

# 16

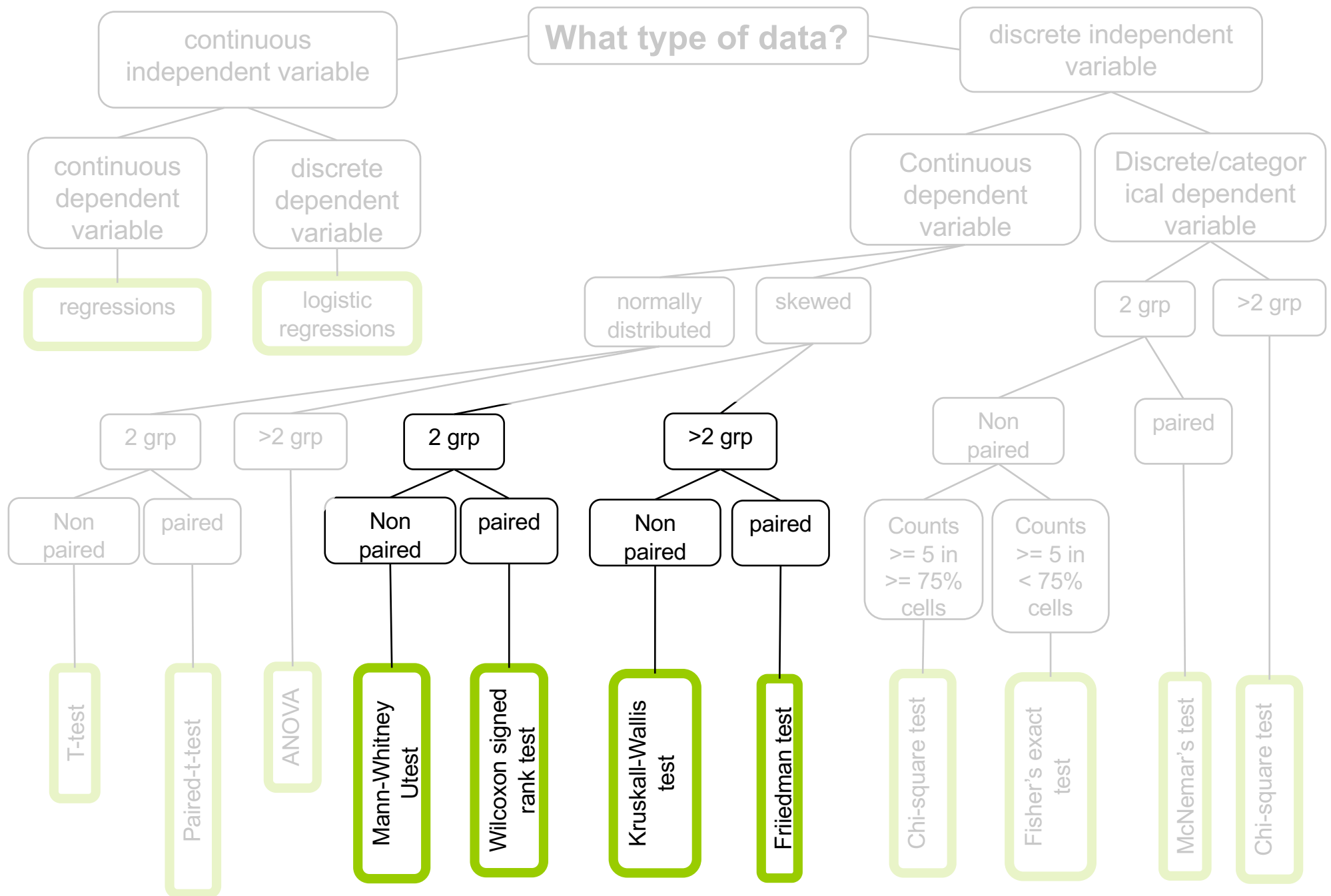
Non-Parametric  
Statistical tests

## Probability and Statistics

COMS10011

Dr. Anne Roudaut  
csxar@bristol.ac.uk

(Thanks S. Massa, Oxford)



# today::

we will look at **four non-parametric tests**

do some recap games about **what tests to use and when**

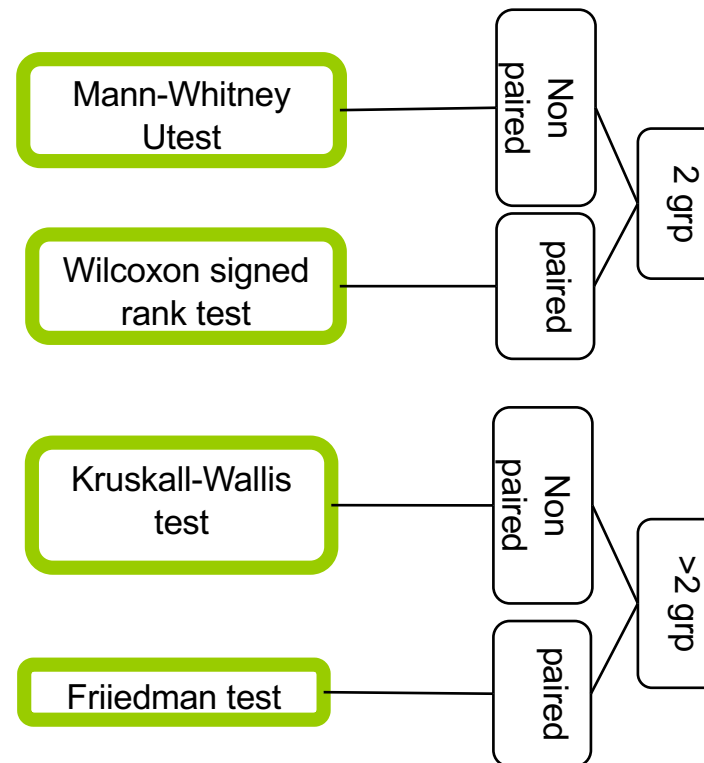
comparing the means of two populations is very important

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.  **$R(\text{drug B}) = 9$**



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

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

... why two values?

15,41	the top gives the 10% critical values = <b>one-tail test</b>
28,50	
26,52	the bottom the 5% ones = <b>two-tail test</b>

$R = 9 < 26.52$  (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  $48.5 > T_{\text{crit}}$  of 34, we can't  
reject the null hypothesis, therefore  
**effect of these seats are not  
significantly different**

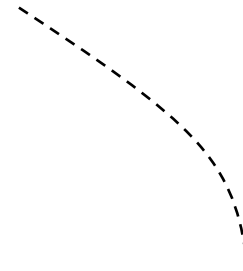


rather simple no?

**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)

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

ANOVA between subject equivalent



# Kruskal Wallis

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

**ANOVA within subject (also called  
repeated measure ANOVA) equivalent**



# Friedman

**practically**

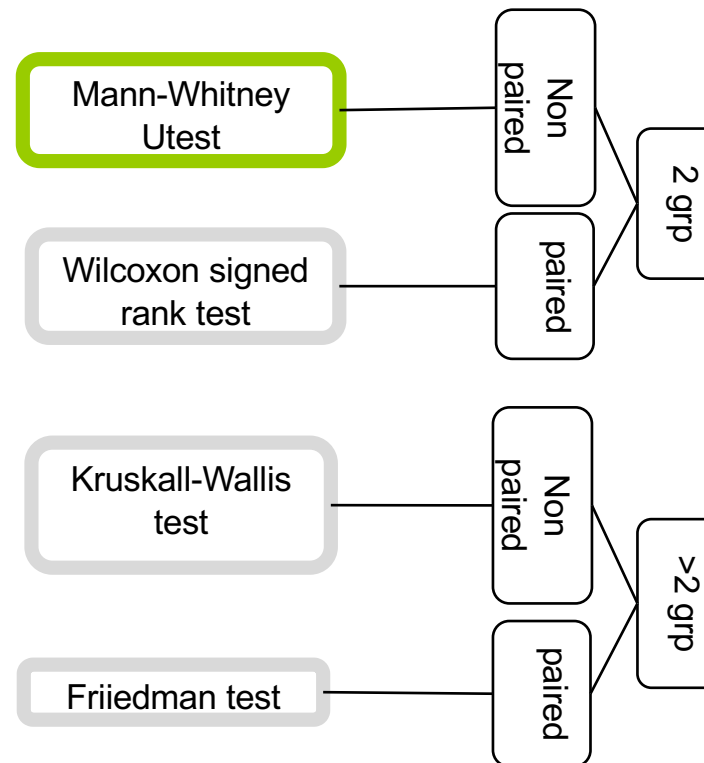
one dataset we know well: **our experiment on reward vs. punishment**

remember we assume the data was normal but it was not

so now we will finally be able to conclude!

	A	B	C
	id	group	score
1	1	A	1
2	2	A	8
3	3	A	5
4	4	A	7
5	5	A	7
6	6	A	8
7	7	A	9
8	8	A	9
9	9	A	7
10	10	A	7
11	11	A	6
12	12	A	8
13	13	A	8
14	14	A	8
15	15	A	6
16	16	A	8
17	17	A	6
18	18	A	8
19	19	A	10
20	20	A	6
21	21	A	6
22	22	A	6
23	23	A	8
24	24	A	8
25	25	A	6
26	26	A	10
27	27	A	6
28	28	A	8
29	29	A	6
30	30	A	10
31	31	A	10
32	32	A	8
33	33	A	6
34	34	A	7
35	35	A	6
36	36	A	5
37	37	A	10
38	38	A	8
39	39	A	7
40	40	A	8
41	41	A	10
42	42	A	6
43	43	A	6
44	44	A	8
45	45	A	8
46	46	A	10
47	47	A	7
48	48	A	8
49	49	B	2
50	50	B	5
51	51	B	6
52	52	B	7
53	53	B	6
54	54	B	8
55	55	B	5

here is our data (chocolate vs. baseline)





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

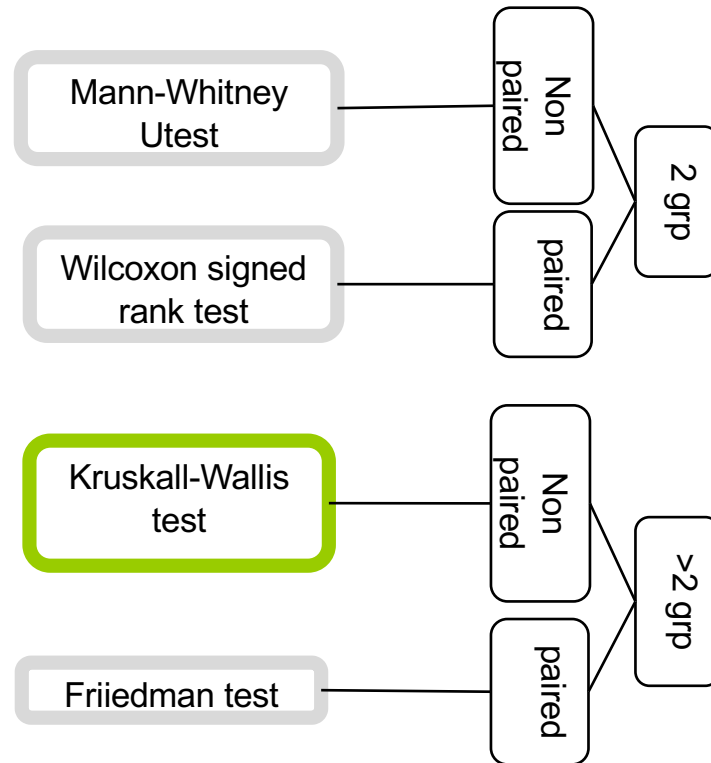
```
dat = read.csv("HCI2018results.csv", header =  
TRUE)
```

```
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

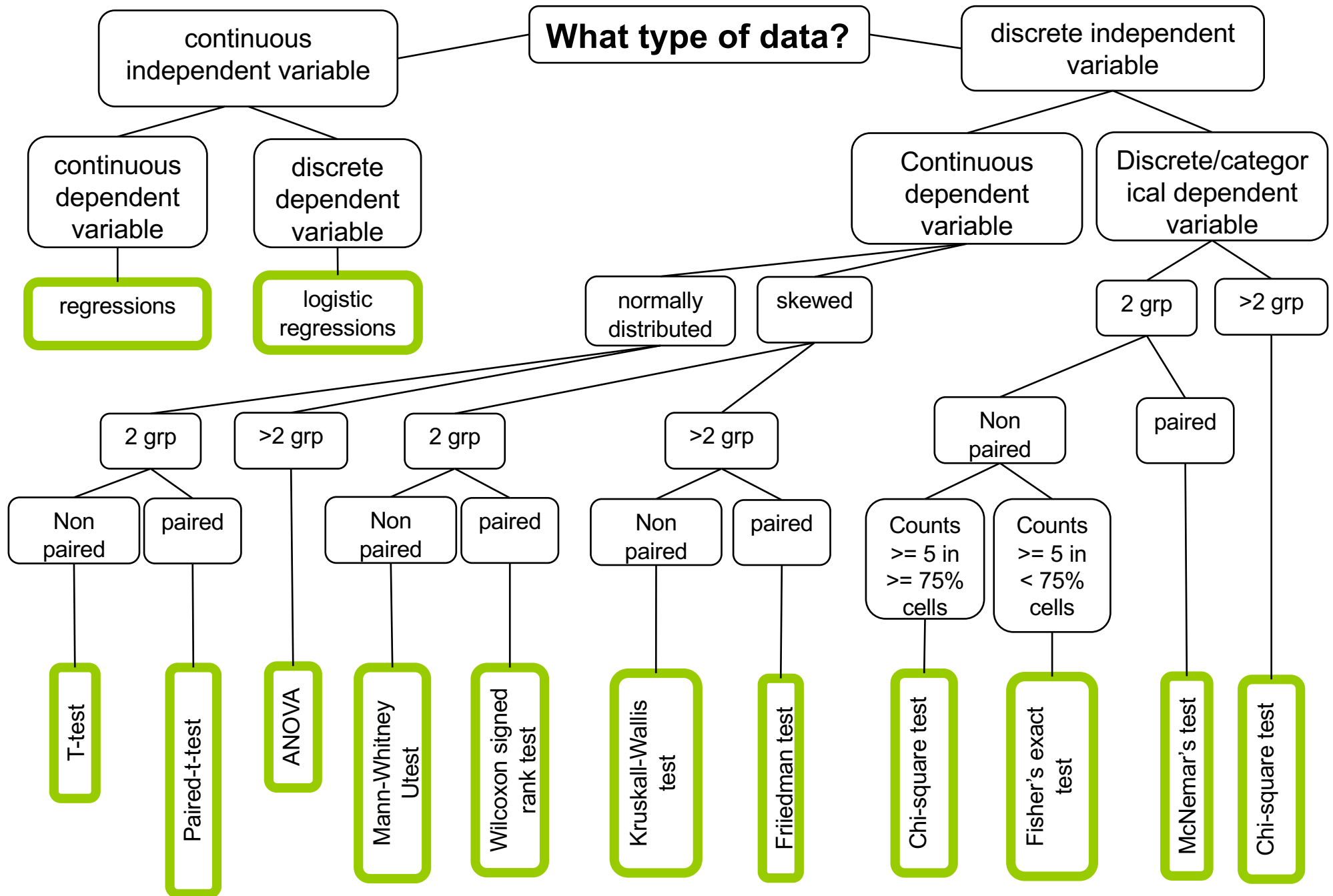


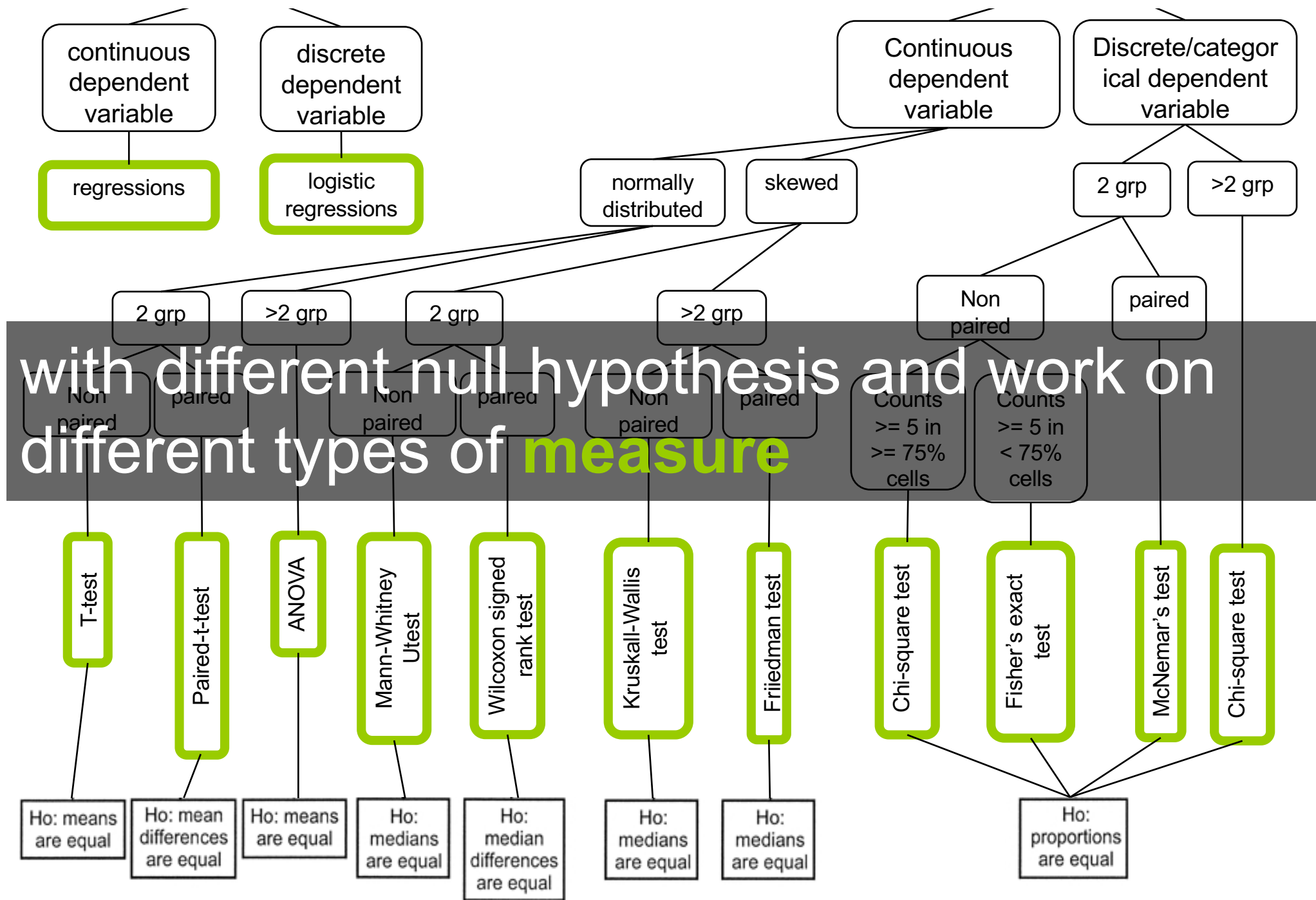
```
#for friedman test (source in GitHub)
dat = read.csv("friedmanExample.csv", header =
TRUE)
friedman.test(dat$count, dat$year, dat$month)
```

**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!





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

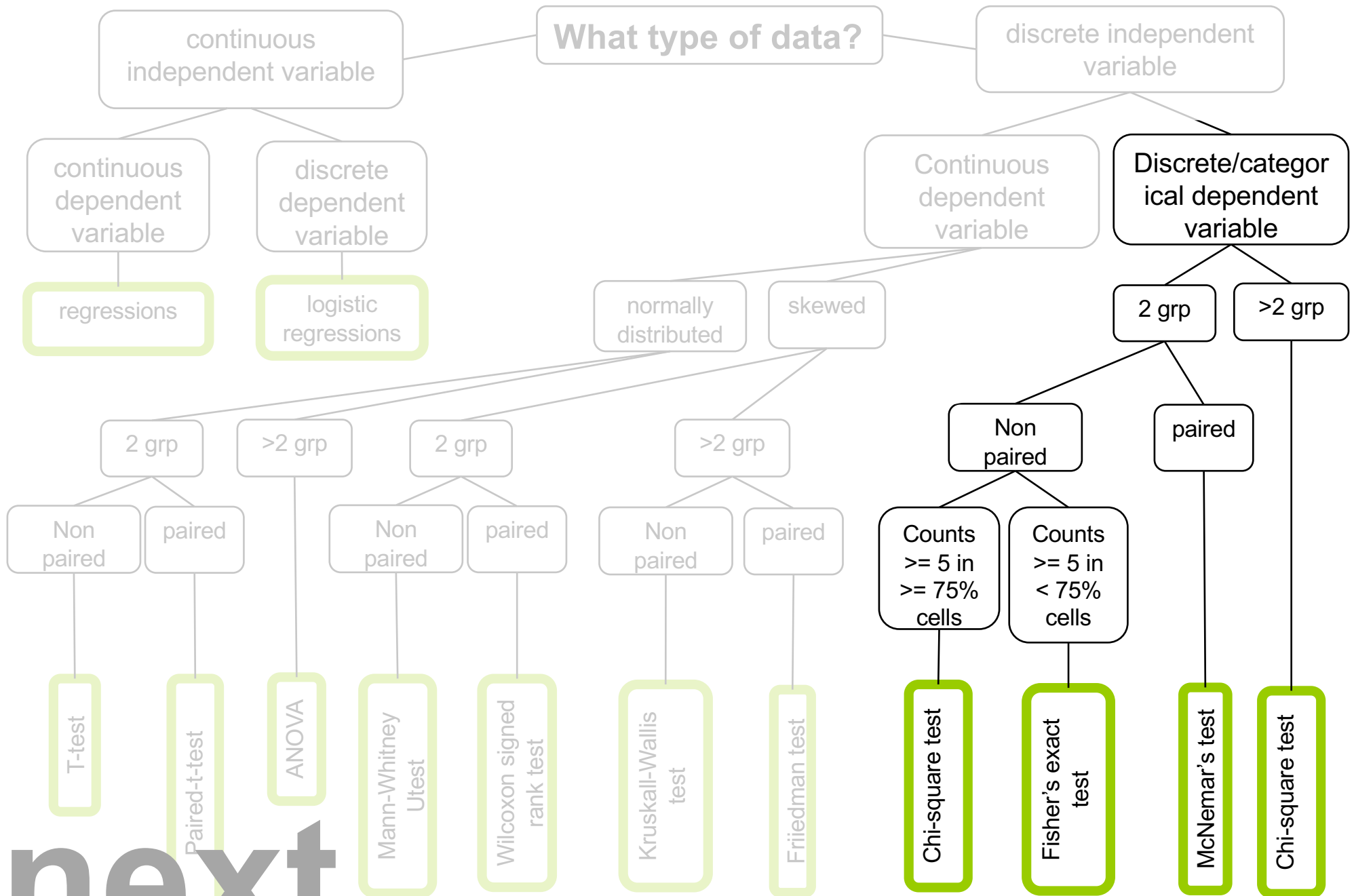


**summary**

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

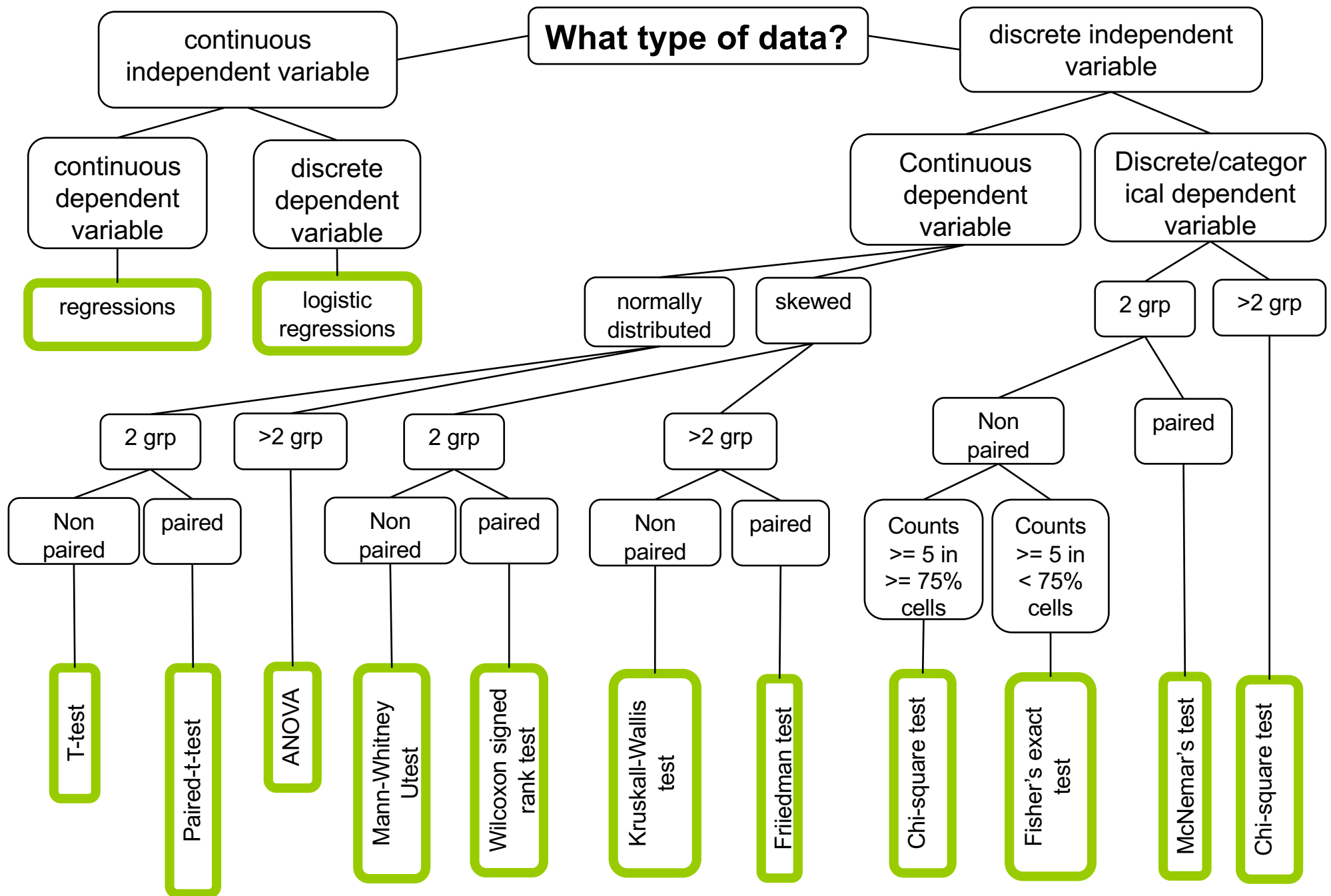




next

end

**quiz to print**



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

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 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 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

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