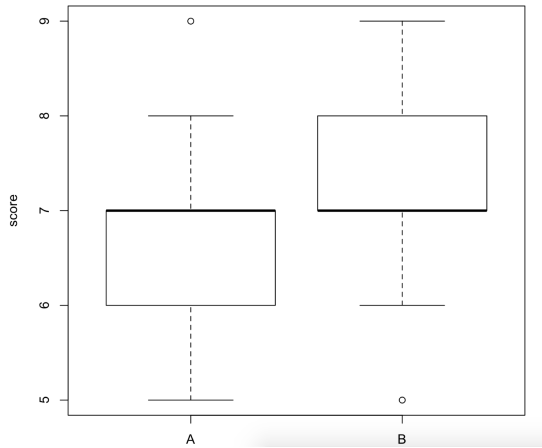

```
dat = read.csv("HCIXP.csv", header = TRUE)
print(dat) # look at the file in R

library(ggplot2)

ggplot(dat, aes(x = dat$id, y = dat$score)) +
geom_bar(stat = 'identity', position = 'dodge')
```



first: does the data look ok?

search for bugs, fatigue effect, learning effect
or outliers (>3 times std) = remove / redo xp


```
plot(score ~ group, data = dat)
```

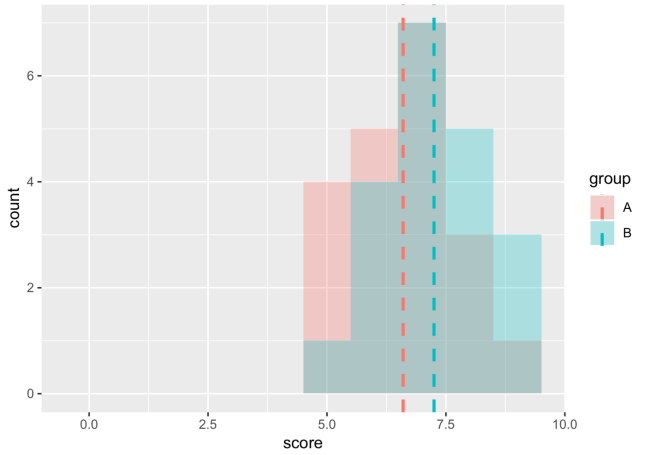


look at histograms


```
# Find the mean of each group
library(plyr)
cdat <- ddply(dat, "group", summarise,
score.mean=mean(score))
cdat
```

```
  group score.mean
1     A       5.60
2     B       7.25
```

```
# Overlaid histograms with means
ggplot(dat, aes(x=score, fill=group)) +
geom_histogram(binwidth=1, alpha=.3, position="identity")
+ geom_vline(data=cdat, aes(xintercept=score.mean,
colour=group), linetype="dashed", size=1) +
expand_limits(x = 0, y = 0)
```



your gut feeling: are these groups different?

are these distributions likely to have happen by chance?
… is this the results of the factor (chocolate)?

use a statistic test


```
# Use a t-test (two-tails, unpaired)
t.test(dat$score[dat$group == "A"], dat$score[dat$group
=="B"], alternative = "two.sided")

        Welch Two Sample t-test

        data:  dat$score[dat$group == "A"] and
        dat$score[dat$group == "B"]
        t = -1.8185, df = 37.982, p-value = 0.07688
        alternative hypothesis: true difference in means is
        not equal to 0
        95 percent confidence interval:-
        1.37361001  0.07361001
        sample estimates: mean of x mean of
        y  6.60      7.25
```

“We could not find any significance differences!”

p-value = 0.07

is is enough to say that the two groups are different?
-> nope, not under significant level of 0.05

can we say that the two groups are same then?
-> nope, can only prove things are different, but not that they are the same

conclude

if p was lower than significance level we could say:

"a student t-test showed significant difference between the two group (two-tailed $t(46)=4.520$, $p < 0.005$)"

otherwise:

"we did not find any significant results"

cannot conclude, no evidences to show that having chocolate rewards improve memorisation

let's go backward a little

1. research question / hypothesis?
2. in(dependant) variables?
3. a. within or between subjects?
   b. counterbalancing?
4. how many repetitions/trials?
5. look at raw data
6. look at distributions
7. a. check for normality
   b. run some stats
8. conclude

research question::

a statement that identifies a phenomenon to be studied

in our xp: I believe that **rewards improve memorization skills**

… suggested by *<insert smart guess>*

hypotheses::

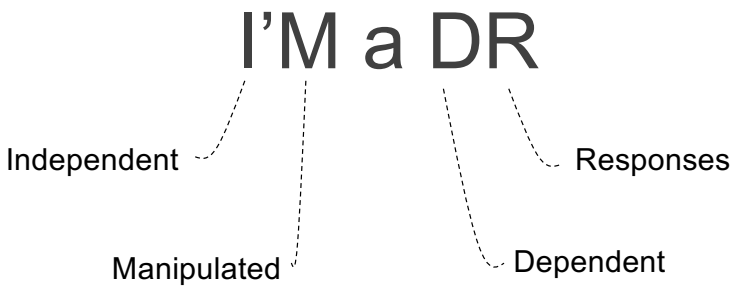
statement of the predicted relationship between at least two experimental variables

provisional answer to a research question

in our xp: **group chocolate will have a higher memorisation score than group with no reward**

(in)dependent variable ::

the **dependent variable** is the event studied and expected to change whenever the **independent variable** is altered



Dependent vs **Independent** variables?

vary A → make A
an **independent variable**

so we want to show that **A causes B**

measure B → make B
a **dependent variable**

in our xp?

independent variable = group type (nothing vs. chocolate)

dependent variable = memorization score

everything else should be a…

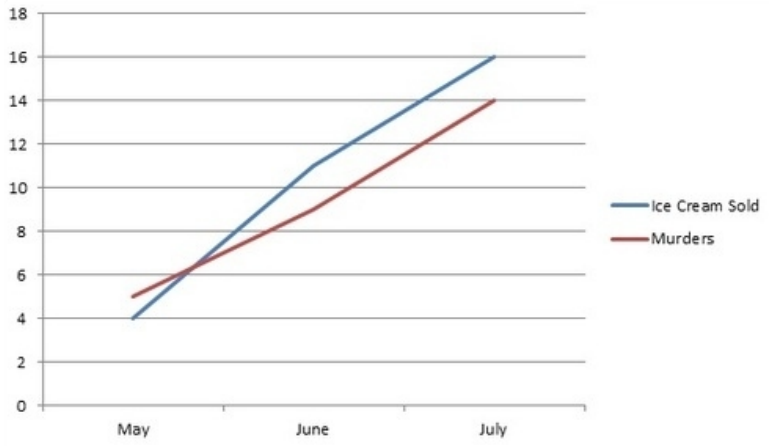
controlled variable ::

the variables that are kept constant to prevent their influence on the effect of the independent variable on the dependent

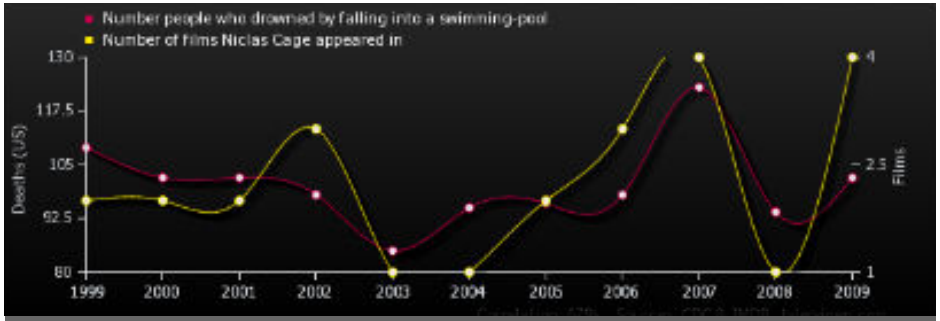
avoid…

confounding variable ::

extraneous variables that **correlates with both** the dependent variable and the independent variable



ice cream consumption leads to murder
counfounding : weather temperature



number of people drowned by falling into a swimming-pool correlates with number of films Nicolas Cage appeared in

this is not about **correlation**

this is about how to show **causality,**
i.e., that **some A causes some B**

in our xp, do we have confounding variables?

yes, it is not greatly designed :s

gender, age, background, what you ate before, if you like chocolate or not, if you are competitive and want the others not to have chocolate, if some of the numbers are familiar to you etc.

what can we do about it?

- avoid them by controlling as much as you can in the environment
- if you cannot, make it an independent variable (e.g. gender)
- some are inherent *noise* (human individuality), use more participants to get *statistical power*

the goal of a quantitative study is to find **a signal** in **a lot of noise**

experimental design:

aims at maximizing your chances of **finding the signal** and not the noise

1. need to absolutely **avoid systematic biases**

(e.g., learning effect, fatigue). They give you **false results!**

2. **avoid random noise.** It makes your results non-significant. Clever experimental design is all about keeping the noise down

e.g. in our xp, I made you **practice before**!

within vs. between?

within = all participants do same

between = participants do only certain conditions

suffer less user variation

statistical power with less participants

no biases from other conditions (e.g. transfer of learning)

within vs. between?

within = all participants do same

between = participants do only certain conditions

in our xp, it had to be **between subjects**
(because of the rewards)

participants did not do all conditions:
½ did the control condition
½ the reward condition

imagine a **within subjects** (test how fast we click an icon):

participants do all conditions:
they start with the trackpad
when finished they do the mouse

is it a good idea?
nope -> learning effect

counterbalancing ::

a method of avoiding confounding among variables
presenting conditions in a different order

one approach to counterbalancing is to use a…

A	B	C
C	A	B
B	C	A

Latin square ::

an $n \times n$ array filled with n different Latin letters, each occurring exactly once in each row and exactly once in each column.

how many trials?

ideally make as much trials as you can to reduce noise but try to keep experiment around 30 min … max 40 min

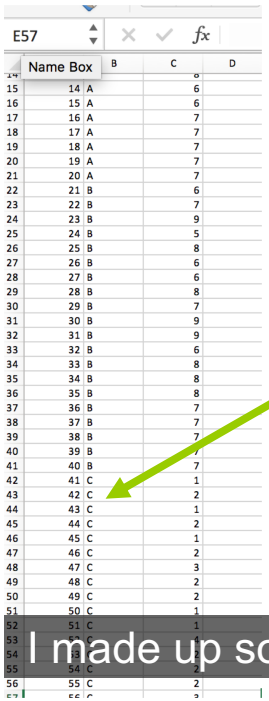
in our xp, we did only one trial because of time constraint, but should have done more to **reduce noises**

1. research question / hypothesis?
2. in(dependant) variables?
3. a. within or between subjects?
   b. counterbalancing?
4. how many repetitions/trials?
5. look at raw data
6. look at distributions
7. a. we will see why check for normality
   b. run some stats
      so far we know t-test
8. conclude

let's complexify a little

in our xp, let's add a 3rd imaginary group

they get a slap if they had the smallest memorisation score (obviously not ethical so let's keep this hypothetical!)

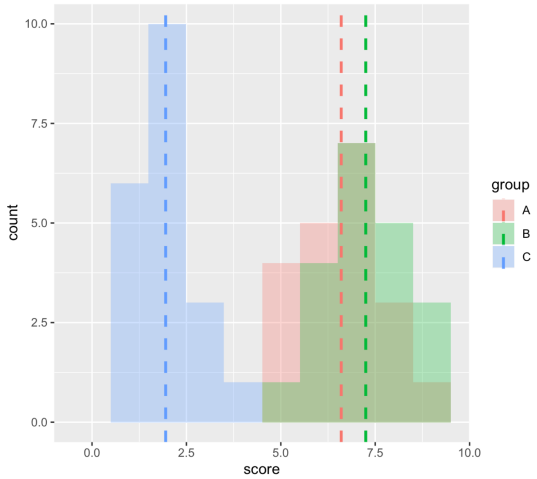


I made up some data

```
# Find the mean of each group
dat = read.csv("HCIXP-anova.csv", header = TRUE)
cdat <- ddply(dat, "group", summarise,
score.mean=mean(score))
cdat
```

```
  group score.mean
1     A       6.60
2     B       7.25
3     C       1.95
```

```
# Overlaid histograms with means
ggplot(dat, aes(x=score, fill=group)) +
geom_histogram(binwidth=1, alpha=.3, position="identity")
+ geom_vline(data=cdat, aes(xintercept=score.mean,
colour=group), linetype="dashed", size=1) +
expand_limits(x = 0, y = 0)
```

your gut feeling: are these groups different?

are these distributions likely to have happen by chance?

can we use t-tests?

(3 tests to compare group 1 with 2, 2 with 3 and 1 with 3)

-> yes but use Bonferoni correction

significance level not 0.05 anymore but 0.05 / number of comparisons performed (here 3) so 0.016


```
# Use a t-test (two-tails, unpaired)

# (we already know A vs B not significative) so we need to
do

t.test(dat$score[dat$group == "A"], dat$score[dat$group ==
"C"], alternative = "two.sided")
```

```
        t = 14.753, df = 34.591, p-value < 2.2e-16
```

```
# and

t.test(dat$score[dat$group == "B"], dat$score[dat$group ==
"C"], alternative = "two.sided")
```

```
        t = 17.054, df = 34.971, p-value < 2.2e-16
```

In both case p_value < 0.016 so we can conclude!

Another test we can use when we have more than two groups to compare is an ANOVA

we have 3 different conditions (or 1 factor with 3 different levels) so we will do a one-way ANOVA

anova::

analyze of variance to compare multiple variables

one-way anova = one variable with multiple levels

two-way anova = two variables with multiple levels


```
# first we run the one-way anova
library(ez)
ezANOVA(dat,id,between=group,dv=score)
```

```
  Effect DFn DFd      F        p p<.05       ges
1  group   2  57 154.8886 9.056612e-24     * 0.8445923
```

ok something is going to be interesting here

```
# second, run the pairwise comparison

pairwise.t.test(dat$score,dat$group, paired=FALSE,
p.adjust.method="bonferroni")
```

```
  A       B
B 0.16    -
C <2e-16  <2e-16
```

here are significant differences

and we don't need to do the Bonferroni correction (already included)

we can write:

"A one-way ANOVA showed a significant effect on time for the variable Group ($F_{2,57}=154.88$, $p < 0.05$)."

and then:

"Post-hoc comparison t-tests (using Bonferoni correction) showed significant difference between the group C and the group A ($p<0.05$) and between group C and group B ($p<0.05$)."

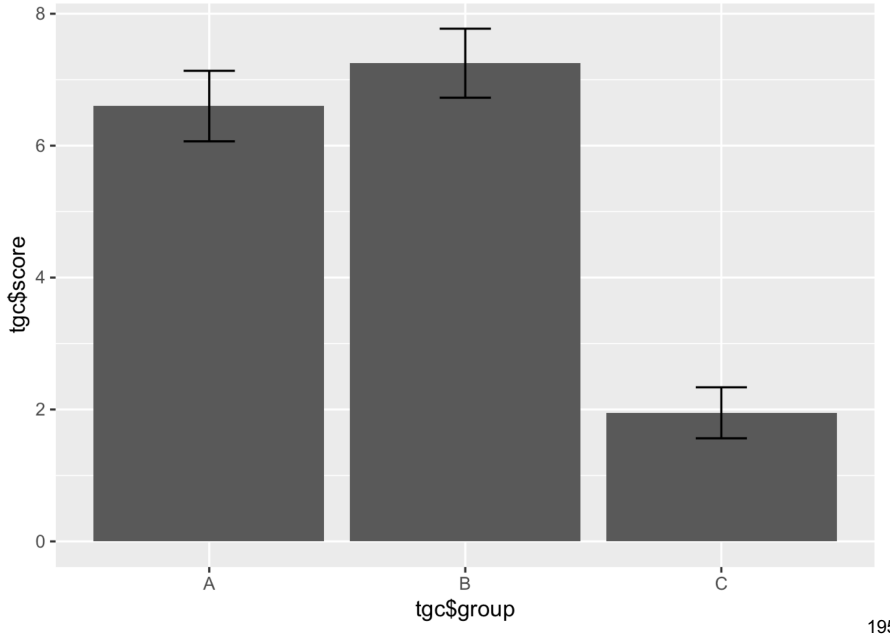
<you could also give means values to give more info>

one last thing you could find useful: how to make a graph with confident interval


```
# first we run the one-way anova
library(Rmisc)
tgc <- summarySE(dat, measurevar="score",
groupvars=c("group"))
tgc
```

```
  group  N score        sd        se        ci
1     A 20  6.60 1.1424811 0.2554665 0.5346976
2     B 20  7.25 1.1180340 0.2500000 0.5232560
3     C 20  1.95 0.8255779 0.1846048 0.3863824
```

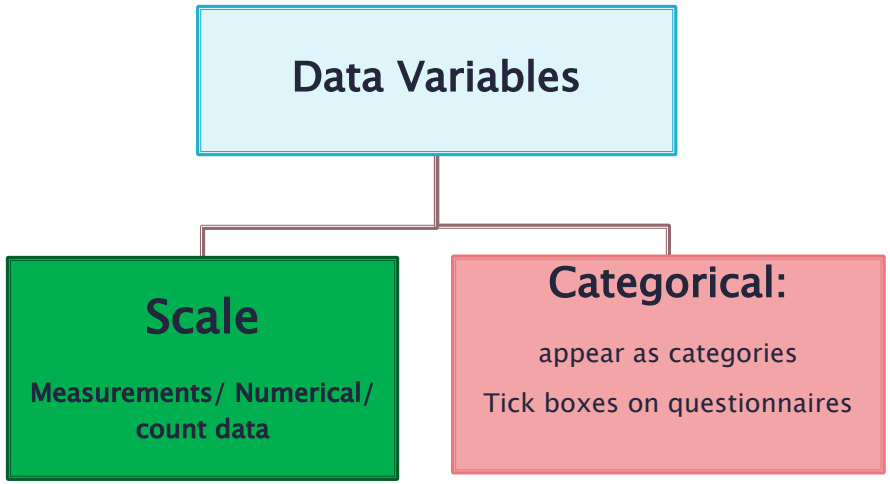
```
ggplot(data = tgc, aes(x = tgc$group, y = tgc$score)) +
geom_bar(stat = 'identity', position = 'dodge') +
geom_errorbar(aes(ymin= tgc$score - ci, ymax= tgc$score  +
ci), width=.2, position=position_dodge(.9))
```

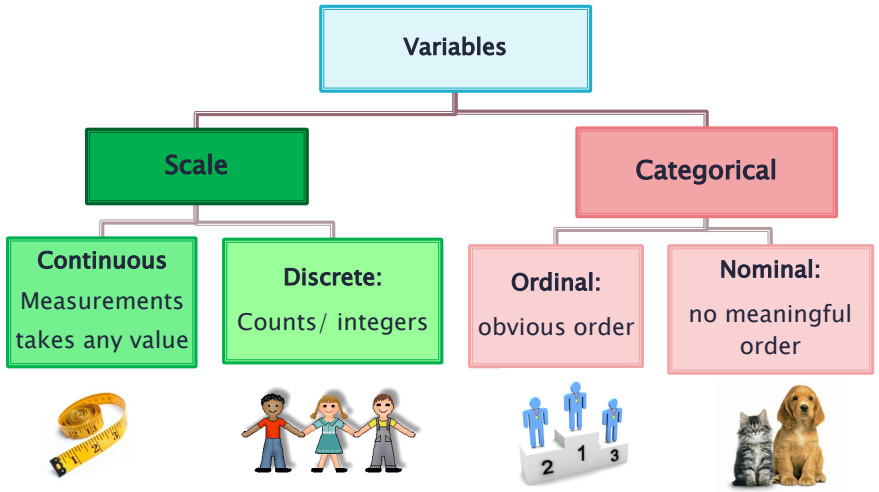



ok we have learned quite a lot so far!

1. research question / hypothesis?
2. in(dependent) variables?
3. 3a within or between subjects?
   - 3b counterbalancing?
4. how many repetitions/trials?
5. look at raw data
6. look at distributions
7. 7a we will see why check for normality
   - 7b run some stats
     T-test if 2 group
     ANOVA if more
8. conclude

let's talk about dependent variables





Q1: What is your favourite subject?

Maths	English	Science	Art	French

Q2: Gender:

Male	Female

Q3: I consider myself to be good at mathematics:

Strongly Disagree	Disagree	Not Sure	Agree	Strongly Agree

Q4: Score in a recent mock GCSE maths exam:

Score between 0% and 100%

Q1: What is your favourite subject? **Nominal**

Maths	English	Science	Art	French

Q2: Gender:

Male	Female

Binary/ Nominal

Q3: I consider myself to be good at mathematics:

Strongly Disagree	Disagree	Not Sure	Agree	Strongly Agree

Ordinal

Q4: Score in a recent mock GCSE maths exam:

Score between 0% and 100%

Scale

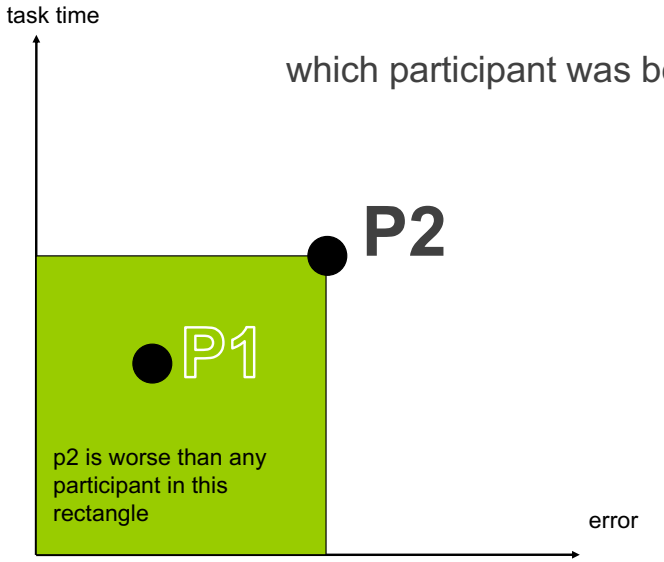
time and error as dependent variables …

imagine an experiment:

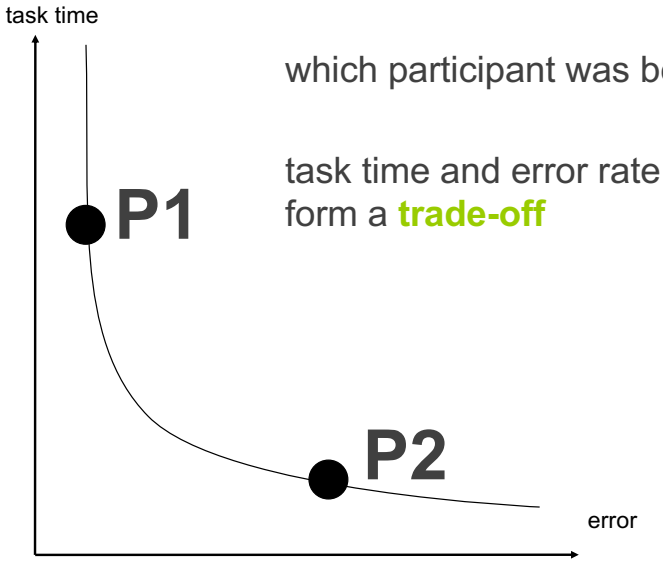
when you hear letter, press the corresponding keys as fast as possible

if there is no penalty for error, participants just slam the keyboard randomly

→ we also **need to consider error rate.**
→ each trial is effectively a (time, error) pair

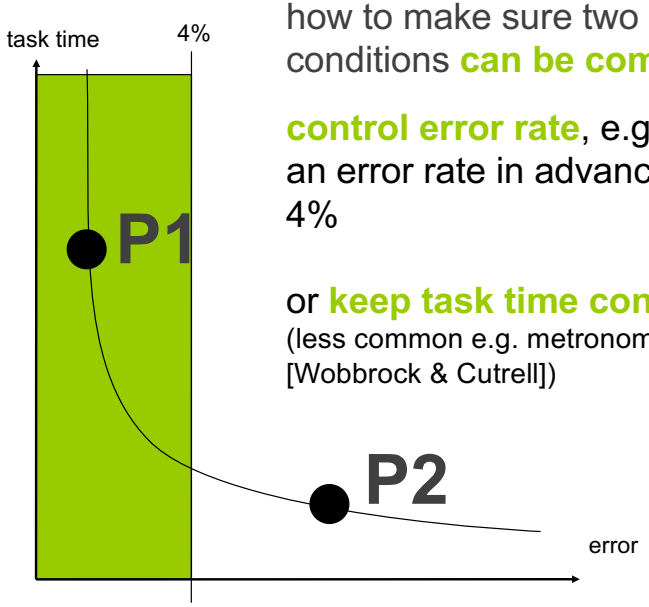


which participant was better?



which participant was better?

task time and error rate form a **trade-off**



how to make sure two conditions **can be compared?**

control error rate, e.g. pick an error rate in advance, say 4%

or **keep task time constant**
(less common e.g. metronome study [Wobbrock & Cutrell])

questionnaires as dependent variables …

Goal is to collect information that is:

Valid

measures the quantity that is supposed to be measured

Reliable

measures the quantity in consistent/reproducible manner

Unbiased

measures the quantity in a way that does not systematically under- or overestimate the true value

Discriminating

can distinguish adequately between respondents for whom the underlying level of the quantity or concept is different

How many cups of coffee or tea do you drink in a day?

No, ask for an answer in only one dimension, separate the question into two

(1) How many cups of coffee do you drink during a typical day?
(2) How many cups of tea do you drink during a typical day?

What brand of computer do you own?
(A) IBM PC
(B) Apple

Avoid hidden assumptions
Make sure to accommodate all possible answers

Make each response a separate dichotomous item
Do you own an IBM PC? (Circle: Yes or No)
Do you own an Apple computer? (Circle: Yes or No)

Or allow for multiple responses
What brand of computer do you own? (Circle all that apply)
Do not own computer
IBM PC
Apple
Other

Have you had pain in the last week?
[ ] Never [ ] Seldom [ ] Often [ ] Very often

Make sure question and answer options match
Reword either question or answer to match

How often have you had pain in the last week?
[ ] Never [ ] Seldom [ ] Often [ ] Very Often

Where did you grow up?

Country

Farm

City

Avoid questions having non-mutually exclusive answers

Design the question with mutually exclusive options

Where did you grow up?

House in the country

Farm in the country

City

Which one of the following do you think increases a person's chance of having a heart attack the most? (Check one.)

[ ] Smoking[ ] Being overweight [ ] Stress

Encourage to consider each possible response to avoid the uncertainty of whether a missing item may represent either an answer that does not apply or an overlooked item

Which of the following increases the chance of having a heart attack?

Smoking:	[ ] Yes	[ ] No	[ ] Don't know
Being overweight:	[ ] Yes	[ ] No	[ ] Don't know
Stress:	[ ] Yes	[ ] No	[ ] Don't know

On a scale from 1 to 5, how fun did you have using our new system?

1. not at all 2. Not really 3.undecided 4. somewhat 5. very much

Avoid biased questions

Design the question with mutually exclusive options

On a scale from 1 to 5, how would you rate your experience with our new system?

1. not fun at all 2. Not really fun 3.undecided 4. somewhat fun 5. very much fun

Rank from 1 to 3 your preference in beverage

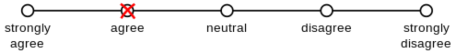
[ ] Tea [ ] Coffee [ ] Orange Jus

Avoid ranking at all cost and rather use Likert scales

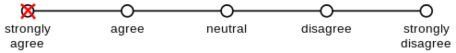
On a scale from 1 to 5 rate how much you like the following beverages

Tea:	1. not at all	2. Not really	3.undecided	4. somewhat	5. very much
Coffee:	1. not at all	2. Not really	3.undecided	4. somewhat	5. very much
Orange jus:	1. not at all	2. Not really	3.undecided	4. somewhat	5. very much

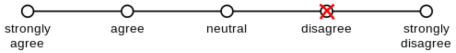
1. Wikipedia has a user friendly interface.



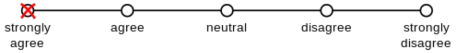
2. Wikipedia is usually my first resource for research.



3. Wikipedia pages generally have good images.



4. Wikipedia allows users to upload pictures easily.



if you want to collect subjective metric such as opinions, use **Likert Scale** = ordinal but treated as **continuous variable**

Likert scale::

psychometric response scale primarily used in **questionnaires** to obtain participant's preferences or degree of agreement with a statement (generally 5pt likert scale, also 7pt)

Agreement

- Strongly Agree
- Agree
- Undecided
- Disagree
- Strongly Disagree

Frequency

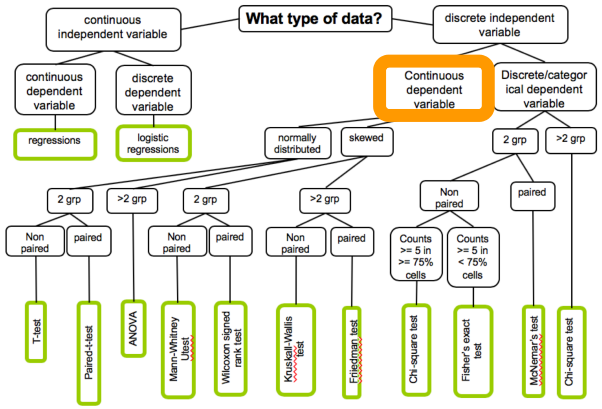
- Very Frequently
- Frequently
- Occasionally
- Rarely
- Never

Importance

- Very Important
- Important
- Moderately Important
- Of Little Importance
- Unimportant

Likelihood

- Almost Always True
- Usually True
- Occasionally True
- Usually Not True
- Almost Never True



Likert are Ordinal but can be treated as **Continuous !!!!!!!**

ok so there are many type of data and so what?

so far we played with data (time, errors, memo)
that tends to **follow curve of normal distribution**
(typical of human performances)

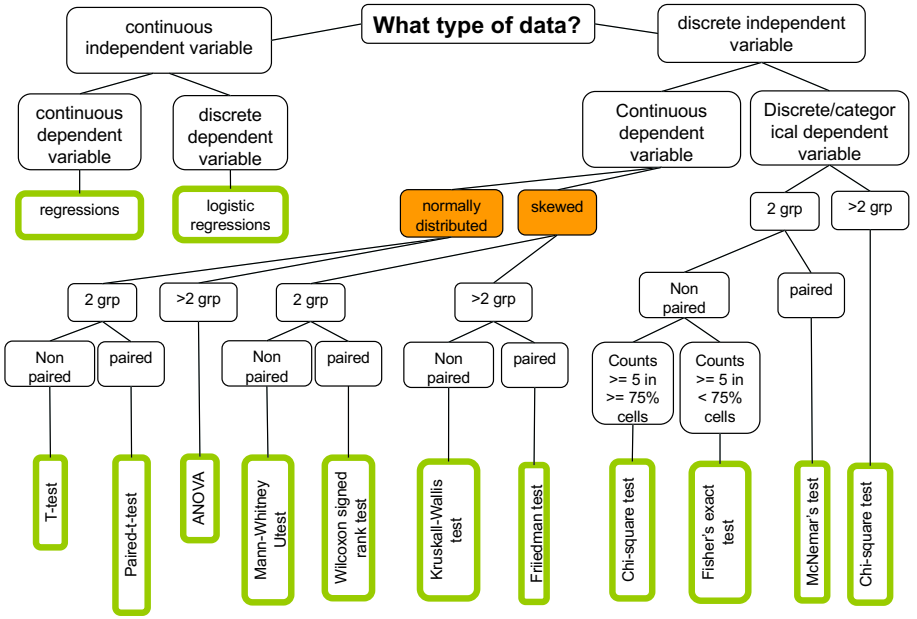
you could also deal with data
that **tends not to follow a normal distribution** (e.g.
Likert scale surveys)

use parametric tests (ttest, anova)

so far we played data (time, errors, memo)
that tends to **follow curve of normal distribution**
(typical of human performances)

you could also deal with data
that **tends not to follow a normal distribution** (e.g.
Likert scale surveys)

use non-parametric tests



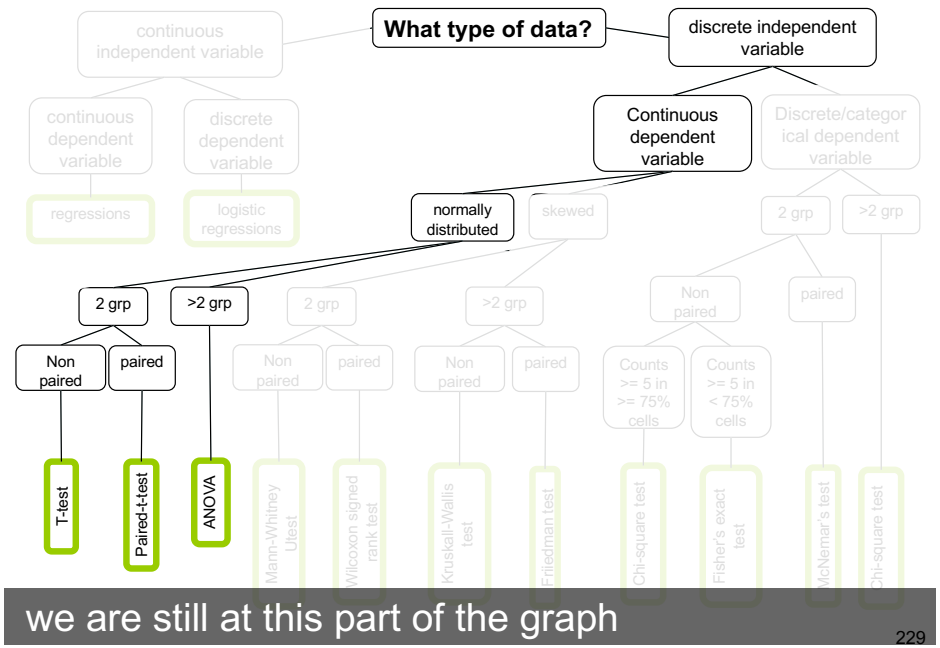
the best thing to do is to test if your data follow a normal distribution or not first before running the stats

… we will look at this in two lectures

1. Explain the eight steps to design and analyze an experiment
2. Explain what is a within or between subject experiment
3. Explain what is a controlled variable or a confounding variable
4. Explain the difference between correlation and causality
5. Identify different types of variables
6. Understand when to use a t-test, when to use an Anova
7. Explain what is a Likert scale in questionnaires
8. Explain when to use non-parametric tests

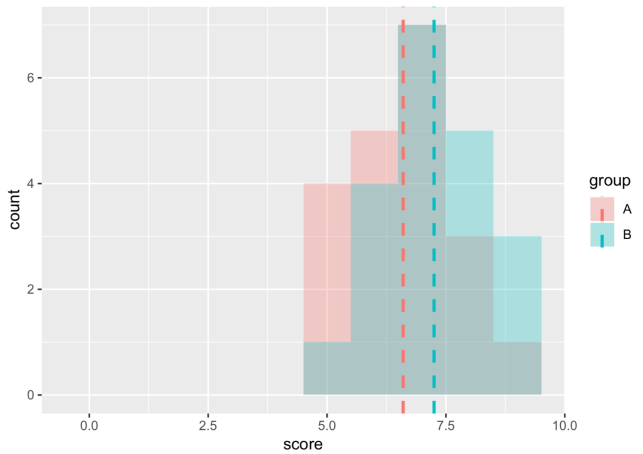
take away

4 parametric tests

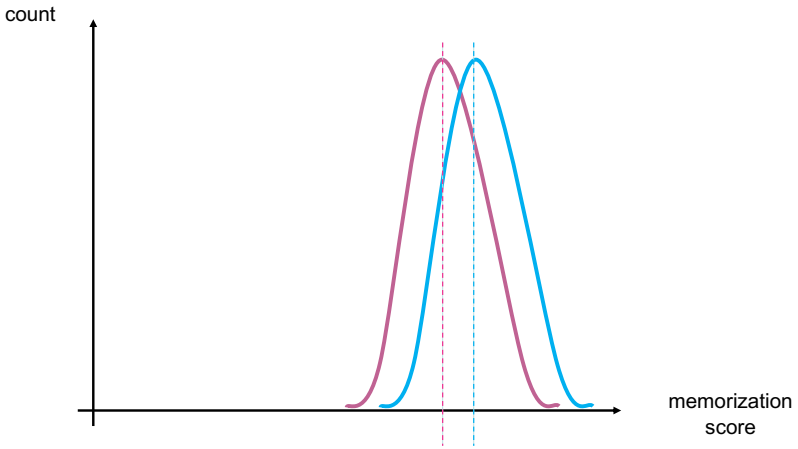


we are still at this part of the graph

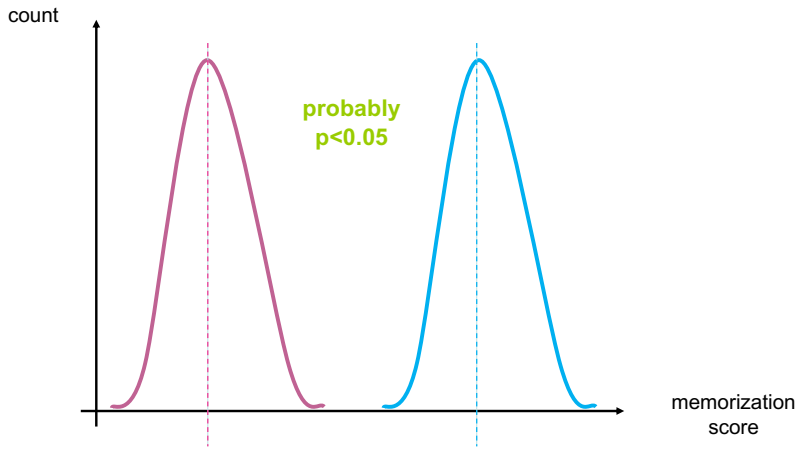
lets go back to our memorization experience



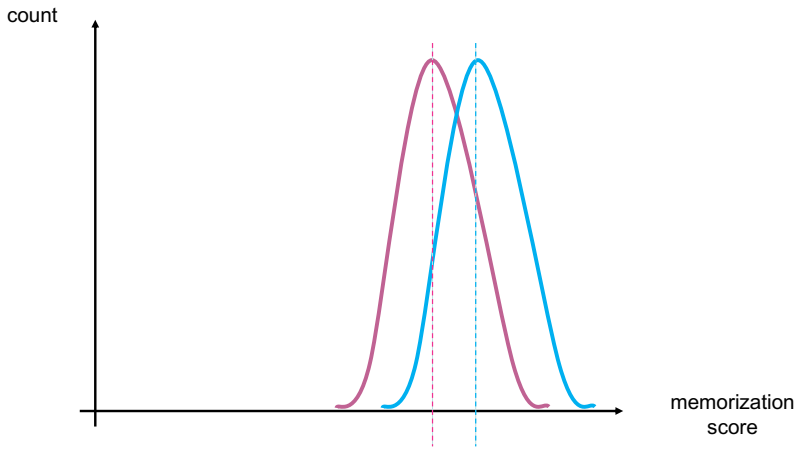
no evidences for chocolate vs. baseline (p>0.05)



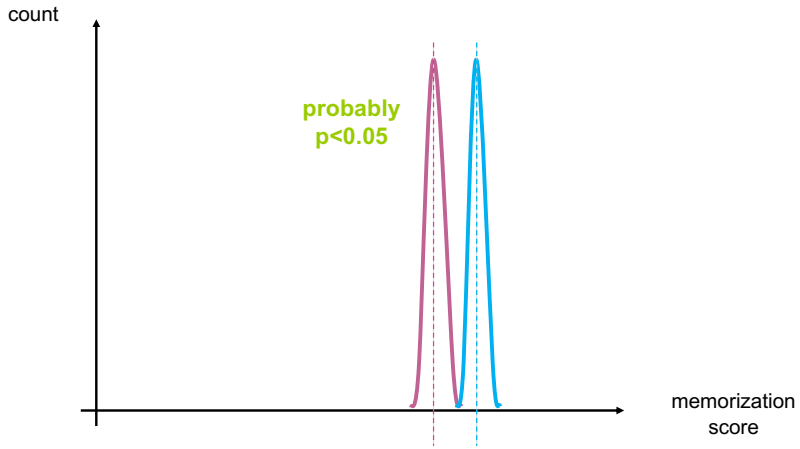
no evidences for chocolate vs. baseline (p>0.05)
(let's just assume these are normally distributed)



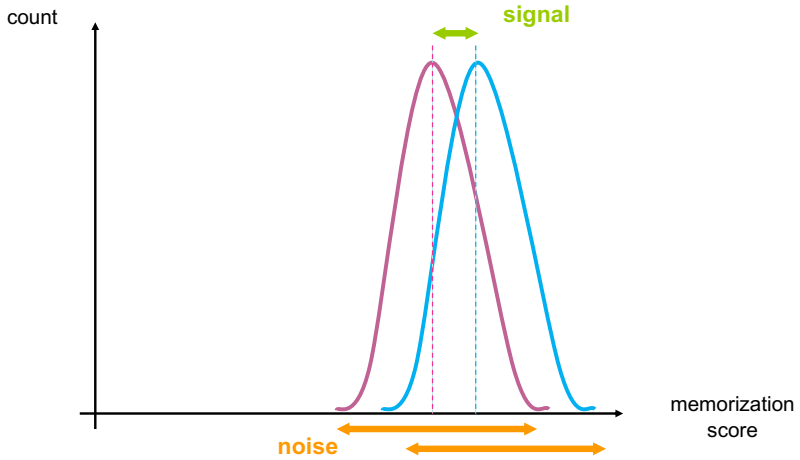
if the mean where further apart, it would increase our chances to have a p<0.05



<without changing the means, what else can we do to these data to make some more different?>



if the distributions where less spread out, it would increase our chances to have a p<0.05

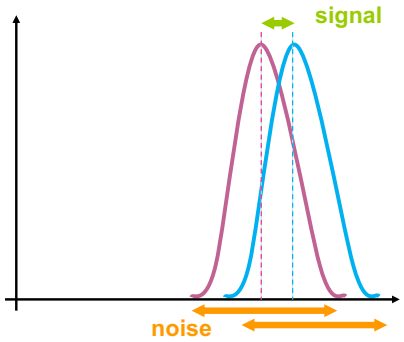


the goal of a study is to find
a signal in **a lot of noise**

any statistical tests ::

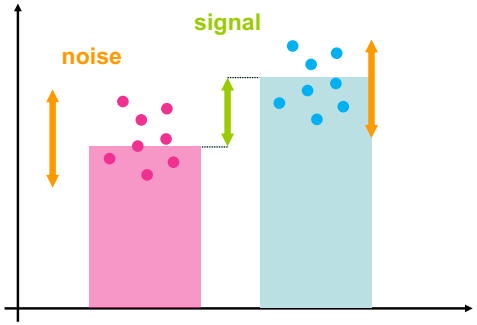
$$\frac{\text{signal}}{\text{noise}}$$

T-tests ::



$$\frac{\text{difference between group means}}{\text{variability of groups}}$$

T-tests ::



$$\frac{\text{difference between group means}}{\text{variability of groups}}$$

T-tests ::

Paired $t = \dfrac{\overline{x1} - \overline{x2}}{s/\sqrt{n}}$

Unpaired $t = \dfrac{\overline{x1} - \overline{x2}}{s\sqrt{\left(\frac{1}{n1} + \frac{1}{n2}\right)}}$

paired t-test

different between group means (to maximize)

$$t = \frac{\overline{x1} - \overline{x2}}{s/\sqrt{n}}$$

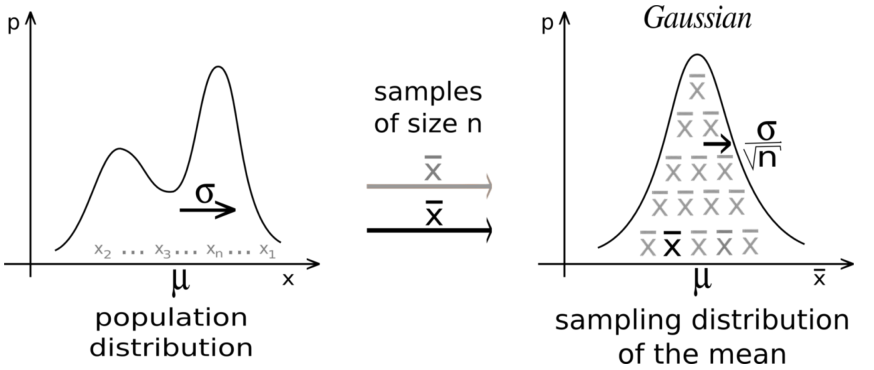
standard deviation of the differences (to minimize)

different between group means (to maximize)

$$t = \frac{\overline{x1} - \overline{x2}}{s/\sqrt{n}}$$

standard error of the mean (to minimize)

but why do we need to divide by $\sqrt{n}$ (n = sample size)?



this comes from the **central limit theorem** you have seen before (lecture 10)

by dividing by $\sqrt{n}$, we add a "**penalty**" for using a sample instead of the entire population

penalty is large when sample is very small

as sample size increases, penalty diminishes …

… infinitely approaching point where sample = the population itself

$$t = \frac{\overline{x1} - \overline{x2}}{s/\sqrt{n}}$$

we get a t-value
(to maximize)

both signal and noise are in the units of your data

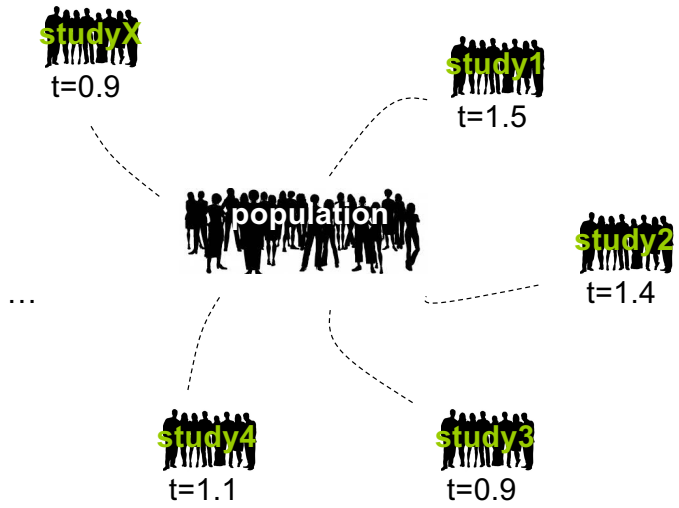
If signal = 6 and noise = 2, your t-value = 3, so the difference is 3 times the size of the standard error

If signal = 6 and noise = 6, your t-value = 1, the signal is at the same scale as the noise

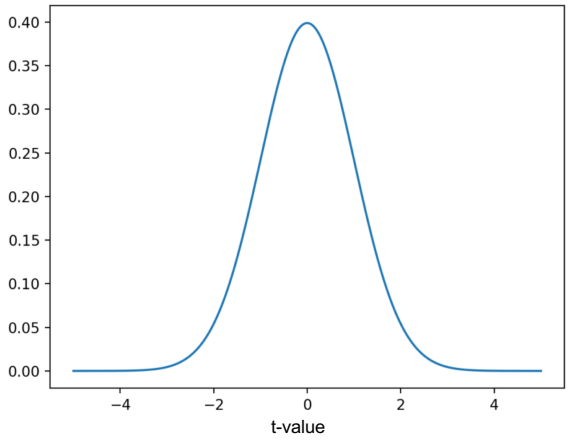
t-values = how distinguishable signal from noise

how do we know our t-value is any good, and how does this related to p-value?

this is where **t-distributions** come in



now let's take all these possible values and …



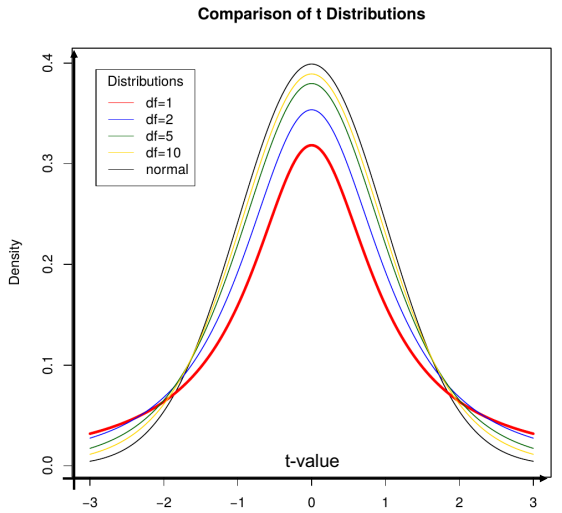
… plot a distribution of them

this type of distribution is a **sampling distribution**

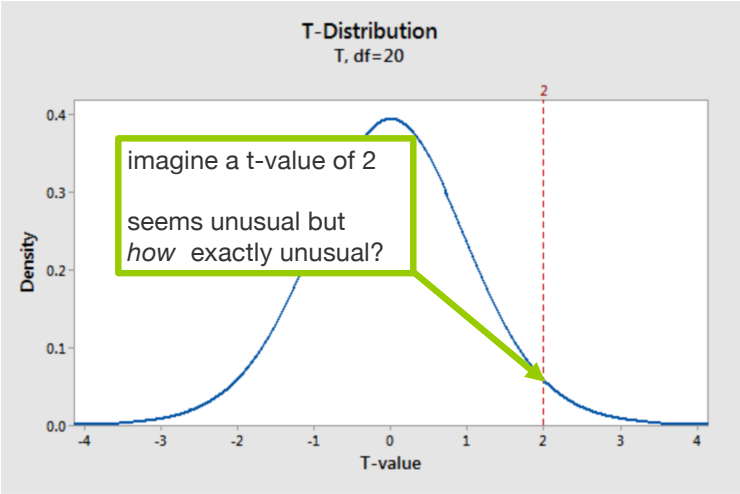
fortunately, the properties of t-distributions are well understood in statistics, so we can plot them without having to collect many samples!

a specific t-distribution is defined by its **degrees of freedom** (DF), a value closely related to sample size (here n-1)

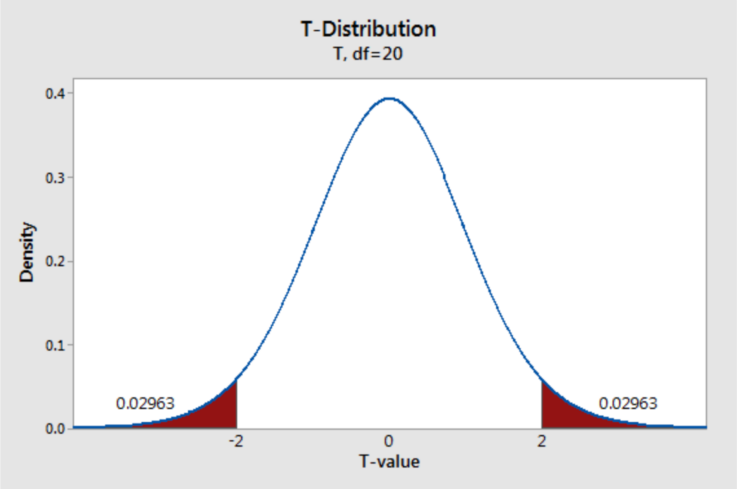
different t-distributions exist for every sample size



t-distributions assume that you draw repeated random samples from a population where the null hypothesis is true. You place the t-value from your study in the t-distribution to determine how consistent your results are with the null hypothesis.



e.g. here a t-distribution (DF =20 which means a sample size of 21). It plots the probability density function (PDF), which describes the likelihood of each t-value.



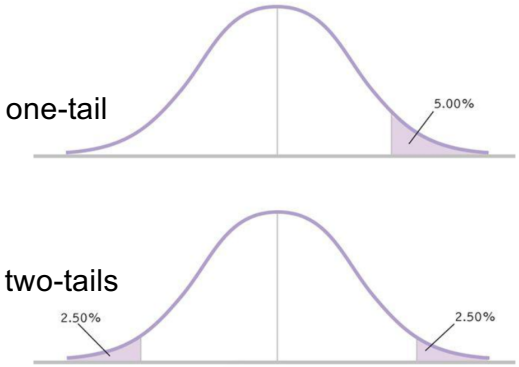
shade the area of the curve with t-values >2 and <-2

each regions has a probability of 0.02963, which sums to a total probability of 0.05926.

when the null hypothesis is true, the t-value falls within these regions nearly 5.9% of the time …

… this is our pvalue!

$$t = \frac{\overline{x1} - \overline{x2}}{s/\sqrt{n}}$$



it also does explain **one-tail vs. two tails** t-tests: one-tail only case about t=2 (not -2 or oppositely), so multiply pvalue by two.

at this point you understand more the **three reasons** of a low p_value (or t_value)

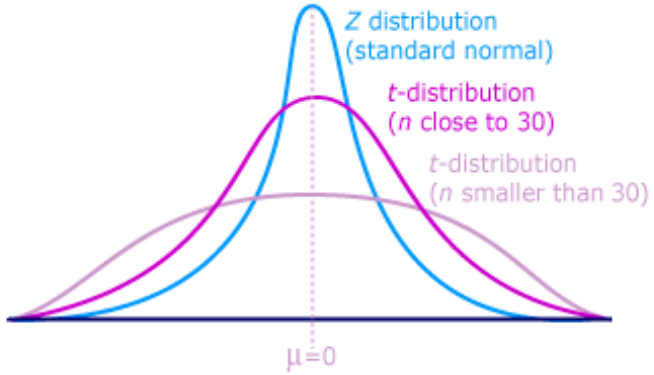
$$t = \frac{\overline{x1} - \overline{x2}}{s/\sqrt{n}}$$

1. difference not large enough
(what you are searching for, your signal is weak)

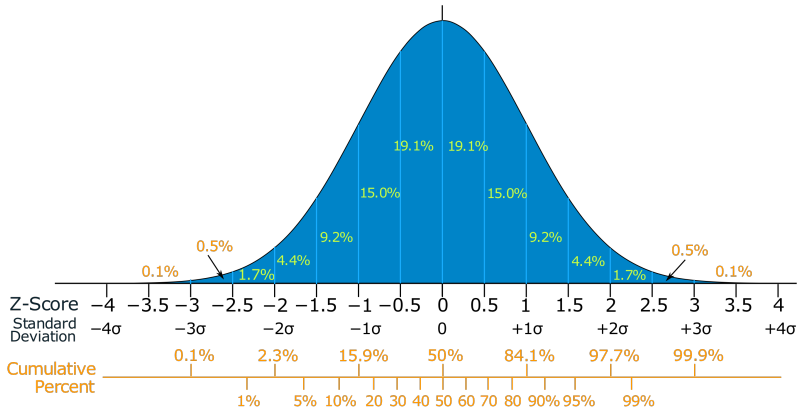
2. too much noise
(could your experimental design introduce noise? Check it)

3. not enough data
(run more participants, gather more trials)

how much data is enough?



the larger the sample size, the more t-distributions become a z-distribution (at around n=30), the less the area under the curve to reach a low p_value.

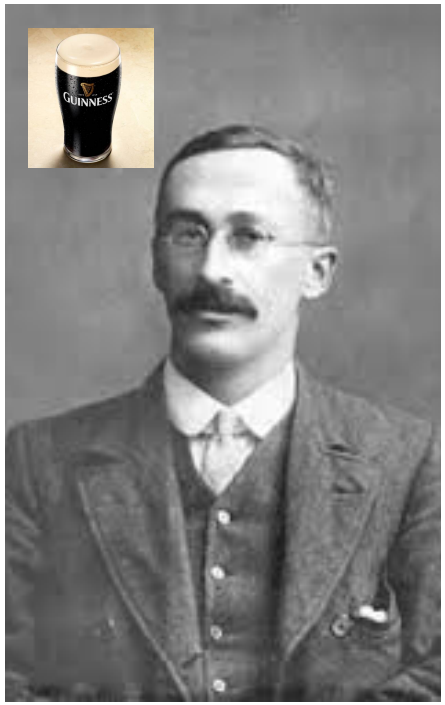


back to using a Z-score (i.e. number of standard deviations from the mean on normal distribution)

so why using a t-test then?

well there are cases when we want to use less sample to speed up the evaluation

this was the case of **William Sealy Gosset** …



Employee of Guiness, Gosset developed a **small sample** method to select the best yielding varieties of barley.

Biometricians like Pearson typically had hundreds of observations.

Guinness allowed him to publish his method under the name “Student” to prevent disclosure of confidential information.

Where do t-distrubitions came from?
https://www.youtube.com/watch?v=NvUDvmrd6fo&feature=youtu.be

now it does not mean that 2 sample is enough … a rule of thumb for a simple within experiment is 12-16 participants for example (and twice more for between experiment).

… we will actually look at how to know if you sample size is good in the lecture 18

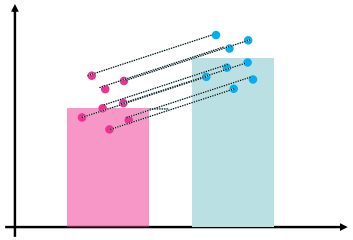
unpaired t-test

T-tests ::

Paired $t = \frac{\overline{x1} - \overline{x2}}{s/\sqrt{n}}$

Unpaired $t = \frac{\overline{x1} - \overline{x2}}{s\sqrt{\left(\frac{1}{n1} + \frac{1}{n2}\right)}}$

similar than paired t-test except this



paired t-test :: divide by $\sqrt{n}$ because data point paired

unpaired t-test :: multiply by $\sqrt{\left(\frac{1}{n1} + \frac{1}{n2}\right)}$

you can do the math: unpaired t-test the denominator (noise) is larger because we add n1 + n2

… thus why harder to reach low pvalue with unpaired t-test

practically

IDs	Before	After	1. difference	Paired T-test example	
1	312	300	12	(In green the initial data, in blue the computation)	
2	242	201	41		
3	340	232	108	steps by steps	
4	388	312	76		
5	296	220	76	1. add new colum to compute the differences between conditions for each participants	
6	254	256	-2	2. compute the mean of the differences (use excel formula =AVERAGE(new column))	56.1111111
7	391	328	63	3. compute the standard deviation of the differences (use formula =STDEV(new column))	34.173983
8	402	330	72	4. compute de standard error of the mean difference (difive 3. by SQRT(n))	11.3913277
9	290	231	59	5. compute t_value, i.e. step 2. divided by step. 4	4.92577449

(excel file in the git hub repository)

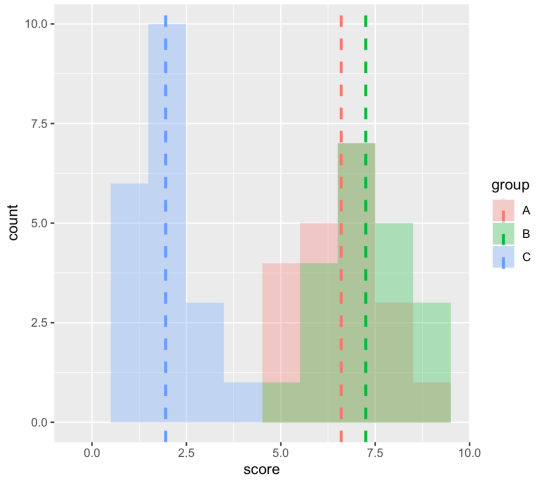
$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

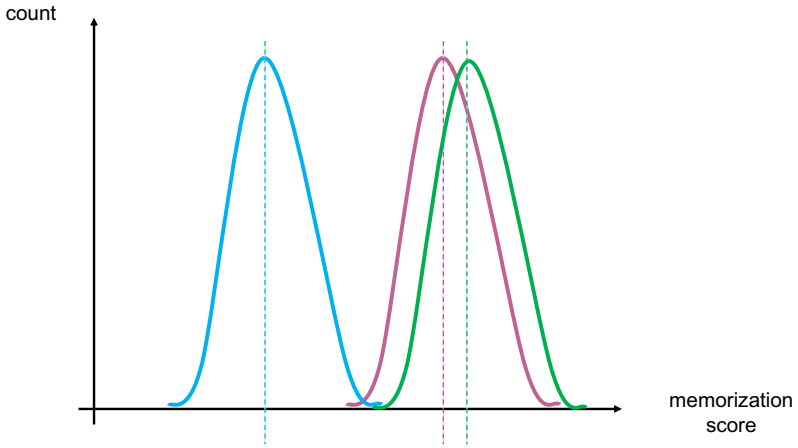
$$s^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_j - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

condA	condB	(xa-meanxa)2	(xb-meanxb)2	Paired T-test example	
134	70	196	961	(In green the initial data, in blue the computation)	
146	118	676	289		
104	101	256	0	step by step	
119	85	1	256	1. compute the mean of condition A (=AVERAGE)	120
124	107	16	36	2. compute the mean of condition B (=AVERAGE)	101
161	132	1681	961	3. compute the sample size in condition A (=COUNT)	12
107	94	169	49	4. compute the sample size in conditon B (=COUNT)	7
83		1369		5. add columns to compute square difference of each x in condition A minus mean condition A	
113		49		6. add columns to compute square difference of each x in condition B minus mean condition B	
129		81		7. compute the sum of 5. (=SUM)	5032
97		529		8. compute the sum of 6. (=SUM)	2552
123		9		9. compute the nominator of the tvalue (1. minus 2.)	19
				10. compute the variance (square of standard variation of the difference) = 7. + 8. divided by (3. + 4. - 2)	446.11765
				11. compute the denominator of the tvalue (SQRT(10. * (1/n1 + 1/n2))	10.045276
				12. compute the tvalue	1.8914364

(excel file in the git hub repository)

anovas



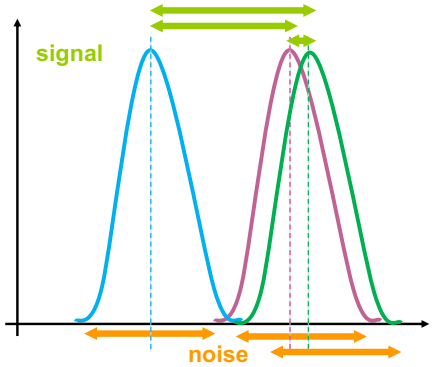


(let's assume again these are normally distributed)

any statistical tests ::

$$\frac{\text{signal}}{\text{noise}}$$

ANOVA ::



$$\frac{\text{difference between group means}}{\text{variability of groups}}$$

ANOVA ::

$$F = \frac{MS_{between}}{MS_{within}}$$

$$MS_{between} \quad \frac{SS_{between}}{df_{between}}$$

$$MS_{within} \quad \frac{SS_{within}}{df_{within}}$$

$$SS_{between} = \sum_{j=1}^{p} n_j(\bar{x}_j - \bar{x})^2$$

$$SS_{within} = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$$

don’t be afraid by this

	Group 1	Group 2	Group 3
n (sample size)	70	70	70
M (mean)	4	3.7	3.4
s^2 (variance)	4.4	5.2	6.1

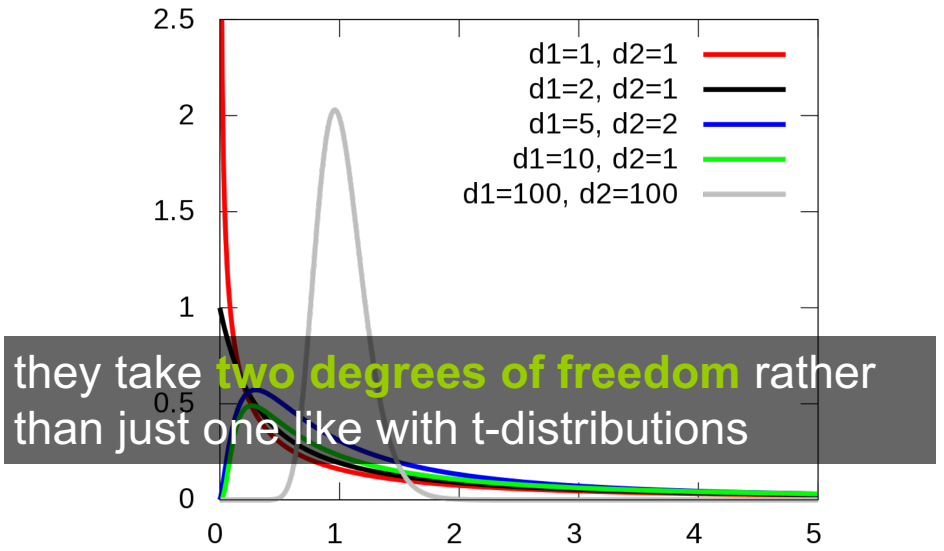
Anova by hand (step by step)		
1. compute th combined sample size N	210	
2. compute the degress of freedom between (dfbetween)	2	(number of groups - 1)
3. compute the degress of freedom within (dfwithin)	207	(n1-1)+(n2-1)+(n3-1)
the nominator		
4. compute the average mean	3.7	
5. compute the SSbetween	12.6	
6. compute the MSbetween (divide by dfbetween)	6.3	
the denominator		
7. compute the SSwithin	1083.3	(I multiply by ni here as the variance formula has a dividor which we don't need here)
8. Compute Mswithing (divide by dfwithin)	5.233333	
9. compute F	1.203822	
10. find p_value	0.302137	(p value NOT < alpha so DO NOT reject Ho)

Sample Variance = $s^2 = \frac{\Sigma(X - \bar{X})^2}{n - 1}$

(excel file in the git hub repository)

how do we know if the Fvalue is anygood?

… this is where **F-distributions** come in



they take **two degrees of freedom** rather than just one like with t-distributions

F Distribution critical values for P=0.05

Denominator (the within df – also called the error)

DF	Numerator DF (the between df) 1	2	3	4	5	7	10	15	20	30	60	120	500	1000
1	161.45	199.50	215.71	224.58	230.16	236.77	241.88	245.95	248.01	250.10	252.20	253.25	254.06	254.19
2	18.513	19.000	19.164	19.247	19.296	19.353	19.396	19.429	19.446	19.462	19.479	19.487	19.494	19.495
3	10.128	9.5522	9.2766	9.1172	9.0135	8.8867	8.7855	8.7028	8.6602	8.6165	8.5720	8.5493	8.5320	8.5292
4	7.7086	6.9443	6.5915	6.3882	6.2560	6.0942	5.9644	5.8579	5.8026	5.7458	5.6877	5.6580	5.6352	5.6317
5	6.6078	5.7862	5.4095	5.1922	5.0504	4.8759	4.7351	4.6187	4.5582	4.4958	4.4314	4.3985	4.3731	4.3691
7	5.5914	4.7375	4.3469	4.1202	3.9715	3.7871	3.6366	3.5108	3.4445	3.3758	3.3043	3.2675	3.2388	3.2344
10	4.9645	4.1028	3.7082	3.4780	3.3259	3.1354	2.9782	2.8450	2.7741	2.6996	2.6210	2.5801	2.5482	2.5430
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7066	2.5437	2.4035	2.3275	2.2467	2.1601	2.1141	2.0776	2.0718
20	4.3512	3.4928	3.0983	2.8660	2.7109	2.5140	2.3479	2.2032	2.1241	2.0391	1.9463	1.8962	1.8563	1.8498
30	4.1709	3.3159	2.9223	2.6896	2.5336	2.3343	2.1646	2.0149	1.9317	1.8408	1.7396	1.6835	1.6376	1.6300
60	4.0012	3.1505	2.7581	2.5252	2.3683	2.1666	1.9927	1.8365	1.7480	1.6492	1.5343	1.4672	1.4093	1.3994
120	3.9201	3.0718	2.6802	2.4473	2.2898	2.0868	1.9104	1.7505	1.6587	1.5544	1.4289	1.3519	1.2804	1.2674
500	3.8601	3.0137	2.6227	2.3898	2.2320	2.0278	1.8496	1.6864	1.5917	1.4820	1.3455	1.2552	1.1586	1.1378
1000	3.8508	3.0047	2.6137	2.3808	2.2230	2.0187	1.8402	1.6765	1.5811	1.4705	1.3318	1.2385	1.1342	1.1096

Example: F for df = 2,207 is 3.0718

also in form of table (here for the dfwithin and dfbetween of our excel example)


```
# first we run the one-way anova
library(ez)
ezANOVA(dat,id,between=group,dv=score)
```

```
  Effect DFn DFd      F        p p<.05       ges
1  group   2  57 154.8886 9.056612e-24     * 0.8445923
```

ok something is going to be interesting here

```
# second, run the pairwise comparison

pairwise.t.test(dat$score,dat$group, paired=FALSE,
p.adjust.method="bonferroni")
```

```
  A       B
B 0.16    -
C <2e-16  <2e-16
```

here are significant differences

and we don't need to do the Bonferroni correction (already included)

this was from last week, R gives us all this numbers

degrees of freedom

degrees of freedom::

the number of values in the final calculation of a **statistic** that are free to vary

a complex notion but here is the intuition …

you have 7 hats, you want to wear one different every day for a week

Monday: 7 choices
Tuesday: 6 choices
Wednesday: 5 choices
Thursday: 4 choices
Friday: 3 choices
Saturday: 2 choices
Sunday: **NO choice**

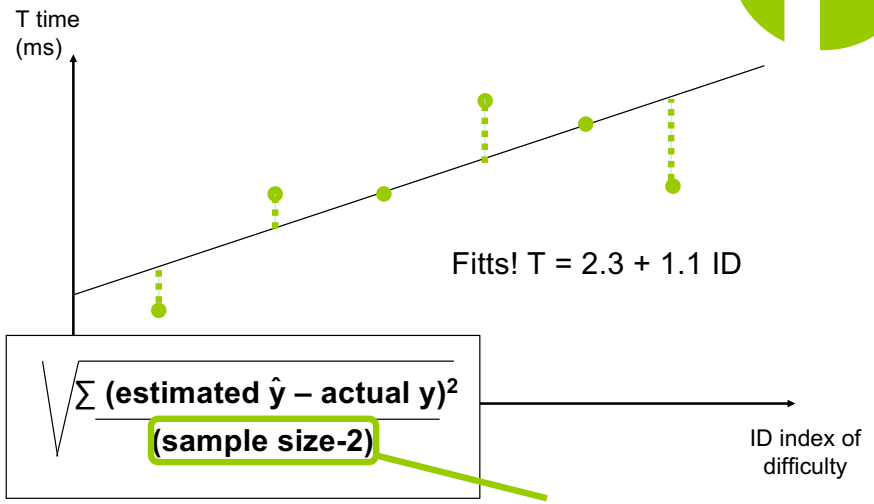
degrees of freedom is 7-1

if we have 7 observations and the mean of these observation (that you need to do t-test or Anova) your degree of freedom is 7-1

because if you know 6 observations you automatically know the 7th one (thanks to the mean)

remember the lecture on regression?

standard error of the estimate



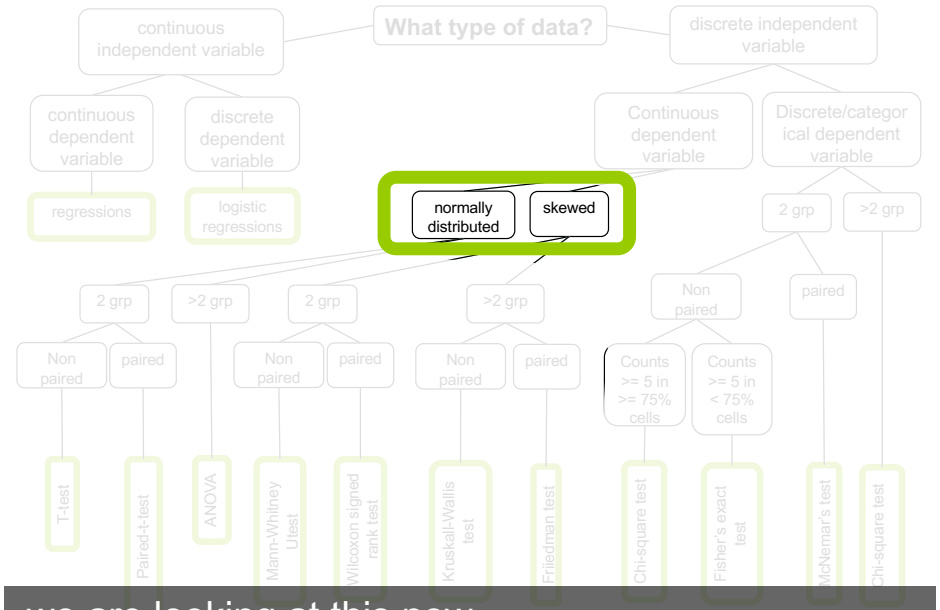
a simple (approximate) way to understand this is that we have two variables, the slope and the intercept of the regression line

that give us extra information, thus the minus 2

1. Explain the general equation of a statistical test (signal/noise)
2. Explain how a t-test is computed and be able to do it by hand
3. Explain the 3 reasons why a t-value can be low (signal too low, noise too high, small sample)
4. Explain how a Anova is computed, and be able to do it by hand
5. Explain what are t-distributions and f-distributions

take away

5 normality tests



we are looking at this now

how to::

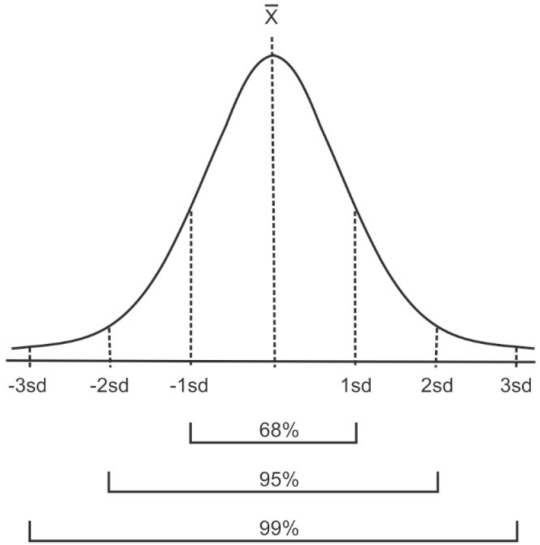
test if data is **normally distributed**

use **alternatives to non-parametric tests** if data not normally distributed

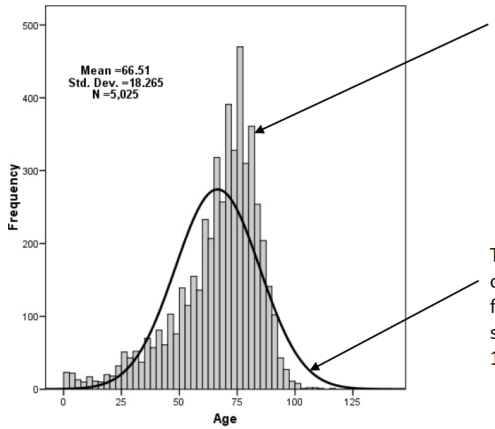
other verifications (assumptions) we need to do on data prior to doing certain statistical tests

assumption of normality

given the **mean** and **standard deviation** of a dataset = a theoretical normal distribution has those proportions



this theoretical normal distribution can then be compared to the actual distribution of the data.



The ***actual*** data distribution that has a mean of 66.51 and a standard deviation of 18.265.

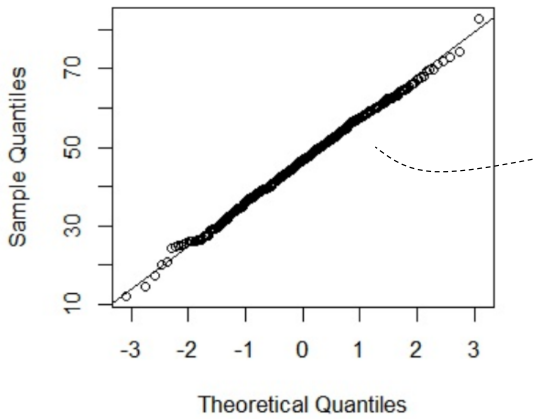
Theoretical normal distribution ***calculated*** from a mean of 66.51 and a standard deviation of 18.265.

<are the actual data statistically different than the computed normal curve? >

several methods to check that, we are only going to look at some of them: **Q-Q probability plots**, **Kolmogorov-Smirnov test** and **Shapiro-Wilks test**

(some others: W/S test, Jarque-Bera test, D'Agostino test)

Quantile Quantile Probability Plots



if the two sets come from a population with the same distribution, the points should fall along a line

it is a plot of the quantiles of the first data set against the quantiles of the second data set

3.89 4.75 6.33 4.75 7.21 5.78 5.80 5.20 7.90

does the following sample comes from a normality distributed population?

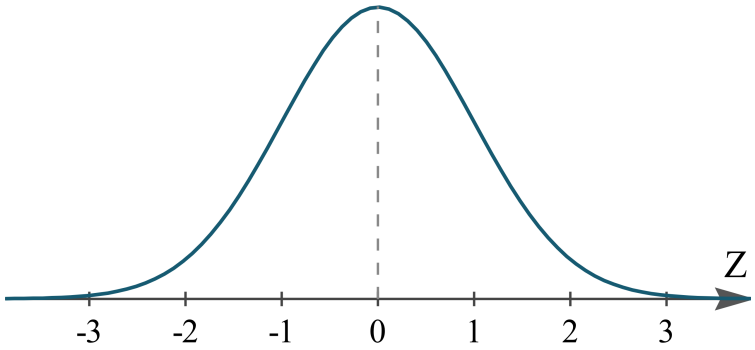
1. order the data:

3.89 4.75 4.75 5.20 5.78 5.80 6.33 7.21 7.90

2. plot these against appropriate quantile from the standard normal distribution

… let's look at this in detail

here a standard normal distribution truncated at -3 and 3

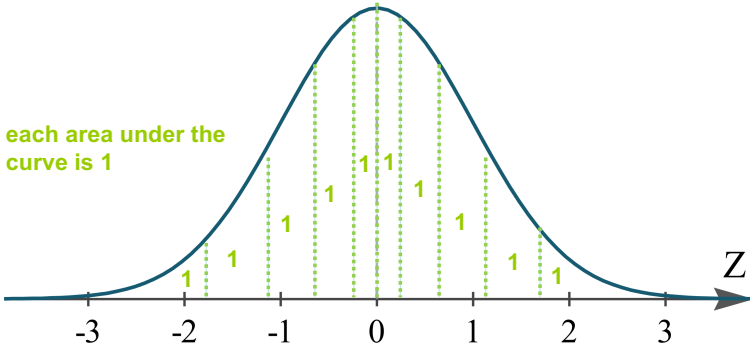


3. search for nine values on the normal distribution = split it in 10 areas

and here our **nine** sample values

3.89 4.75 4.75 5.20 5.78 5.80 6.33 7.21 7.90

here a standard normal distribution truncated at -3 and 3



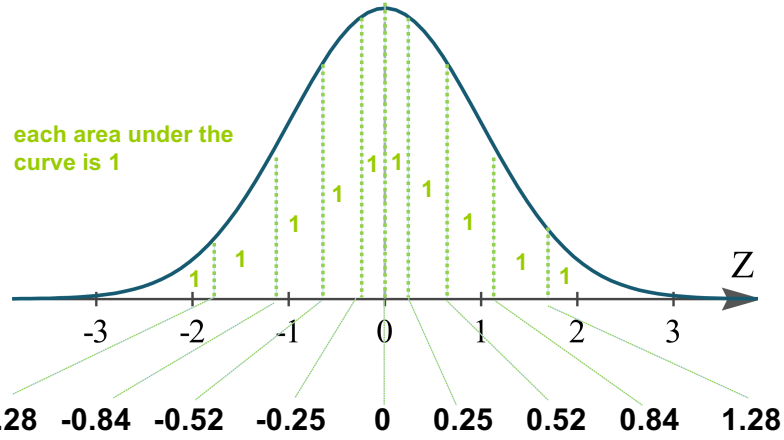
3. search for nine values on the normal distribution = split it in 10 areas

4. find the values that makes that happen

and here our **nine** sample values

3.89 4.75 4.75 5.20 5.78 5.80 6.33 7.21 7.90

here a standard normal distribution truncated at -3 and 3



and here our **nine** sample values

3.89 4.75 4.75 5.20 5.78 5.80 6.33 7.21 7.90

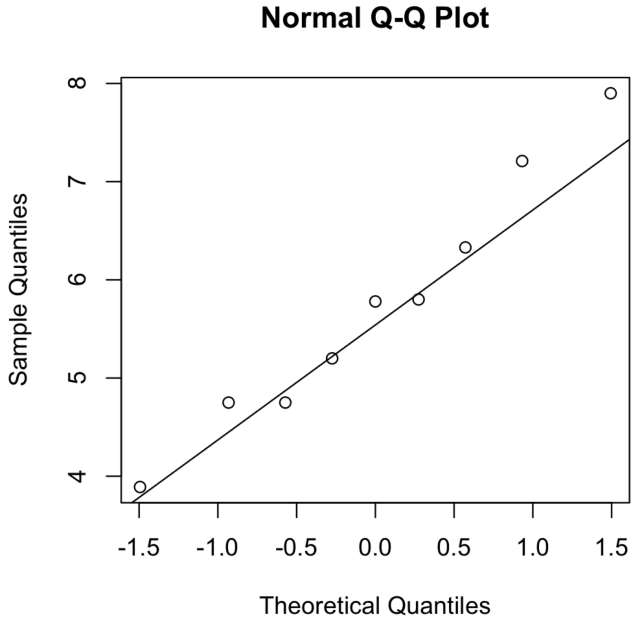
5. plot smallest value in our sample of size nine against what we expect to get as the smaller value in a sample of the same size from the standard normal distribution … we do that for all pairs

here what we would expect to get

-1.28 -0.84 -0.52 -0.25 0 0.25 0.52 0.84 1.28

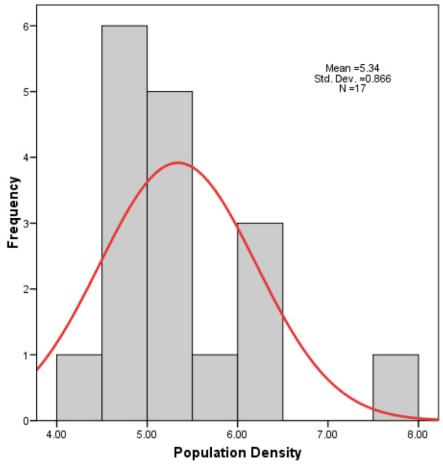
and here our **nine** sample values

3.89 4.75 4.75 5.20 5.78 5.80 6.33 7.21 7.90



```
data <- c(3.89, 4.75, 4.75, 5.20, 5.78, 5.80,
6.33, 7.21, 7.90)

# generate expected from normal distribution
u=seq(0.1,0.9,by=0.1)
expected <- qnorm(u)
print(expected)

[1] -1.2815516 -0.8416212 -0.5244005 -
0.2533471  0.0000000  0.2533471  0.5244005
[8]  0.8416212  1.281551

plot(expected,data)

#or simpler:
qqnorm(data)
qqline(data)
```


Q-Q plots are great **graphical tools for large sample** but not **very useful for small sample size**, e.g. this is the histogram of the last example: data do not 'look' normal, but they are not statistically different than normal.



statistical tests for normality are more precise since actual probabilities are calculated

Kolmogorov-Smirnov

works best for data sets with n > 50

not sensitive to problems in the tails

Shapiro-Wilks

works best for data sets with n < 50

doesn't work well if several values are same

Kolmogorov-Smirnov test

$$D_n = \max_x |F_{\text{exp}}(x) - F_{\text{obs}}(x)|$$

cumulative distribution function observed

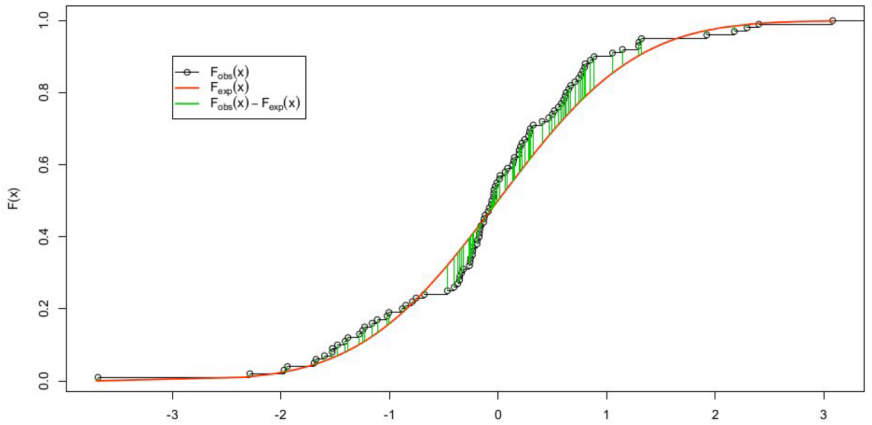
cumulative distribution function expected

can generate a **p-value**

```
 0.16 -0.68 -0.32 -0.85  0.89 -2.28  0.63  0.41  0.15  0.74
 1.30 -0.13  0.80 -0.75  0.28 -1.00  0.14 -1.38 -0.04 -0.25
-0.17  1.29  0.47 -1.23  0.21 -0.04  0.07 -0.08  0.32 -0.17
 0.13 -1.94  0.78  0.19 -0.12 -0.19  0.76 -1.48 -0.01  0.20
-1.97 -0.37  3.08 -0.40  0.80  0.01  1.32 -0.47  2.29 -0.26
-1.52 -0.06 -1.02  1.06  0.60  1.15  1.92 -0.06 -0.19  0.67
 0.29  0.58  0.02  2.18 -0.04 -0.13 -0.79 -1.28 -1.41 -0.23
 0.65 -0.26 -0.17 -1.53 -1.69 -1.60  0.09 -1.11  0.30  0.71
-0.88 -0.03  0.56 -3.68  2.40  0.62  0.52 -1.25  0.85 -0.09
-0.23 -1.16  0.22 -1.68  0.50 -0.35 -0.35 -0.33 -0.24  0.25
```

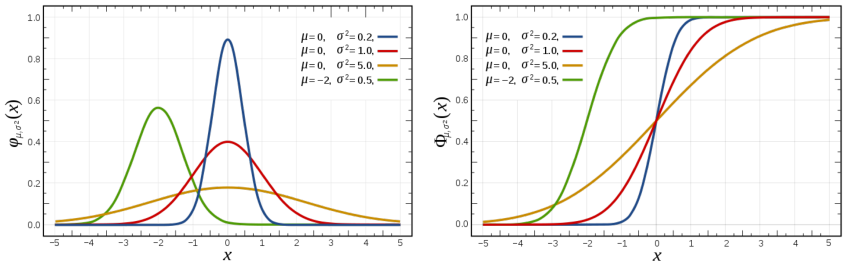
does the following sample of n=100 comes from a normality distributed population?

intuitively, we search for the maximum absolute distance between our data cumulative distribution function and the normal cumulative distribution function



so far we looked at **probability density function:** represents probability that the variate has the value x

another way to look at this is the **cumulative distribution function**: represents probability that the variable takes a value less than or equal to x




```
 0.16 -0.68 -0.32 -0.85  0.89 -2.28  0.63  0.41  0.15  0.74
 1.30 -0.13  0.80 -0.75  0.28 -1.00  0.14 -1.38 -0.04 -0.25
-0.17  1.29  0.47 -1.23  0.21 -0.04  0.07 -0.08  0.32 -0.17
 0.13 -1.94  0.78  0.19 -0.12 -0.19  0.76 -1.48 -0.01  0.20
-1.97 -0.37  3.08 -0.40  0.80  0.01  1.32 -0.47  2.29 -0.26
-1.52 -0.06 -1.02  1.06  0.60  1.15  1.92 -0.06 -0.19  0.67
 0.29  0.58  0.02  2.18 -0.04 -0.13 -0.79 -1.28 -1.41 -0.23
 0.65 -0.26 -0.17 -1.53 -1.69 -1.60  0.09 -1.11  0.30  0.71
-0.88 -0.03  0.56 -3.68  2.40  0.62  0.52 -1.25  0.85 -0.09
-0.23 -1.16  0.22 -1.68  0.50 -0.35 -0.35 -0.33 -0.24  0.25
```

does the following sample of n=100 comes from a normality distributed population?

1. order the data:

```
-3.68 -2.28 -1.97 -1.94 -1.69 -1.68 -1.60 -1.53 -1.52 -1.48
-1.41 -1.38 -1.28 -1.25 -1.23 -1.16 -1.11 -1.02 -1.00 -0.88
-0.85 -0.79 -0.75 -0.68 -0.47 -0.40 -0.37 -0.35 -0.35 -0.33
-0.32 -0.26 -0.26 -0.25 -0.24 -0.23 -0.23 -0.19 -0.19 -0.17
-0.17 -0.17 -0.13 -0.13 -0.12 -0.09 -0.08 -0.06 -0.06 -0.04
-0.04 -0.04 -0.03 -0.01  0.01  0.02  0.07  0.09  0.13  0.14
 0.15  0.16  0.19  0.20  0.21  0.22  0.25  0.28  0.29  0.30
 0.32  0.41  0.47  0.50  0.52  0.56  0.58  0.60  0.62  0.63
 0.65  0.67  0.71  0.74  0.76  0.78  0.80  0.80  0.85  0.89
 1.06  1.15  1.29  1.30  1.32  1.92 2.18   2.29  2.40  3.08
```

2. compute the empirical distribution function

$$F_{\text{obs}}(-3.68) = \frac{1}{100}, \quad F_{\text{obs}}(-2.28) = \frac{2}{100}, \ldots, \quad F_{\text{obs}}(3.08) = 1$$

F_{obs}

```
0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.10
0.11 0.12 0.13 0.14 0.15 0.16 0.17 0.18 0.19 0.20
0.21 0.22 0.23 0.24 0.25 0.26 0.27 0.28 0.29 0.30
0.31 0.32 0.33 0.34 0.35 0.36 0.37 0.38 0.39 0.40
0.41 0.42 0.43 0.44 0.45 0.46 0.47 0.48 0.49 0.50
0.51 0.52 0.53 0.54 0.55 0.56 0.57 0.58 0.59 0.60
0.61 0.62 0.63 0.64 0.65 0.66 0.67 0.68 0.69 0.70
0.71 0.72 0.73 0.74 0.75 0.76 0.77 0.78 0.79 0.80
0.81 0.82 0.83 0.84 0.85 0.86 0.87 0.88 0.89 0.90
0.91 0.92 0.93 0.94 0.95 0.96 0.97 0.98 0.99 1.00
```


3. for each observation xi from the data, compute:

$$F_{\text{exp}}(x_i) = P(Z \leq x_i)$$

(in this case, the expected distribution function is standard normal so use the normal table)

F_{exp}

```
0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.10
0.11 0.12 0.13 0.14 0.15 0.16 0.17 0.18 0.19 0.20
0.21 0.22 0.23 0.24 0.25 0.26 0.27 0.28 0.29 0.30
0.31 0.32 0.33 0.34 0.35 0.36 0.37 0.38 0.39 0.40
0.41 0.42 0.43 0.44 0.45 0.46 0.47 0.48 0.49 0.50
0.51 0.52 0.53 0.54 0.55 0.56 0.57 0.58 0.59 0.60
0.61 0.62 0.63 0.64 0.65 0.66 0.67 0.68 0.69 0.70
0.71 0.72 0.73 0.74 0.75 0.76 0.77 0.78 0.79 0.80
0.81 0.82 0.83 0.84 0.85 0.86 0.87 0.88 0.89 0.90
0.91 0.92 0.93 0.94 0.95 0.96 0.97 0.98 0.99 1.00
```


now we have two tables Fobs and Fexp …

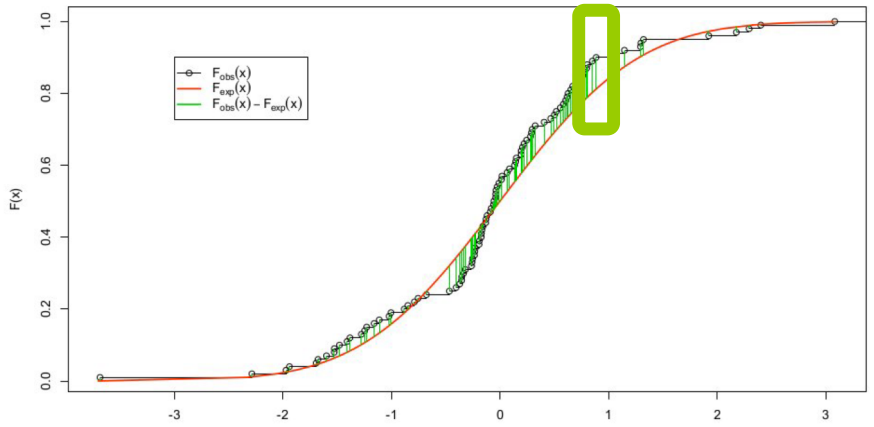
4. lets compute the absolute difference between the two and find the highest value

```
0.01  0.02  0.01  0.01  0.02  0.02  0.02  0.03  0.04  0.04
0.04  0.05  0.04  0.04  0.05  0.06  0.07  0.07  0.08  0.08
0.09  0.09  0.09  0.06 -0.04 -0.05 -0.05 -0.04 -0.03 -0.04
-0.03 -0.04 -0.03 -0.02 -0.01  0.00  0.01  0.01  0.02  0.03
0.04  0.05  0.03  0.04  0.05  0.06  0.06  0.06  0.07  0.08
0.09  0.10  0.11  0.12  0.11  0.12  0.12  0.13  [illegible]  0.11
0.12  0.12  0.13  0.14  0.15  0.16  0.17  0.18  0.19  [illegible]
0.18  0.12  0.11  0.10  0.11  0.10  0.08  0.09  [illegible]  0.09
0.09  0.09  0.10  0.10  0.10  0.11  0.11  0.12  0.13  0.11
0.06  0.06  0.04  0.05  0.06 -0.02 -0.03 -0.02 -0.01  0.00
```

$$D_n = \max_x |F_{\text{exp}}(x) - F_{\text{obs}}(x)|$$

this is the D searched

we have calculated the maximum absolute distance between expected and observed distribution functions



5. at 95% level the critical value is approximately given by

$$D_{\text{crit},0.05} = \frac{1.36}{\sqrt{n}}$$

we have a sample size of n = 100 so Dcrit = 0.136

and 0.19 > **0.136**

$$D_{\text{crit},0.05} = \frac{1.36}{\sqrt{n}}$$

there is a plethora of **tables / sampling distributions** that are established and are the basis of all statistic tests

n	α 0.01	α 0.05	α 0.1	α 0.15	α 0.2
1	0.995	0.975	0.950	0.925	0.900
2	0.929	0.842	0.776	0.726	0.684
3	0.828	0.708	0.642	0.597	0.565
4	0.733	0.624	0.564	0.525	0.494
5	0.669	0.565	0.510	0.474	0.446
6	0.618	0.521	0.470	0.436	0.410
7	0.577	0.486	0.438	0.405	0.381
8	0.543	0.457	0.411	0.381	0.358
9	0.514	0.432	0.388	0.360	0.339
10	0.490	0.410	0.368	0.342	0.322
11	0.468	0.391	0.352	0.326	0.307
12	0.450	0.375	0.338	0.313	0.295
13	0.433	0.361	0.325	0.302	0.284
14	0.418	0.349	0.314	0.292	0.274
15	0.404	0.338	0.304	0.283	0.266
16	0.392	0.328	0.295	0.274	0.258
17	0.381	0.318	0.286	0.266	0.250
18	0.371	0.309	0.278	0.259	0.244
19	0.363	0.301	0.272	0.253	0.237
20	0.356	0.294	0.264	0.246	0.231
25	0.320	0.270	0.240	0.220	0.210
30	0.290	0.240	0.220	0.200	0.190
35	0.270	0.230	0.210	0.190	0.180
40	0.250	0.210	0.190	0.180	0.170
45	0.240	0.200	0.180	0.170	0.160
50	0.230	0.190	0.170	0.160	0.150
OVER 50	$\frac{1.63}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.07}{\sqrt{n}}$

so 0.19 > **0.136** so null hypothesis rejected

H0: the samples come from a normal distribution

conclusion: the data do not come from a normal distribution

note KS is different than other tests we saw where we looked for a value below a critical level to reject the null, here it is the opposite (the larger the results the less likely is H0 so we reject it)

what if Dn < Dcrit?

here is a tricky bit … remember lecture on hypothesis testing, we cannot prove that two things are equal so we are going to **assume** that the normality is met

which is why we call this **assumption of normality**

more example on GitHub repository or at http://www.real-statistics.com/tests-normality-and-symmetry/statistical-tests-normality-symmetry/kolmogorov-smirnov-test/

```
#sorting a table
y <-c(0.16,-0.68,-0.32,-0.85,0.89,-2.28,0.63,0.41,0.15,0.74,1.30,-0.13,0.80,-
0.75,0.28,-1.00,0.14,-1.38,-0.04,-0.25,-0.17,1.29,0.47,-1.23,0.21,-0.04,0.07,-
0.08,0.32,-0.17,0.13,-1.94,0.78,0.19,-0.12,-0.19,0.76,-1.48,-0.01,0.20,-1.97,-
0.37,3.08,-0.40,0.80,0.01,1.32,-0.47,2.29,-0.26,-1.52,-0.06,-
1.02,1.06,0.60,1.15,1.92,-0.06,-0.19,0.67,0.29,0.58,0.02,2.18,-0.04,-0.13,-0.79,-
1.28,-1.41,-0.23,0.65,-0.26,-0.17,-1.53,-1.69,-1.60,0.09,-1.11,0.30,0.71,-0.88,-
0.03,0.56,-3.68,2.40,0.62,0.52,-1.25,0.85,-0.09,-0.23,-1.16,0.22,-1.68,0.50,-0.35,-
0.35,-0.33,-0.24,0.25)
ysorted <- sort(y)

X <- rnorm(100)
p = ecdf(X) #cumulative distribution function
X = sort(X) # trick to get fexp
fexp <- p(X)
fobs <- p(ysorted)
KS = max(abs(fexp-fobs))
```



```
#or easier
ks.test(X,y)

Two-sample Kolmogorov-Smirnov test


data:  X and y
D = 0.19, p-value = 0.05410262
alternative hypothesis: two-sided

#note that if you run the code you will
have different D (because of the random
rnorm generation) but likely that your
pvalue will always be above 0.05
```


Kolmogorov-Smirnov works well with **sample size > 50**
but when the sample is smaller Shapiro-Wilks works best

Shapiro-Wilks test

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

x(i) is the ith order statistic

SS (sum of squared difference)

$$(a_1, \cdots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m^T)^{1/2}}, \textit{where } m = (m_1, \cdots, m_n)^T$$

m1, ..., mn are the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution, and **V** is the covariance matrix of those order statistics.

can generate a **p-value**

a bit more beefy but let's go steps by steps …

3.83 3.16 4.70 3.97 2.03 2.87 3.65 5.09

does the following sample comes from a normality distributed population?

1. order the data:

2.03 2.87 3.16 3.65 3.83 3.97 4.70 5.09

2. divide them in two

2.03 2.87 3.16 3.65 | **3.83 3.97 4.70 5.09**

2.03 2.87 3.16 3.65 ← **3.83 3.97 4.70 5.09**

3. compute di the differences between both

3.06
1.83
0.81
0.18

4. multiply each of these by ai

good new we have shapiro-wilk table

n =	2	3	4	5	6	7	8	9	10	11	12	13	14
a1	0.7071	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.5888	0.5739	0.5601	0.5475	0.5359	0.5251
a2			0.1677	0.2413	0.2806	0.3031	0.3164	0.3244	0.3291	0.3315	0.3325	0.3325	0.3318
a3					0.0875	0.1401	0.1743	0.1976	0.2141	0.2260	0.2347	0.2412	0.2460
a4							0.0561	0.0947	0.1224	0.1429	0.1586	0.1707	0.1802
a5									0.0399	0.0695	0.0922	0.1099	0.1240
a6											0.0303	0.0539	0.0727
a7													0.0240

…

di		ai		
3.06	*	0.6052	=	1.851912
1.83	*	0.3164	=	0.579012
0.81	*	0.1743	=	0.141183
0.18	*	0.0561	=	0.010098

↓

total: **2.582205**

5. Divide the total by SS

$$W = \frac{(\sum_{i=1}^{[n/2]} a_i (x_{(n+1-i)} - x_{(i)}))^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{(2.582205)^2}{6.782549963} = 0.98307903$$

6. from the reference table of W (another table yeah!), Wcrit(n=8 at 0.05)=0.818

and 0.983>Wcrit

0.983>Wcrit, so we cannot reject null hypothesis, so we **assume the data follows a normal distribution**

otherwise (if <) we could rejected the null hypothesis and conclude with 95% confidence that that the data are not normally distributed

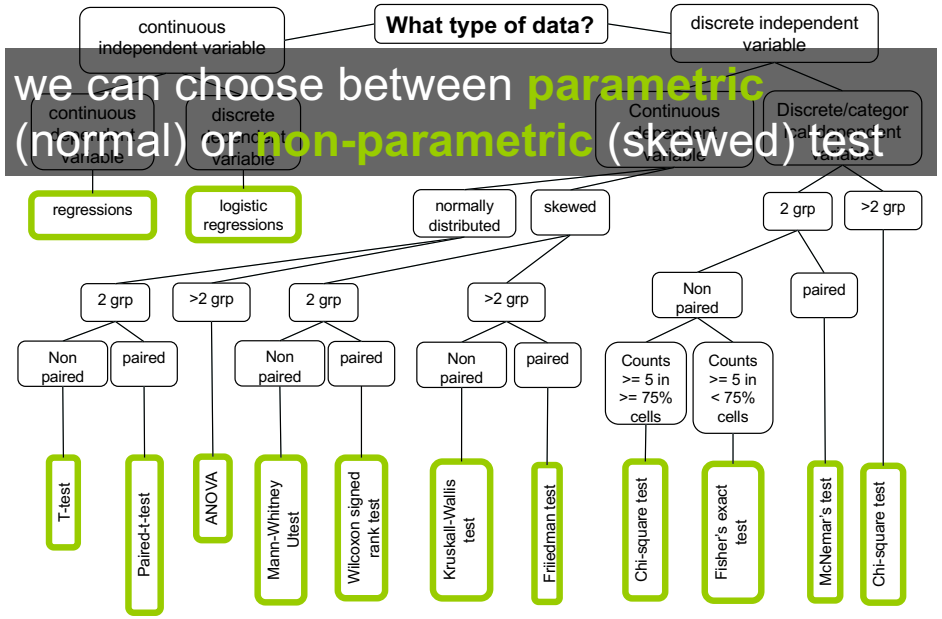
note we search for value below a critical level to reject the null, this is quite different from the results using the Kolmogorov-Smirnov test where this is the opposite

more example on GitHub repository or at http://www.real-statistics.com/tests-normality-and-symmetry/statistical-tests-normality-symmetry/shapiro-wilk-test/

```
y <-c(3.83, 3.16, 4.70, 3.97, 2.03, 2.87,
3.65, 5.09)
shapiro.test(y)
```

```
Shapiro-Wilk normality test

data:  y
W = 0.98317, p-value = 0.9769
```


ok so we know how to check our data, now what?

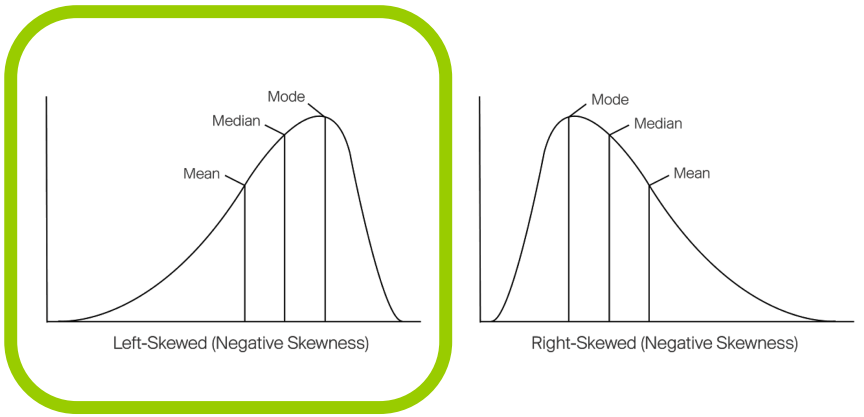


we can choose between **parametric** (normal) or **non-parametric** (skewed) test

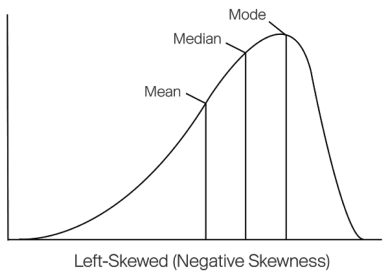
but if your data is not normally distributed you could also try to make it normal using **transformations**

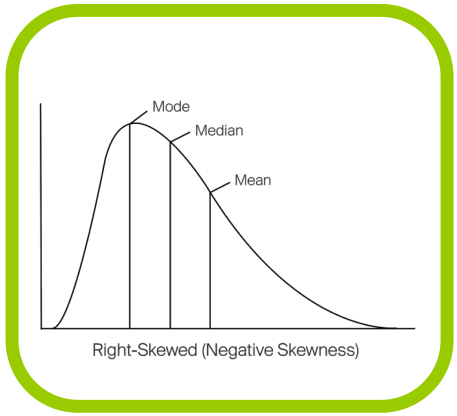
... more generally because parametric tests are more robust than non-parametric ones

transformations



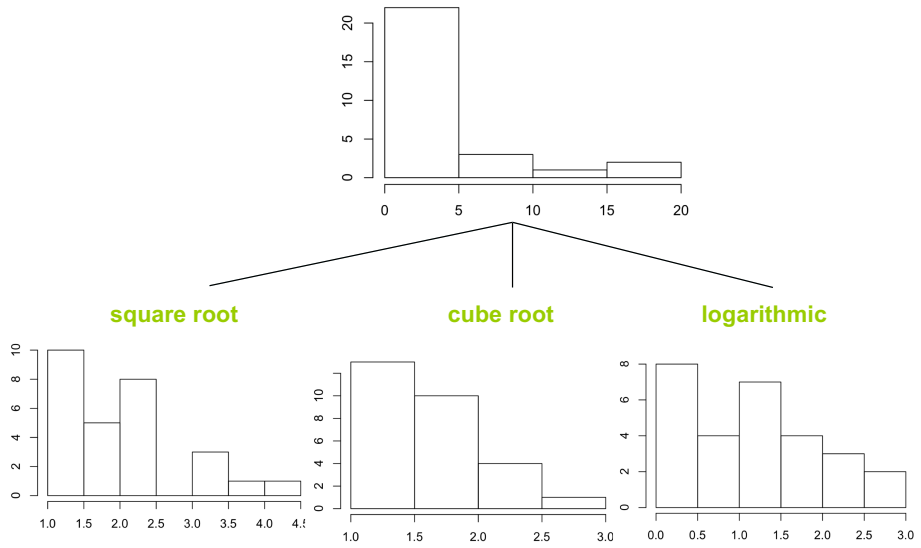
common transformations for left skewed::
square root, cube root, log





common transformations for right skewed::
square, cube root and logarithmic


```
y <-c(1.0, 1.2, 1.1, 1.1, 2.4, 2.2, 2.6,
4.1, 5.0, 10.0, 4.0, 4.1, 4.2, 4.1, 5.1,
4.5, 5.0, 15.2, 10.0, 20.0, 1.1, 1.1, 1.2,
1.6, 2.2, 3.0, 4.0, 10.5)
hist(y)
qqnorm(y)
qqline(y)

y_sqrt = sqrt(y) #cube root
y_cub = sign(y) * abs(y)^(1/3) #square root
y_log = log(y) #logarithm

# you can now try
qqnorm(y_log)
qqline(y_log)
```



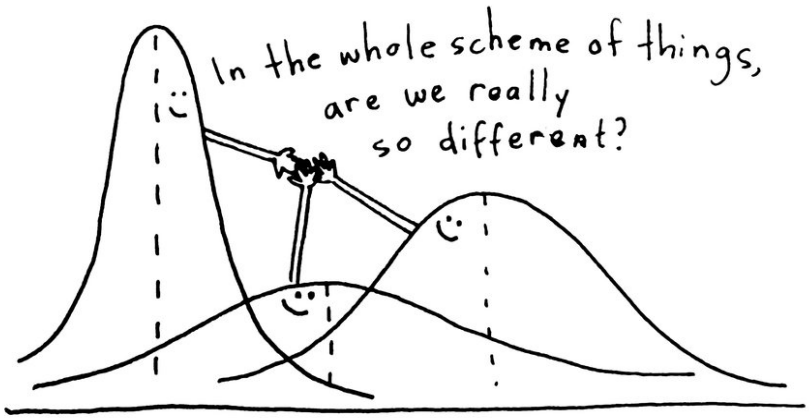
sometimes it does not work and rather than trying a complex transformations it id better to use non parametric tests

also skewed distributions could come from **outliers** so make sure to get rid of them!

other assumptions

assumption of normality is not the only thing we need to check, particularly when performing ANOVAs

also need to check the **assumption of homogeneity**



ANOVAs do not work well in this case

we can only do ANOVA if distribution are homogeneous and we can do this with the Levene's test


```
# first we run the one-way anova
dat = read.csv("HCIXP-anova.csv", header =
TRUE)
library(ez)
ezANOVA(dat,id,between=group,dv=score)

  Effect DFn DFd        F            p
p<.05       ges
1  group   2  57 154.8886 9.056612e-24     *
0.8445923
```

```
$`Levene's Test for Homogeneity of Variance
  DFn DFd      SSn  SSd        F         p
p<.05
1   2  57 1.433333 29.3 1.394198 0.2563608
```


the levene's test checks for **homogeneity of variances** (null hypothesis is that all variances are equal)

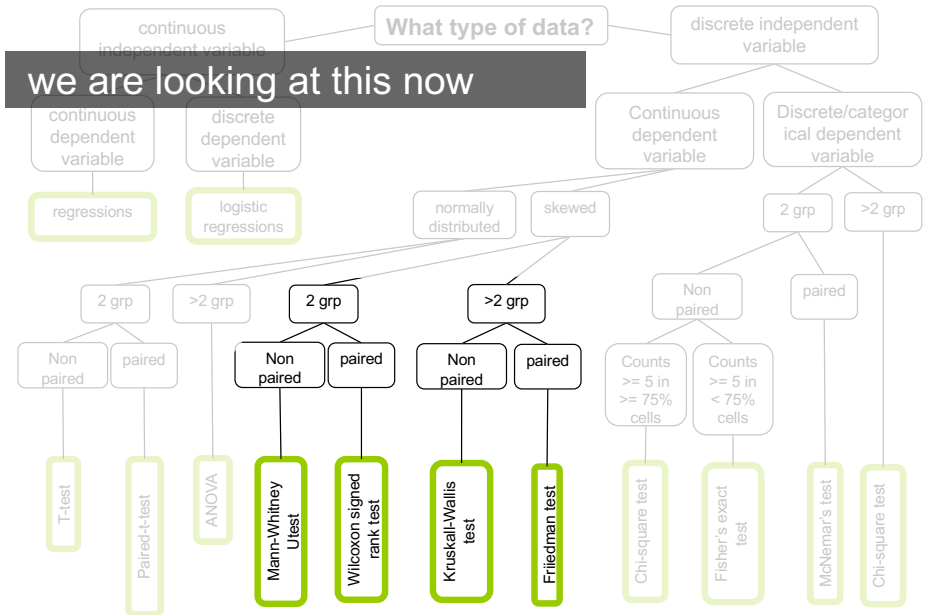
we won't go in detail with this test but the most important is this:

> if p-value < 0.05 means variances not equal and parametric tests such as ANOVA **are not suited** (need non-parametric tests)

1. Give the names of tests we can use to check normality and explain their differences and when to use them
2. I will not ask you to them by hand in the exam
3. Explain what to do if the data are not normal (either transforming the data or using non-parametric tests)
4. Explain what is the goal of a test of homogeneity of variance and what to do if the variances are not equal

take away

6 non-parametric tests



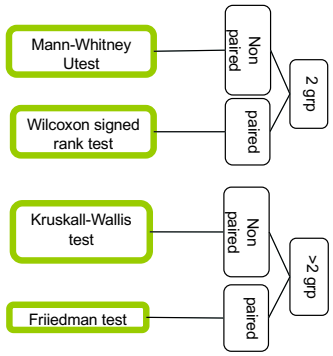
comparing the means of two populations is very important

in the last lecture we saw what we can do if we assume that the samples are normally distributed

for large sample sizes, we can invoke the **central limit theorem** to claim that data are approximately normal

but in some cases the data are **NOT normal**, and the sample size is too small to invoke the CLT

four non-parametric tests are very robust: the significance level is known regardless of the distribution of the data, but nothing is perfect: what **you gain in robustness you lose in power**.



Mann Whitney by hand
pdf in GitHub repository
http://www.real-statistics.com/non-parametric-tests/mann-whitney-test/

unpaired t-test equivalent

rank sum test (Mann Whitney)

received drug A	9	9.50	9.75	10	13	9.50
(different sets of participants for each)						
received drug B	11.50	12	9	11.50	13.25	13

1. rank the observations according to their size relative to the whole sample.

	9	9	9.50	9.50	9.75	10	11.50	11.50	12	13	13	13.25
rank	1	2	3	4	5	6	7	8	9	10	11	12
modified rank	1.5	1.5	3.5	3.5	5	6	7.5	7.5	9	10.5	10.5	12

(when ties – average the rank)

2. add up the ranks for the observations which came from smaller group. The sum of ranks in sample 2 is now determinate, since the sum of all the ranks equals N(N + 1)/2 where N is the total number of observations.

our statistic R is $$R_1 - \frac{n_1(n_1+1)}{2}$$

	9	9	9.50	9.50	9.75	10	11.50	11.50	12	13	13	13.25
modified rank	1.5	1.5	3.5	3.5	5	6	7.5	7.5	9	10.5	10.5	12

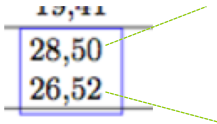
here we have the same sample size for each group so we can take any, e.g. **R (drug B) = 9**

3. we then look in the critical table

		larger sample size, n_2						
		4	5	6	7	8	9	10
smaller sample size n_1	4	12,24 11,25	13,27 12,28	14,30 12,32	15,33 13,35	16,36 14,38	17,39 15,41	18,42 16,44
	5		19,36 18,37	20,40 19,41	22,43 20,45	23,47 21,49	25,50 22,53	26,54 24,56
	6			28,50 26,52	30,54 28,56	32,58 29,61	33,63 31,65	35,67 33,69
	7				39,66 37,68	41,71 39,73	43,76 41,78	46,80 43,83
	8					52,84 49,87	54,90 51,93	57,95 54,98
	9						66,105 63,108	69,111 66,114
	10							83,127 79,131

rows and columns correspond to the sizes of the smaller and larger samples, respectively.

… why two values?



the top gives the 10% critical values = **one-tail test**

the bottom the 5% ones = **two-tail test**

R = 9 < 26.52 (let's say we do a two tails)

so we **reject the null hypothesis** and conclude that the two groups are significantly different

note that the critical value table only goes up to n =10

for larger samples we can use normal approximation

$$z = \frac{R - \mu}{\sigma},$$
$$\mu = \frac{1}{2} n_x (n_x + n_y + 1),$$
$$\sigma = \sqrt{\frac{n_x n_y (n_x + n_y + 1)}{12}}.$$

we then compare with the normal table, e.g. for two-tailed test at 0.05 reject null if $|z| > 1.96$


```
#wilcox.test do both paired (Mann whitney test)
and unpaired, so paired = TRUE would run the
Wilcoxon sign rank test, otherwise the Mann
Whitney (sometime called Wilcoxon sum rank test)

y1<- c(9,9.50, 9.75, 10,13, 9.50)
y2<- c(11.50,12,9,11.50,13.25, 13)
wilcox.test(y1,y2,paired=FALSE)
```

```
data:  y1 and y2
W = 9, p-value = 0.1705
alternative hypothesis: true location shift is not
equal to 0
```

Wilcoxon by hand
pdf in GitHub repository
http://www.real-statistics.com/non-parametric-tests/wilcoxon-signed-ranks-test/

paired t-test equivalent

signed rank test (Wilcoxon)

very quite similar but this time our data are paired (each participants made the two conditions so we have two data points per participants)

example: we measured the effect of two car seats on level of discomfort, here are the differences for 19 participants

-0.525, 0.172, -0.577, 0.200, 0.040, -0.143, 0.043, 0.010, 0.000, -0.522, 0.007, -0.122, -0.040, 0.000, -0.100, 0.050, -0.575, 0.031, -0.060

1. rank the observations **by absolute values** and removing the zeros

0.007	0.010	0.031	0.040	-0.040	0.043	0.050	-0.060	-0.100
1	2	3	4.5	4.5	6	7	8	9

-0.122	-0.143	0.172	0.200	-0.522	-0.525	-0.575	-0.577
10	11	12	13	14	15	16	17

2. we then compute R+ (sum of ranks for only positive differences) and R- (sum of ranks for negative differences)

R+ = 48.5
R- = 104.5

3. We take the min of the two (call this T)

T = 48.5

4. we then compare with appropriate table

n	P = 0.10	P = 0.05
5	2	-
6	2	0
7	3	2
8	5	3
9	8	5
10	10	8
11	14	10
12	17	13
13	21	17
14	26	21
15	30	25
16	36	29
17	41	34
18	47	40
19	53	46
20	60	52
21	67	58
22	75	65
23	83	73
24	91	81
25	100	89

we computed T = 48:5

since we dropped two values (zeros) our sample size is 19-2=17.

we found the critical value of 34 at the 5% level.

since 48.5 > Tcric of 34, we can't reject the null hypothesis, therefore **effect of these seats are not significantly different**

rather simple no?

Kruskal Wallis and Friedman, which are the non-parametric ANOVA equivalent, work on a very similar principles but for more groups depending if they are paired or not (within or between)

Kruskal Wallis by hand
pdf in GitHub repository
http://www.real-statistics.com/one-way-analysis-of-variance-anova/kruskal-wallis-test/

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{g} \frac{\bar{r}_{i\cdot}^2}{n_i} - 3(N+1)$$

ANOVA between subject equivalent

Kruskal Wallis

Friedman by hand
pdf in GitHub repository
http://www.real-statistics.com/anova-repeated-measures/friedman-test/

$$Q = \frac{12n}{k(k+1)} \sum_{j=1}^{k} \left(\bar{r}_{\cdot j} - \frac{k+1}{2} \right)^2$$

ANOVA within subject (also called repeated measure ANOVA) equivalent

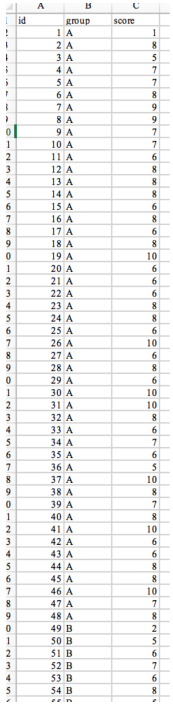
Friedman

practically

one dataset we know well: **our experiment on reward vs. punishment**

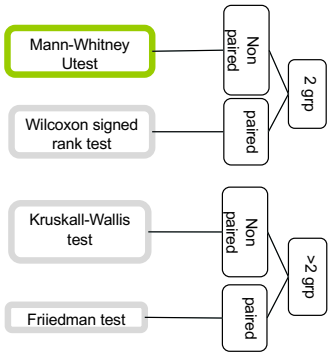
remember we assume the data was normal but it was not

so now we will finally be able to conclude!



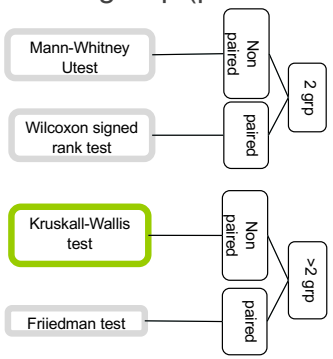
id	group	score
1	A	1
2	A	8
3	A	5
4	A	7
5	A	7
6	A	8
7	A	9
8	A	9
9	A	7
10	A	7
11	A	6
12	A	8
13	A	8
14	A	8
15	A	6
16	A	8
17	A	6
18	A	8
19	A	10
20	A	6
21	A	6
22	A	6
23	A	8
24	A	8
25	A	6
26	A	10
27	A	6
28	A	8
29	A	6
30	A	10
31	A	10
32	A	8
33	A	6
34	A	7
35	A	6
36	A	5
37	A	10
38	A	8
39	A	7
40	A	8
41	A	10
42	A	6
43	A	6
44	A	8
45	A	8
46	A	10
47	A	7
48	A	8
49	B	2
50	B	5
51	B	6
52	B	7
53	B	6
54	B	8

here is our data (chocolate vs. baseline)




```
#wilcox.test do both paired (Mann whitney test)
and unpaired

dat = read.csv("HCI2018results.csv", header =
TRUE)


wilcox.test(dat$score[dat$group == "A"],
dat$score[dat$group =="B"],paired=FALSE)

Wilcoxon rank sum test with continuity correction


data:  dat$score[dat$group == "A"] and
dat$score[dat$group == "B"]
W = 1290, p-value = 0.6408
alternative hypothesis: true location shift is not
equal to 0
```


now let's add the hypothetical group (punishment)



```
dat = read.csv("HCI2018results.csv", header =
TRUE)
kruskal.test(score ~ group, data = dat)
```

```
data:  score by group
Kruskal-Wallis chi-squared = 44.77,
df = 2, p-value = 1.898e-10
```

```
pairwise.wilcox.test(dat$score, dat$group,
p.adjust.method = "bonferroni")
```

```
  A       B
B 1       -
C 1.6e-09 2.6e-09
```


here turns out we get the same tendencies than with parametric tests, i.e. there is no evidences of significant effect of chocolate reward on memorization

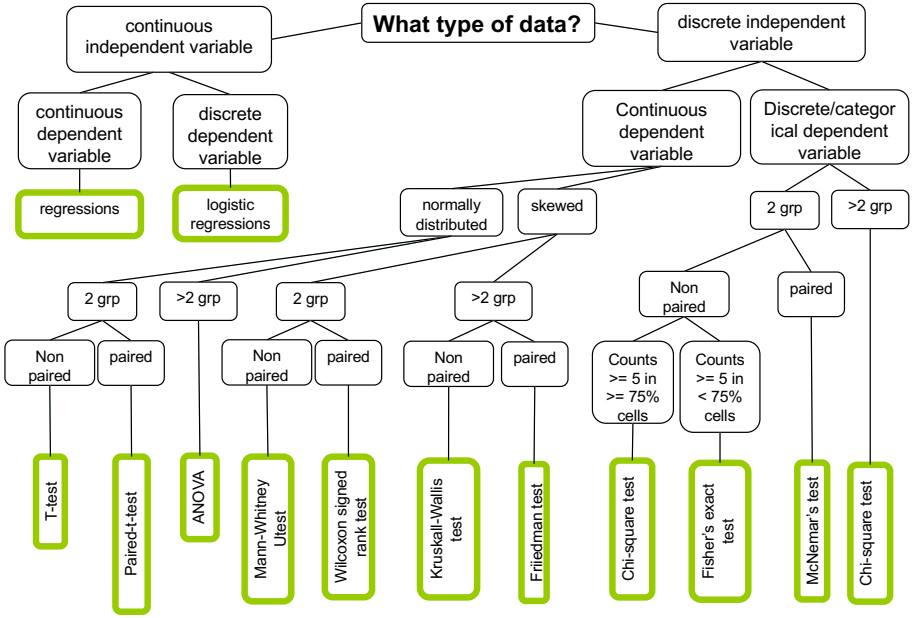
but there is an effect of punishment

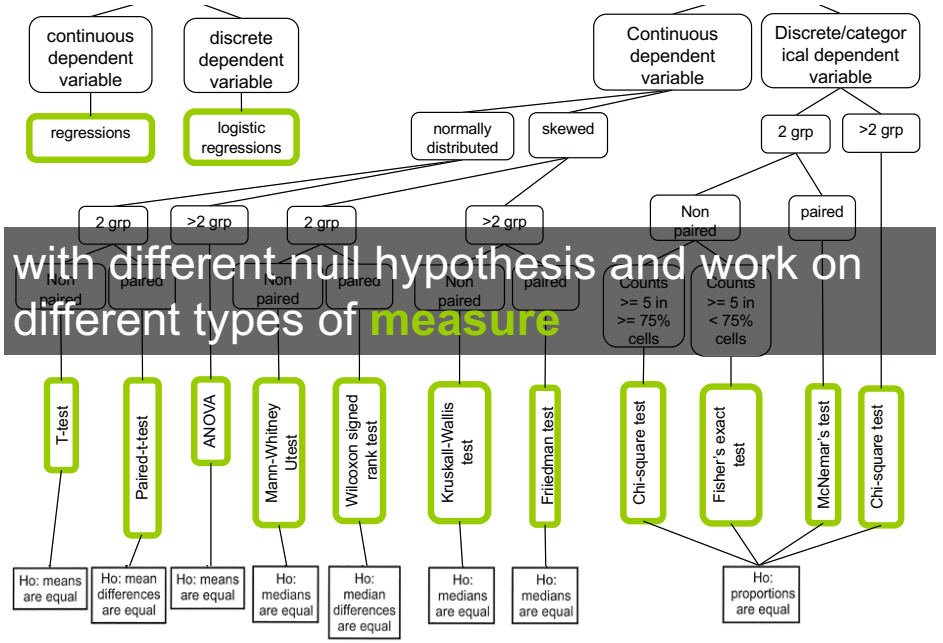

```
#for friedman test (source in GitHub)
dat = read.csv("friedmanExample.csv", header =
TRUE)
friedman.test(dat$count, dat$year, dat$month)
```

```
data:  dat$count, dat$year and dat$month
Friedman chi-squared = 7.6, df = 2, p-value =
0.02237
```

```
# note there is a real drop in statistical power
when using a Friedman test. There are methods that
enable post-hoc tests but the power is such that
obtaining significance is well nigh impossible.
The best you can do is to present a boxplot of the
data (dependent ~ group).
```


ok so now you know almost all the tests needed!





with different null hypothesis and work on different types of **measure**

the most important is not that you do these by hands but that you understand the intuition behind them and more importantly **when to use them**

quiz

20 participants were asked to write text using two different keyboard layouts (A and B). Half of the participants started the task on the A layout and then the B and the other half of the participants started the task on the B layout and then the A. The number of words typed per minute was collected for each participant and layout. Choose the most appropriate procedure to decide which layout allow participants to type the fastest. Assumption normality and homogeneity are verified.

Paired T-test
Unpaired T-test
One-Way Anova (between)
Repeated Anova (within)

Mann Whitney
Wilcoxon
Kruskal Wallis
Friedman

Linear regression
Kolmogorov-Smirnov
Shapiro-Wilk

20 participants were asked to write text using two different keyboard layouts (A and B). Half of the participants started the task on the A layout and then the B and the other half of the participants started the task on the B layout and then the A. The number of words typed per minute was collected for each participant and layout. Choose the most appropriate procedure to decide which layout allow participants to type the fastest. Assumption normality and homogeneity are verified.

Paired T-test
Unpaired T-test
One-Way Anova (between)
Repeated Anova (within)

Mann Whitney
Wilcoxon
Kruskal Wallis
Friedman

Linear regression
Kolmogorov-Smirnov
Shapiro-Wilk

40 participants were randomized to two groups. One group received a drug to decrease hair loss and the other group received a placebo (a pill of sugar). At the end of the program, the percentage hair loss for each patient was recorded. Choose the most appropriate procedure to decide if there is a relationship between the use of the drug and the percentage of hair loss. Assumption normality and homogeneity are verified.

Paired T-test
Unpaired T-test
One-Way Anova (between)
Repeated Anova (within)

Mann Whitney
Wilcoxon
Kruskal Wallis
Friedman

Linear regression
Kolmogorov-Smirnov
Shapiro-Wilk

40 participants were randomized to two groups. One group received a drug to decrease hair loss and the other group received a placebo (a pill of sugar). At the end of the program, the percentage hair loss for each patient was recorded. Choose the most appropriate procedure to decide if there is a relationship between the use of the drug and the percentage of hair loss. Assumption normality and homogeneity are verified.

Paired T-test
Unpaired T-test
One-Way Anova (between)
Repeated Anova (within)

Mann Whitney
Wilcoxon
Kruskal Wallis
Friedman

Linear regression
Kolmogorov-Smirnov
Shapiro-Wilk

A study attempted to find out if the age of an animal had any relationship to their athletic ability. The researchers took the data of 104 cheetahs, calculating their age and running a test to measure their speed. Choose the most appropriate procedure to decide if the age has any relationship with the run speed.

Paired T-test
Unpaired T-test
One-Way Anova (between)
Repeated Anova (within)

Mann Whitney
Wilcoxon
Kruskal Wallis
Friedman

Linear regression
Kolmogorov-Smirnov
Shapiro-Wilk

A study attempted to find out if the age of an animal had any relationship to their athletic ability. The researchers took the data of 104 cheetahs, calculating their age and running a test to measure their speed. Choose the most appropriate procedure to decide if the age has any relationship with the run speed.

Paired T-test
Unpaired T-test
One-Way Anova (between)
Repeated Anova (within)

Mann Whitney
Wilcoxon
Kruskal Wallis
Friedman

Linear regression
Kolmogorov-Smirnov
Shapiro-Wilk

20 participants were asked to type of their phone touchscreen in four different postures (sitting, lying down, standing and running). The number of words typed per minute was collected for each participant and postures. Choose the most appropriate procedure to decide which posture allow participants to type the fastest. Assumption normality and homogeneity are verified.

Paired T-test
Unpaired T-test
One-Way Anova (between)
Repeated Anova (within)

Mann Whitney
Wilcoxon
Kruskal Wallis
Friedman

Linear regression
Kolmogorov-Smirnov
Shapiro-Wilk

20 participants were asked to type of their phone touchscreen in four different postures (sitting, lying down, standing and running). The number of words typed per minute was collected for each participant and postures. Choose the most appropriate procedure to decide which posture allow participants to type the fastest. Assumption normality and homogeneity are verified.

Paired T-test
Unpaired T-test
One-Way Anova (between)
Repeated Anova (within)

Mann Whitney
Wilcoxon
Kruskal Wallis
Friedman

Linear regression
Kolmogorov-Smirnov
Shapiro-Wilk

20 participants were asked to run as fast as possible using two different pairs of shoes. Their speed was collected for each pairs of shoes. Choose the most appropriate procedure to decide which shoes allow participants to run the fastest. Assumption normality is verified but not the assumption of homogeneity.

Paired T-test
Unpaired T-test
One-Way Anova (between)
Repeated Anova (within)

Mann Whitney
Wilcoxon
Kruskal Wallis
Friedman

Linear regression
Kolmogorov-Smirnov
Shapiro-Wilk

20 participants were asked to run as fast as possible using two different pairs of shoes. They tested both pairs of shoes and each time their speed was collected. Choose the most appropriate procedure to decide which shoes allow participants to run the fastest. Assumption normality is verified but not the assumption of homogeneity.

Paired T-test
Unpaired T-test
One-Way Anova (between)
Repeated Anova (within)

Mann Whitney
Wilcoxon
Kruskal Wallis
Friedman

Linear regression
Kolmogorov-Smirnov
Shapiro-Wilk

20 participants were asked to type of their phone touchscreen in four different postures (sitting, lying down, standing and running). They were asked to rate their comfort for each posture using a Likert Scale questionnaire. Choose the most appropriate procedure to decide which posture was most comfortable.

Paired T-test
Unpaired T-test
One-Way Anova (between)
Repeated Anova (within)

Mann Whitney
Wilcoxon
Kruskal Wallis
Friedman

Linear regression
Kolmogorov-Smirnov
Shapiro-Wilk

20 participants were asked to type of their phone touchscreen in four different postures (sitting, lying down, standing and running). They were asked to rate their comfort for each posture using a Likert Scale questionnaire. Choose the most appropriate procedure to decide which posture was most comfortable.

Likert scale – special case of ordinal data that can be considered as continuous variable!

Paired T-test
Unpaired T-test
One-Way Anova (between)
Repeated Anova (within)

Mann Whitney
Wilcoxon
Kruskal Wallis
Friedman

Linear regression
Kolmogorov-Smirnov
Shapiro-Wilk

A study has gathered 10000 observations of computer performances (speed) in three different room of varying temperature (15, 25 and 35 degrees Celsius). Choose the most appropriate procedure to decide if the data follows a normal distribution.

Paired T-test Unpaired T-test One-Way Anova (between) Repeated Anova (within)	Mann Whitney Wilcoxon Kruskal Wallis Friedman	Linear regression Kolmogorov-Smirnov Shapiro-Wilk

A study has gathered 10000 observations of computer performances (speed) in three different room of varying temperature (15, 25 and 35 degrees Celsius). Choose the most appropriate procedure to decide if the data follows a normal distribution.

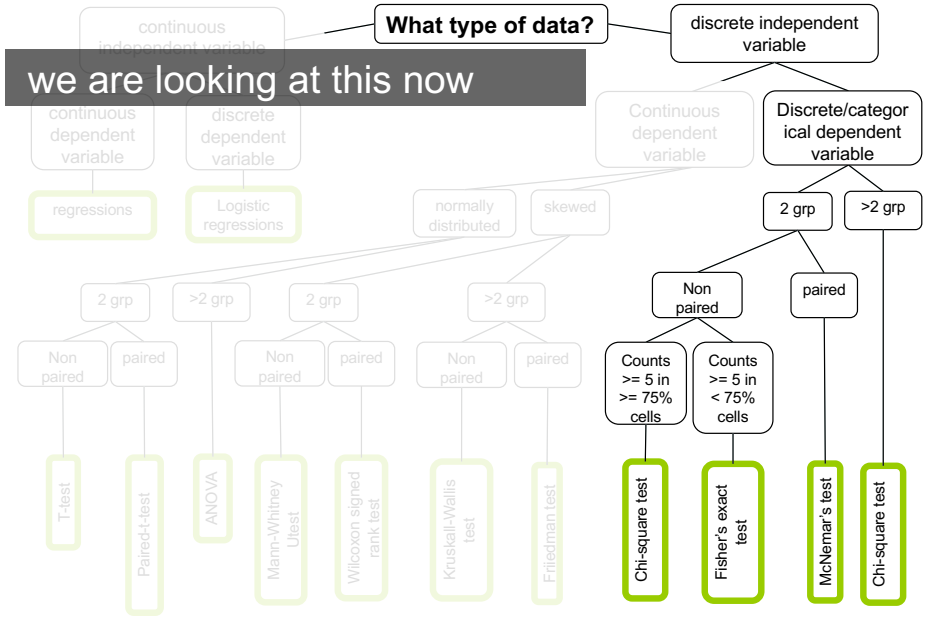
(because more than 50 observations!)

Paired T-test Unpaired T-test One-Way Anova (between) Repeated Anova (within)	Mann Whitney Wilcoxon Kruskal Wallis Friedman	Linear regression **Kolmogorov-Smirnov** Shapiro-Wilk

1. Give the four non parametric tests seen today
2. Explain the basis of Mann Whitney and Wilcoxon test, aka that they use ranks rather than mean
3. I won't ask you to do it by hand in the exam
4. Be able to read a design protocol and figure out what statistic tests to use

take away

7 others (categorical, anova++)



we are looking at this now

until now, we did statistical test using means or medians but the assumptions for means have eliminated certain types of variables (e.g. gender)

mean not appropriate measure of central tendency for nominal (categorical type) data

Chi-square can do do!

there are two types of Chi-square tests:

goodness of fit test (for one variable only)
contingency table test (for two variables at a time)

goodness of fit

looks to **see if a single variable fits some hypothesised probability distribution**

e.g. in a population of students, there would be an equal number of students who like or dislike brussels sprouts

in fact we don't even have to go 50/50, we may theorize that only 1/4 (25%) will like them (because they are disgusting!)

	# of persons
like BP	11
Dislike BP	139
	150 (total)

	# of persons	%expected	# expected
like BP	11	25%	37.5 (25% of 150)
Dislike BP	139	75%	112.5 (75% of 150)
	150 (total)	100%	150 (total)

observed cases

expected cases

$$X^2 = \sum \frac{(o-e)^2}{e}$$

$$= \frac{(11-37.5)^2}{37.5} + \frac{(139-112.5)^2}{112.5}$$

$$= 24.96$$

now, like with all the test we have seen, we look into a table, here the Chi-square table)

Critical values of the Chi-square distribution with *d* degrees of freedom

	Probability of exceeding the critical value						
d	0.05	0.01	0.001	*d*	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

degree of freedom DF = number of group - 1

24.96 > 3.841 so we **reject the null hypothesis**

our theory of 25% Brussel Sprout lovers does not hold

	# of persons	%expected	# expected
like BP	25	25%	25 (25% of 100)
Dislike BP	75	75%	75 (75% of 100)
	100 (total)	100%	100 (total)

observed cases

expected cases

$$X^2 = \sum \frac{(o-e)^2}{e}$$

$$= \frac{(25-25)^2}{25} + \frac{(75-75)^2}{75} = 0$$

if data was perfect fit (pvalue would be = 1)
… cannot reject null (thus cannot conclude)

let's see if our theory holds with a raise of hand

who like Brussel sprout?

who dislike Brussel sprout?


```
table = c(11,139)

       male female Sum
sport    26      3  29
family   24     22  46
Sum      50     25  75

chisq.test(tulip, p = c(1/4, 3/4))

Chi-squared test for given probabilities


data:  tulip
X-squared = 24.969, df = 1, p-value = 5.826e-07
```

this example is fairly simple but Chi-square also work with more data, e.g. 30% prefer eating chicken for Christmas dinner, 50% prefer turkey, 10% prefer vegetarian option, 10% prefer other types of meat

… problem sheet 5 will be about that

contingency tables

public opinion surveys tend to show there is a relationship between gender and *something*, e.g. preference in sport car vs. family car (public opinion surveys are very stereotypical!)

so here we have **two variables/groups**: gender (female or male) and car preference (sport or family)

	male	female
sport	26	3
family	24	22

we do a **Chi-square contingency table test** to prove preference of car is related to or dependant upon gender

but we don't have "expected value" here so we first need to calculate them for each cell

$$E_{ij} = \frac{R_i C_j}{N}$$

where R = row, C= column, N = total,
for ith row and jth column

	male	female	Sum
sport	26	3	29
family	24	22	46
Sum	50	25	75

1. we compute the sums in all direction

2. for each cell, multiplying that cells row and column totals and dividing by our total sample size

e.g. case (sport, male)= (29 * 50) / 75
e.g. case (family, male= (46 * 50) / 75
…

Obs.	male	female
sport	26	3
family	24	22

Exp.	male	female
sport	19.3	0.9
family	30.6	15.3

4. use same Chi-square formula than before

5. Calculate degree of freedom as DF =(number of rows−1)∗(number of columns−1) (here = 1)

6. Use the Chi-square table to conclude!

```
table = matrix(c(26, 24, 3, 22), ncol=2)
colnames(table) = c('male', 'female')
rownames(table) = c('sport','family')
addmargins(table)
```

```
       male female Sum
sport    26      3  29
family   24     22  46
Sum      50     25  75
```

```
chisq.test(table,correct=FALSE) #must use correct=FALSE
for a 2 by 2 table otherwise = TRUE
```

```
Pearson's Chi-squared test

data:  table
X-squared = 11.244, df = 1, p-value = 0.0007986
```

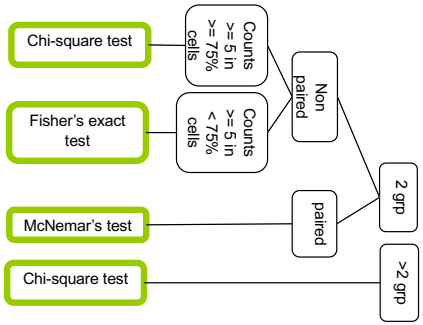

we reject the null and conclude that car type preference is dependant of gender

Chi-square contingency test::

if you have two categorical variables, and you'd like to determine whether the variables are independent (sometimes called a test of independence)

H0: the 2 categorical variables are independent (no relationship between the variables)

in R




```
> library(MASS)       # load the MASS package
> tbl = table(survey$Smoke, survey$Exer)
> tbl                 # the contingency table

        Freq None Some
  Heavy    7    1    3
  Never   87   18   84
  Occas   12    3    4
  Regul    9    1    7

>chisq.test(tbl) # or fisher.test(tbl) if Counts >= 5 in <
75% cells


        Pearson's Chi-squared test

data:  table(survey$Smoke, survey$Exer)
X-squared = 5.4885, df = 6, p-value = 0.4828
```

```
# mcnemar example on presidential Approval Ratings:
Approval of the President's performance in office in two
surveys, one month apart, for a random sample of 1600
voting-age Americans.

Performance <- matrix(c(794, 86, 150, 570), nrow = 2,
dimnames = list("1st Survey" = c("Approve", "Disapprove"),
"2nd Survey" = c("Approve", "Disapprove")))
Performance

            2nd Survey
1st Survey   Approve Disapprove
  Approve        794        150
  Disapprove      86        570


mcnemar.test(Performance)
```


that was easy and we are done with Chi-square!

now I would like to briefly come back to ANOVA for a bit
(oh not again!)

statistic tests on multiple variables

until now, we did statistical test on one independent variable (with multiple conditions or groups) and one dependant variable

e.g. effect of **chocolate, baseline, punishment (IV)** on **memorization score (DV)**

now it is possible to do tests for multiple IVs and multiple DVs (e.g. with CHI-square contingency table we look at two IVs).

however doing so decrease the power of your experiment (because you run more tests)

so it only works with powerful tests based on ANOVA (i.e. continuous variable and assumption of normality and homogeneity assumed)

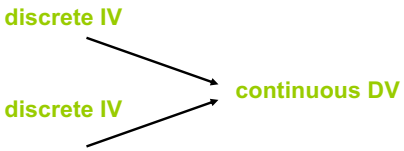
e.g. **two-ways ANOVA, MANOVA, ANVOVA**

discrete IV → continuous DV

one-way ANOVA::

compare the effect of **one discrete independent variables**, having 2 or more levels on **one dependant variable**

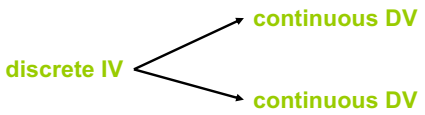
e.g. effect of alcohol consumption (none, 2-pints, 4-pints) on attractiveness ratings



two-way ANOVA::

compare the effect of **two discrete independent variables**, each of them having 2 or more levels on **one dependant variable**

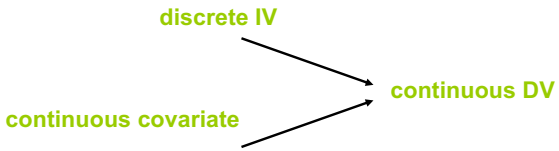
e.g. effect of gender (female, male) and alcohol consumption (none, 2-pints, 4-pints) on attractiveness ratings



one-way MANOVA::

(multivariate analysis of variance) compare the effect of **one independent variable**, having 2 or more levels on **two dependant variables**

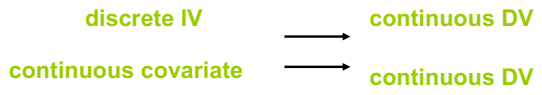
e.g. effect of different memorization enhancing drugs (placebo, drug A, drug B) on memorization skills and emotional ratings (to find the sweet spot for a drug that enhance skills without depressing people!)



one-way ANCOVA::

(analysis of **co**variance) compare the effect of **one independent variable**, having 2 or more levels and **one continuous covariate** on **one dependant variable**

e.g. effect of phone sizes (iphone 4, iphone 5, iphone 6, iphone 7) on the amplitude of phone movements made when texting **given the measure of the participants hand width (covariate)**



one-way MANCOVA::

(multivariate analysis of covariance) compare the effect of **one independent variable**, having 2 or more levels and **one continuous covariate** on **two dependant variable**

… you can even two a two-way MANCOVA (but your experience might not have much statistical power because there are too many tests to perform)

although these tests exist, my advice would be to keep the experimental design as simple as possible as you can, analysis will be easier and more powerful

… we will look at power next time with a guest lecturer: Luluah Al-Barrack

two ways ANOVA practically

```
# two-ways anova in R (I added a gender column in our
chocolate vs. reward vs. punishment file)
library(ez)
dat = read.csv("HCI2018resultsTwoWays.csv", header = TRUE)
ezANOVA(dat,id,between=.(group,gender),dv=score)
```

```
        Effect DFn DFd       F            p p<.05        ges
1        group   2  54 72.7776 4.709005e-16     * 0.72939818
2       gender   1  54  1.4112 2.400561e-01       0.02546778
3 group:gender   2  54  0.8064 4.517654e-01       0.02900052
```

Like with one way with have the effect for each IV

As well as the **interaction** between both

```
# note here our two IV are between but we could have them
both within or a combination of within and between we
would write:
ezANOVA(dat,id,within=group,between=gender,dv=score)
```


we can write (only the significant one)

A two-way ANOVA showed a significant effect on IV1 (Fdf = F_value, p<0.05), on IV2 (Fdf = F_value, p<0.05) and on the interaction IV1 x IV2 (Fdf = F_value, p<0.05)

and then you can do your post-hoc comparison tests (although there are more to do!) and conclude

1. Be able to give the CHI-square formula (goodness of fit and contingency table)
2. Calculate a CHI-square by hand on an example with a single variable and conclude
3. Explain what is the different between goodness of fit and contingency table methods
4. Explain what is a two-way ANOVA, MANOVA and ANCOVA and be able to explain the differences between them

take away

8

(by Luluah Albarrak)

effect size and power effect

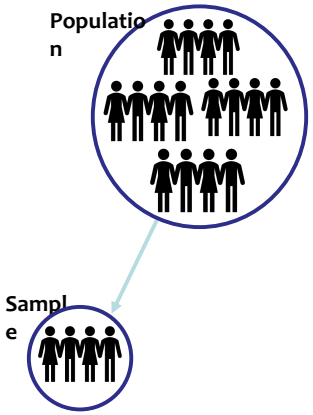
Estimation, Type I & II errors, power, effect size

Random Sample

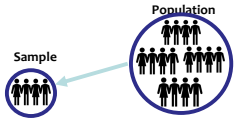
Population: is a set of all units of interest.

Sample: a subset of the population.

Random sample: a sample collected in such a way that every member of the population is equally likely to be selected.



Random Sample



The **goal** is to make estimates and predictions about a population based on information from a sample. In particular, we want to estimate the population mean μ , and the population standard deviation σ.

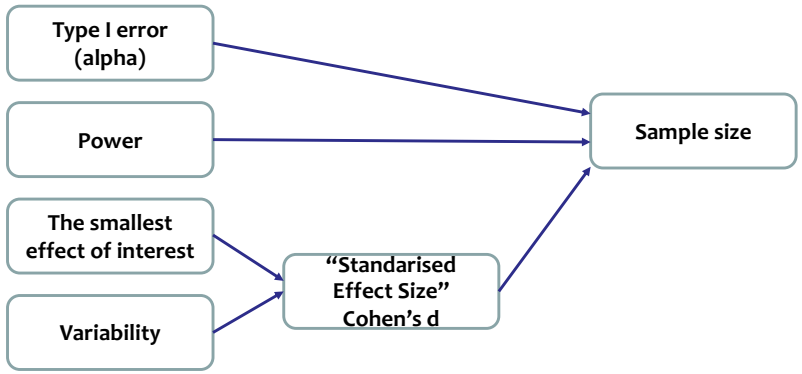
Sample Size Calculation

The main aim of a sample size calculation is to determine the number of participants needed to detect a scientifically relevant treatment effect.

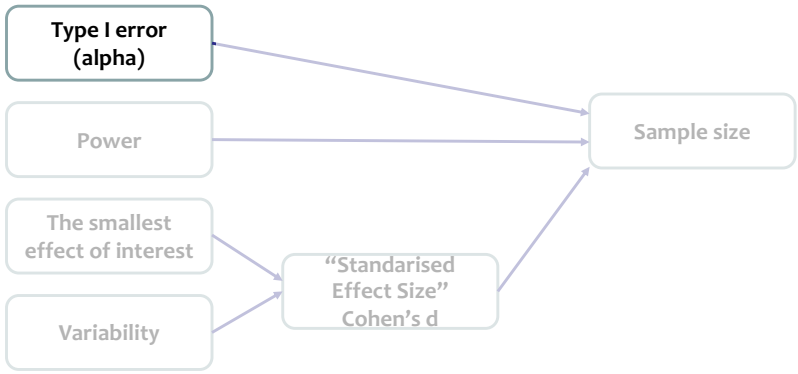
Sample size is too small → one may not be able to detect an important existing effect.

Sample size is too large → waste of time, resources and money.

Components of sample size calculations

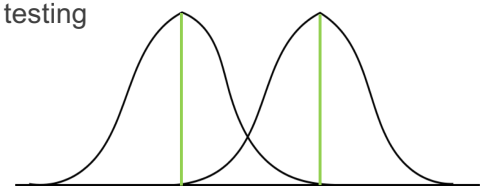


Components of sample size calculations



Type I & Type II Errors

Hypothesis testing



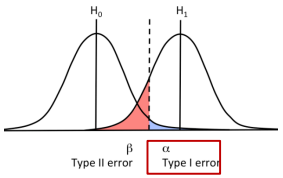
H_o: $\mu_1 - \mu_2 = 0$ H_A: $\mu_1 - \mu_2 \neq 0$

When we conduct a test of any hypothesis regardless of the test used we make one of two possible decisions:

Reject the null (H_o) in favor of the alternative (H_A) **OR** **Fail to reject the null hypothesis (H_o)**

Type I Errors

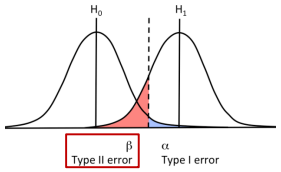
	Truth about the population/reality	
TEST DECISION	H_o TRUE	H_a TRUE
Reject the Null Hypothesis	TYPE I ERROR (α)	Power ($1-\beta$) CORRECT DECISION
Fail to Reject the Null Hypothesis	CORRECT DECISION	TYPE II ERROR (β)



Type I error: the chance of detecting a statistically significant difference when there is no real difference between treatments (The risk of a false positive result).

Type II Errors

	Truth about the population/reality	
TEST DECISION	H_o TRUE	H_a TRUE
Reject the Null Hypothesis	TYPE I ERROR (α)	Power ($1-\beta$) CORRECT DECISION
Fail to Reject the Null Hypothesis	CORRECT DECISION	TYPE II ERROR (β)



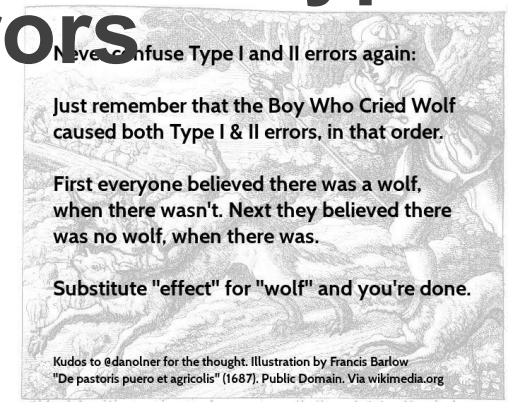
Type II error: the chance of not detecting a significant difference when there really is a difference (The risk of a false negative result).

Type I & Type II Errors

We choose P(Type I Error) = α
Typically α = .05 or .01 or .10

However we do NOT directly choose
β = P(Type II Error)

Type I & Type II Errors



Question

It has been shown many times that on a certain memory test, recognition is substantially better than recall. However, the probability value for the data from your sample was 0.12, so you were unable to reject the null hypothesis that recall and recognition produce the same results. What type of error did you make?

Type I or Type II Error?

Question

It has been shown many times that on a certain memory test, recognition is substantially better than recall. However, the probability value for the data from your sample was 0.12, so you were unable to reject the null hypothesis that recall and recognition produce the same results. What type of error did you make?

Type I or Type II Error?

Type II Error

Question

In the population, there is no difference between men and women on a certain test. However, you found a difference in your sample. The probability value for the data was 0.03, so you rejected the null hypothesis. What type of error did you make?

Type I or Type II Error?

Question

In the population, there is no difference between men and women on a certain test. However, you found a difference in your sample. The probability value for the data was 0.03, so you rejected the null hypothesis. What type of error did you make?

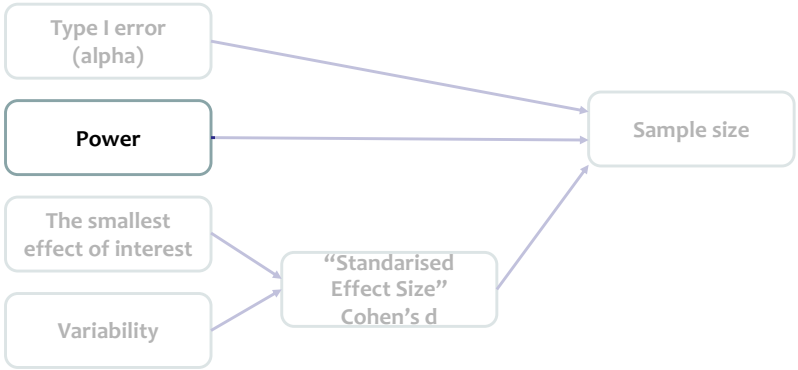
Type I or Type II Error?

Type I Error

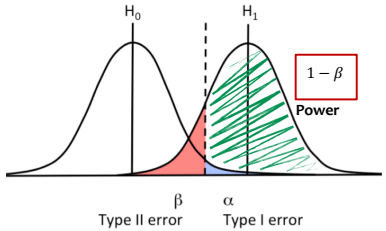
How p-value affects sample size?

As the p-value **decreases**, the necessary sample size **increases**.

Components of sample size calculations



Power



The **power** of a hypothesis test is the probability of making the correct decision if the alternative hypothesis is true. That is, the **power** of a hypothesis test is the probability of rejecting the null hypothesis H_0 when the alternative hypothesis H_A is true.

Power

Higher power is better (the closer the power is to 1.0 or 100%).

The ideal power is considered to be 80% → we are accepting that one in five times (20%) we will miss the difference.

Power

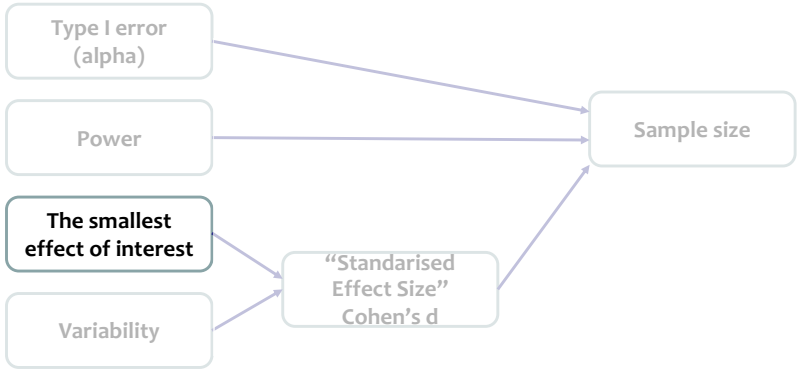
To increase power (and hence decrease type 2 error rate):

Increase the sample size	**Decrease the standard deviation of the sample**	**Increase α**	**Consider a larger effect size**

How power affects sample size?

As the power **increases,** the necessary sample size **increases**.

Components of sample size calculations



Smallest Effect of Interest

The smallest effect of interest is the minimal difference between the studied groups that the investigator wishes to detect.

Smallest Effect of Interest

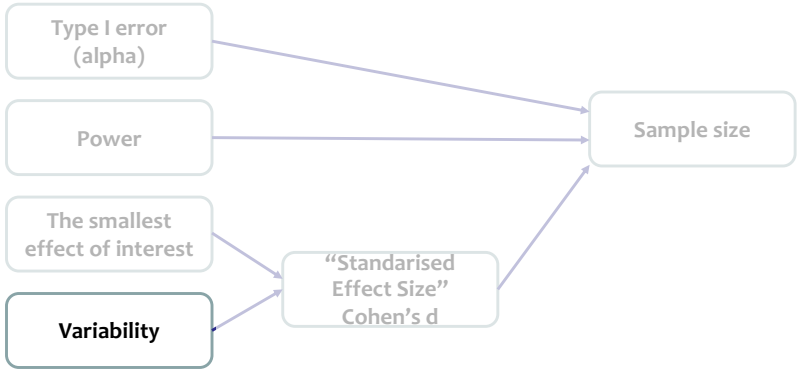
For **continuous** outcome variables, the minimal scientifically relevant difference is a **numerical** difference. For example, if body weight is the outcome of a trial, an investigator could choose a difference of 5 kg as the minimal scientifically relevant difference.

For **binary** outcome variables, the minimal difference is expressed in **rates**. For example, in the case of studying the effect of a drug on weight loss (yes/no), an investigator could choose a difference of 10% between the treatment group and control group as the minimal scientifically relevant difference .

How effect of interest affects sample size?

As the effects of interest between the study groups **increases**, the necessary sample size **decreases**.

Components of sample size calculations



Variability

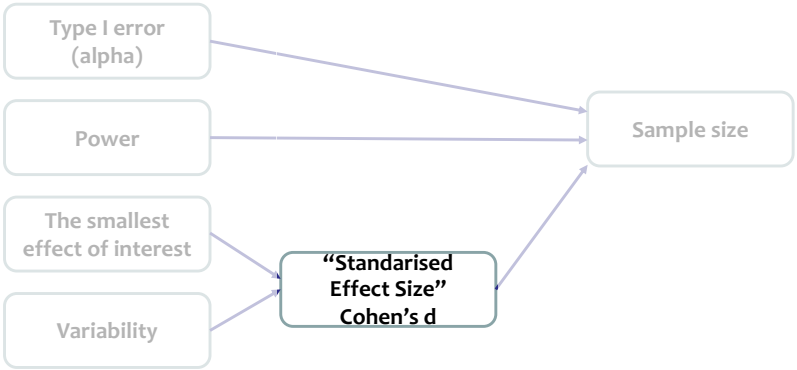
Sample size calculation is based on using the population variance of a given outcome variable that is estimated by means of the standard deviation (SD) in case of a continuous outcome.

Because the variance is usually an unknown quantity, investigators often use an estimate obtained from a pilot study or use information from a previous study.

How variance affects sample size?

As the variance **increases**, the necessary sample size **increases**.

Components of sample size calculations



Effect Size

The smallest effect of interest and the variability are combined and expressed as a multiple of the SD of the observations; known as the standardised difference.

The standardised difference is also referred to as the effect size.

$$\text{Effect Size} = \frac{\text{Difference between the means in the two treatment groups}}{\text{Standard Deviation}}$$

Effect Size

Effect size is a way of quantifying the difference between two or more groups, or a measure of the difference in the outcomes of the experimental and control groups.

For example, if one group has a new treatment and the other has not (control group), then the effect size is a measure of the effectiveness of the treatment.

Effect Size

A statistically significant result does not mean it is substantive in effect. For example, two treatments could be shown to be significantly different, but their clinical effects may be so small as to be unimportant.

Cohen's d Effect Size

Depending on the type of study, effect size is estimated with different measures.

The most straightforward effect size measure is the difference between two means.

Cohen's d is a good example of a standardized effect size measurement.

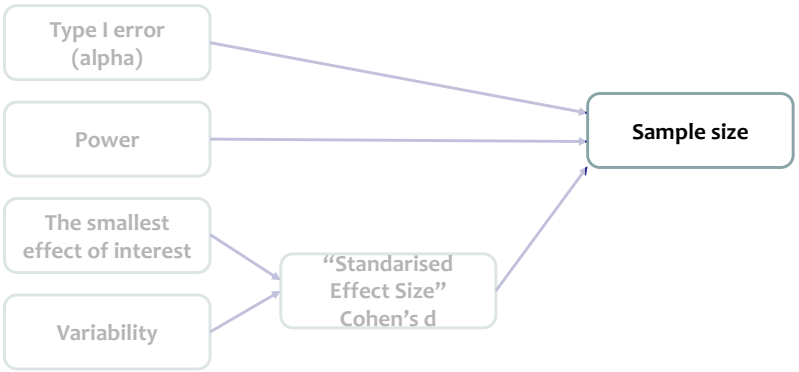
Cohen (1988) proposed a simple categorisation of small, moderate and large effect size.

$$d = \frac{\mu_1 - \mu_2}{SD_{pooled}}$$

Effect	Cohen's d
Small	0.20
Medium	0.50
Large	0.80

Sample Size Estimation

Components of sample size calculations



Sample Size for Continuous Data

There are several methods used to calculate the sample size depending on the type of data or study design.

The sample size for continuous data when comparing two means (independent) is calculated using the following formula:

$$n = \frac{2\left(Z_{\alpha} + Z_{1-\beta}\right)^2 \sigma^2}{\delta^2}$$

The number, n, is the sample size required in each group.

Example

In a sample of inactive overweight men aged between 50 and 60, suppose we wish to compare the mean blood pressure of Group 1 who underwent a calorie-controlled diet to Group 2 undertook the exercise-training programme.

Let,

μ_1=$mean$ blood pressure of group 1
μ_2=$mean$ blood pressure of group 2

Example

Then,

H_0: $\mu_1 = \mu_2$
H_a: $\mu_1 \neq \mu_2$

Suppose researchers would like to have a power of 80% to detect a difference of 5 mmHg between these two population means at the α=.05 level. The standard deviation (based on data in a published paper) would be approximately 20 mmHg.

What samples sizes (n_1) and (n_2) should they use?

Example

Then,

H_0: $\mu_1 = \mu_2$
H_a: $\mu_1 \neq \mu_2$

Suppose researchers would like to have a power of 80% to detect a difference of 5 mmHg between these two population means at the α=.05 level. The standard deviation (based on data in a published paper) would be approximately 20 mmHg.

What samples sizes (n_1) and (n_2) should they use?

Example

$\delta^2 = (\mu_1 - \mu_2)^2 = (5)^2 = 25$
β=.20 , $\alpha = .05$
SD= 20

α-error	5%	1%
2-sided	1.96	2.5758
1-sided	1.65	2.33

Power	80%	85%	90%	95%
Value	0.8416	1.0364	1.2816	1.6449

$n = \frac{2(Z_\alpha + Z_{1-\beta})^2 \sigma^2}{\delta^2} = \frac{2(1.96+0.84)^2 (20)^2}{25} =$ 250.88=251

Hence we would use equal samples sizes of 251 for n_1 and n_2.

Solving in R

Install pwr package

```
> pwr.t.test(d=.25, sig.level=.05, power=.80, type='two.sample')

     Two-sample t test power calculation

              n = 252.1275
              d = 0.25
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

R

Sample Size for Categorical Data

The sample size for categorical data when comparing two proportions is calculated using the following formula:

$$n = \left(Z_{\alpha} + Z_{1-\beta}\right)^2 \frac{p1(1-p1)+p2(1-p2)}{(p1-p2)^2}$$

The number, n, is the sample size required in each group.

Example

A new treatment has been developed for patients who've had a heart attack. It is known that 10% of people who've suffered from a heart attack die within one year. It is thought that a reduction in deaths from 10% to 5% would be clinically important to detect.

Let,

P_1 = proportion of deaths in placebo group = 0.1

P_2 = proportion of deaths in treatment group = 0.05.

Example

Then,

H_0: $P_1 = P_2$

H_a: $P_1 \neq P_2$

It is thought that a reduction in deaths from 10% to 5% would be clinically important to detect. Using α = 0.05 and β = 0.10, What samples sizes (n_1) and (n_2) should they use ?

Example

Then,

H_0: $P_1 = P_2$

H_a: $P_1 \neq P_2$

It is thought that a reduction in deaths from 10% to 5% would be clinically important to detect. Using α = 0.05 and β = 0.10, What samples sizes (n_1) and (n_2) should they use ?

Example

P1= 0.10

P2= 0.05

α-error	5%	1%
2-sided	1.96	2.5758
1-sided	1.65	2.33

$$n = \left(Z_\alpha + Z_{1-\beta}\right)^2 \frac{p1(1-p1)+p2(1-p2)}{(p1-p2)^2}$$

Power	80%	85%	90%	95%
Value	0.8416	1.0364	1.2816	1.6449

$$= (1.96 + 1.28)^2 \frac{0.1\,(0.9)+0.05(0.95)}{(0.1-0.05)^2} = 10.5\,\frac{.09+.048}{.0025} = 579.6 = 580$$

580 patients would be needed in each group to be 90% sure of being able to detect a reduction in mortality of 5% as significant at the 5% level.

Solving in R

```
>
>
> ES.h(.10,.05)  # calculate Cohen's h using arcsine transformation
[1] 0.1924743
> pwr.2p.test(h=.19,sig.level=.05, power=.90)

     Difference of proportion power calculation for binomial distribution (arcsine transformation)

              h = 0.19
              n = 582.1285
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: same sample sizes
```

R

Other Outcome Types

In many studies, the outcomes may not be continuous or categorical. In these cases, the details of calculation differ, but using the four aforementioned components, persist through calculations with other types of outcomes.

Summary

Definitions

α	Type I error :The risk of a false positive result. i.e. the chance of detecting a statistically significant difference when there is no real difference between treatments.
β	Type II error :The risk of a false negative result i.e. the chance of not detecting a significant difference when there really is a difference.
Power ($1-\beta$)	Power The chance of not getting a false negative result. i.e. the chance of spotting a difference as being statistically significant if there really is a difference.
The smallest effect of interest	The minimal difference between the groups that the investigator considers scientifically plausible and clinically relevant.
Variance	The variability of the outcome measure, expressed as the standard deviation.

Take Away

1. Explain the components of sample size calculations and how they affect the sample size
2. Explain the difference between Type I error, Type II error and power.
3. Explain effect size.
4. Calculate sample size for continuous and categorical data

9 P_hacking

ego depletion story

1998

ROY Baumeister

do people have a **limited amount** of **will power**?

1998

so he conducted a study …

the results show that humans **do have a limited pool of self-control**

once we have had to **resist temptation** it is a lot **harder to di it again**

1998

they called this

Ego Depletion::

has had huge influence on psychological research, been incorporated into dieting tactics, training techniques, and even adverting used today

ads telling how we deserve a product, causing mental fatigue and frustration, leading us to buy

but today it seems that **this phenomenon does not exist at all**

1998

ROY Baumeister's hypothesis:

self-control is **a limited resource**, and it takes **energy and motivation to maintain restraint**

every time use your self-control, your draw on that strength and it takes some time for you to recover it

1998

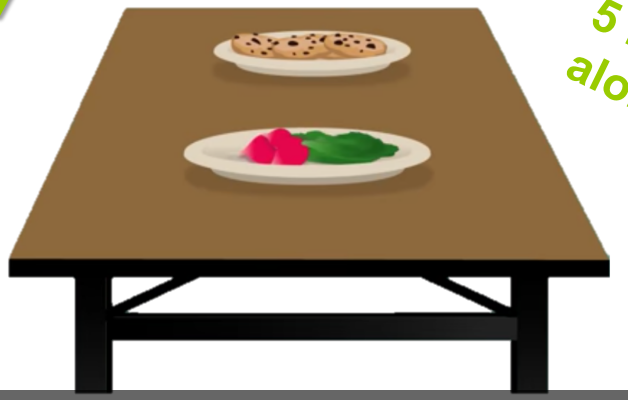
ROY Baumeister's task:

two act of self-control back-to-back

1998

cooked delicious cookies and invited 67 students with empty stomach to sit in the baking room

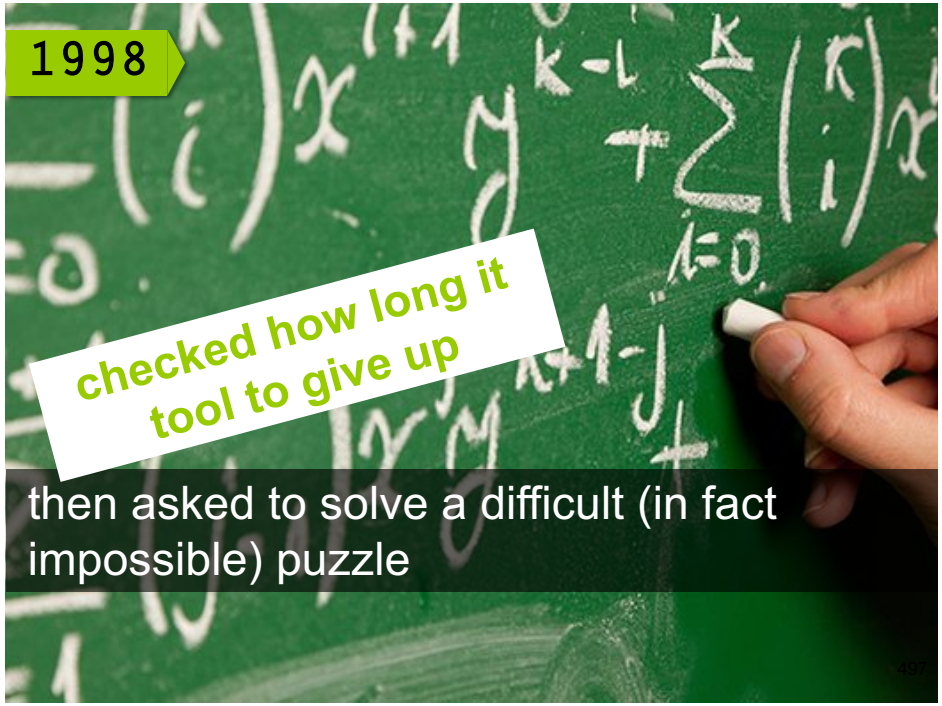
1998



5 min alone

half the students told to **eat the cookies**
half the students **the radishes**

1998



1998

would resisting the cookies make it harder for participants to keep trying?

those who ate **radishes** = **8 minutes** of trying
those who ate **cookies** = **19 minutes** of trying

<brainstorming: could you already see something wrong with their studies?>

1998

confounding variable::

eating cookies (high in sugar) could be the reason of the longer effort provided

1998

would resisting the cookies make it harder for participants to keep trying?

those who ate **radishes** = **8 minutes** of trying
those who ate **cookies** = **19 minutes** of trying
3rd group with no cookies encounter = 21 minutes

and of course all these comparison with $p<0.05$ = strong evidences for ego depletion (who later became a subfield of psychology with many more studies done to confirm it)

2007

in 2007, researchers figured out what *seemed* to be happening biologically:

as people used up their self-control, their blood sugar levels were dropping

they made subjects watching emotional videos without showing emotions / while others did not have to hold back

2007

2010

evidences even more in 2010 when group of researchers led by Martin Hagger, examined **83 published studies** on ego depletion to conclude that **effect was real**

2012

but in 2012 researchers **casted some doubts**

e.g. subjects did not have to drink lemonade to replenish their will power, **tasting it** was enough

e.g. subjects who **believe in willpower** could affect their performance

2014

in 2014, researchers tried to replicate the original studies and **could not find the effect**

they also looked at the meta analysis of 2010 (the 83 papers) and found a lot of issues, e.g. they re run the analysis with newer methods … and the **ego depletion effect disappeared**

this started a wave a concern about **replicability** hitting the field of psychology

2016

in 2016, the Association for Psychological Science opened a Registered Replication Report on ego depletion:

one official experiment would be conducted by researchers in many different labs

Baumeister (original study) even helped design the experiment and Martin Hagger (led of the meta analysis) lead the project

2016

on 24 labs (different language and culture)

only **2 found the effect**

and **1 group found the opposite effect**

what does it mean?

could argue that it is just this particular task where ego depletion does not show up but that willpower is still a limited resource

or may be ego depletion only happens under very specific circumstances … if at all

this has created a lot of **movements in the scientific community** and a lot of researches is working on it to find ways to better analyze and report studies

p hacking

p crisis

replication crisis

watch this
https://www.youtube.com/watch?v=42QuXLucH3Q&t=420s

mmm does it mean that all I have learned in this lecture is wrong?

no fortunately there are a many things you can do to analyse stats and report data correctly as well as minimise these problems …

good statistics

fishing = gathering as many data as you can, then try to find something statistically significant in it and report it = **NO**

rather have clear hypothesis to start with (remember one hypothesis = logical sentence that can be directly tested) and **just test for your hypothesis**

1 research question / hypothesis?

2 in(dependant) variables?

3a within or between subjects?

3b counterbalancing?

4 how many repetitions/trials?

5 look at raw data

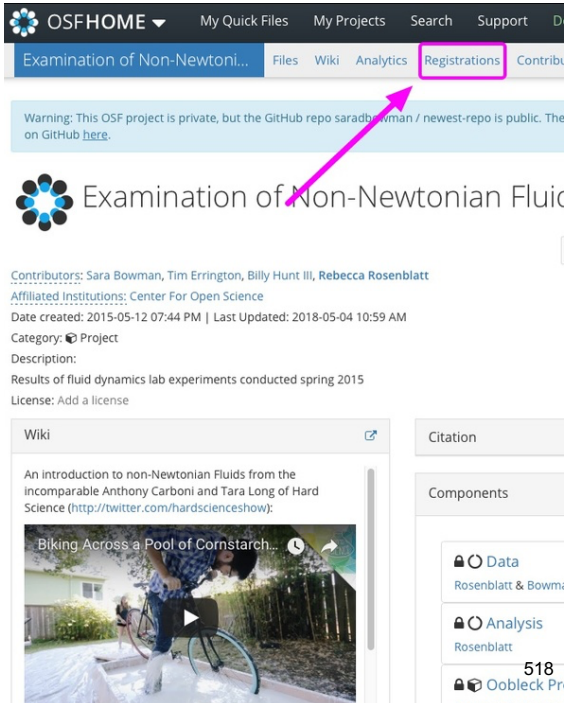
6 look at distributions

7a check for normality

7b run some stats

8 conclude

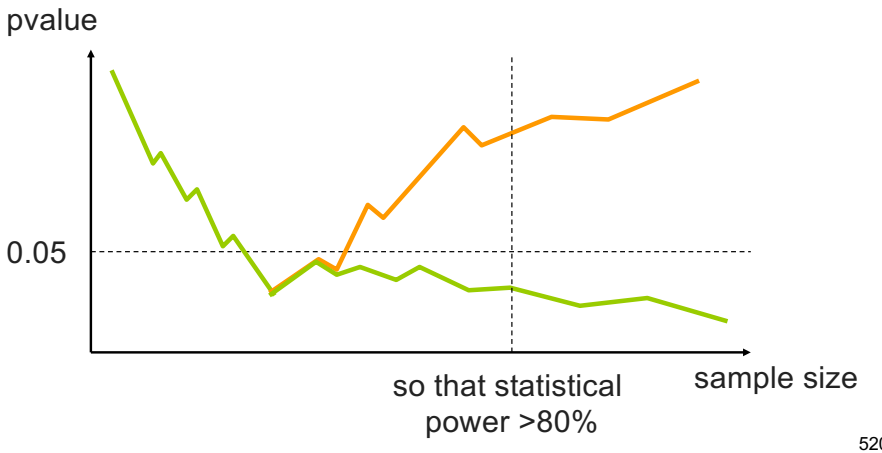
you could also decide to **pre-register your study** (more and more frequent in scientific venues)



and of course design your experiment to remove as much noise as possible

i.e. pilot it!

increase sample size or better compute the required sample size



3 report p value with effect size

e.g. a new hair loss shampoo is statistically better than existing shampoo

but does not say that subjects who took it only grew 5 hairs more than control group …

effect size matters more!

4 report confident intervals and non misleading grap

confident intervals::

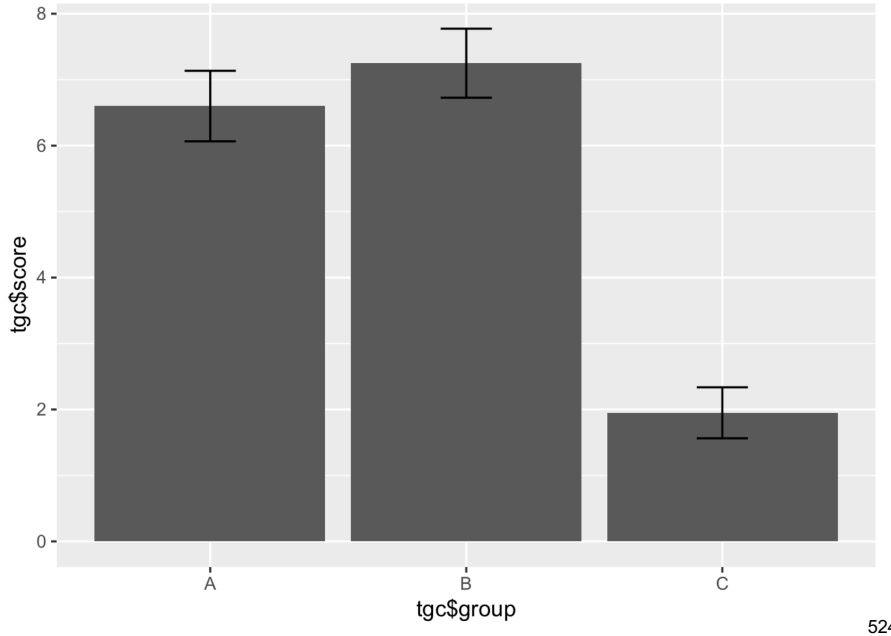
a 95% confidence interval is a range of values that you can be 95% certain contains the true mean of the population

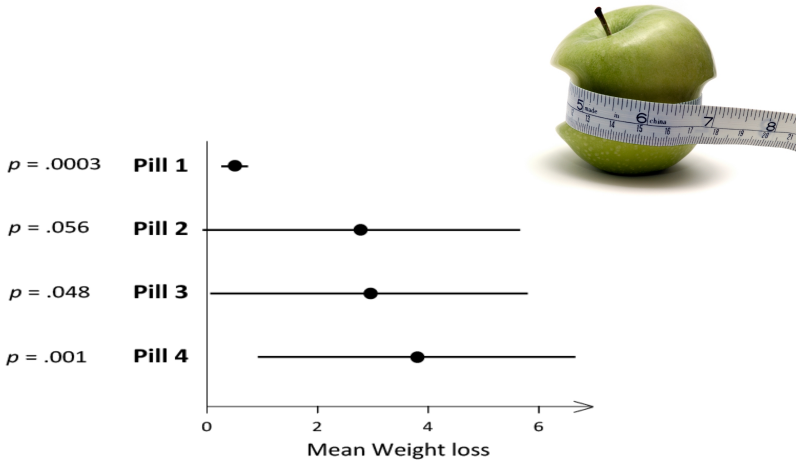
a range of plausible values for the mean (values outside relatively implausible)


```
# first we run the one-way anova
library(Rmisc)
tgc <- summarySE(dat, measurevar="score",
groupvars=c("group"))
tgc
```

```
  group  N score        sd        se        ci
1     A 20  6.60 1.1424811 0.2554665 0.5346976
2     B 20  7.25 1.1180340 0.2500000 0.5232560
3     C 20  1.95 0.8255779 0.1846048 0.3863824
```

```
ggplot(data = tgc, aes(x = tgc$group, y = tgc$score)) +
geom_bar(stat = 'identity', position = 'dodge') +
geom_errorbar(aes(ymin= tgc$score - ci, ymax= tgc$score  +
ci), width=.2, position=position_dodge(.9))
```

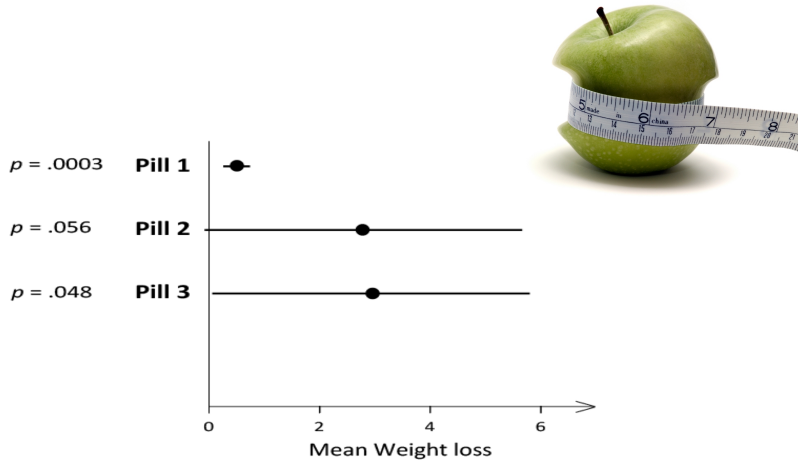





Error bars are 95% CIs
p-values are based on a null hypothesis of no effect

Adapted from (Ziliak and McCloskey, 2009)

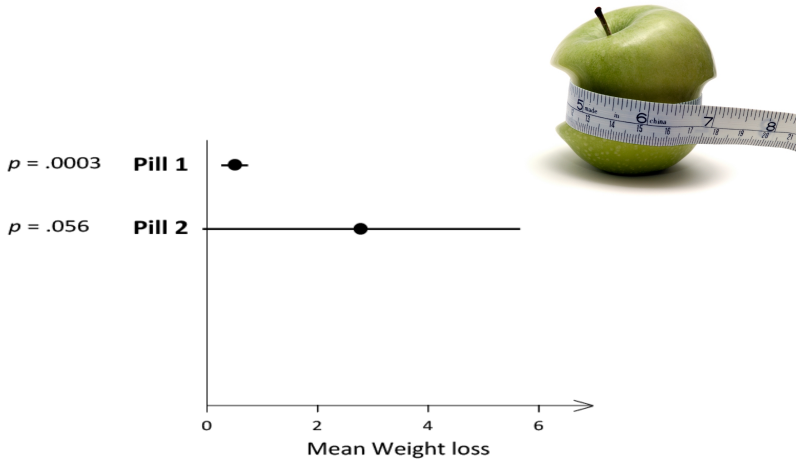
which weight-loss pill would you recommend?



Error bars are 95% CIs
p-values are based on a null hypothesis of no effect

Adapted from (Ziliak and McCloskey, 2009)

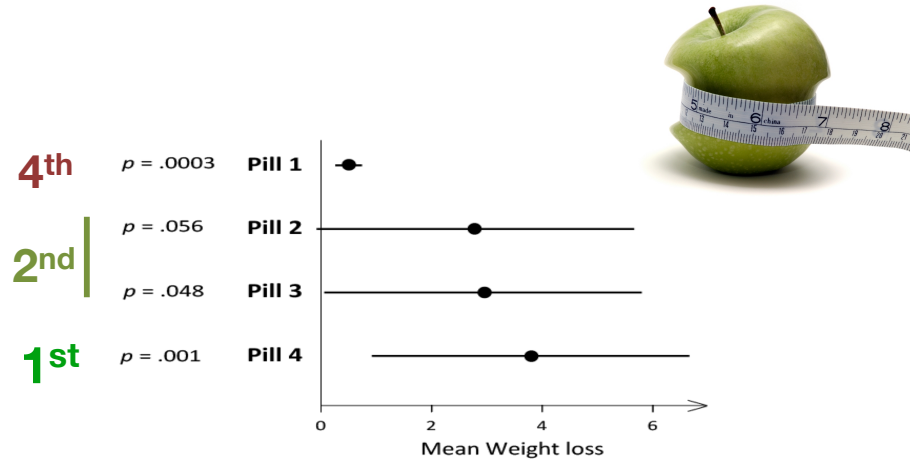
which weight-loss pill would you recommend?



Error bars are 95% CIs
p-values are based on a null hypothesis of no effect

Adapted from (Ziliak and McCloskey, 2009)

which weight-loss pill would you recommend?



Error bars are 95% CIs
p-values are based on a null hypothesis of no effect

Adapted from (Ziliak and McCloskey, 2009)

which weight-loss pill would you recommend?

Error bars are 95% CIs
p-values are based on a null hypothesis of no effect

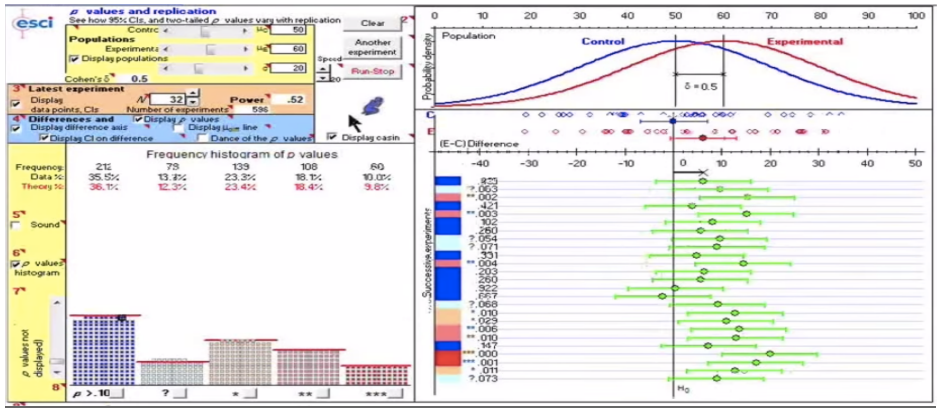
Adapted from (Ziliak and McCloskey, 2009)

which weight-loss pill would you recommend?

"Statistical significance is perhaps the least important attribute of a good experiment; it is never a sufficient condition for claiming that a theory has been usefully corroborated, that a meaningful empirical fact has been established, or that an experimental report ought to be published" (Likken, 1968)

"We have the duty of communicating our conclusions in intelligible form, in recognition of the right of other free minds to utilize them in making their own decisions" (Fisher, 1955)

"no confidence interval should be interpreted as a a significance test" (Schmidt and Hunter, 1997)



Geoff Cumming's dance of p-values

https://www.youtube.com/watch?v=ez4DgdurRPg

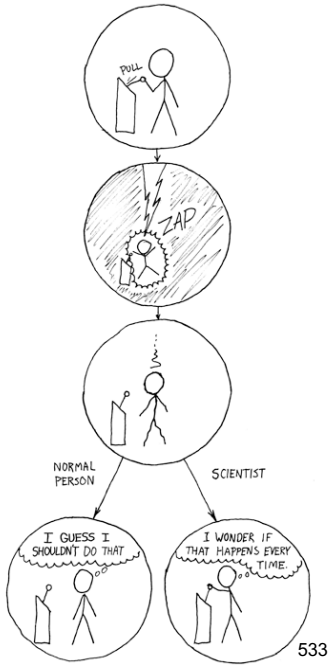
spot misleading graphs

https://www.youtube.com/watch?v=E91bGT9BjYk

replicate!

and fight publication standards,

e.g. in certain fields there are conferences to publish replication studies



rather use artificial data

certain tasks are **very sensitive to human variability** (e.g. ego depletion on will power but also anything that related to preferences)

tasks involving participants but relying on motor skills (e.g. tapping on a key) **suffer less from human variability**

or **use data without involving human** (e.g. algorithms comparisons)

remember this is an active field, always look up for new statistical methods

e.g. at the moment there is a strong tendency to push for **Bayesian testing,** although it also has drawbacks

need **prior data**

simple for AB testing but **could become quickly complex**

unclear **how it compares** to pvalue testing

(still some research to do on this so keep your eyes open!)

for the curious: tutorial on GitHub to do a simple comparison of two groups with Bayesian methods

be ethical

i.e. moral principles that govern a person's behaviour or the conducting of an activity

why are you doing a study, intrinsically because you want to learn something, not just publishing

of course be also ethical with your study design

research goes wrong (Standford Experiment)
… use ethical boards (in each university)

1. Explain what is the replication crisis
2. Give the steps seen in class to avoid phacking and do good statistics
3. Understand that this is a hot topic of research and know that you need to keep your eye open if you ever encounter stats later in your career

take away

end

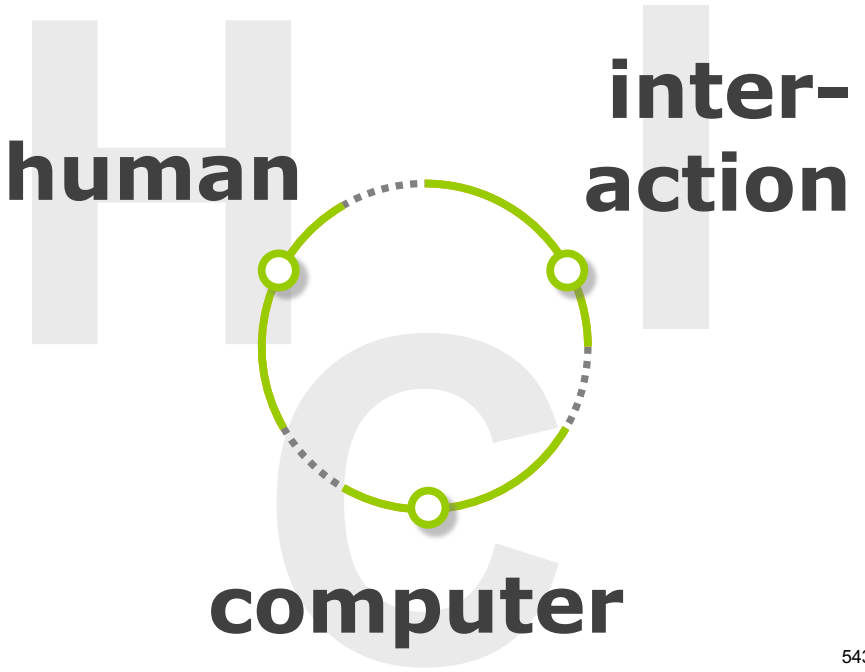
to go further

www.biglab.co.uk

BristolIG lab
(Youtube)

example of what we do
https://www.youtube.com/watch?v=liPzZIex54M

BIG Bristol Interaction Group



to go further:

year 1: **Probability and statistic**

year 2: **CS and society** (with introduction to HCI)

year 3 (currently year 2): **HCI**

year 4: **Interactive Devices (pre-req HCI)**

curriculum