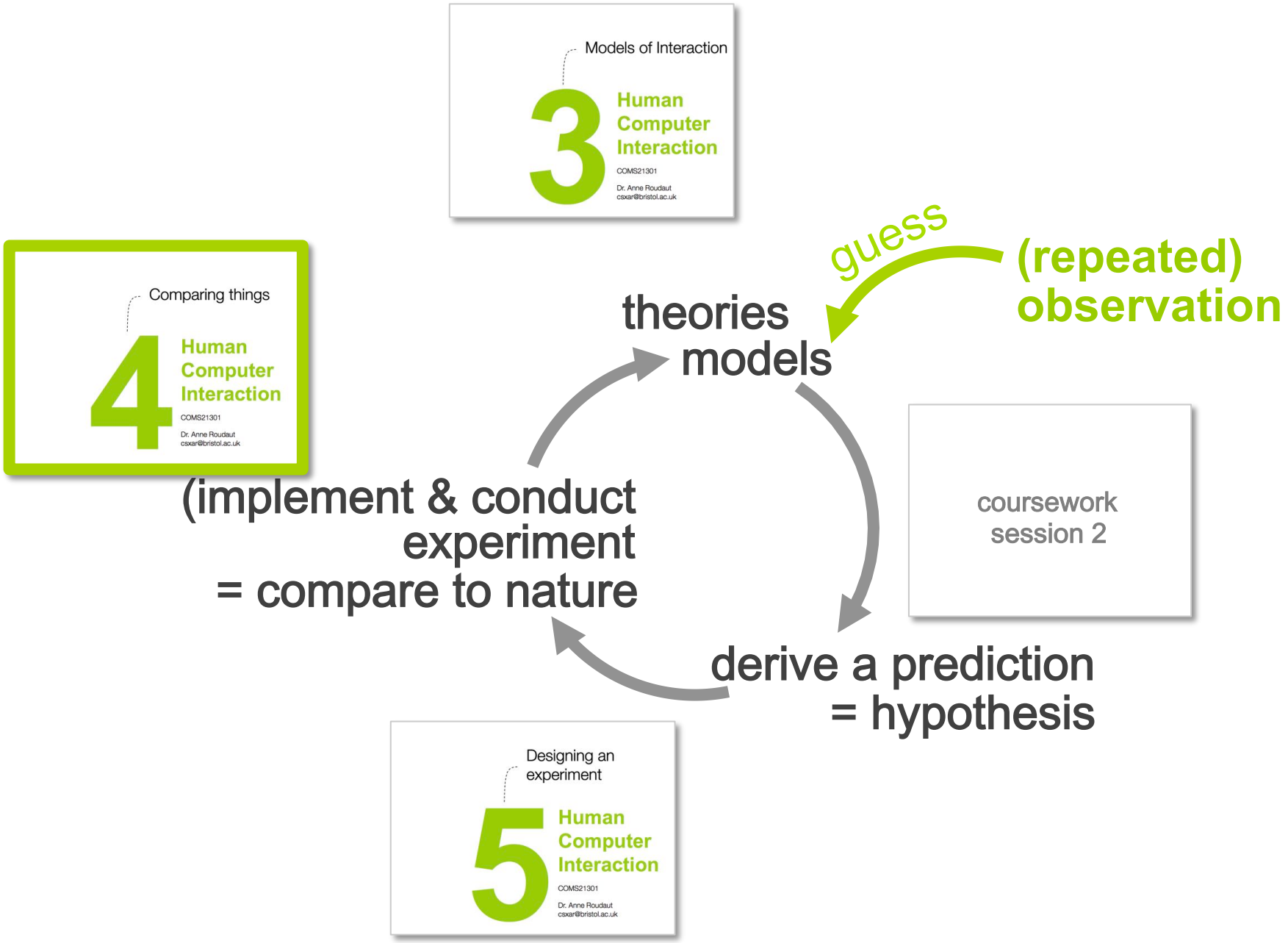
Comparing things
and statistical tests

4

Human Computer Interaction

COMS21301

Dr. Anne Roudaut
csxar@bristol.ac.uk


Models of Interaction
3
Human Computer Interaction
COMS21301
Dr. Anne Roudaut
csxar@bristol.ac.uk
Comparing things
4
Human Computer Interaction
COMS21301
Dr. Anne Roudaut
csxar@bristol.ac.uk
guess
(repeated) observation
theories models
derive a prediction = hypothesis
(implement & conduct experiment = compare to nature
coursework session 2
Designing an experiment
5
Human Computer Interaction
COMS21301
Dr. Anne Roudaut
csxar@bristol.ac.uk


let’s start with a contrived example…

you own a die. You have played with it a lot and from some game you **have a long log of the numbers you have rolled with it.**

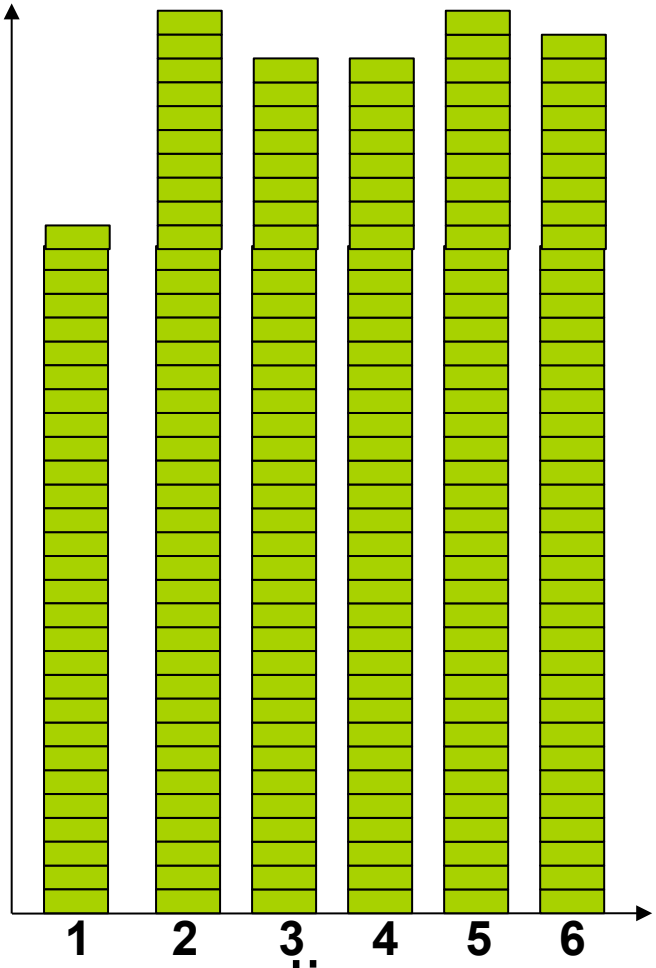
one day, you come home and it seems like someone has moved your die. You get concerned that someone might have taken your (beloved) die and instead **replaced it with an identical looking die.**

you inspect the die long and hard and it *looks* the same. Still you are worried. **How do you verify that it is the same?**

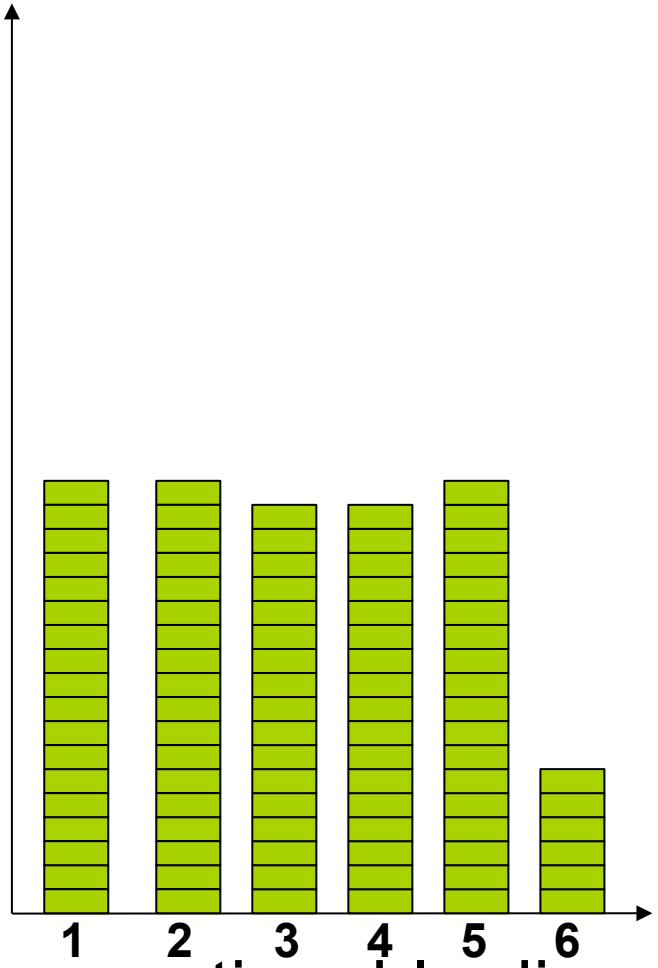
<30 sec brainstorming>

ok, so **you roll the “questionable” die**,
a bunch of times.

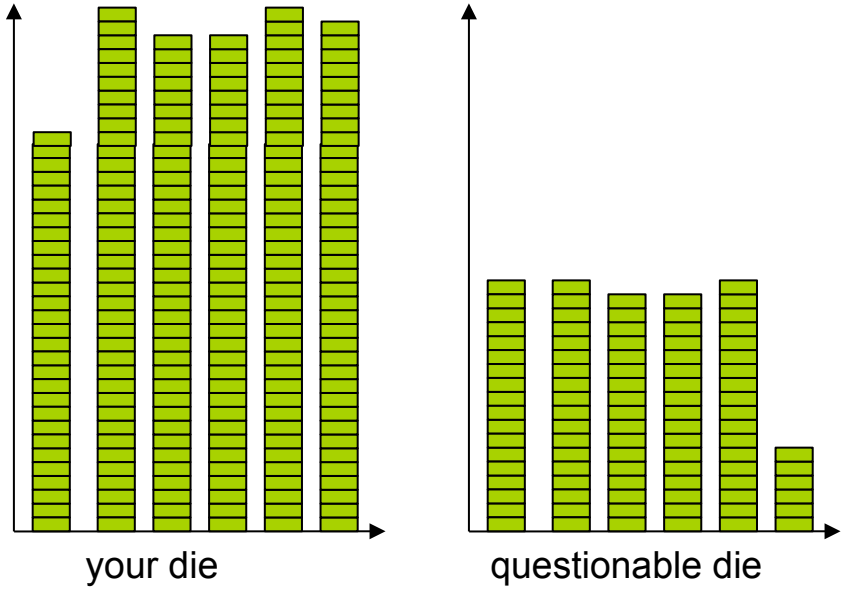
here is what you see…



your die



questionable die

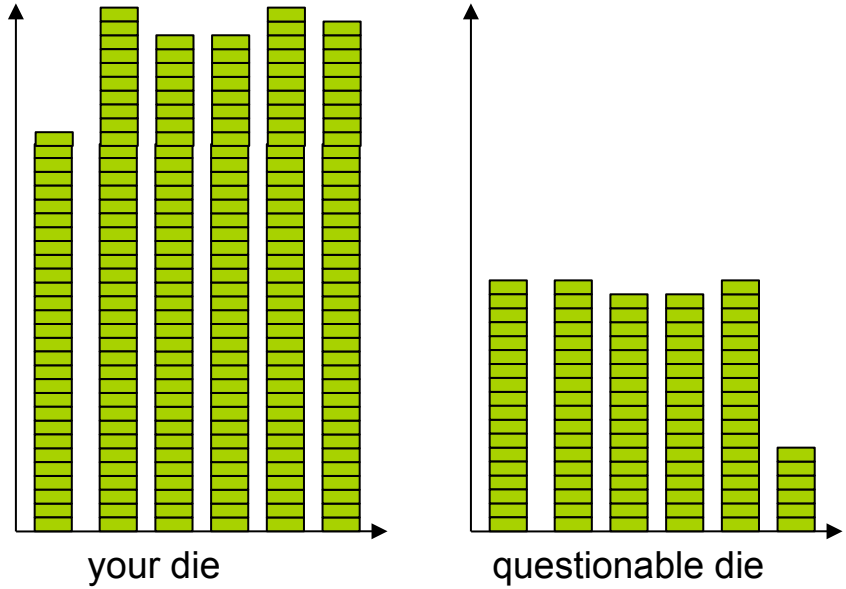


[ ] this is still the original die

[ ] someone has replaced my die

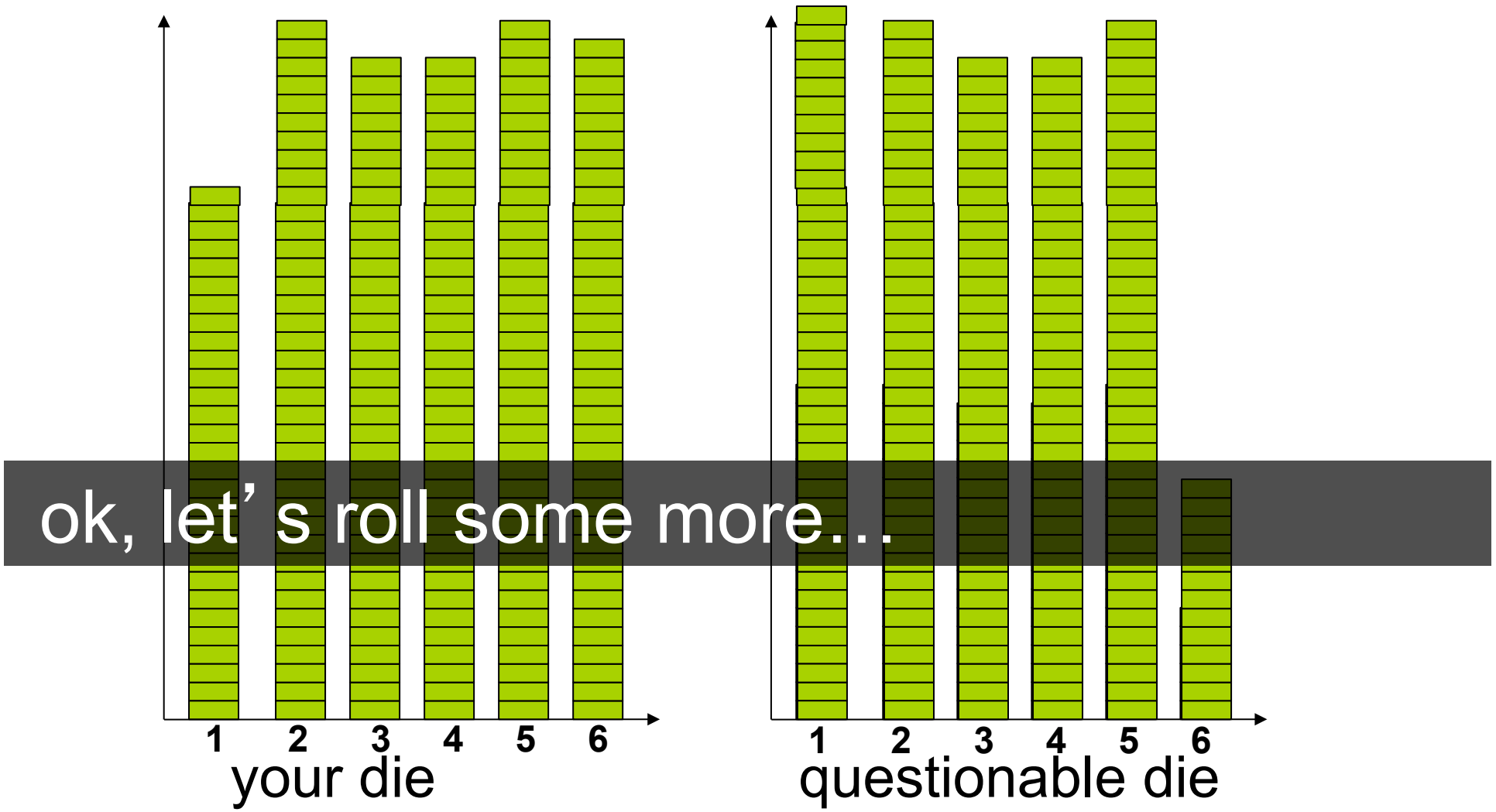
[x] could be the original or a replacement die

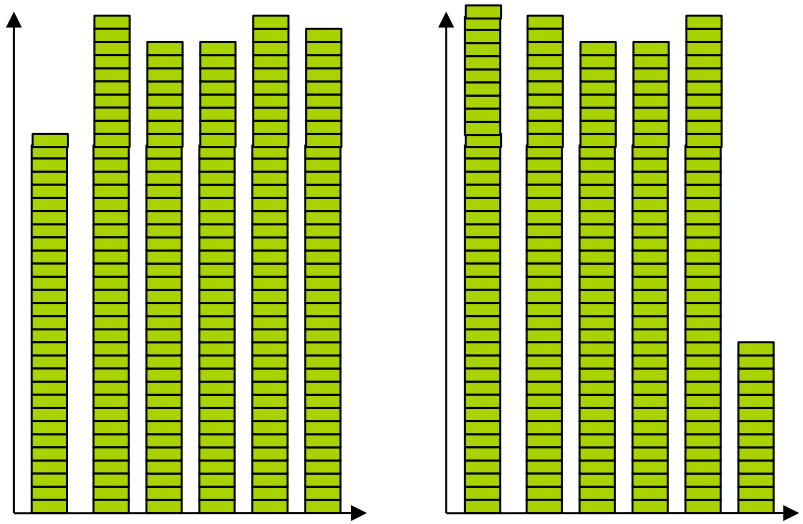
<let's vote>



the distribution looks different from the die you know.

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


ok, let’s roll some more…
1 2 3 4 5 6
your die
1 2 3 4 5 6
questionable die

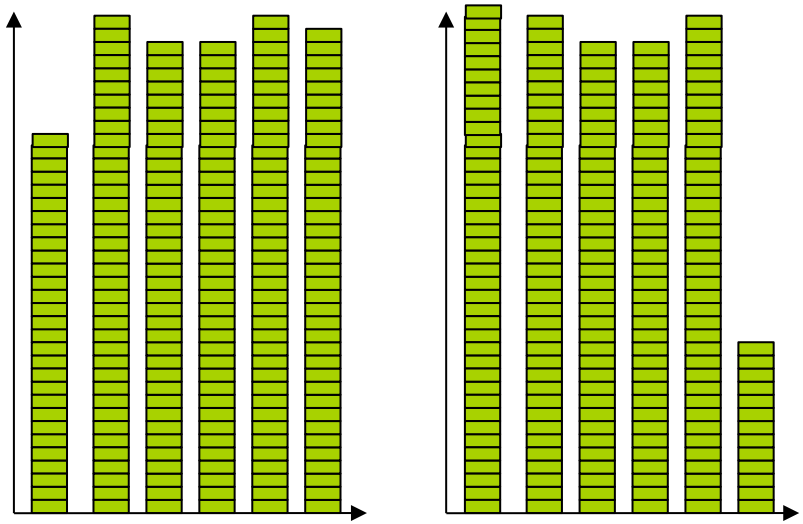


[ ] this is probably still the original die

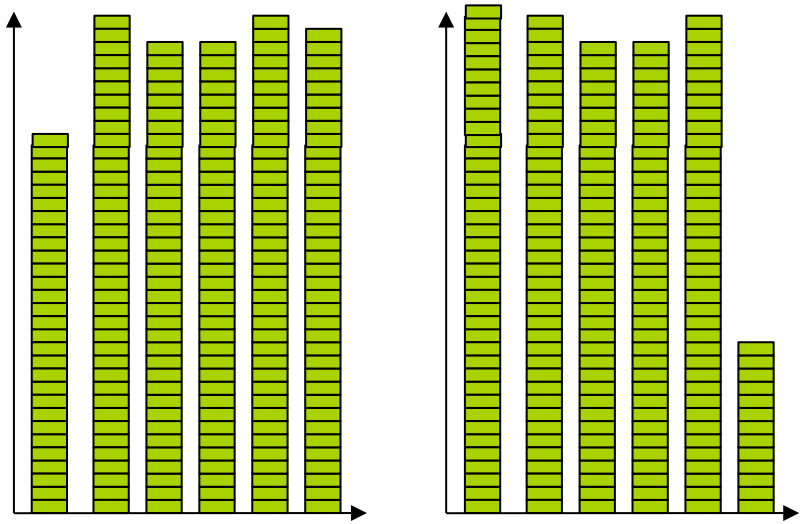
[x] someone probably has replaced my die

[x] could be the original or a replacement die

<let's vote>



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



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

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

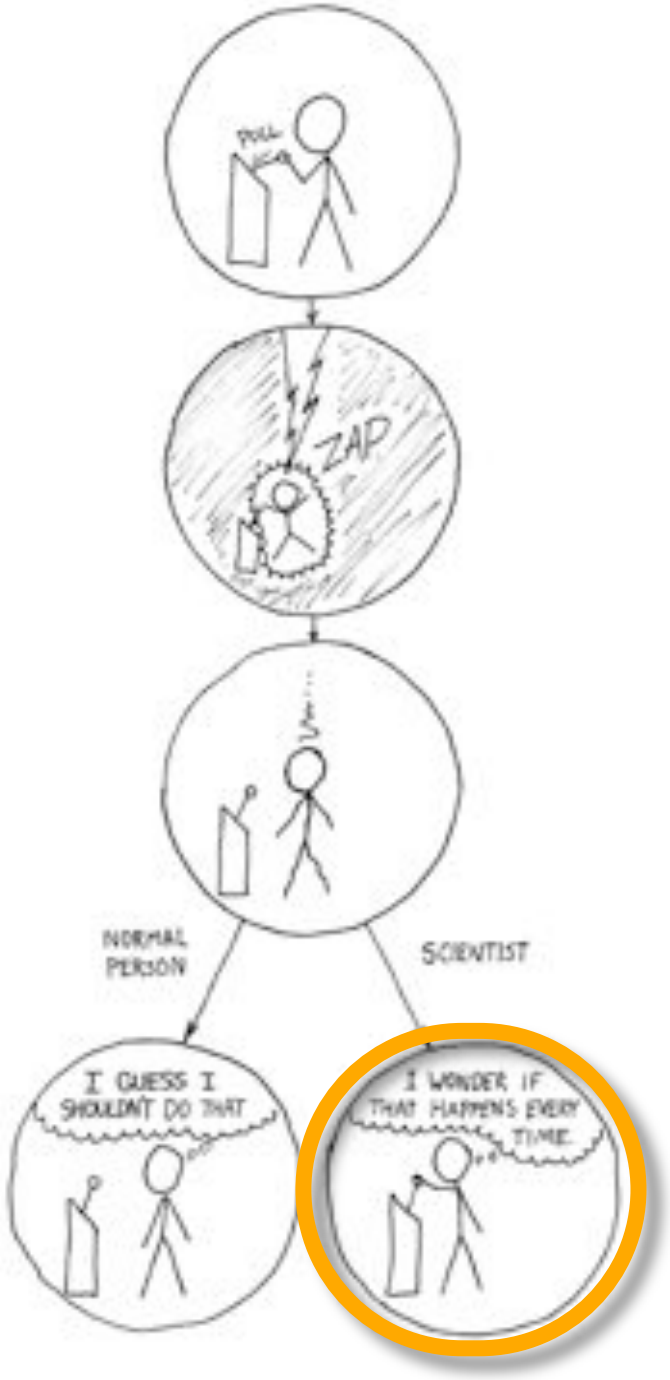
you are **not sure.**

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



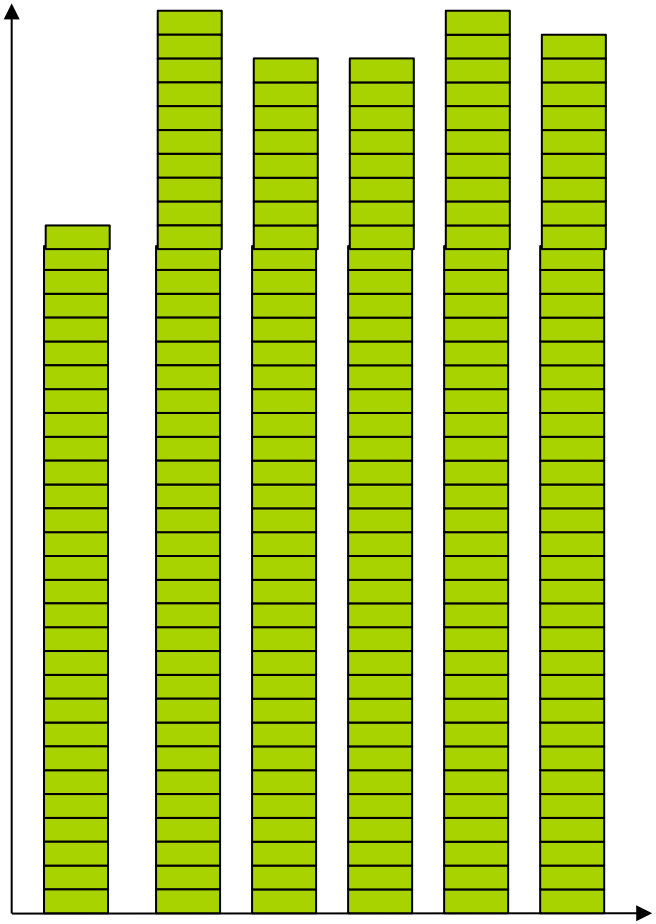
the good news:

(if you have many samples of original die)
→ with # of rolls, your confidence increases →
you can be **arbitrarily sure**

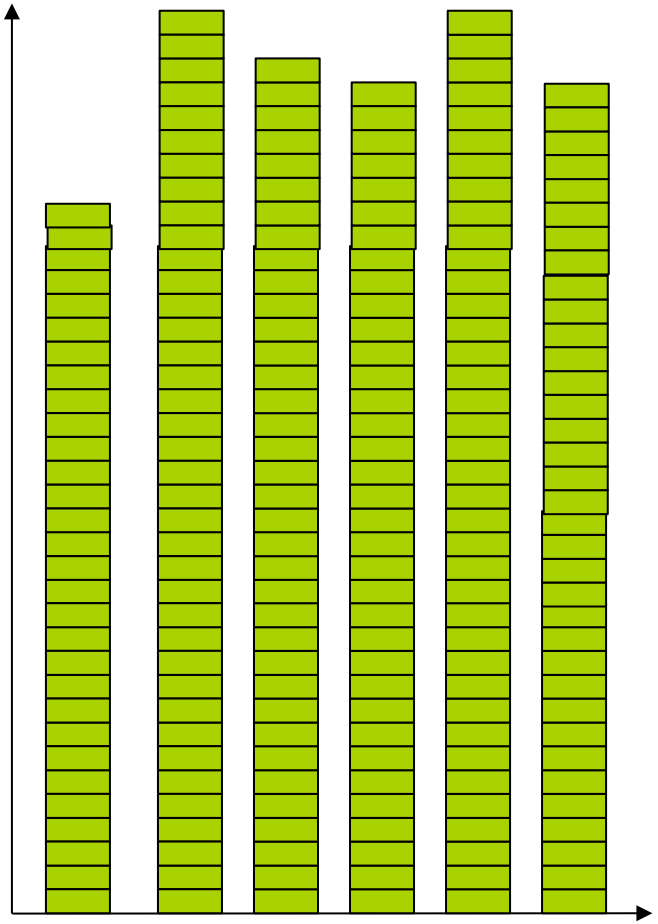
another round

ok, you get your original die back

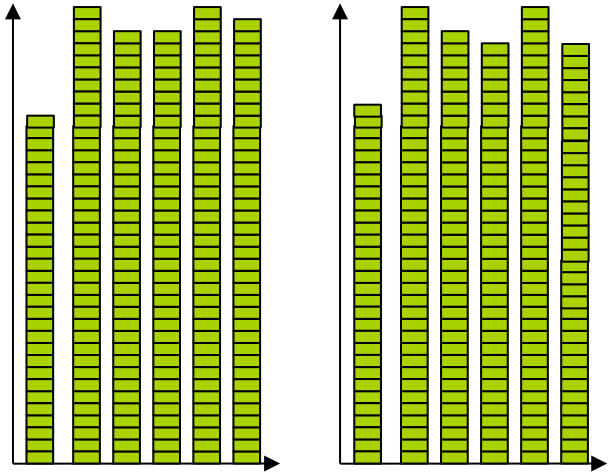
a week later the same thing **happens again.**
again, **you roll** the questionable die…



your die



questionable die

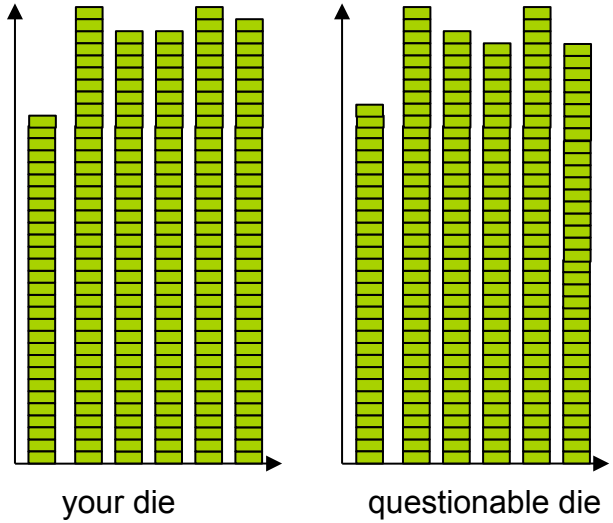


[ ] this is still the original die

[ ] someone has replaced my die

[ ] could be the original or a replacement die

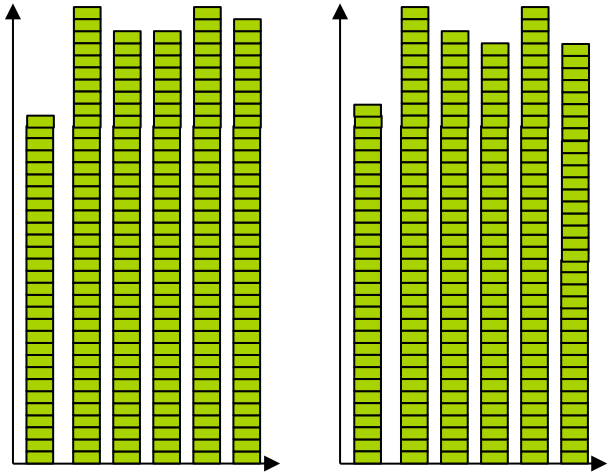
<let's vote>



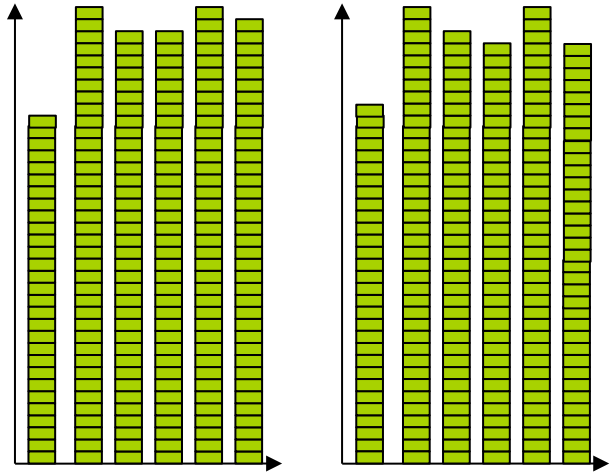
[ ] this is still the original die

[ ] someone has replaced my die

[x] could be the original or a replacement die



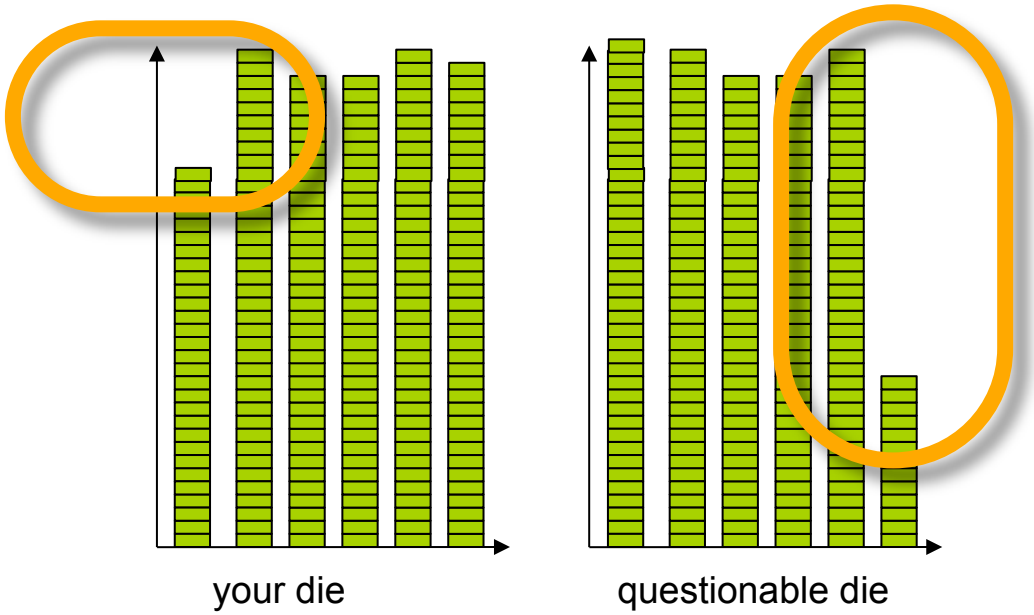
it could be your original die, or one that just happens to **behave the same.** a very, very, good copy maybe.



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

you **cannot compute** the odds.

that’s strange! why not?



this seems **unlikely…**

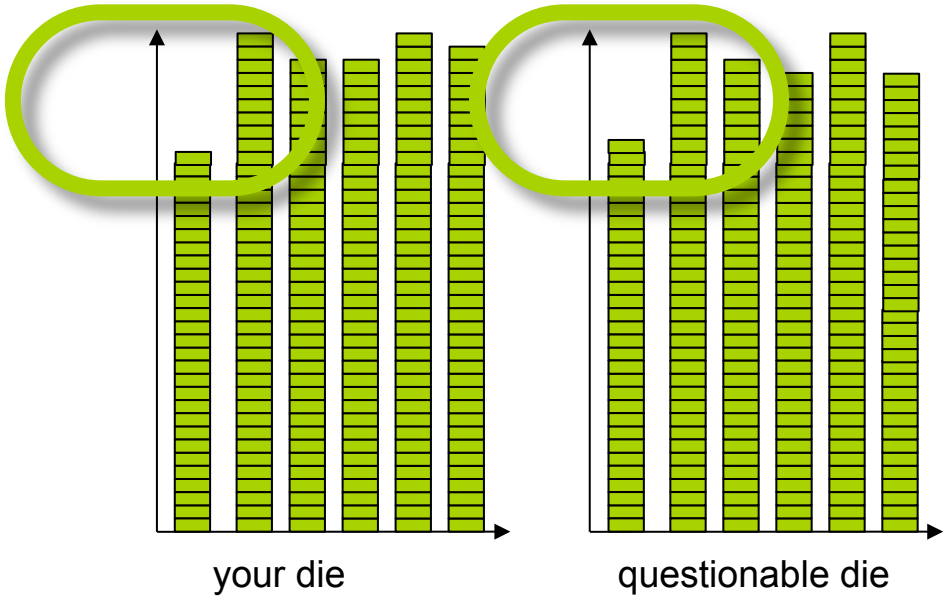
in both cases, there are two explanations

1.same die

2.different dice

…thus this **must be true**

…now, in the other case



this **does seem not unlikely…**

in both cases, there are two explanations

1.same die

2.different dice

we still have **two possible explanations**

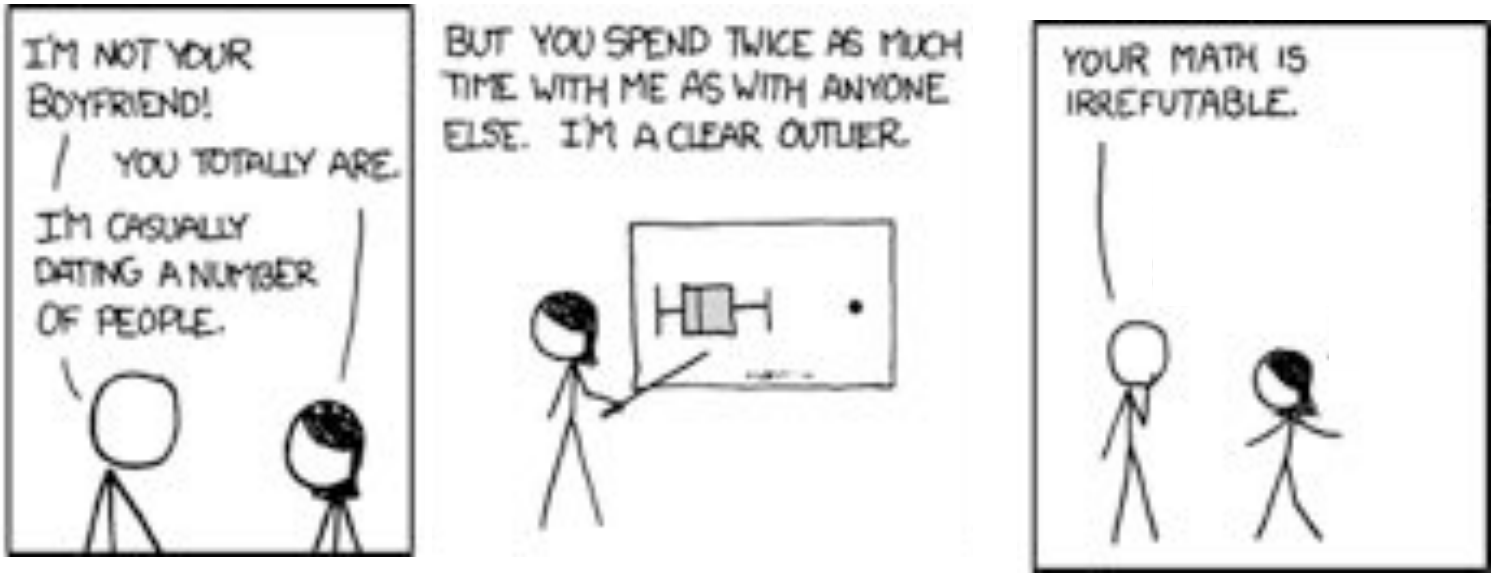
→ we **cannot conclude** anything

…thus… **no thus**

let’s use the stats

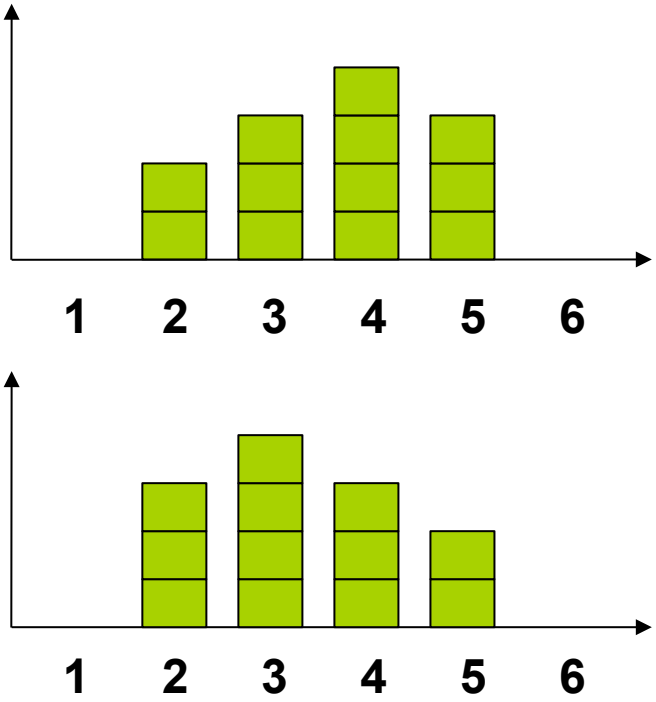
statistical significance ::

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


I'M NOT YOUR BOYFRIEND!
YOU TOTALLY ARE.
I'M CASUALLY DATING A NUMBER OF PEOPLE.
BUT YOU SPEND TWICE AS MUCH TIME WITH ME AS WITH ANYONE ELSE. I'M A CLEAR OUTLIER.
YOUR MATH IS IRREFUTABLE.

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

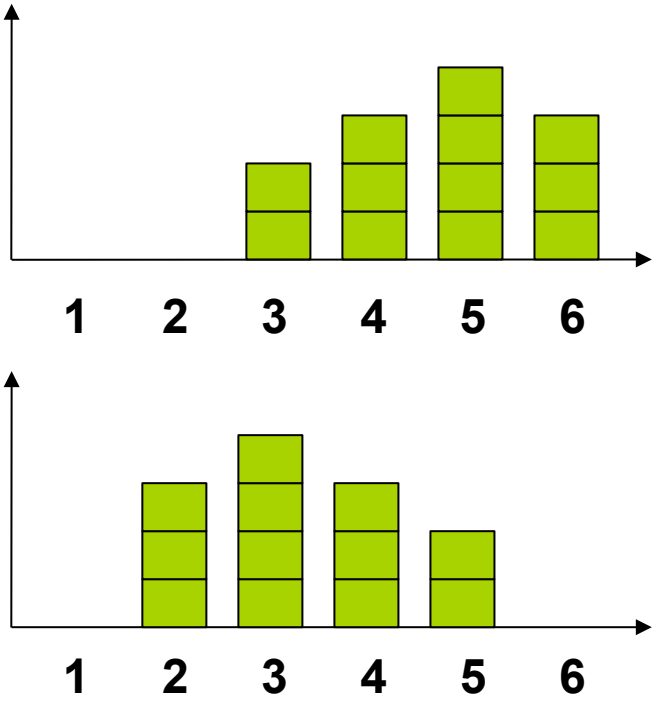
I show you pairs of distributions,
you tell me if the differences are **"statistically different"**



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

<30sec brainstorming>

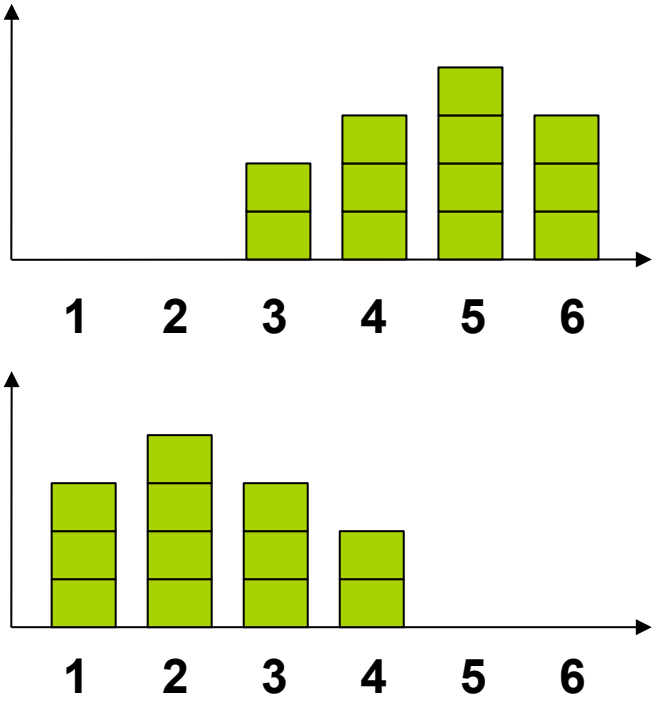
TTEST(A1:A12,B1:B12,2,2) = 0.4548



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

<30sec brainstorming>

TTEST(A1:A12,B1:B12,2,2) = 0.1423



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

<30sec brainstorming>

TTEST(A1:A12,B1:B12,2,2) = 0.00097

A	B
2	2
2	2
3	2
3	3
3	3
4	3
4	3
4	4
4	4
5	4
5	5
5	5

= TTEST(A1:A12,B1:B12,2,2)

(student's) t-test

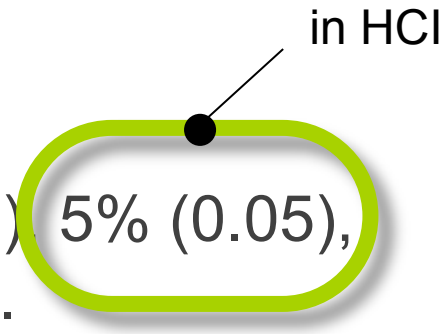
return a **p-value**

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

significance level ::

If a test of significance gives a **p-value lower** than the significance level, the **null hypothesis is thus rejected.** Such results are informally referred to as 'statistically significant'.

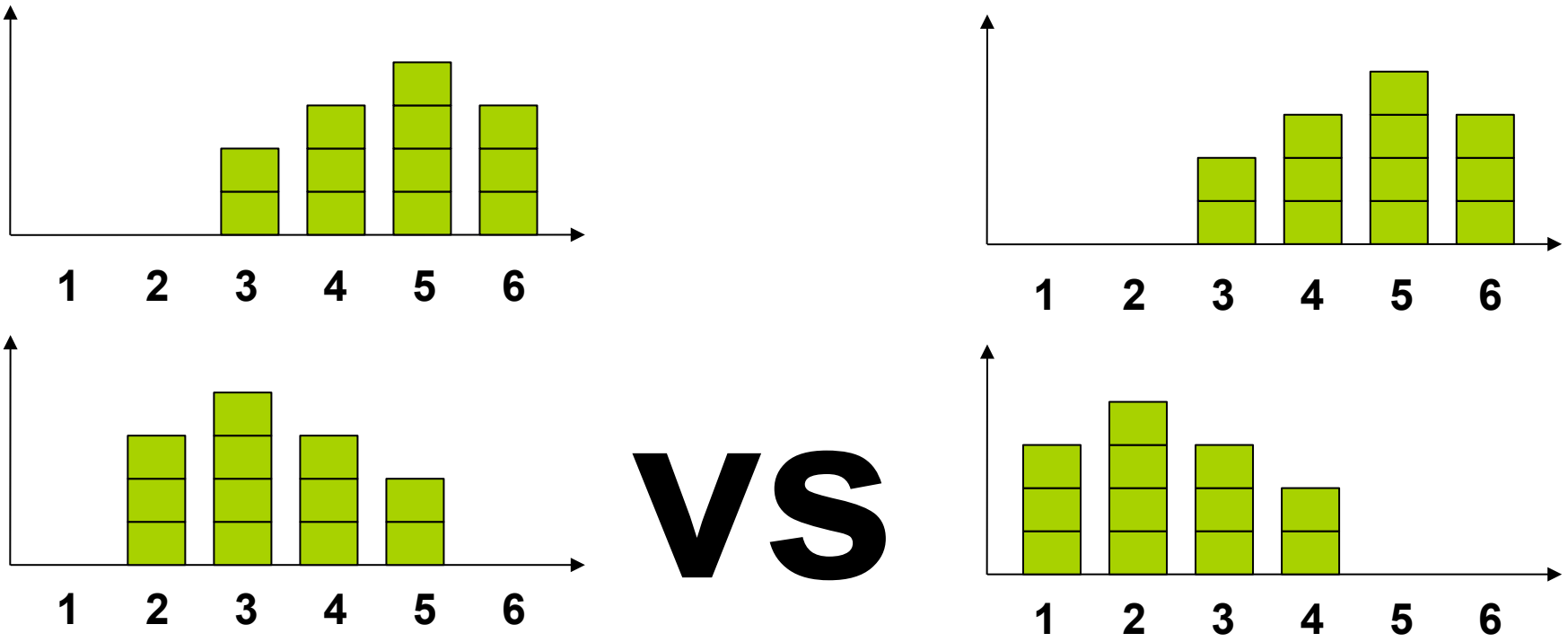
Popular levels of significance are 10% (0.1), 5% (0.05), 1% (0.01), 0.5% (0.005), and 0.1% (0.001).



if someone argues that "there's only **one chance in a thousand** this could have happened by coincidence,"
a 0.001 level of statistical significance is being implied

the lower the significance level,
the stronger the **evidence required.**

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

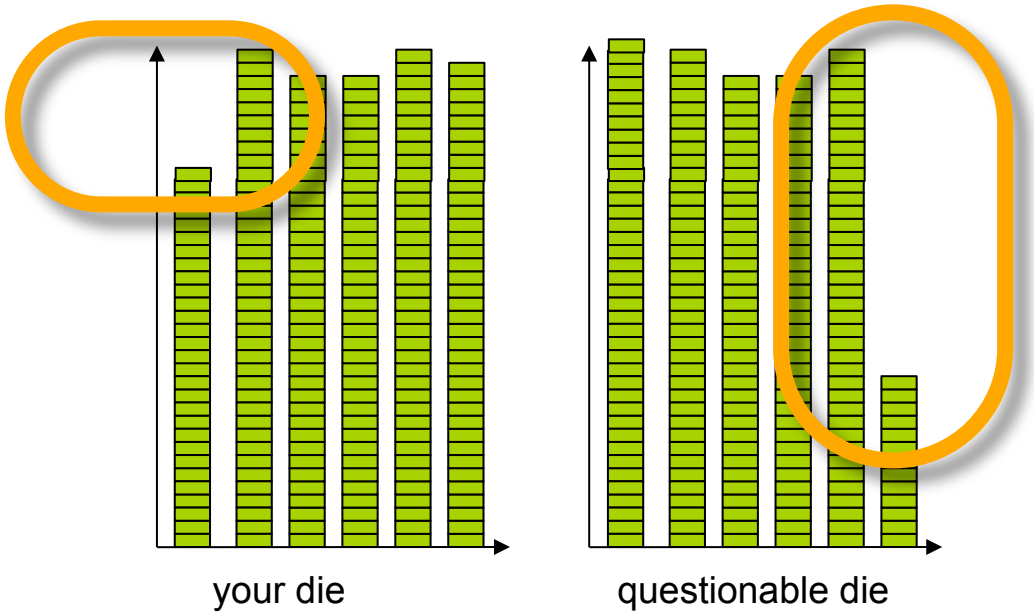


null hypothesis: both data sets are from same mechanism

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…**

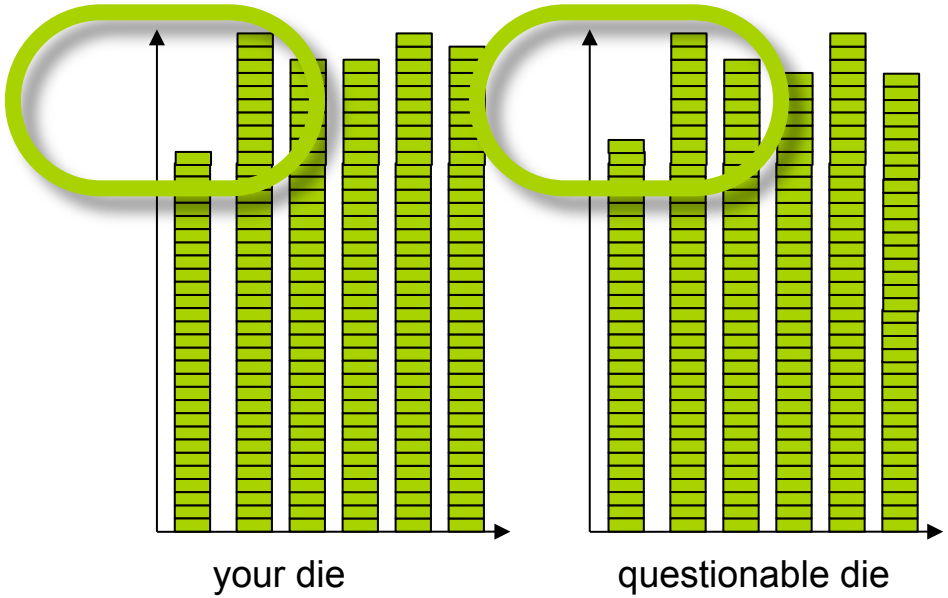
in both cases, there are two explanations

1.same die

2.different dice

…thus this **must be true**

…now, in the other case



this **does seem not unlikely…**

in both cases, there are two explanations

1.same die

2.different dice

we still have **two possible explanations**

→ we **cannot conclude** anything

…thus… **no thus**

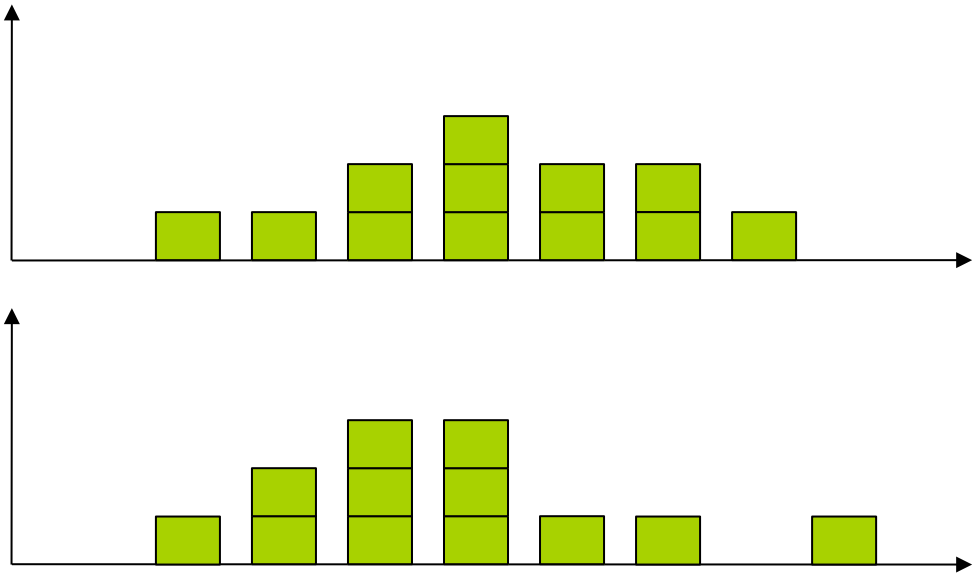
a classic screw-ups

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**”?

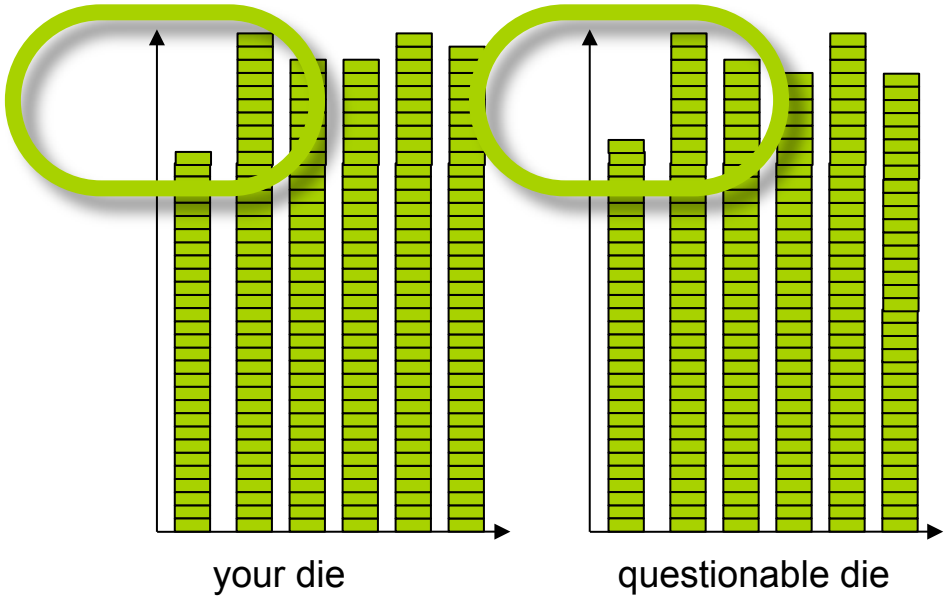


wrong!

significant difference → mechanisms different

no significant difference → **nothing**

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



this **does seem not unlikely…**

in both cases, there are two explanations

1.same die

2.different dice

…thus… **no thus**

we’ve had this: **you cannot**

how to report non-significant results:

“our test did not **find** a significant difference”

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?

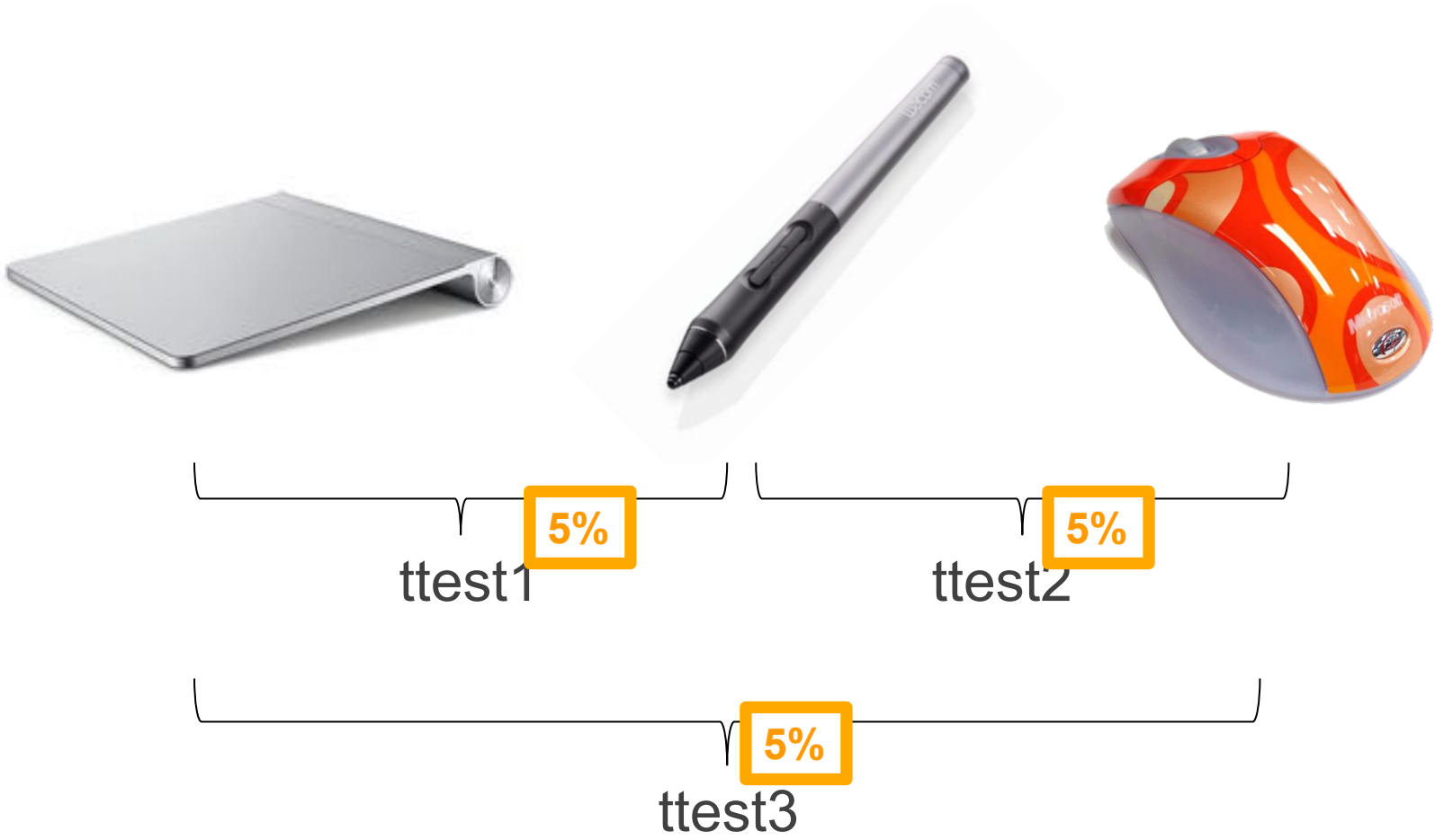
<30sec brainstorming>

ttest1

ttest2

ttest3

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**

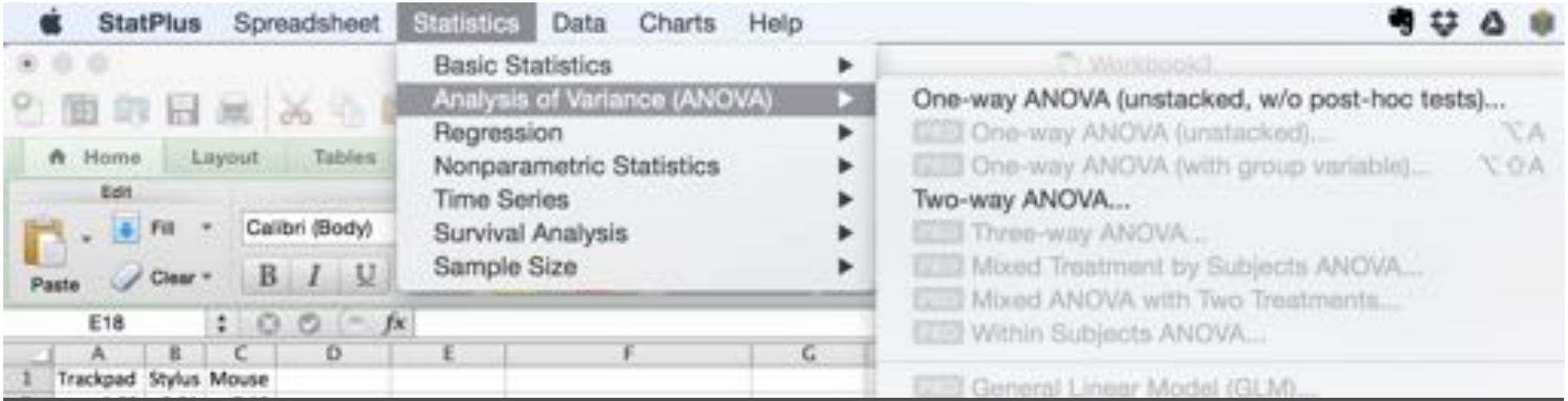
in our example we would need to use **0.05/3** as a significant threshold instead of 0.05

Anova::

analyze of variance to compare multiple variables

one-way anova = one variable with multiple levels

two-way anova = two variables with multiple levels



run data analysis with excel, statplus for mac users

also many other ways: R, SPSS, Matlab, Stata etc.

Analysis of Variance (One-Way)

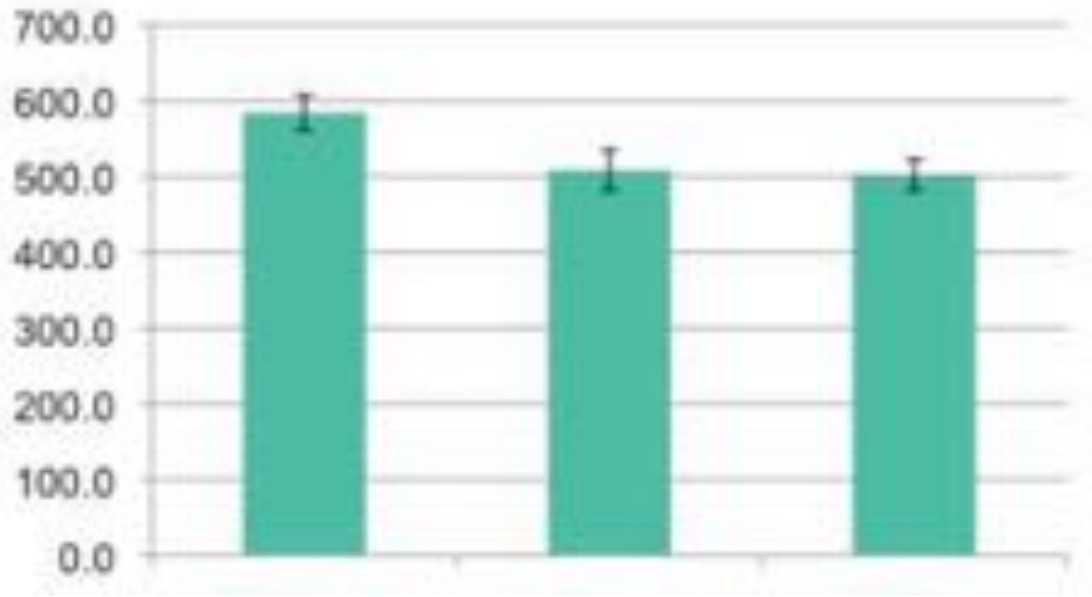
Summary

Groups	*Sample size*	*Sum*	*Mean*	*Variance*
1	5	9.	1.8	0.2
3	5	9.	1.8	0.7
5	5	28.	5.6	1.3

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*p-level*	*F crit*
Between Groups	48.13333	2	24.06667	32.81818	0.00136%	3.88529
Within Groups	8.8	12	0.73333			
Total	56.93333	14				

"An one-way ANOVA showed a significant effect on time for the variable Input device ($F_{2,12}$=32.81818, $p < 0.05$)."



so the anova tells us the choice of input device affects the time but how? By itself the test **doesn't say *which* differences are significant.**

For a statistical test of this question, we can use the Tukey HSD (Honestly Significant Difference) test, which is also sometimes called the **Tukey post-hoc test**.

(don't need to adjust your significance level for it)

Tukey HSD Test
for Post-ANOVA Pair-Wise Comparisons
in a One-Way ANOVA

After performing a one-way analysis of variance, enter the values outlined in red, then click the Calculate button:

	Sample			
	A	B	C	D
Mean	2.02879	2.97871	6.1051	
n	10	10	10	

n = number of measures per sample. If there are only three samples in the analysis, leave the D entries blank.

MS_{error} = 0.47611

df_{error} = 27

Clear Calculate

$HSD_{.05}$ = 0.77

$HSD_{.01}$ = 0.98

Pair-Wise Comparisons via Tukey HSD Test

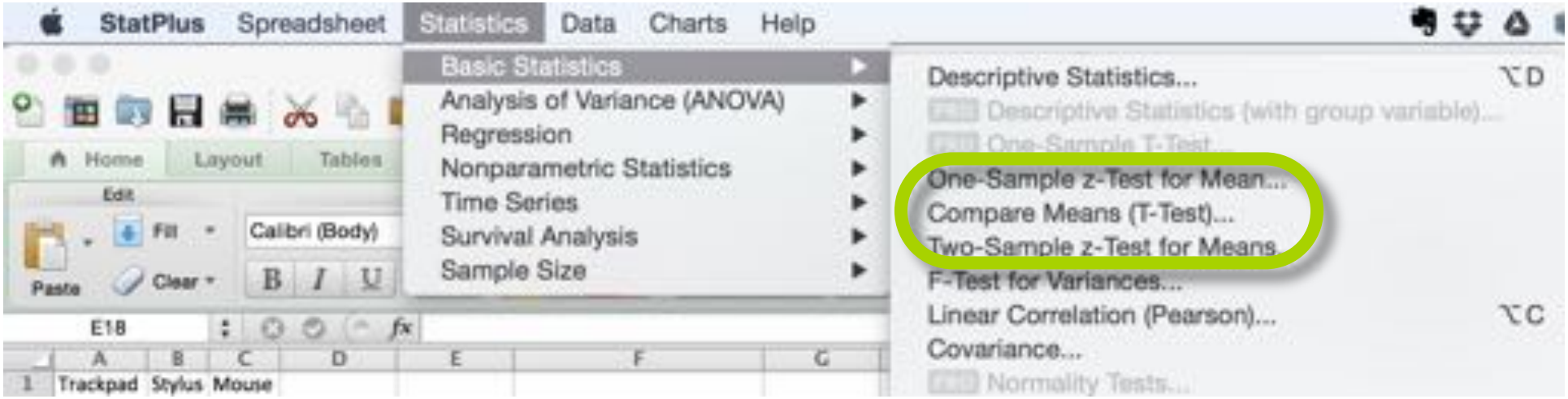
	B	C	D
A	P<.05	P<.01	
B		P<.01	
C			

n/s = non-significant

enter here the results from the anova

Tukey post-hoc test online

http://faculty.vassar.edu/lowry/hsd.html



btw excel more detailed ttest: **run data analysis** with excel, **statplus** for mac users

also many other ways: R, SPSS, Matlab, Stata etc.

Comparing Means [Paired two-sample t-test]

Descriptive Statistics

VAR	*Sample size*	*Mean*	*Variance*
Trackpad	10	2.02879	0.30966
Stylus	10	2.97871	0.38647

Summary

Degrees Of Freedom	9	*Hypoth sized Mean Difference*	0.
Test Statistics	3.61197	*Pooled Variance*	0.34807

Two-tailed distribution

p-level	0.00564	*t Critical Value (5%)*	2.26216

One-tailed distribution

p-level	0.00282	*t Critical Value (5%)*	1.83311
Pearson Correlation Coefficient	0.00647		

G-criterion

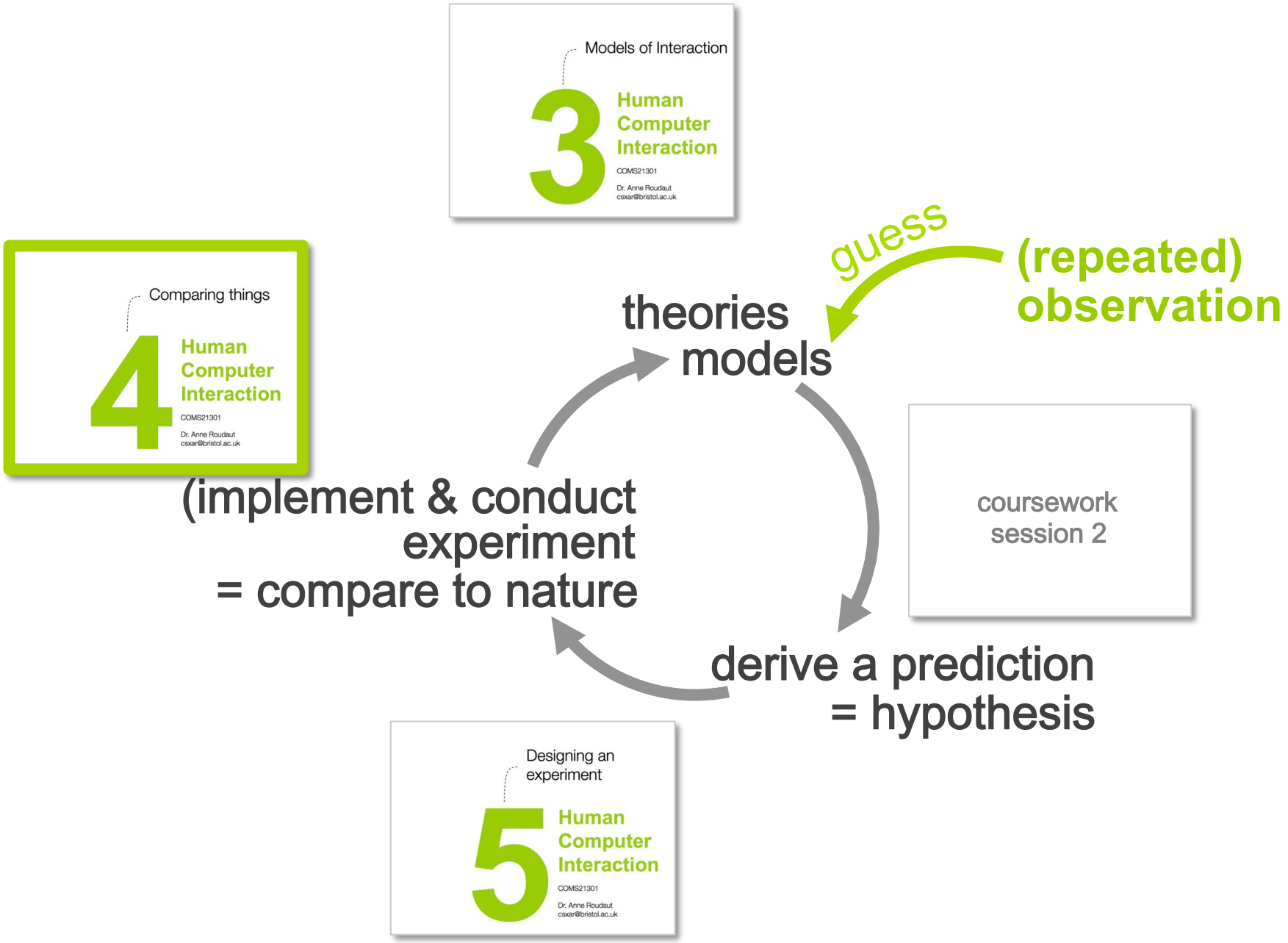
Test Statistics	0.57633	*p-level*	0.00054
Critical Value (5%)	0.25		

Pagurova criterion

	3.60034	*p-level*	0.99791
	0.44483	*Critical Value (5%)*	0.0636

will explain that next time

"A paired student t-test showed significant difference between Trackpad and Stylus (two-tailed =3.61197, df=9, p = 0.005)"


Models of Interaction
3
Human Computer Interaction
COMS21301
Dr. Anne Roudaut
csxar@bristol.ac.uk
Comparing things
4
Human Computer Interaction
COMS21301
Dr. Anne Roudaut
csxar@bristol.ac.uk
guess
(repeated) observation
theories models
derive a prediction = hypothesis
(implement & conduct experiment = compare to nature
coursework session 2
Designing an experiment
5
Human Computer Interaction
COMS21301
Dr. Anne Roudaut
csxar@bristol.ac.uk

end