13

Designing an Experiment (in 2 parts)

Probability and Statistics

COMS10011

Dr. Anne Roudaut
csxar@bristol.ac.uk

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?

3a within or between subjects?

3b counterbalancing?

4 how many repetitions/trials?

5 look at raw data

6 look at distributions

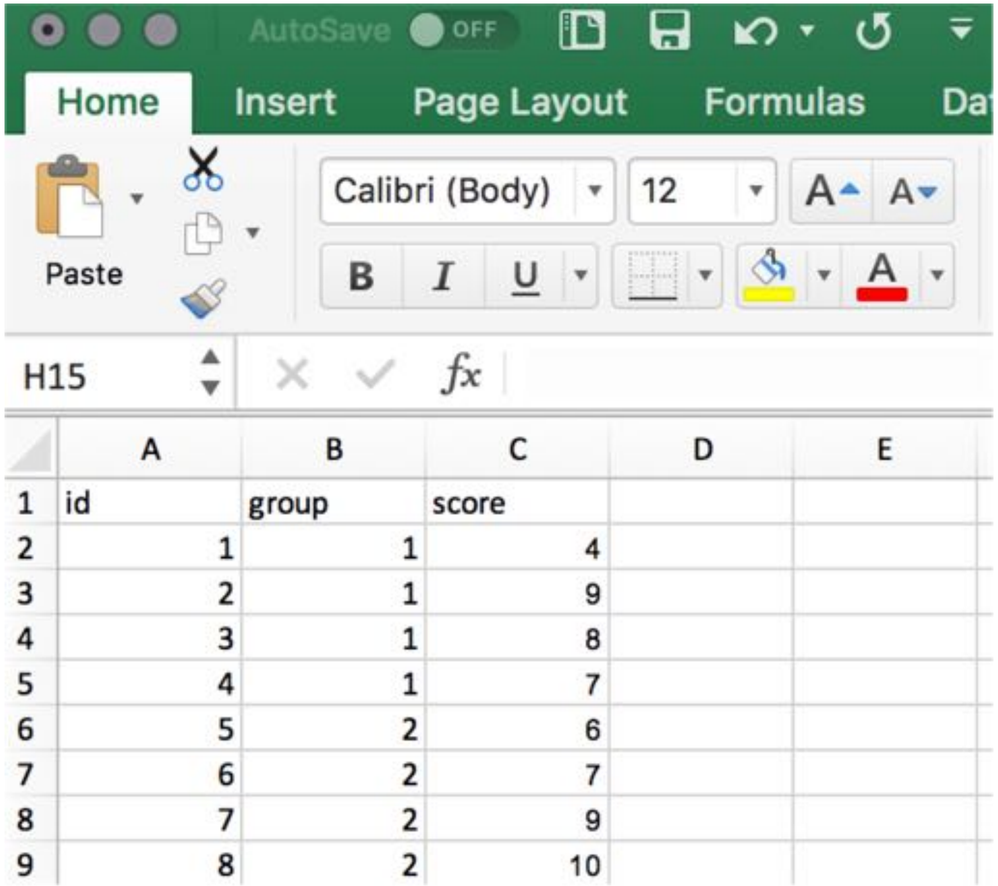
7a check for normality

7b run some stats

8 conclude

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)

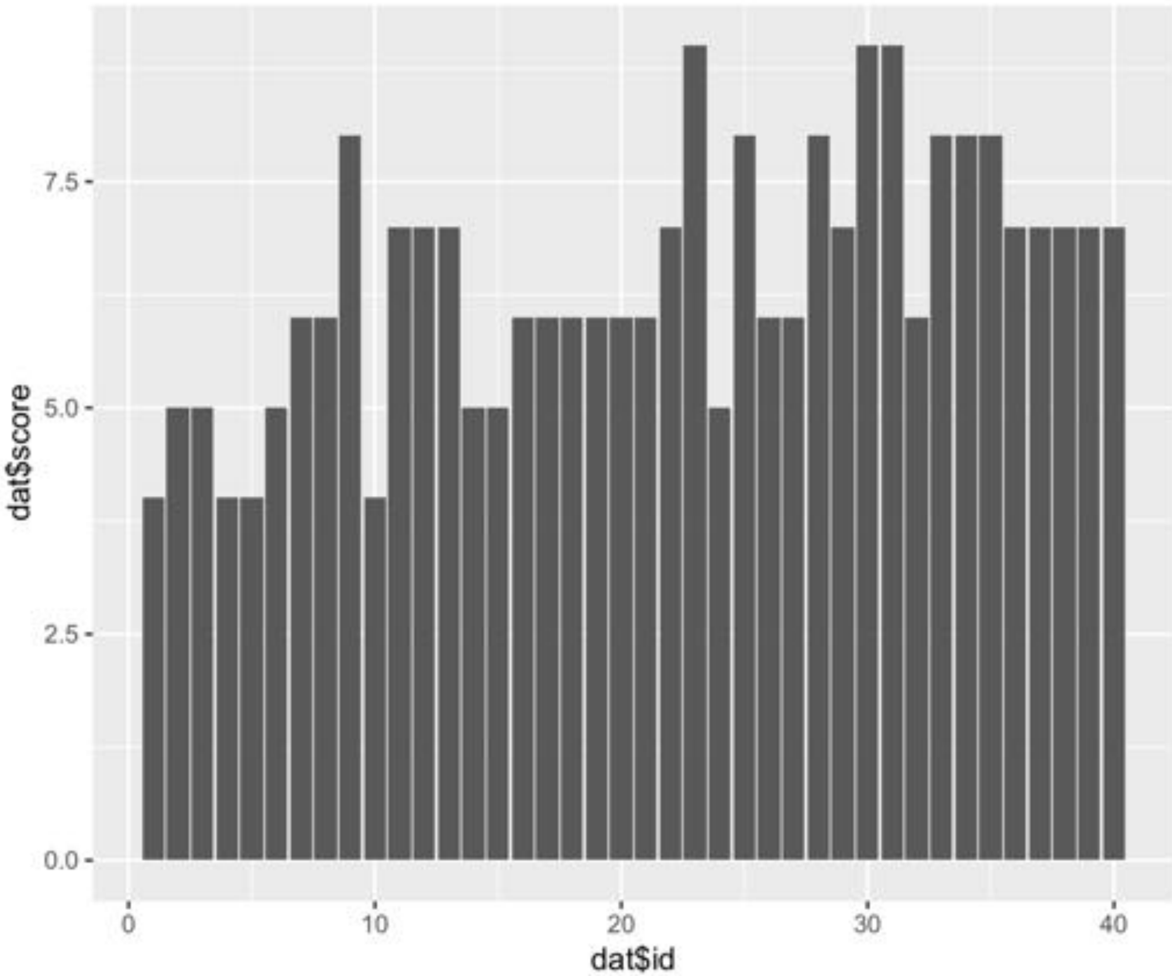
you can open csv with excel, text file an many other software

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
```

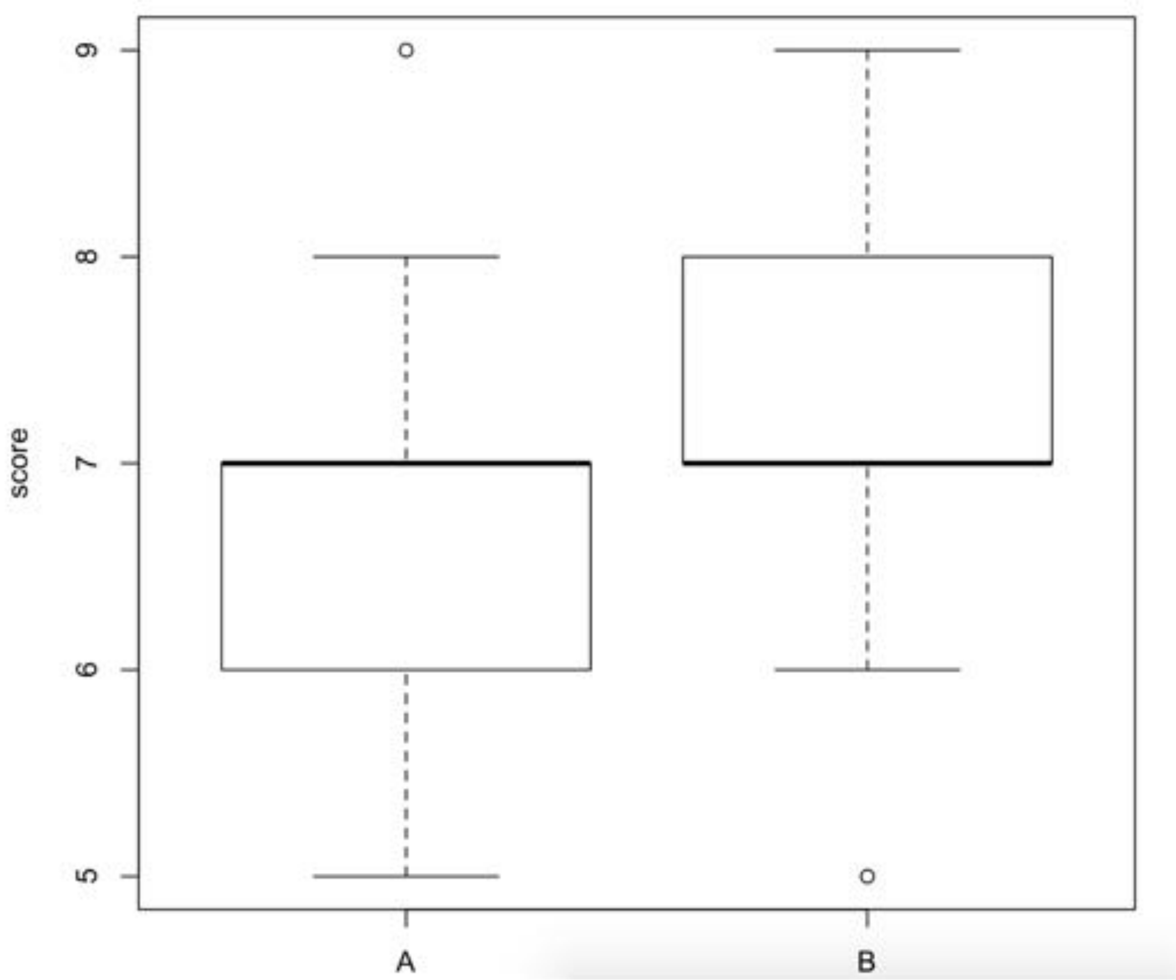
```r
dat = read.csv("HCIXP.csv", header = TRUE)
print(dat) # look at the file in R

library(ggplot2)

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



first: does the data look ok?

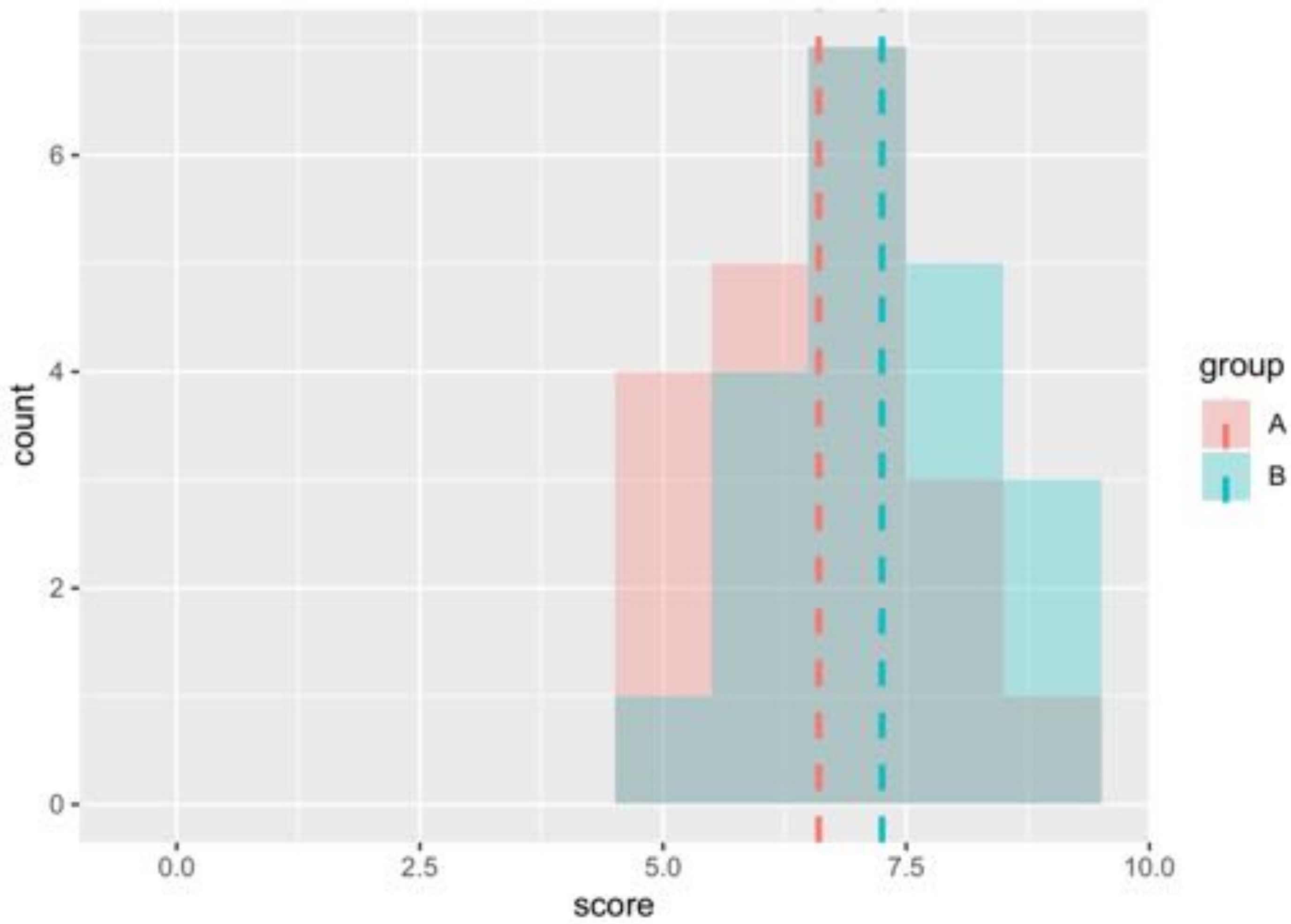
search for bugs, fatigue effect, learning effect
or outliers (>3 times std) = remove / redo xp

```
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

8

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?

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

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

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)

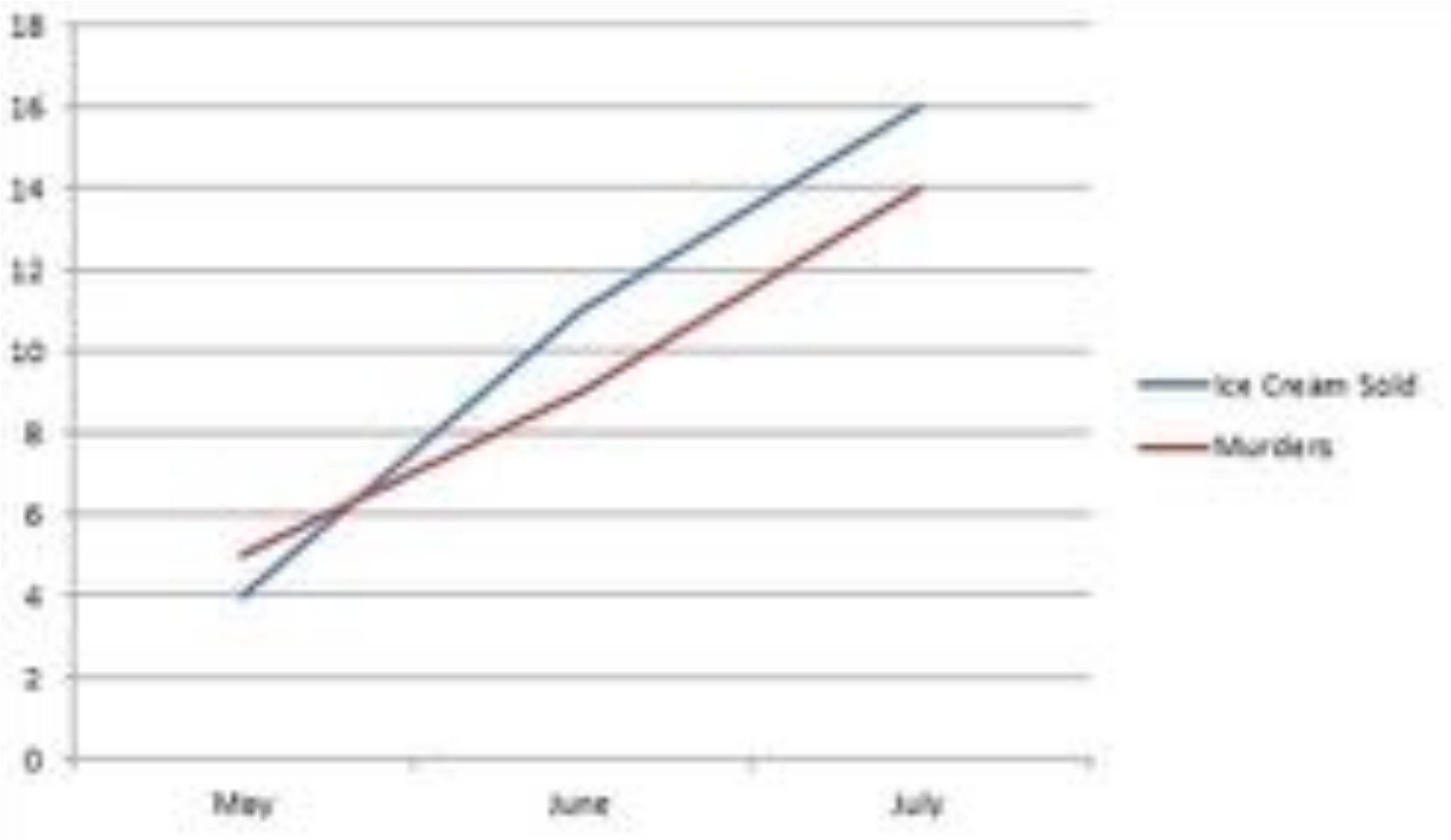
dependent variable = memorization score everything else should be a…

controlled variable ::

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

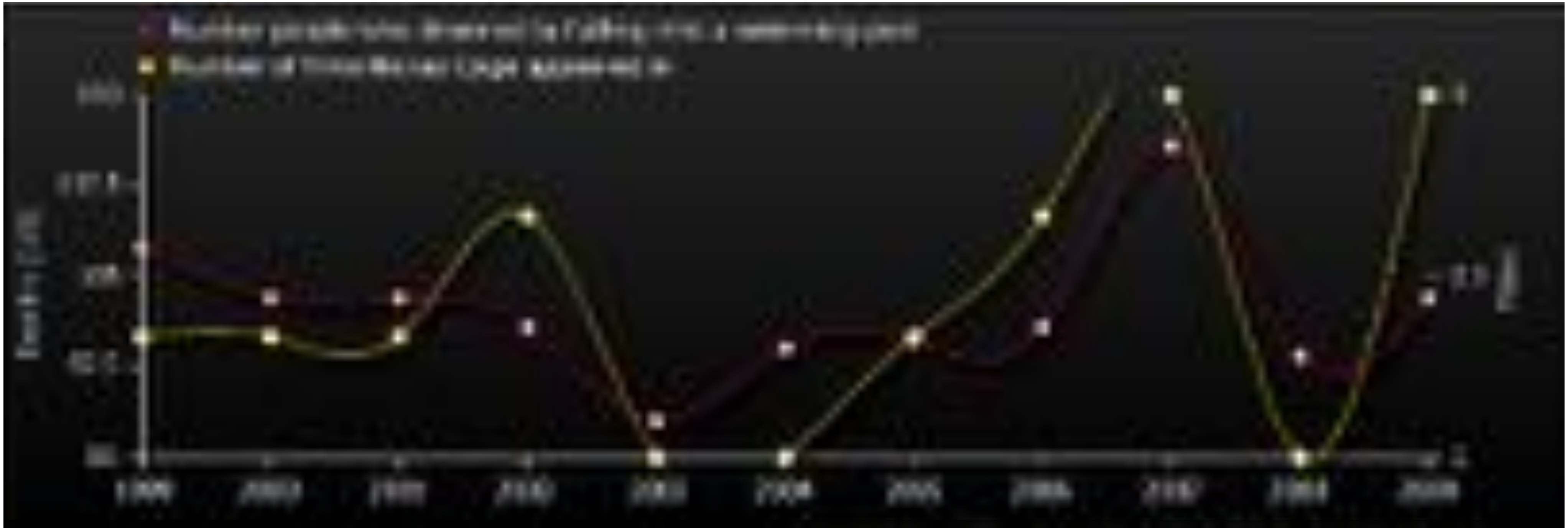
confounding variable ::

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

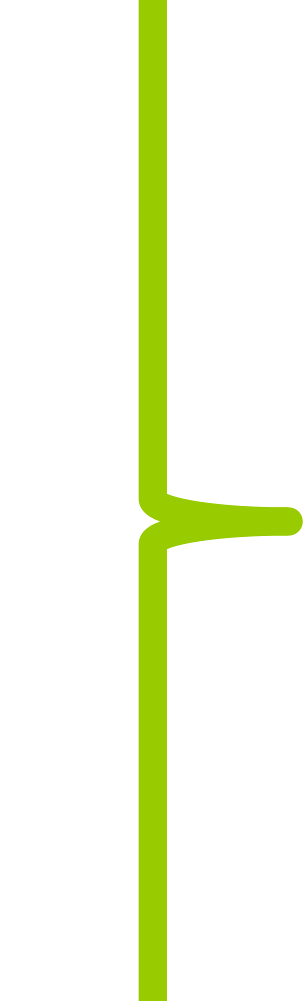
presenting conditions in a different order

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**

summary

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

we will see why

7b run some stats

so far we know t-test

8 conclude

end of part one

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 we will see why

check for normality

7b run some stats

so far we know t-test

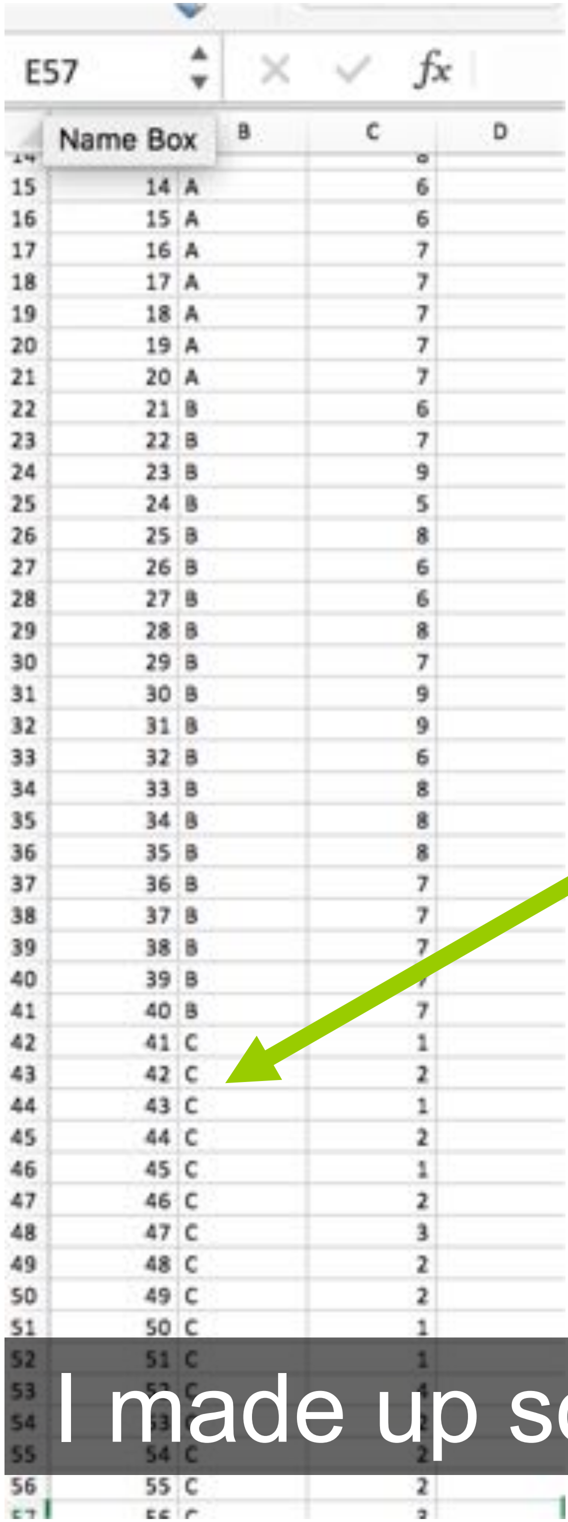
8 conclude

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!)

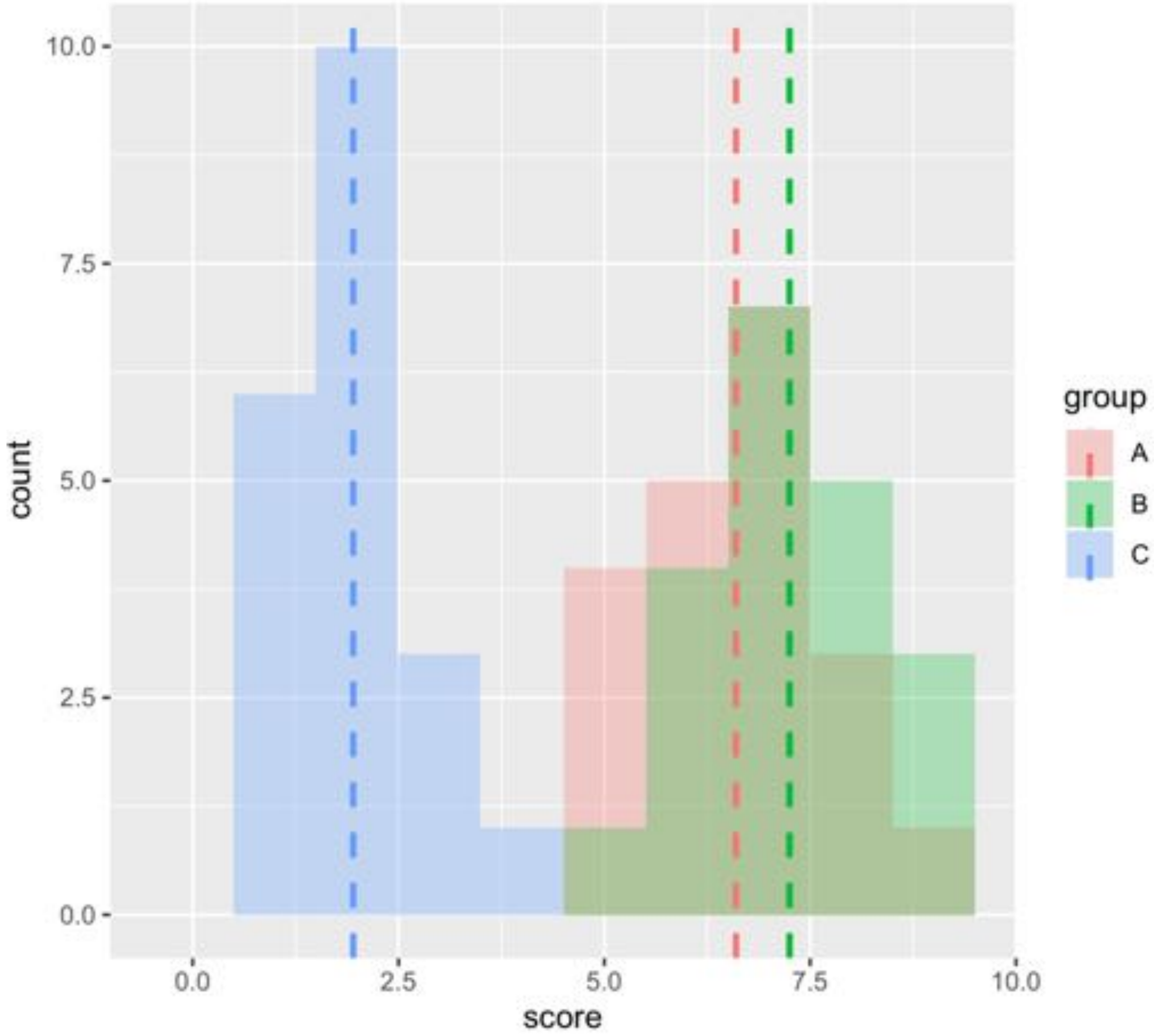
I made up some data



	A	B	C	D
14	[illegible]	[illegible]	8	
15	14	A	6	
16	15	A	6	
17	16	A	7	
18	17	A	7	
19	18	A	7	
20	19	A	7	
21	20	A	7	
22	21	B	6	
23	22	B	7	
24	23	B	9	
25	24	B	5	
26	25	B	8	
27	26	B	6	
28	27	B	6	
29	28	B	8	
30	29	B	7	
31	30	B	9	
32	31	B	9	
33	32	B	6	
34	33	B	8	
35	34	B	8	
36	35	B	8	
37	36	B	7	
38	37	B	7	
39	38	B	7	
40	39	B	7	
41	40	B	7	
42	41	C	1	
43	42	C	2	
44	43	C	1	
45	44	C	2	
46	45	C	1	
47	46	C	2	
48	47	C	3	
49	48	C	2	
50	49	C	2	
51	50	C	1	
52	51	C	1	
53	52	C	4	
54	53	C	2	
55	54	C	2	
56	55	C	2	
57	56	C	3	

Group C: “slap”

```r
# 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

```
# 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 the significant differences

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 ($p<0.05$)."

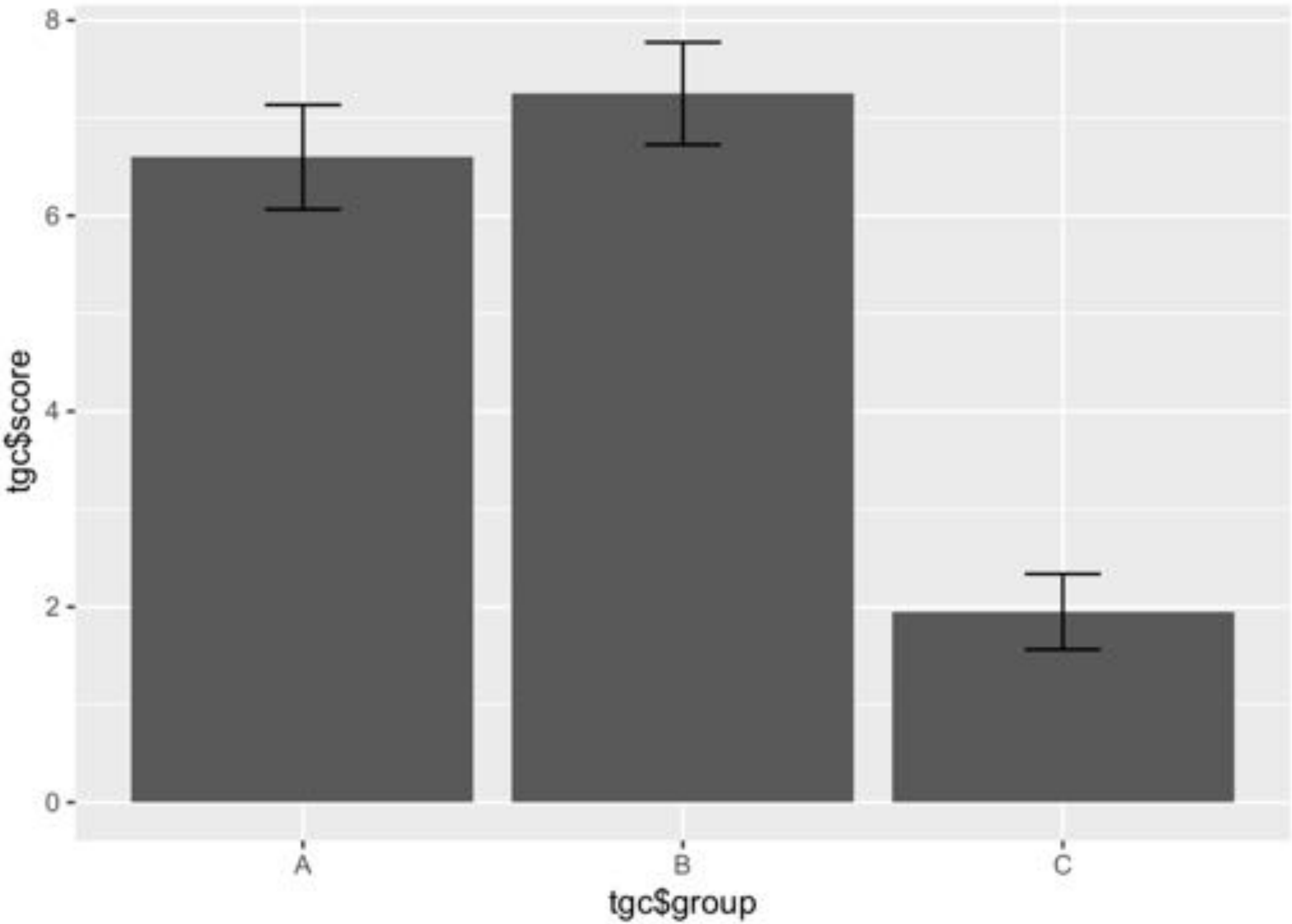
<you could also give means values to give more info>

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))
```



8
6
4
2
0
tgc$score
A
B
C
tgc$group

ok we have learned quite a lot so far!

1 research question / hypothesis?

2 in(dependent) variables?

3a within or between subjects?

3b counterbalancing?

4 how many repetitions/trials?

5 look at raw data

6 look at distributions

7a we will see why
check for normality

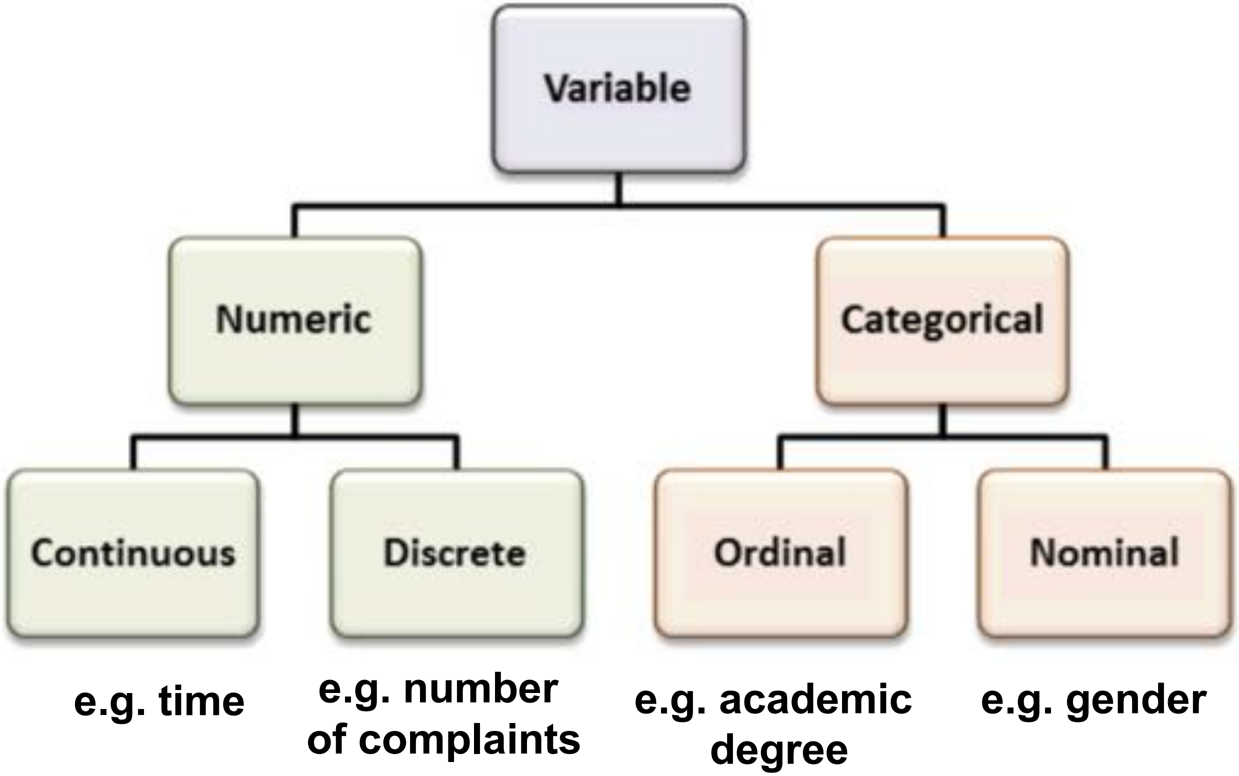
7b run some stats
T-test if 2 group
ANOVA if more

8 conclude


quick chocolate break! you all deserve it!

let’s talk about dependent variables

there are many type of dependent variables you can collect



opinions/surveys
as dependent
variables ...

if you want to collect subjective metric such as opinions, use **Likert Scale**

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

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

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

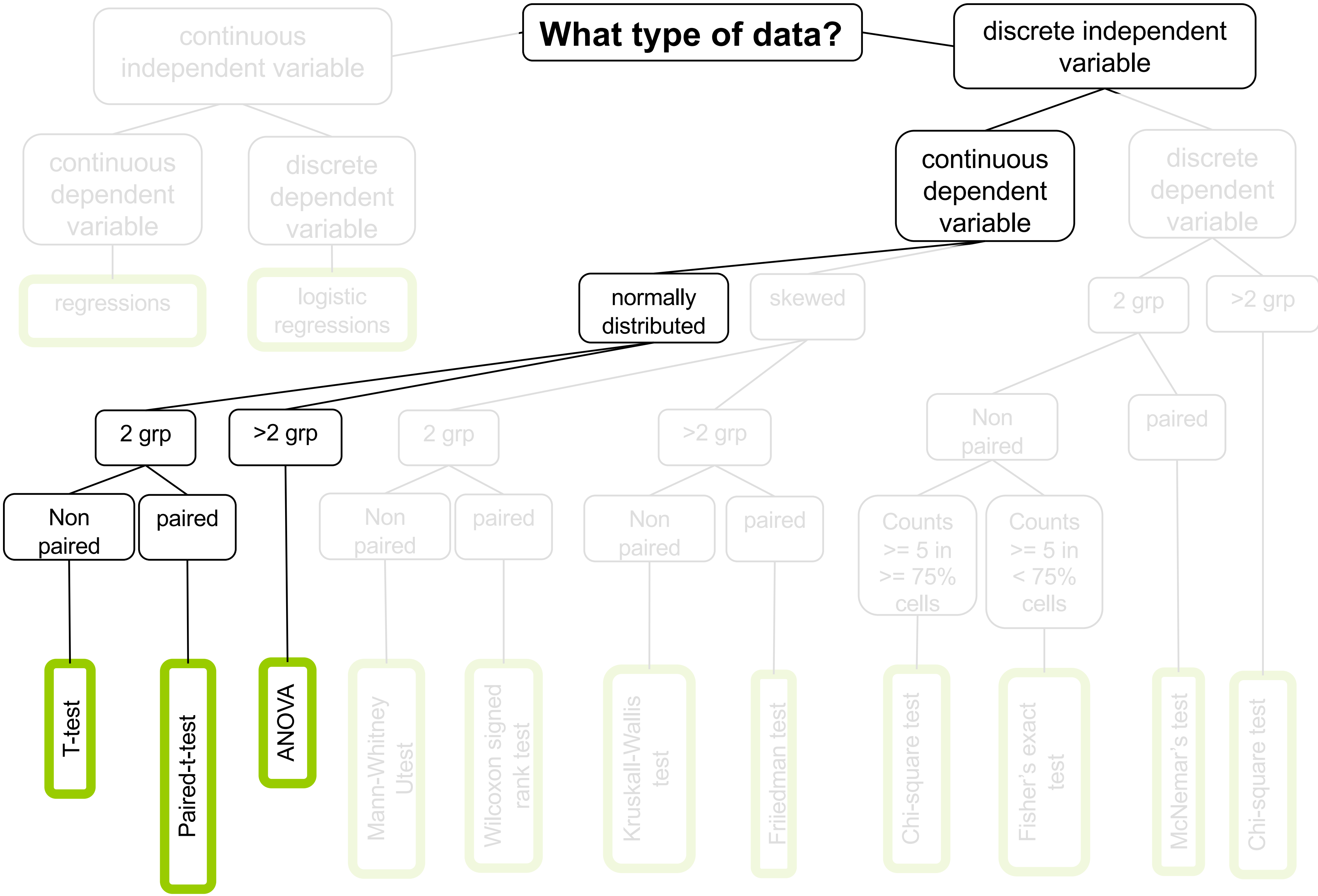

bad news: we cannot use Ttest and Anova

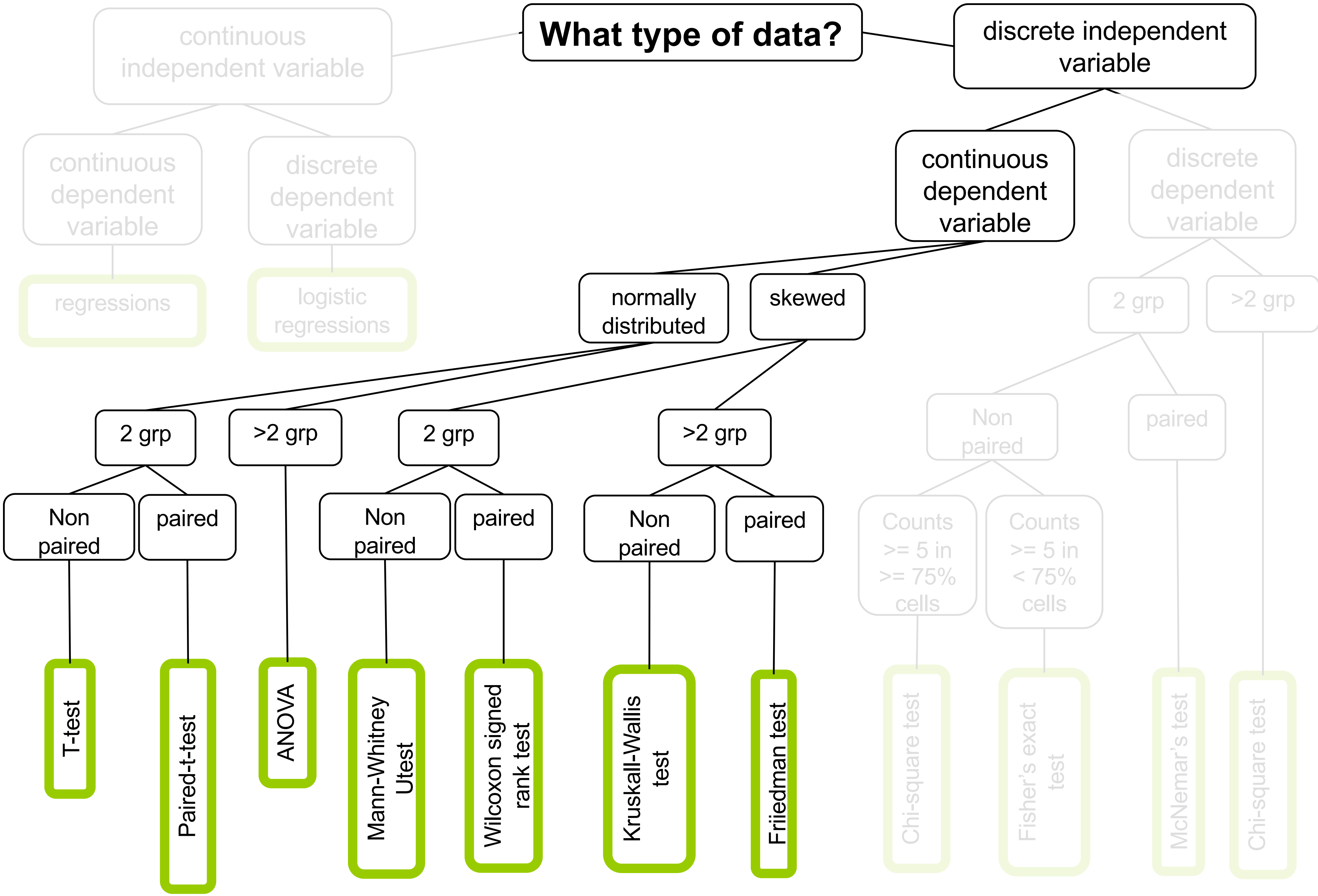
use parametric tests (ttest, anova)

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

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

use non-parametric tests


What type of data?
continuous independent variable
discrete independent variable
continuous dependent variable
discrete dependent variable
regressions
logistic regressions
continuous dependent variable
discrete dependent variable
normally distributed
skewed
2 grp
>2 grp
2 grp
>2 grp
2 grp
>2 grp
Non paired
paired
Non paired
paired
Non paired
paired
Non paired
paired
Counts >= 5 in >= 75% cells
Counts >= 5 in < 75% cells
T-test
Paired-t-test
ANOVA
Mann-Whitney Utest
Wilcoxon signed rank test
Kruskall-Wallis test
Friedman test
Chi-square test
Fisher's exact test
McNemar's test
Chi-square test



What type of data?
continuous independent variable
discrete independent variable
continuous dependent variable
discrete dependent variable
regressions
logistic regressions
continuous dependent variable
discrete dependent variable
normally distributed
skewed
2 grp
>2 grp
2 grp
>2 grp
Non paired
paired
Non paired
paired
Non paired
paired
T-test
Paired-t-test
ANOVA
Mann-Whitney Utest
Wilcoxon signed rank test
Kruskall-Wallis test
Friedman test
2 grp
>2 grp
Non paired
paired
Counts >= 5 in >= 75% cells
Counts >= 5 in < 75% cells
Chi-square test
Fisher's exact test
McNemar's test
Chi-square test

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 2 weeks

summary

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
8. conclude

design the experiment in such way that the results will be **easy to analyze**

be sure you will be **able to perform the statistical** analysis

there are many R tutorials online!

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 type 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

end