12

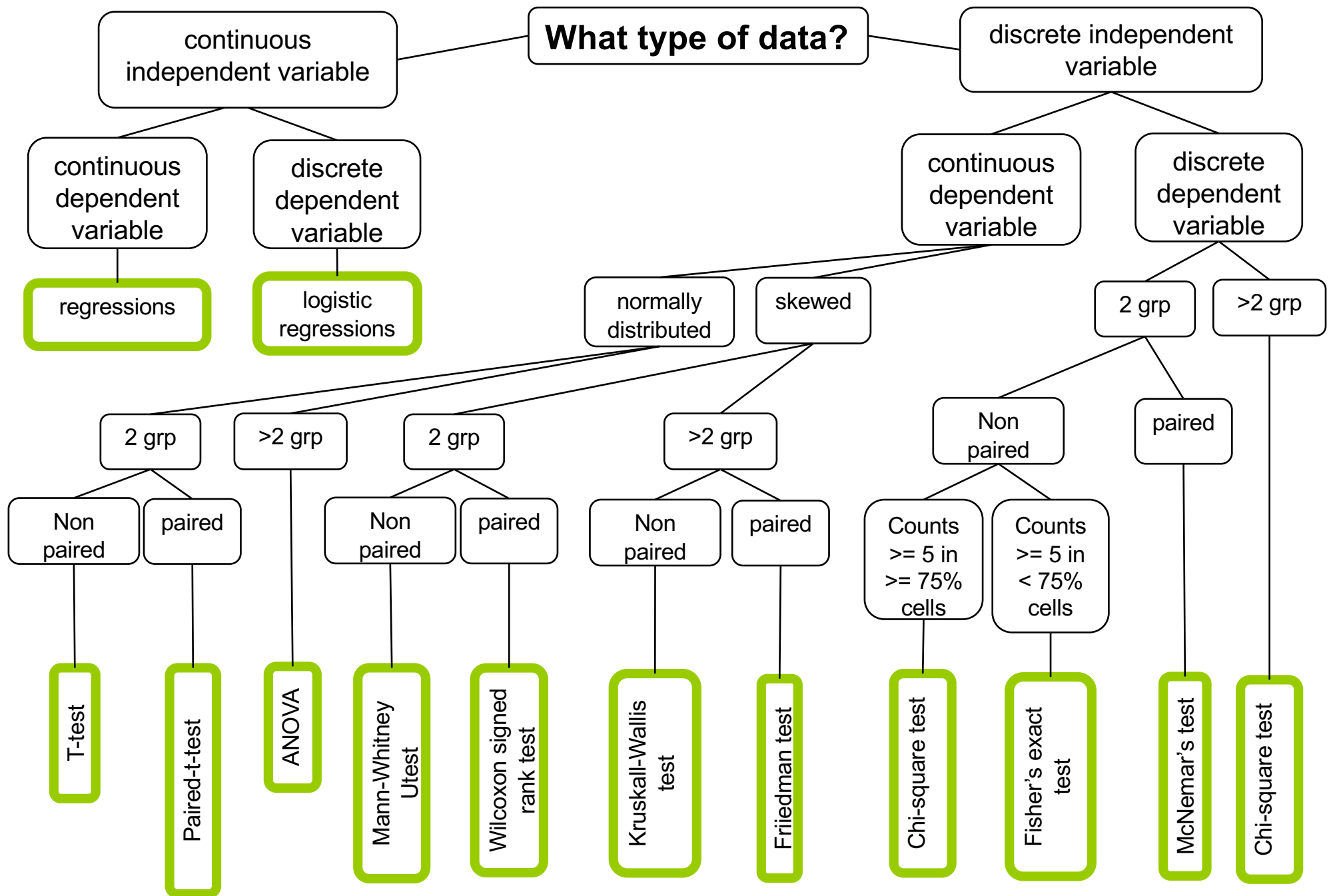**Comparing things**
and hypothesis testing
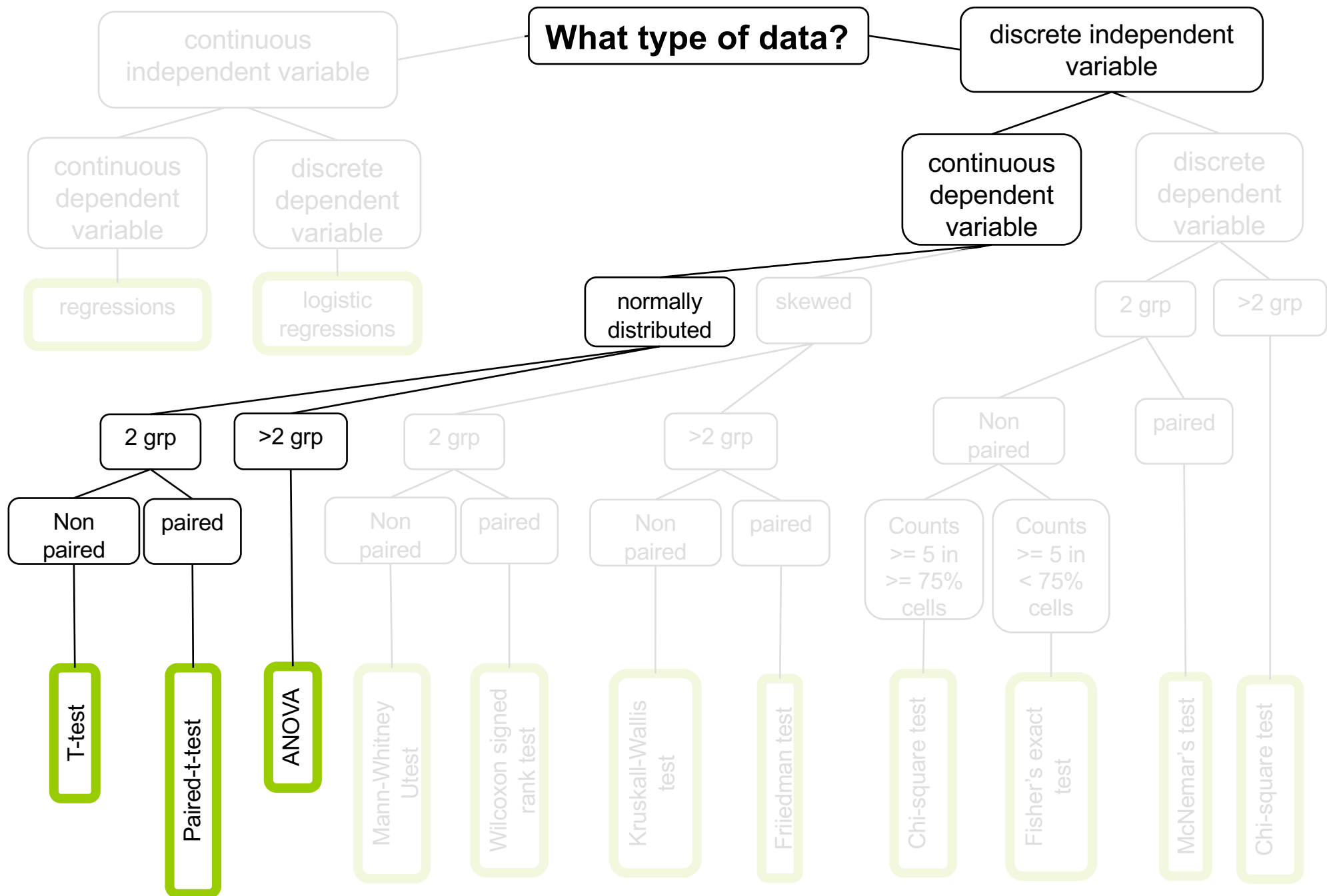
Probability and Statistics

COMS10011

Dr. Anne Roudaut
csxar@bristol.ac.uk

What type of data?

continuous independent variable
- **continuous dependent variable** → regressions
- **discrete dependent variable** → logistic regressions

discrete independent variable
- **continuous dependent variable**
  - **normally distributed**
    - **2 grp**
      - Non paired → T-test
      - paired → Paired-t-test
    - **>2 grp** → ANOVA
  - **skewed**
    - **2 grp**
      - Non paired → Mann-Whitney U test
      - paired → Wilcoxon signed rank test
    - **>2 grp**
      - Non paired → Kruskall-Wallis test
      - paired → Friiedman test
- **discrete dependent variable**
  - **2 grp**
    - **Non paired**
      - Counts >= 5 in >= 75% cells → Chi-square test
      - Counts >= 5 in < 75% cells → Fisher's exact test
    - **paired** → McNemar's test
  - **>2 grp** → Chi-square test

What type of data?

continuous independent variable

discrete independent variable

continuous dependent variable — **continuous dependent variable**

discrete dependent variable

regressions

logistic regressions

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

Friiedman test

Chi-square test

Fisher's exact test

McNemar's test

Chi-square test

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
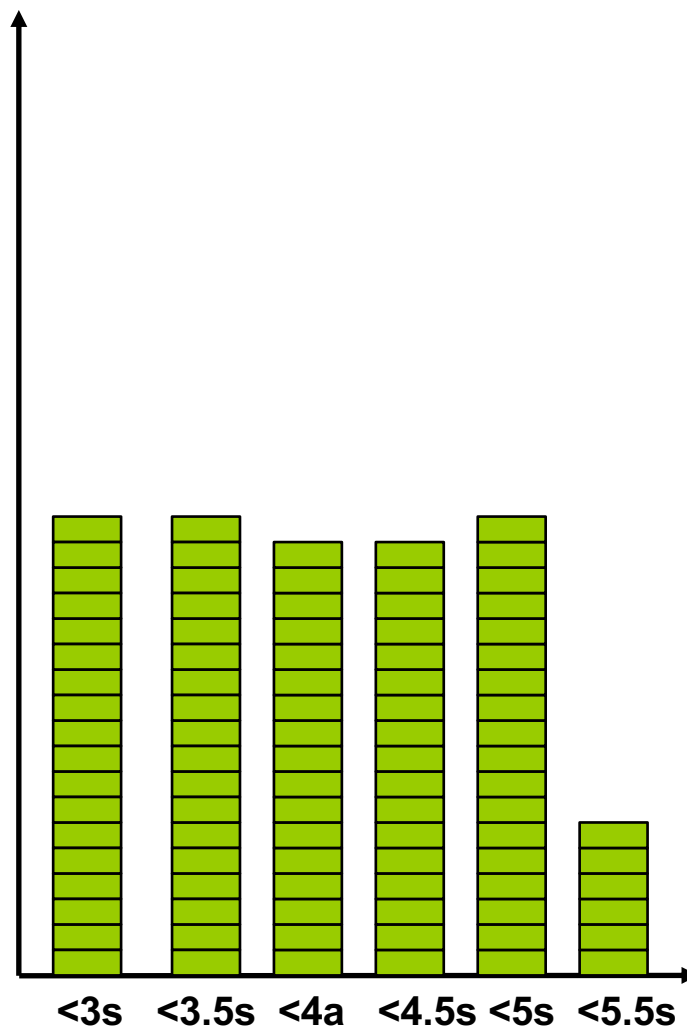
questionable shoes
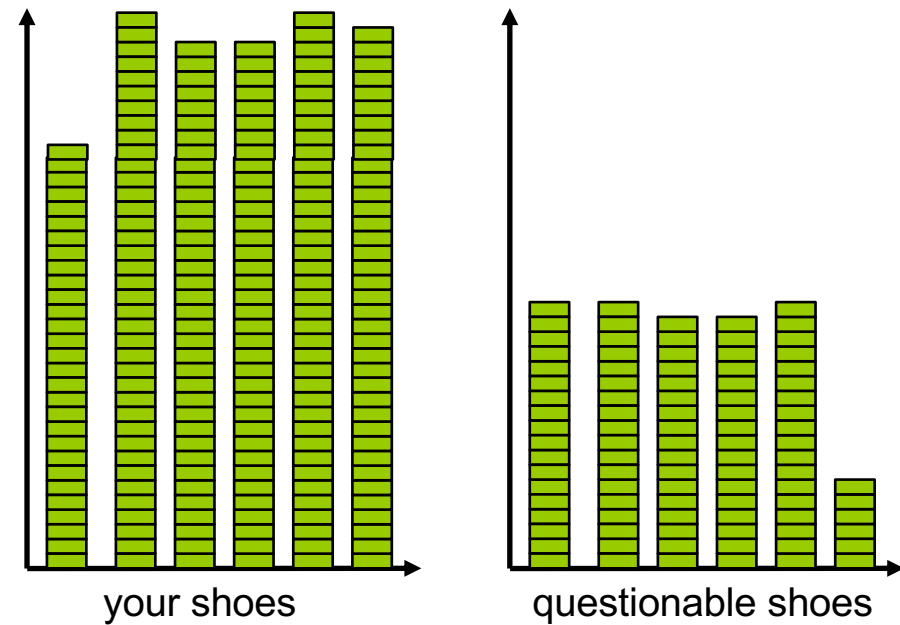
[ ] this is still the original shoes

[ ] someone has replaced my shoes

[x] could be the original or replacement shoes
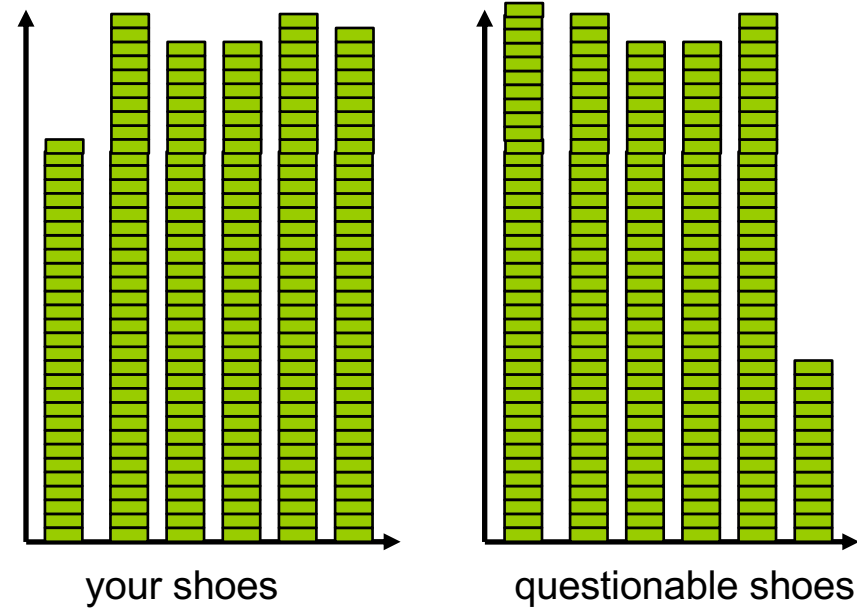
<let's vote>

your die   questionable die
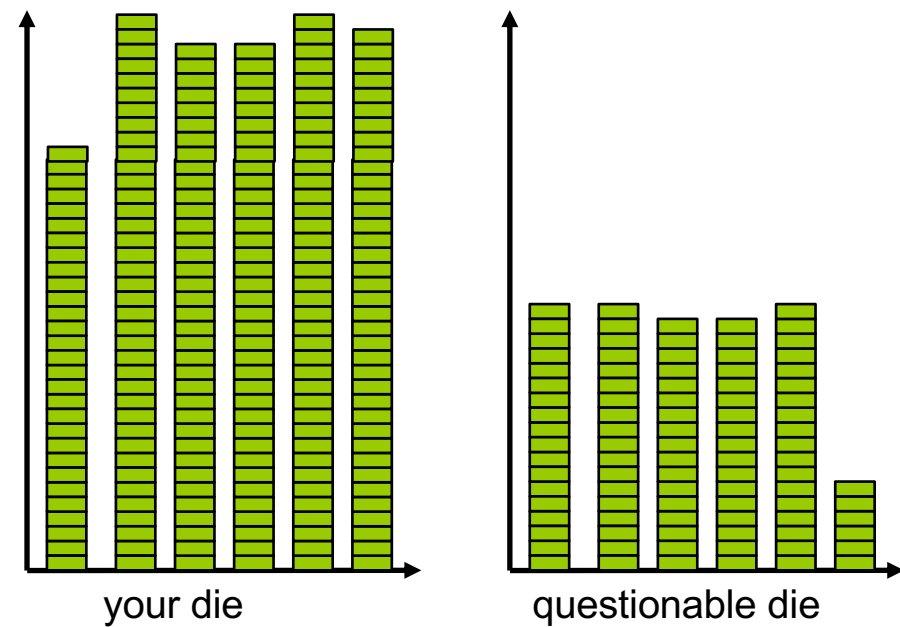
the distribution looks different from the shoes you know

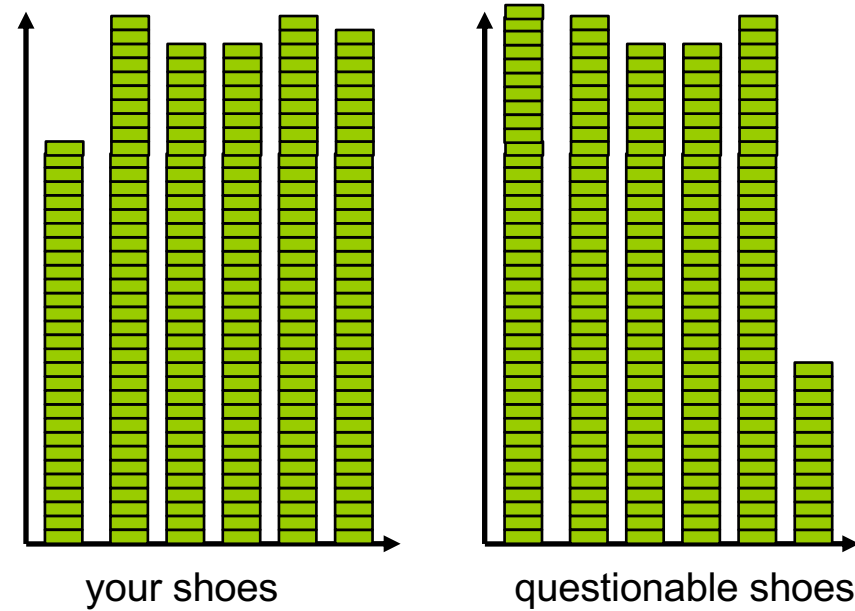while it is **possible** that it is the same shoes, it seems somewhat **unlikely**

ok, let's run some more…

your shoes

<3s <3.5s <4a <4.5s <5s <5.5s

questionable shoes

<3s <3.5s <4a <4.5s <5s <5.5s
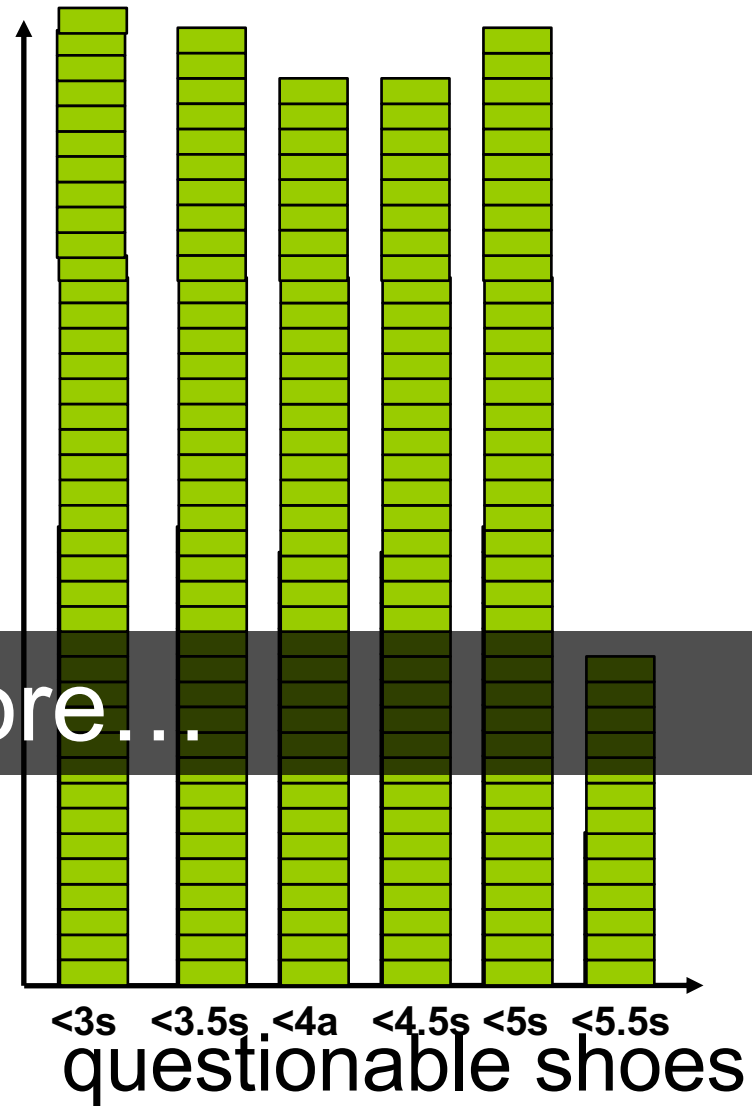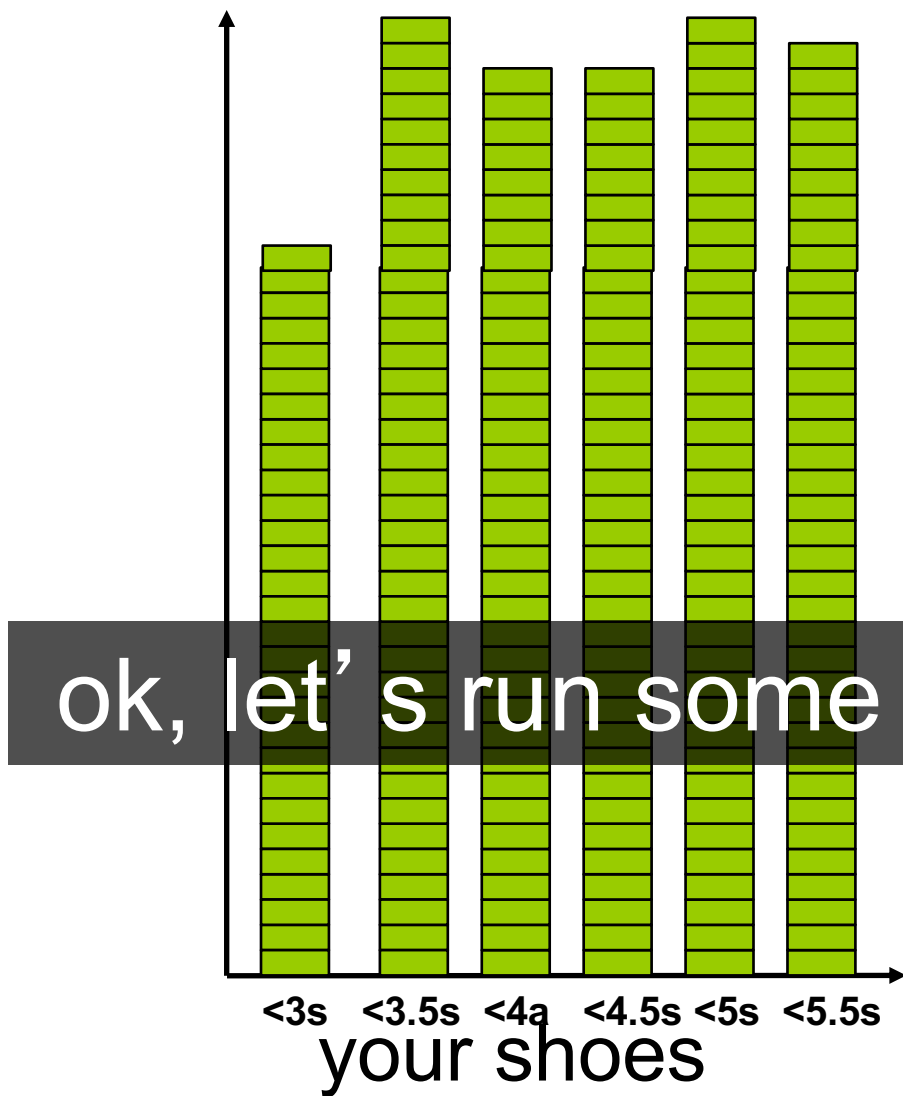
[ ] this is probably still the original shoes

[x] someone probably has replaced my shoes

[x] could be the original or replacement shoes

<let's vote>

your shoes   questionable shoes

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

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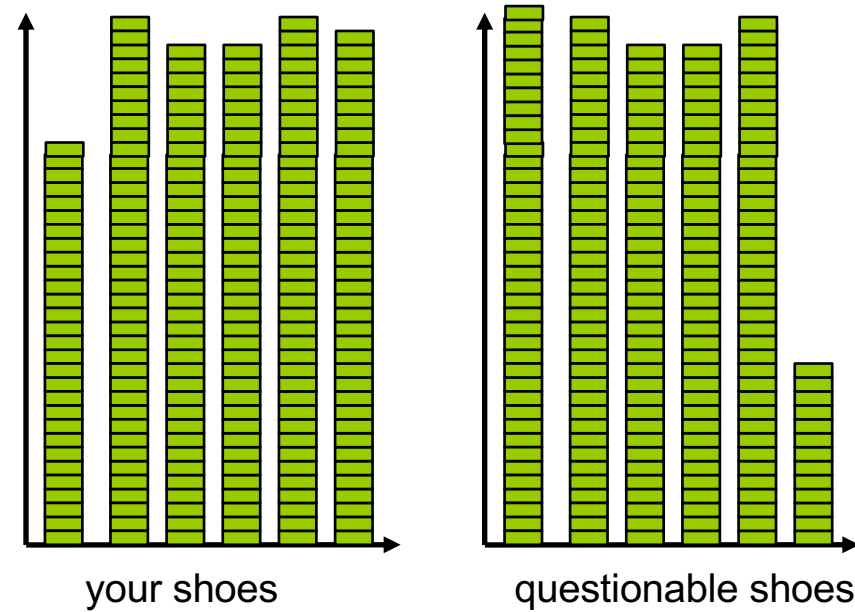
again, the distribution could have happened by chance, but it seems **even more unlikely.** This is **probably not** your shoes

 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
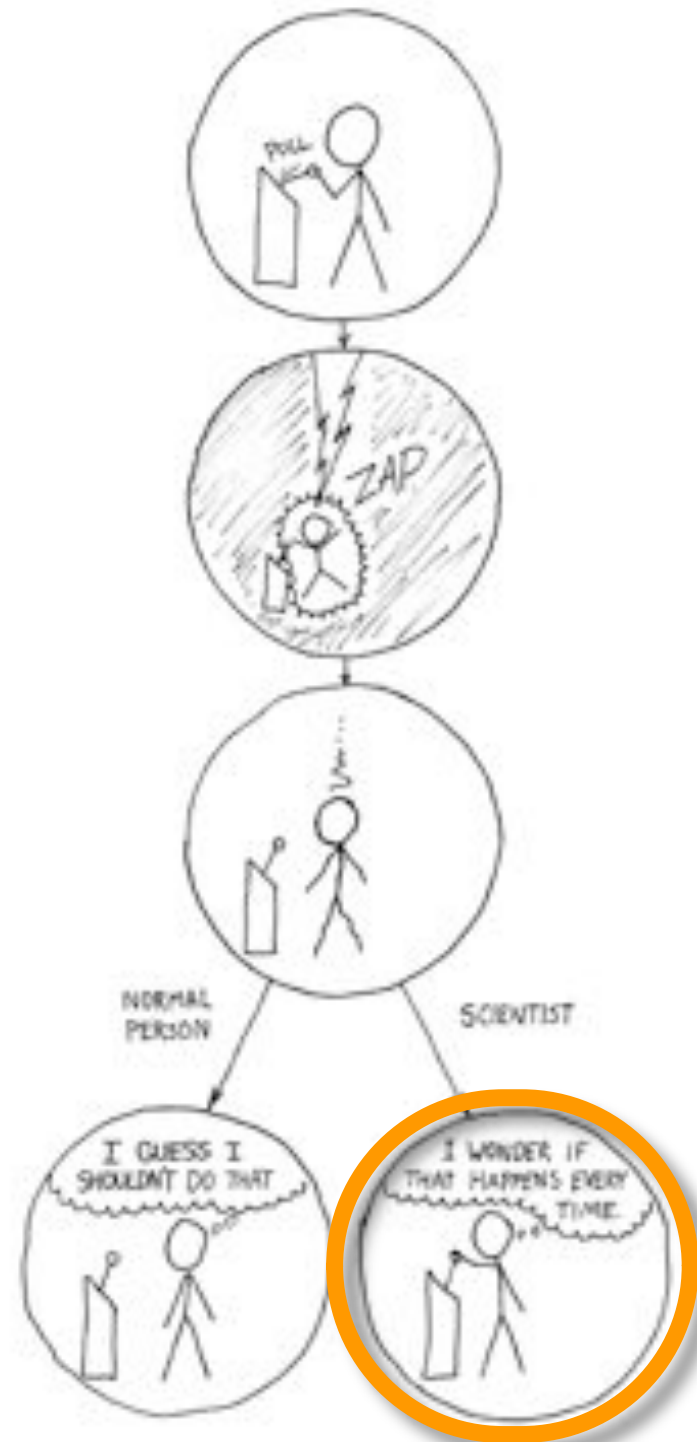
the good news:

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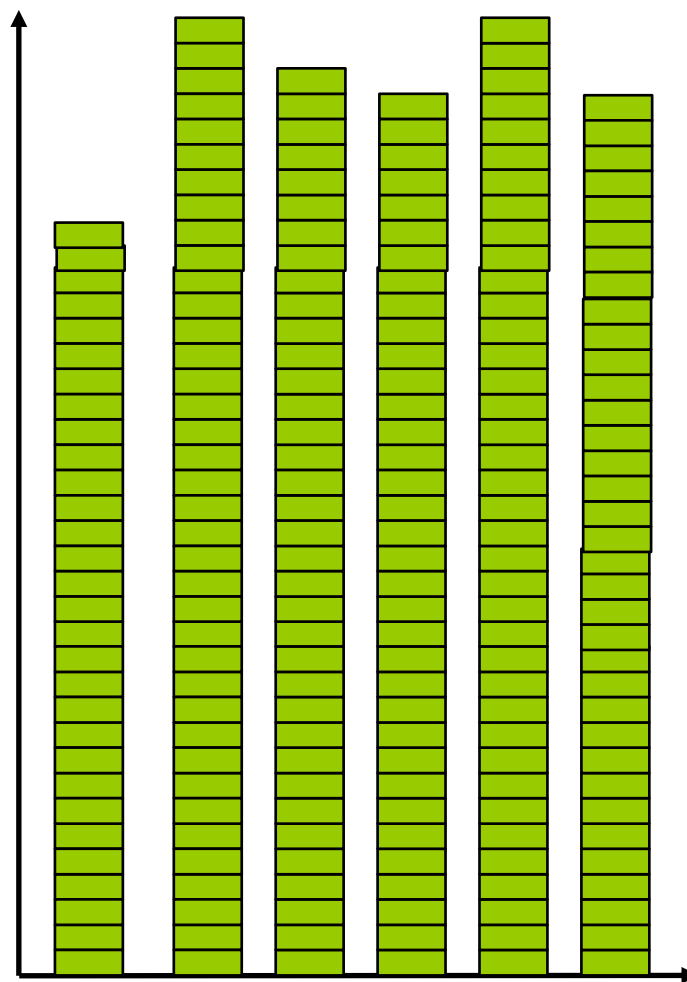
→ you can be **arbitrarily sure**

another round
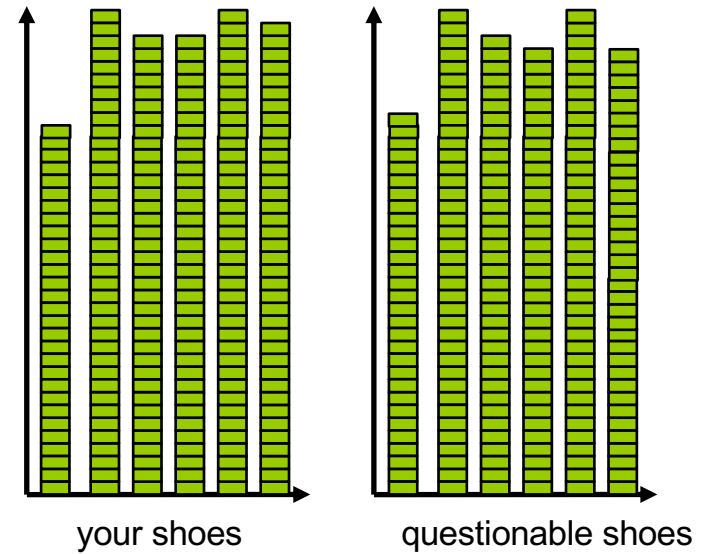
ok, you get your original shoes back

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

your shoes   questionable shoes
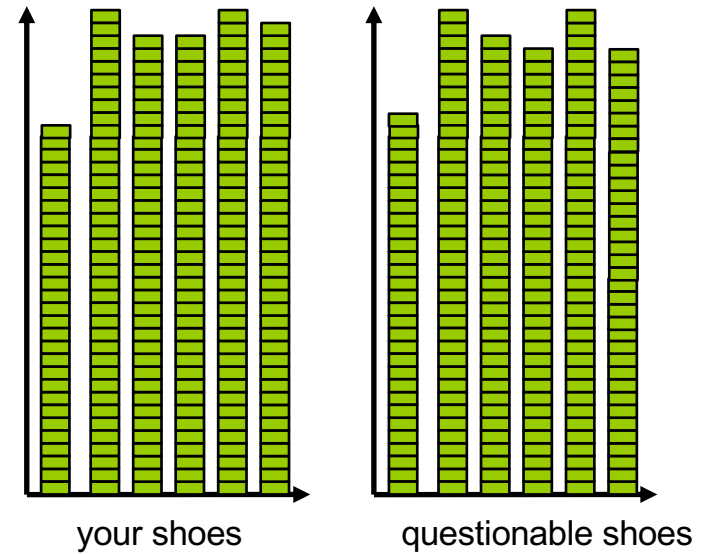
your shoes   questionable shoes

[ ] this is still the original shoes

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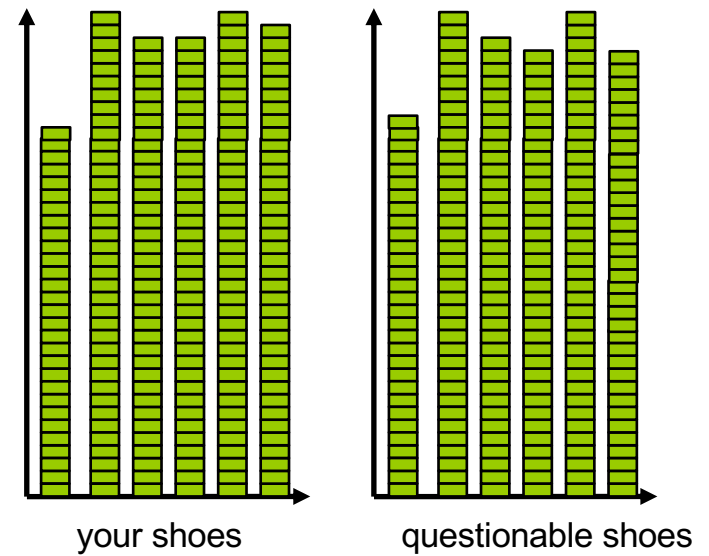
[ ] could be the original or replacement shoes

\<let's vote\>

your shoes   questionable shoes

[ ] this is still the original shoes

[ ] someone has replaced my shoes

[x] could be the original or replacement shoes
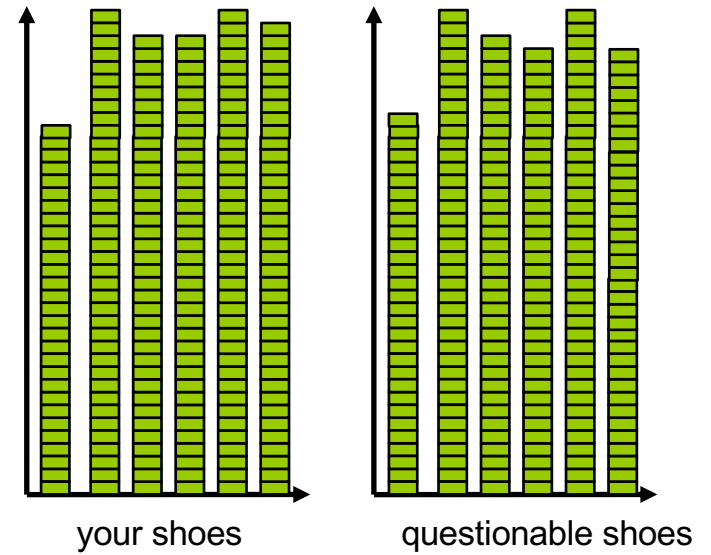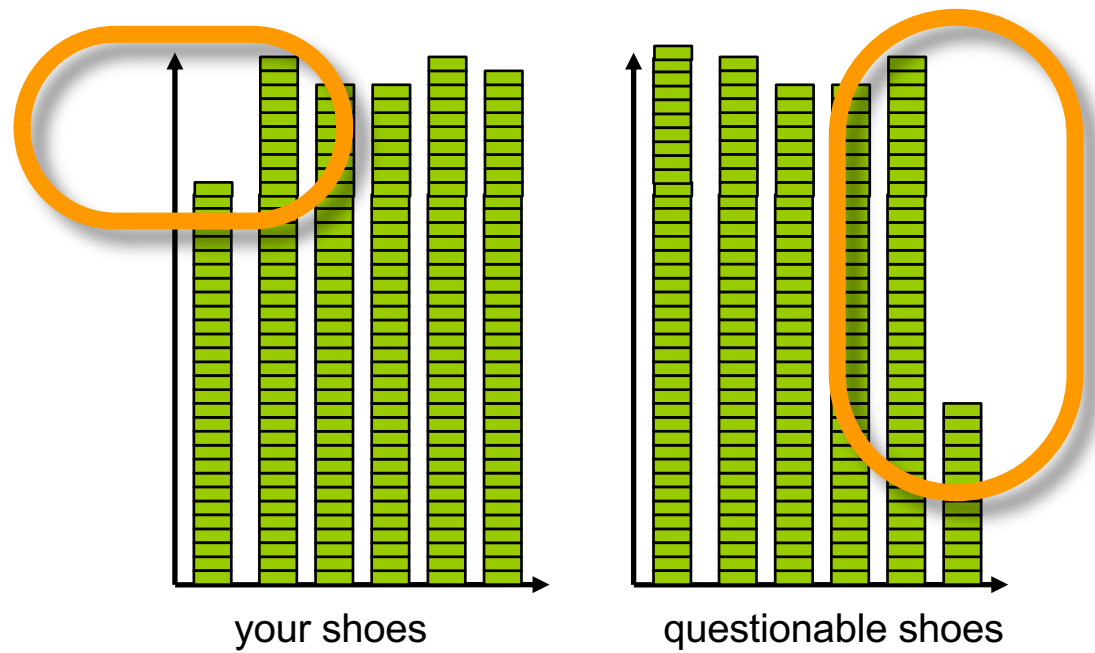
your shoes   questionable shoes

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.
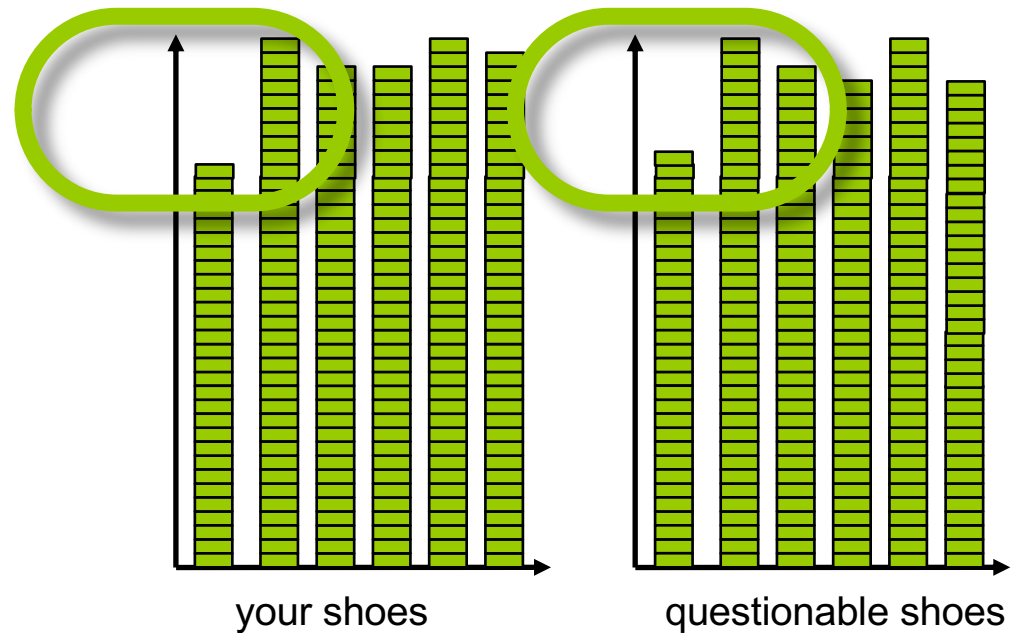
that's strange! why not?

your shoes · questionable shoes

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

your shoes   questionable shoes

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**
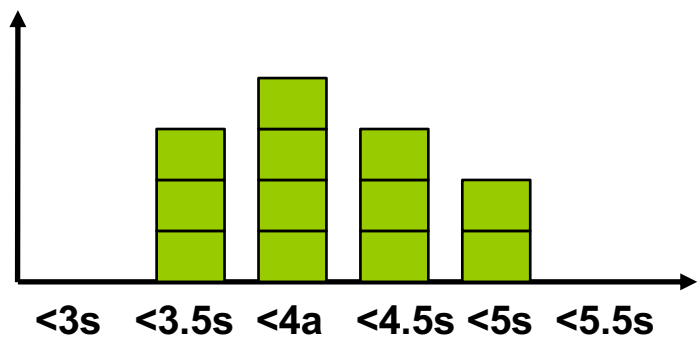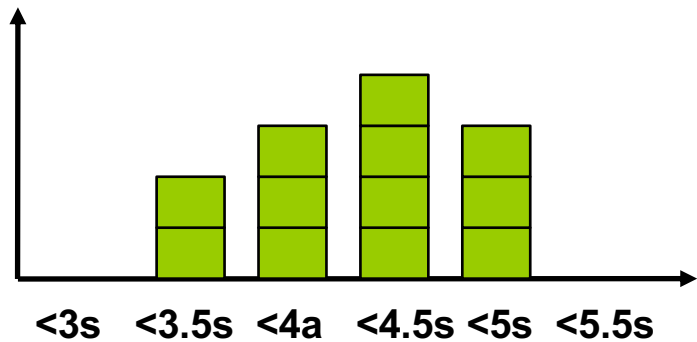→ we **cannot conclude** anything

let's use stats

statistical significance ::

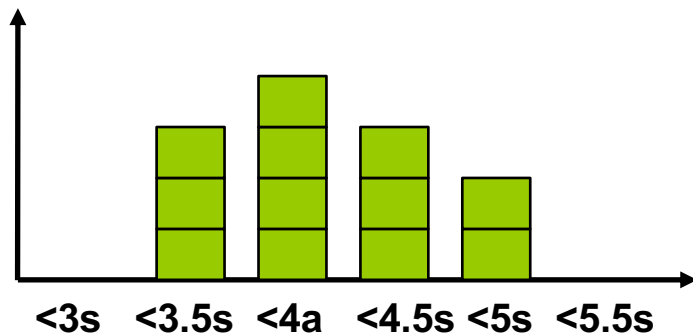a result is called statistically significant if it is **unlikely to have occurred by chance**

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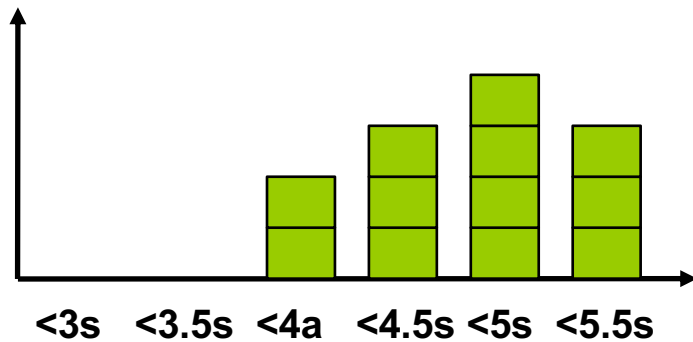
still could have happened **by chance**
(14% dissimilar)

<30sec brainstorming>

TTEST pvalue = 0.1423
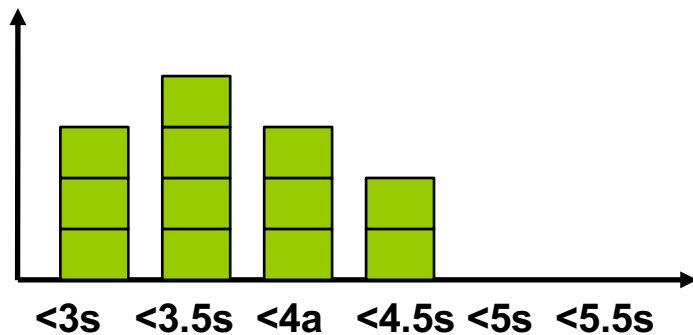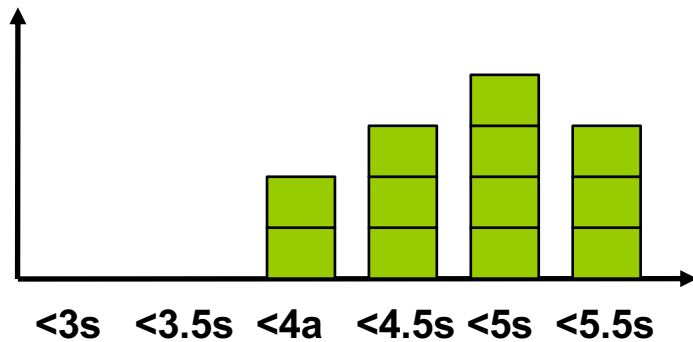
unlikely to have happened **by chance**
(0.1% dissimilar)

<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…**
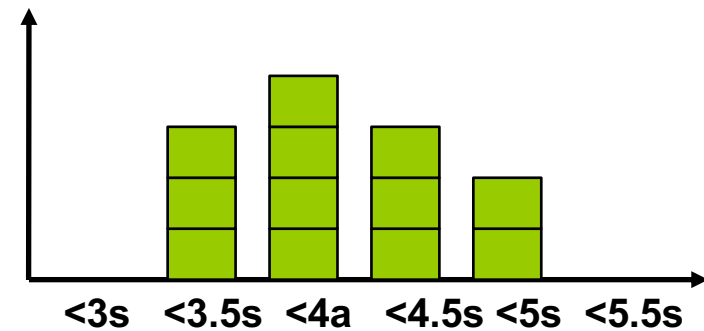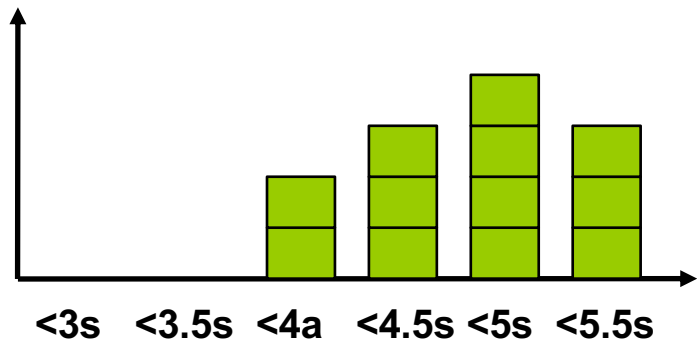
null hypothesis: both data sets are from same mechanism

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).**

your shoes

questionable shoes

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

your shoes

questionable shoes

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**

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**"?
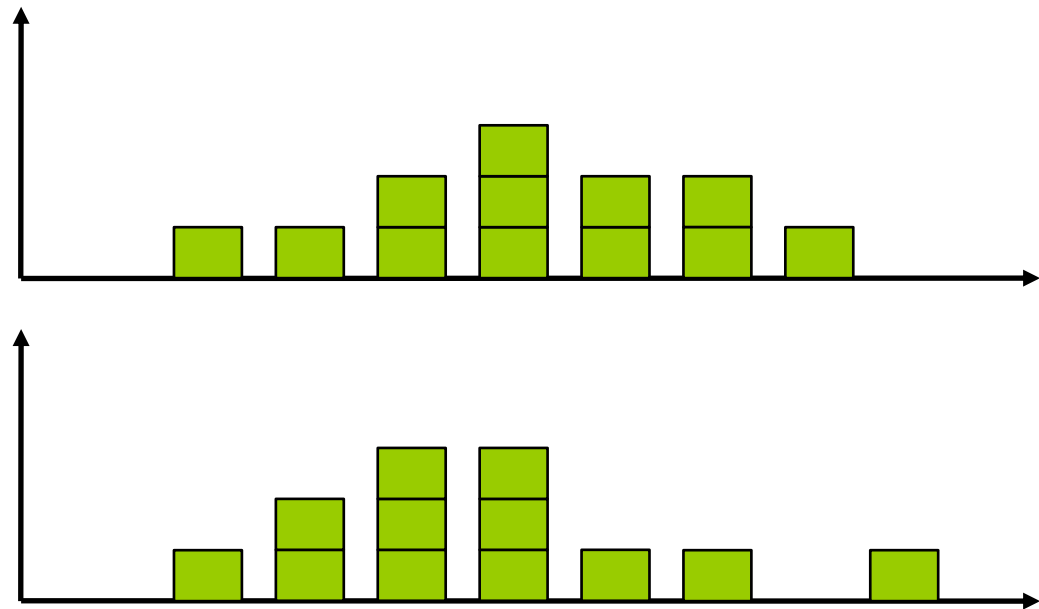
nope!

significant difference → mechanisms different
no significant difference → **nothing**

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

your shoes

questionable shoes
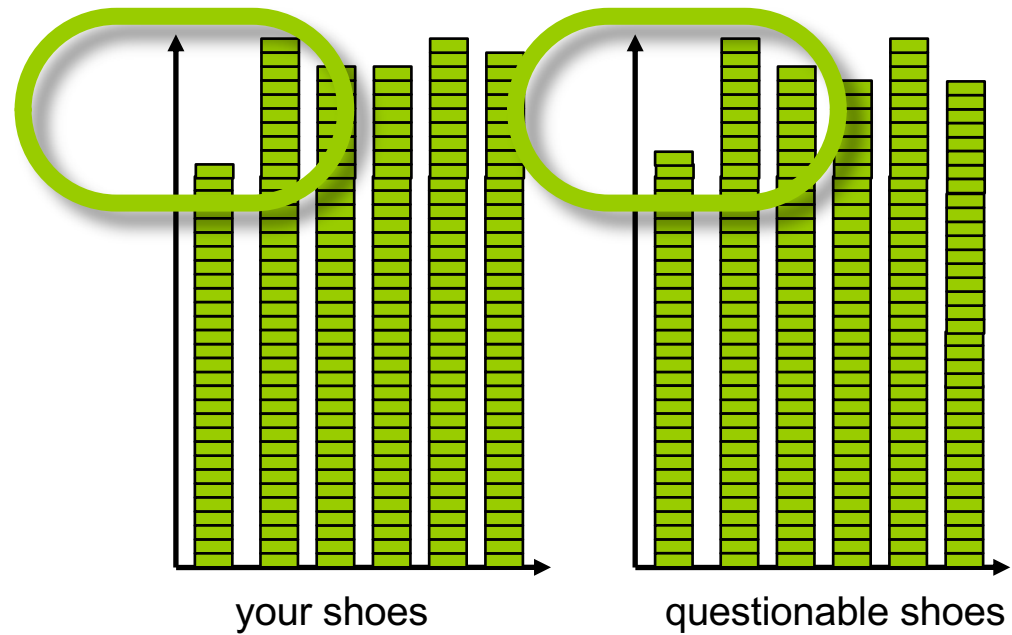
in both cases, there are two explanations

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

```
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

"A paired 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!)

```
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>

ttest1   ttest2

ttest3

a simple solution would be to do this …

5% ttest1

5% ttest2

5% ttest3
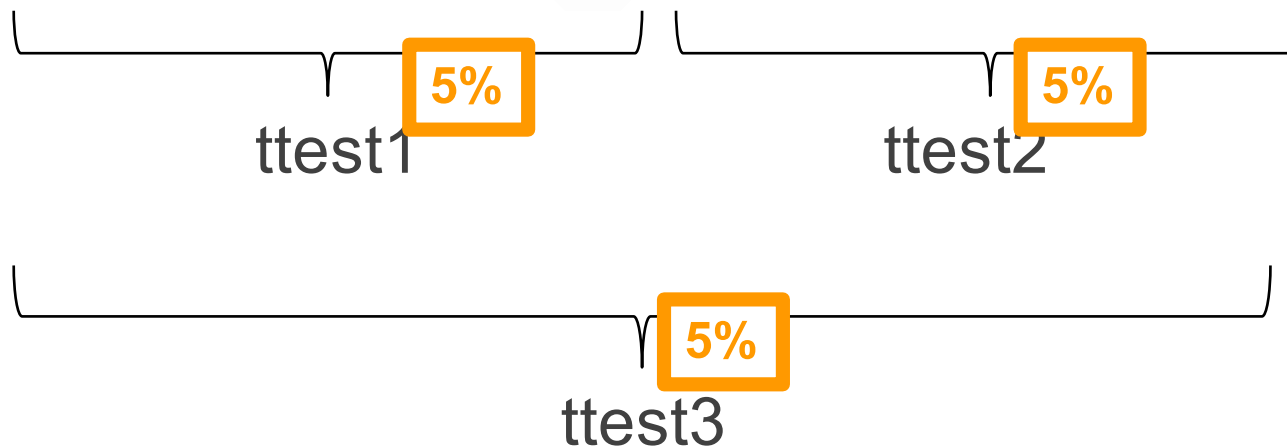
a simple solution would be to do this …

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
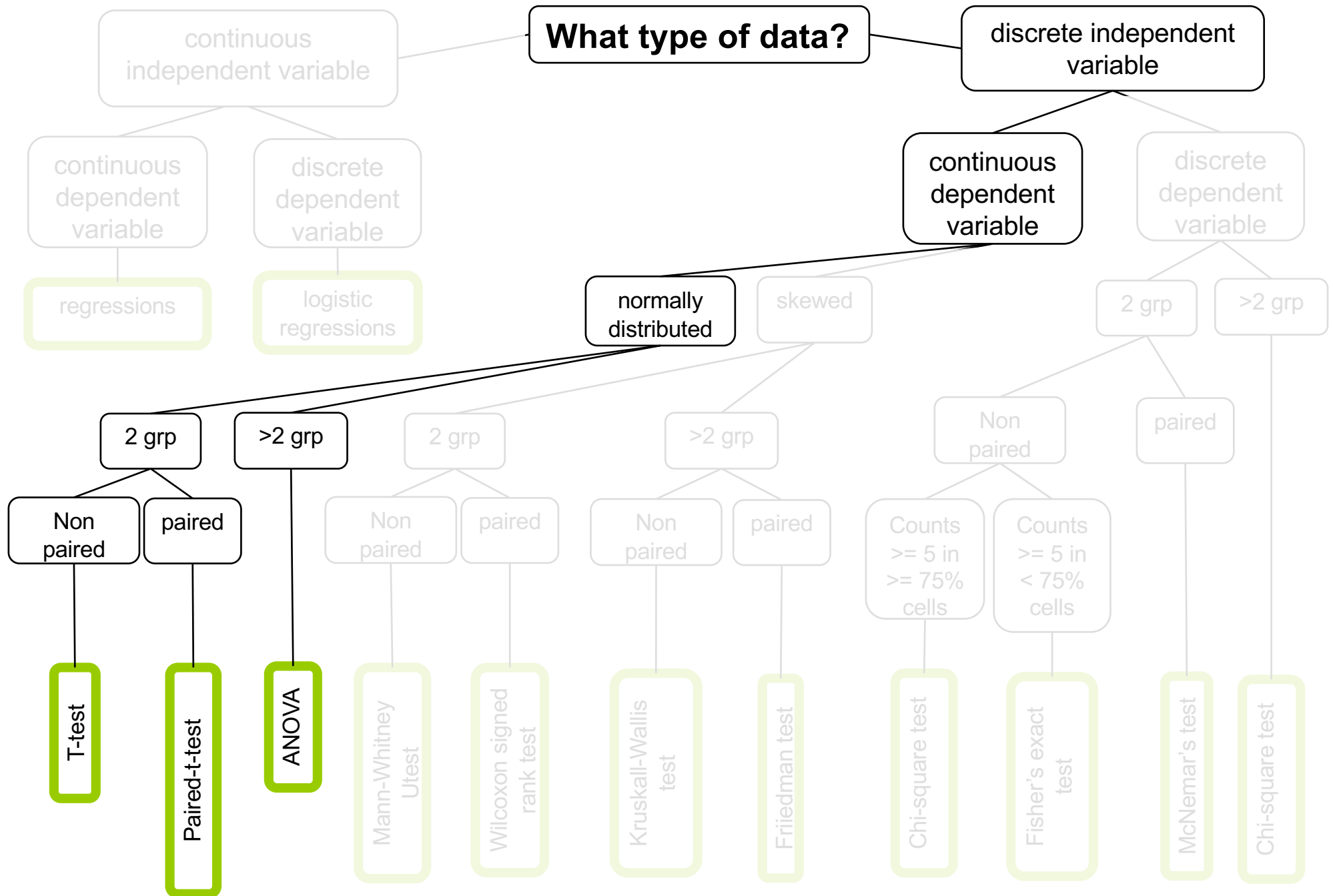
or you could also use an

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 few weeks

What type of data?

continuous independent variable

discrete independent variable

continuous dependent variable

discrete dependent variable

regressions

logistic regressions

normally distributed

skewed

2 grp

>2 grp

2 grp

>2 grp

Non paired

paired

2 grp

>2 grp

Non paired

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

Friiedman test

Chi-square test

Fisher's exact test

McNemar's test

Chi-square test

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

exam

end