17

Non-Parametric

Statistical tests

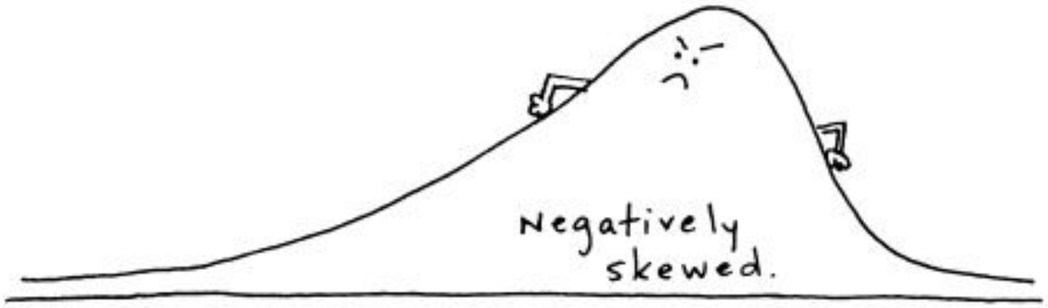
Probability and Statistics

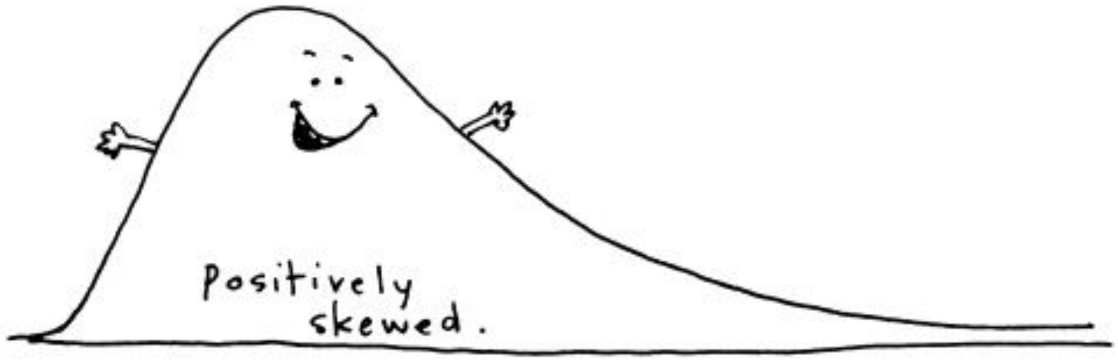
COMS10011

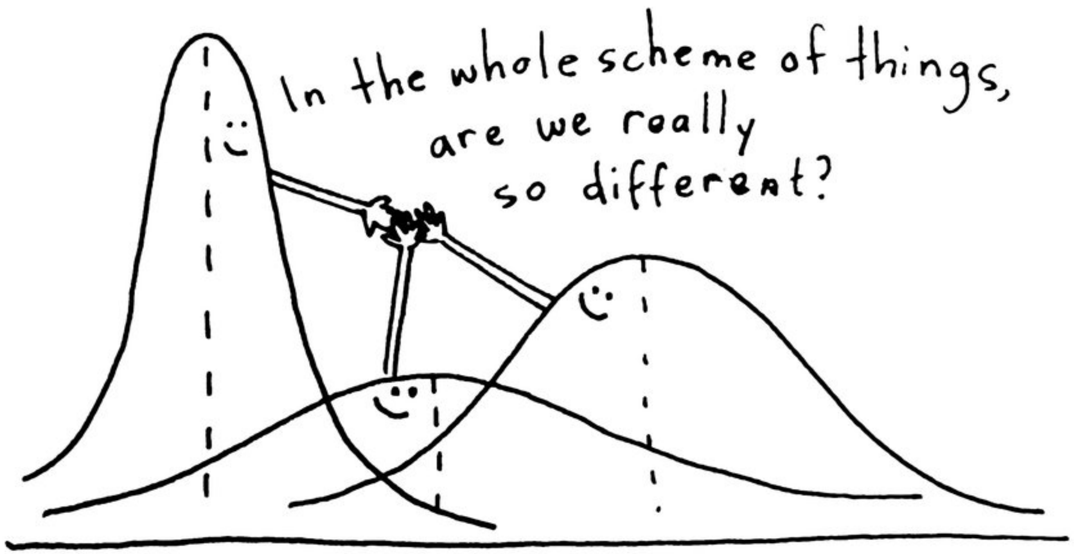
Dr. Anne Roudaut
csxar@bristol.ac.uk

(Thanks S. Massa, Oxford)

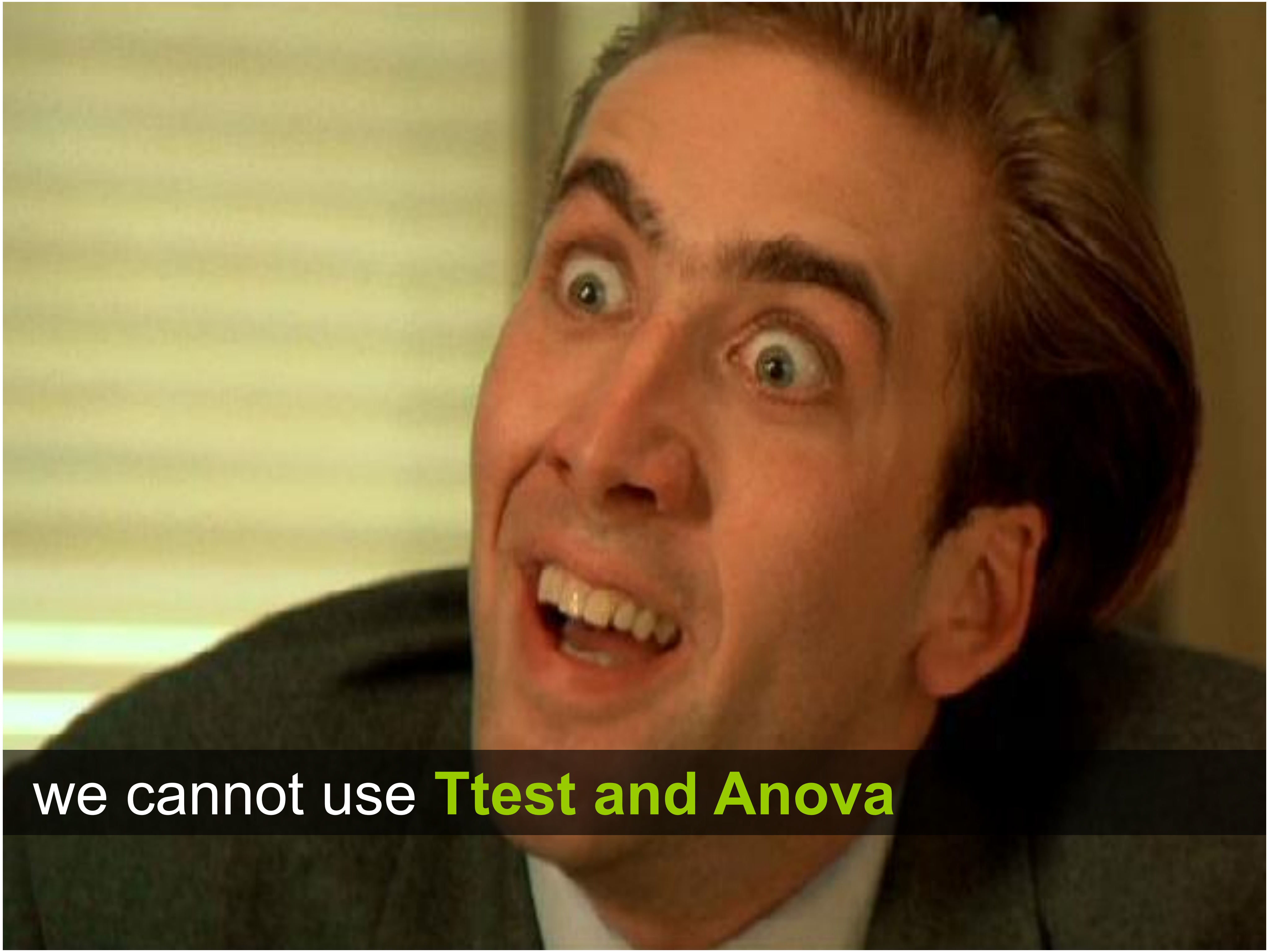
but if we have distributions like this … (assumption **normality** non verified)



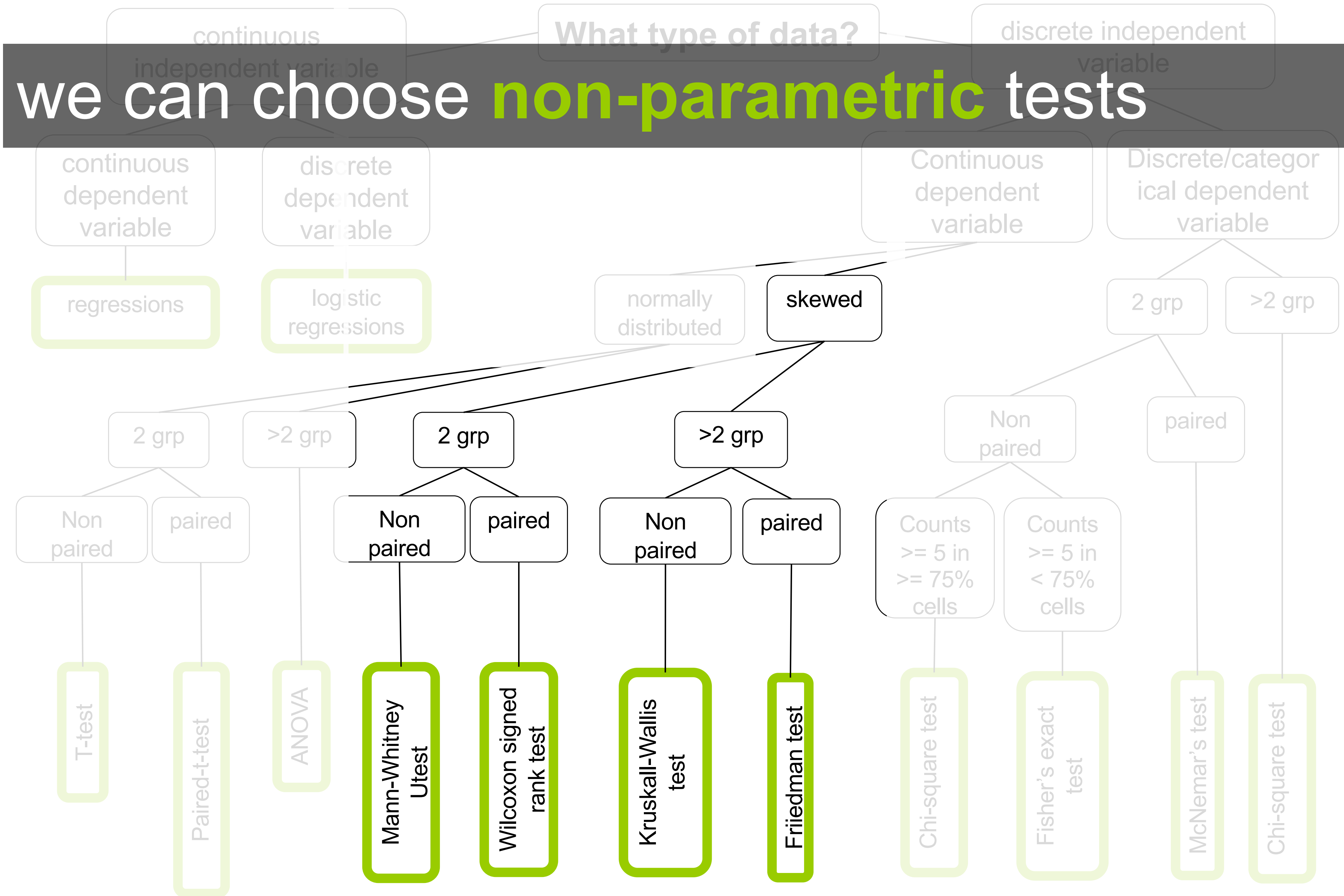




… or that (assumption **homogeneity** non verified)


we cannot use Ttest and Anova

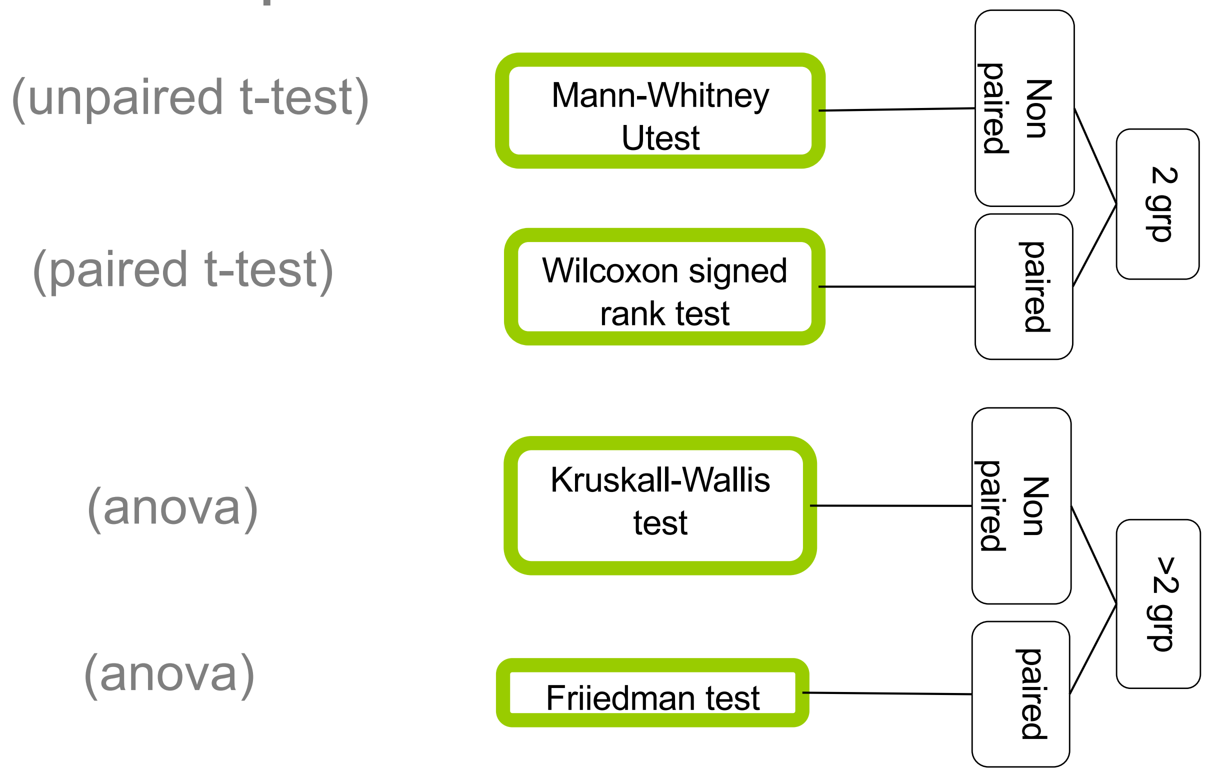
we can choose **non-parametric** tests



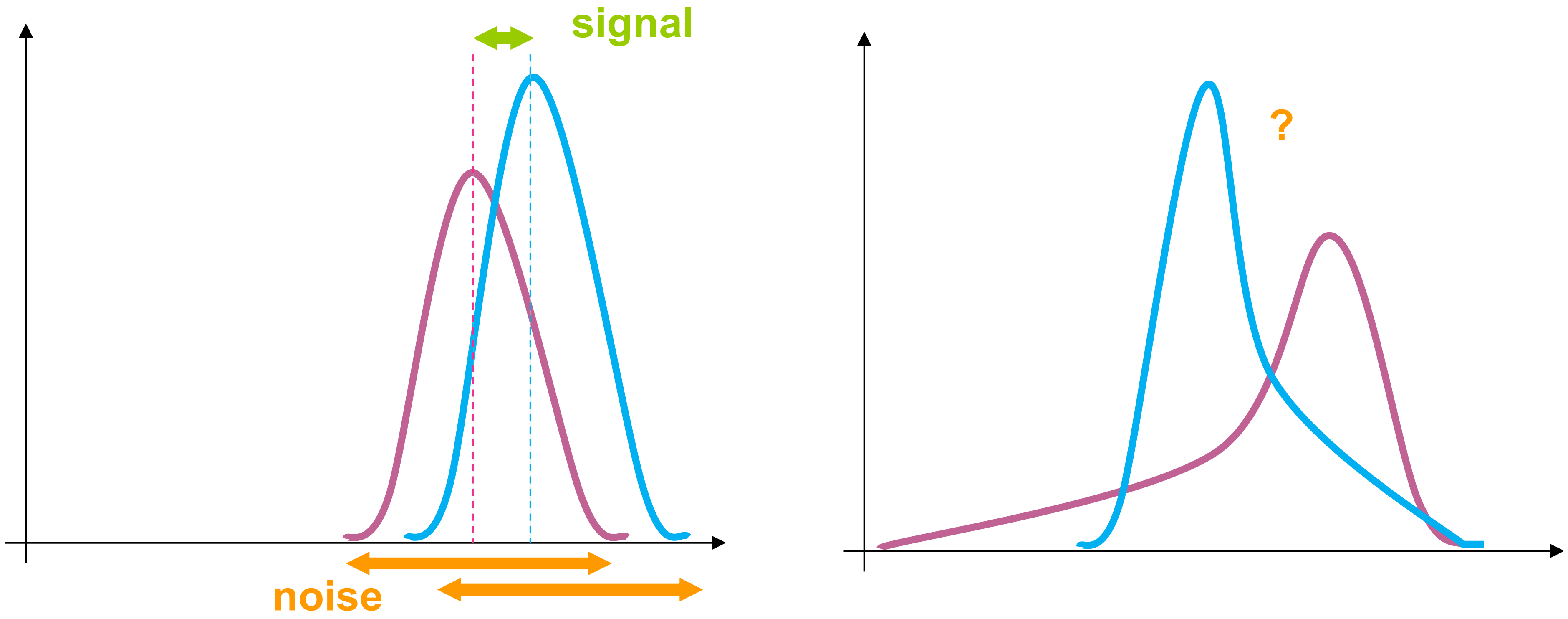
today::

we will look at **four non-parametric tests** four non-parametric tests are very robust (i.e. skewed and non-homogeneous data ok) but nothing is perfect: what **you gain in robustness you lose in power**.

parametric equivalent



so parametric tests used mean and variance, what do we do now?



most tests use **ranking** … let’s look at one example

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.

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

	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

R1 = 30 (n1 = 6)

R2 = 48 (n2=6)

here we have the same sample size for each group so we can take any, e.g $\boxed{\textbf{R (drug B) = 9}}$ and **R (drug A) = 17**

we keep the min

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?

28,50
26,52

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

```
#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)

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

R+ = 48.5
R- = 104.5

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

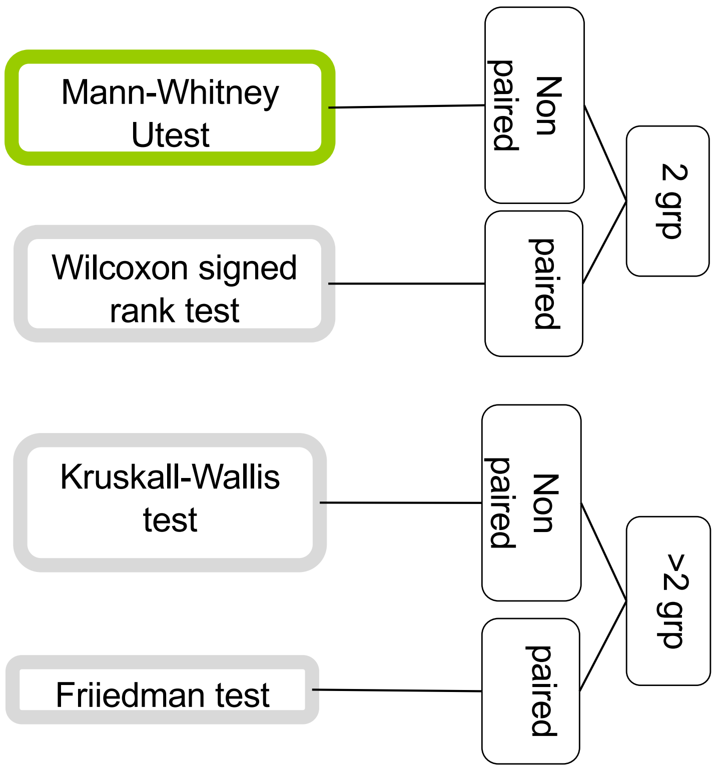
one dataset we know well: **our experiment on reward vs. punishment**

remember we assumed the data was normal but it was not

so now we will finally be able to conclude!

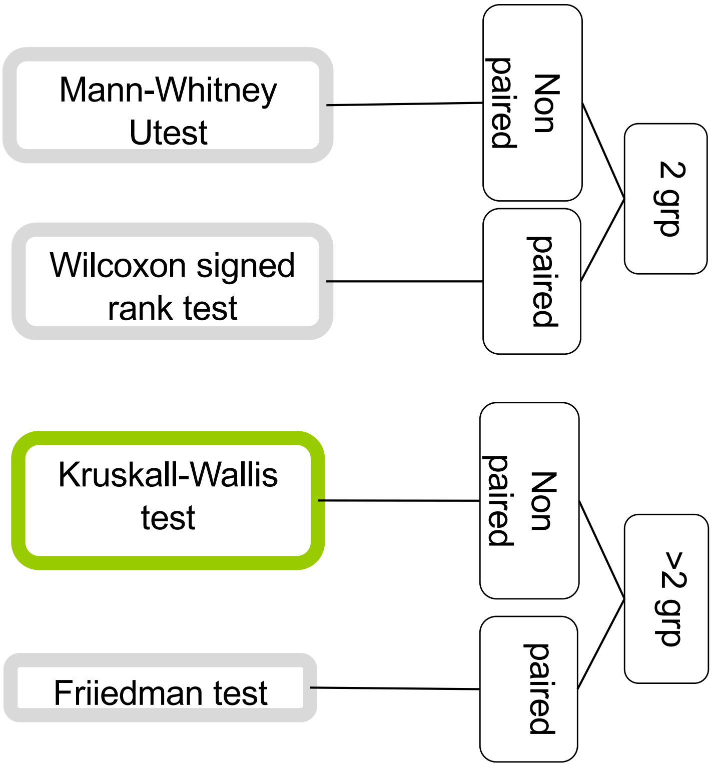
id	group	score
1	A	1
2	A	8
3	A	5
4	A	7
5	A	7
6	A	8
7	A	9
8	A	9
9	A	7
10	A	7
11	A	6
12	A	8
13	A	8
14	A	8
15	A	6
16	A	8
17	A	6
18	A	8
19	A	10
20	A	6
21	A	6
22	A	6
23	A	8
24	A	8
25	A	6
26	A	10
27	A	6
28	A	8
29	A	6
30	A	10
31	A	10
32	A	8
33	A	6
34	A	7
35	A	6
36	A	5
37	A	10
38	A	8
39	A	7
40	A	8
41	A	10
42	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


W = 1290, p-value = 0.6408
```

now let's add the hypothetical group (punishment)

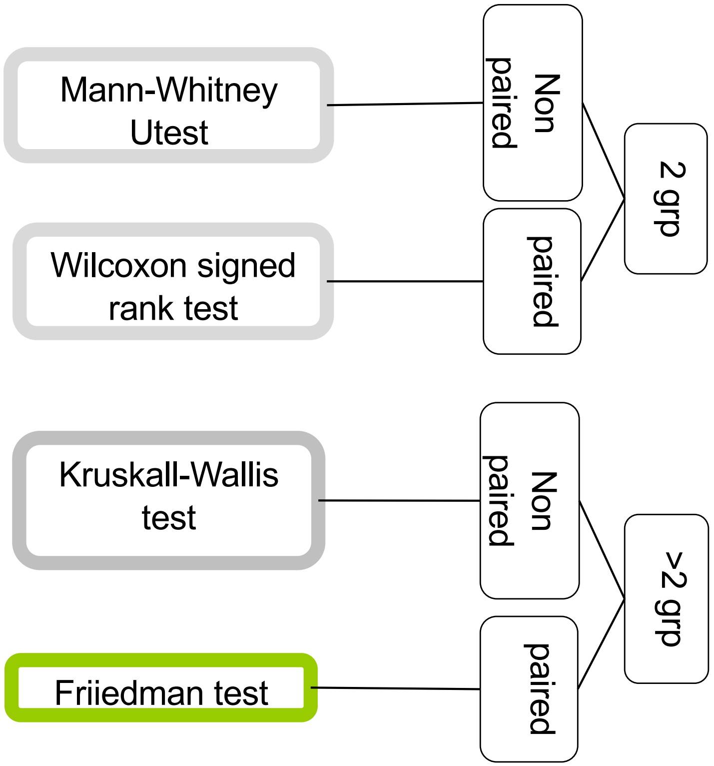


```
dat = read.csv("HCI2018results.csv", header =
TRUE)
kruskal.test(score ~ group, data = dat)


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

just so you know how to do it



```
#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).
```

example from scratch

biggest cause disputes in UK

do you put milk in your cup of tea before or after the boiling water?

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

H = participants will prefer the taste of tea when the milk is put after the boiling water

IV = One cup is made with milk before

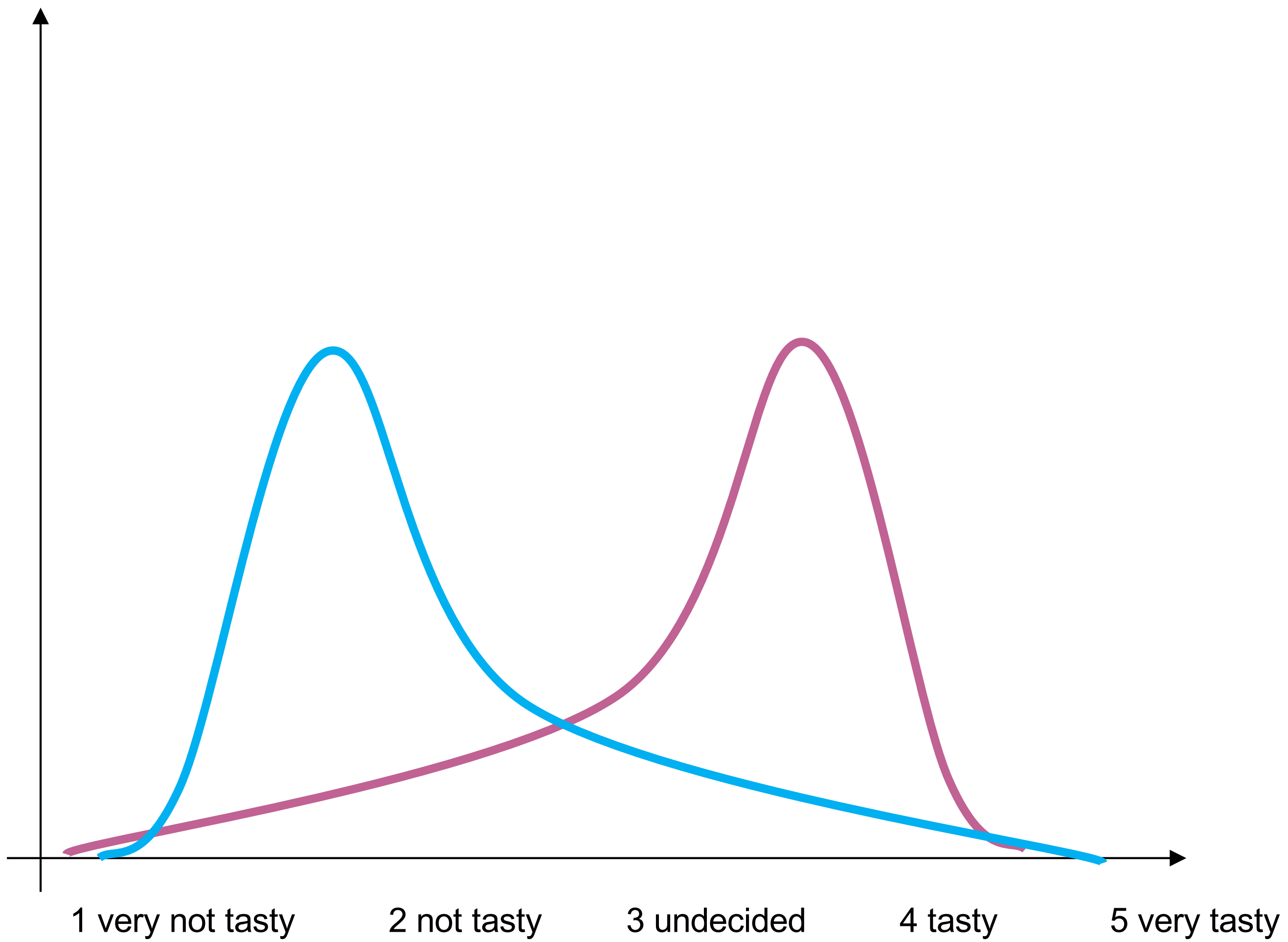
One cup is made with milk after

DV = tastiness

On a scale of 1 to 5 rate the tastiness of this cup?

1 very not tasty 2 not tasty 3 undecided 4 tasty 5 very tasty

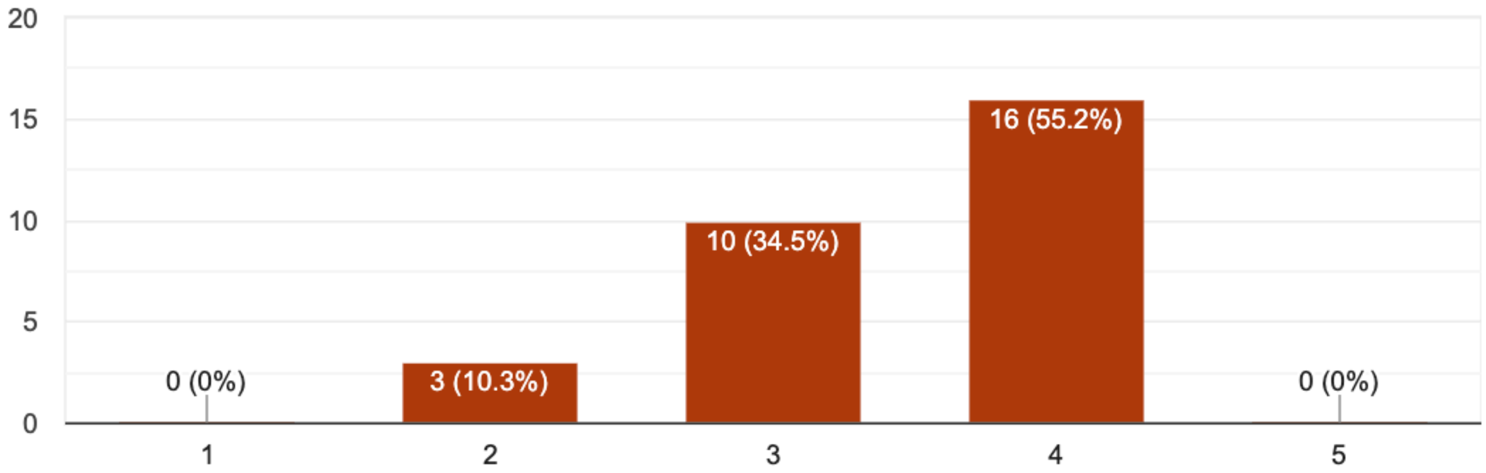
Between subjects with one trial only



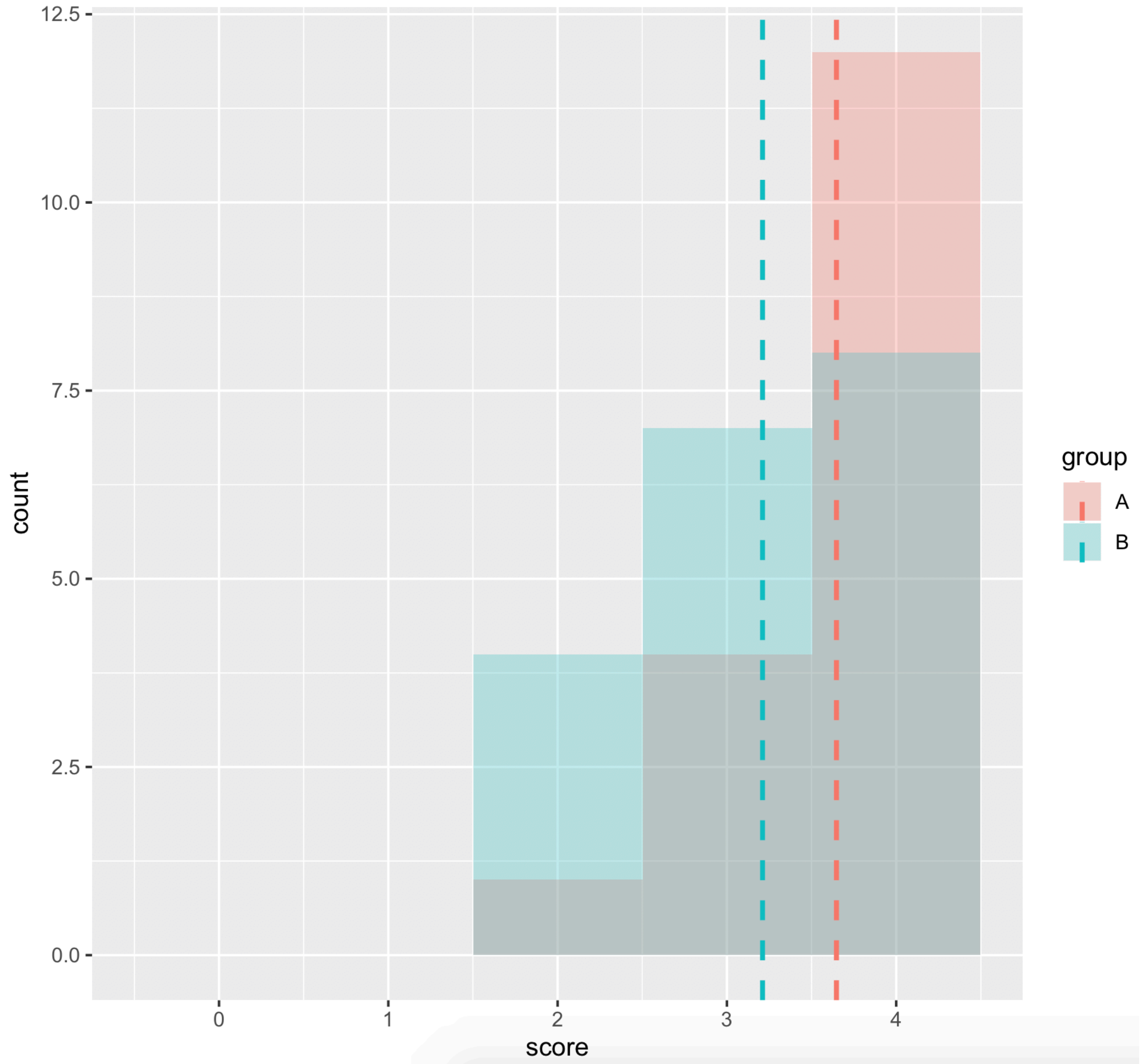
Likert are most often skewed

On a scale of 1 to 5, how would your rate the taste of the tea?

29 responses



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

  group score.mean
1     A   3.647059
2     B   3.210526

# Overlaid histograms with means
library(ggplot2)
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)
```



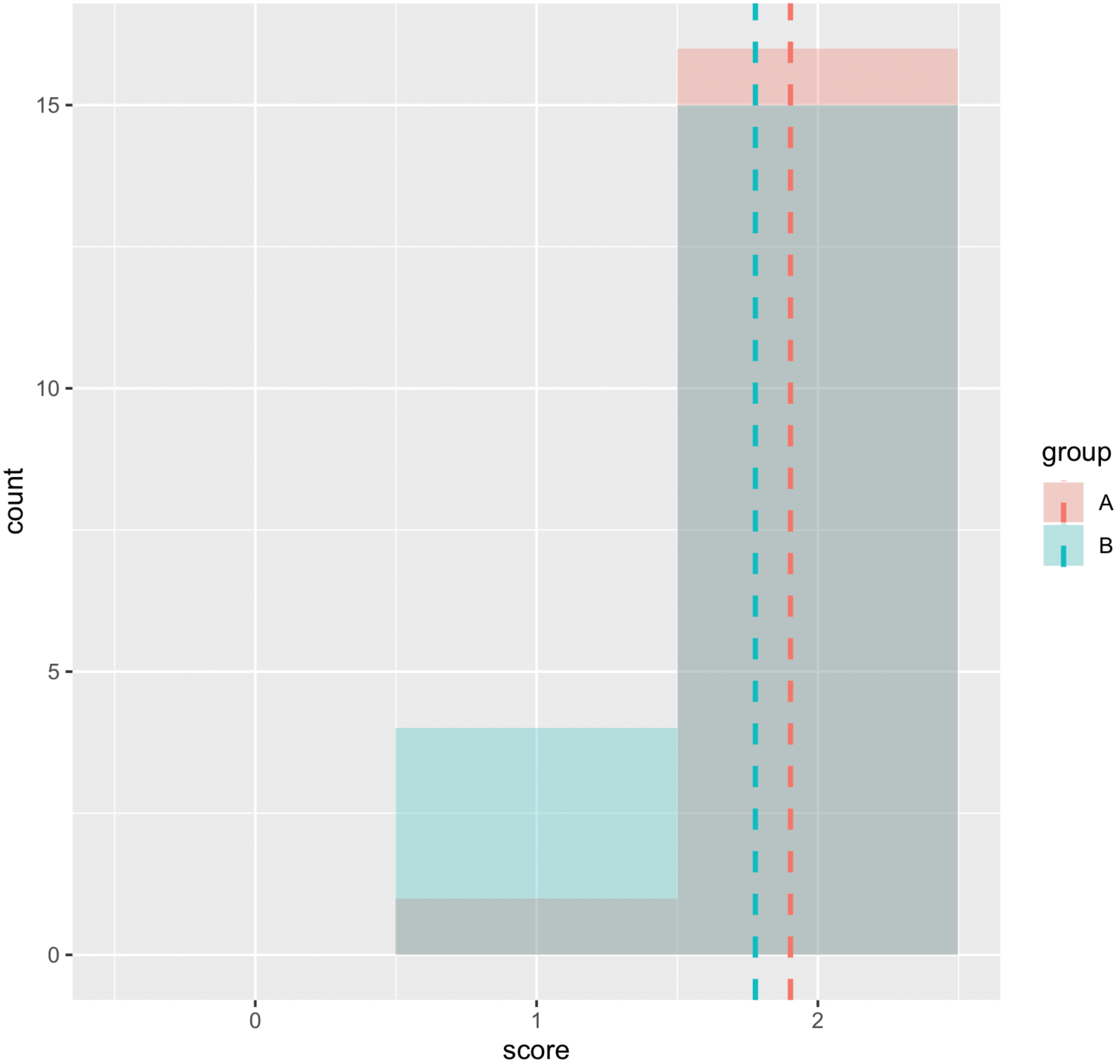
count
12.5
10.0
7.5
5.0
2.5
0.0
0
1
2
3
4
score
group
A
B

lest try to transform this … with square root

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

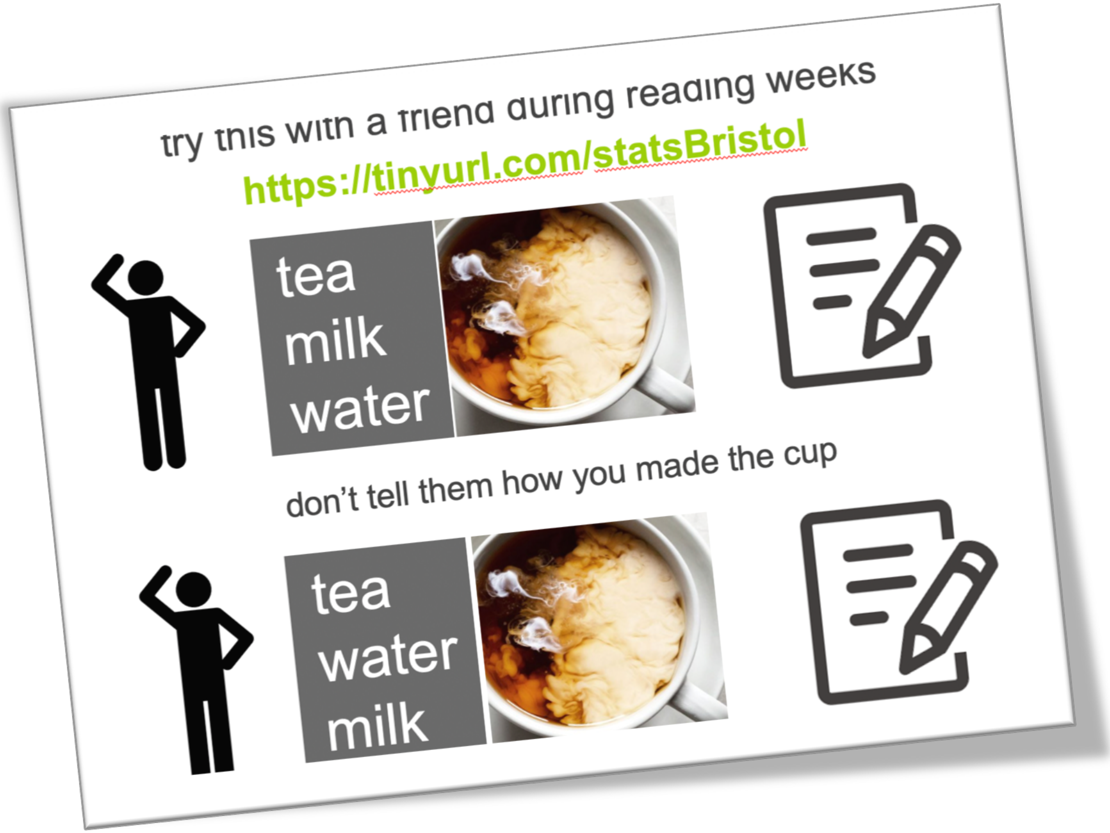
```
  group score.mean
1     A   1.902495
2     B   1.777958
```

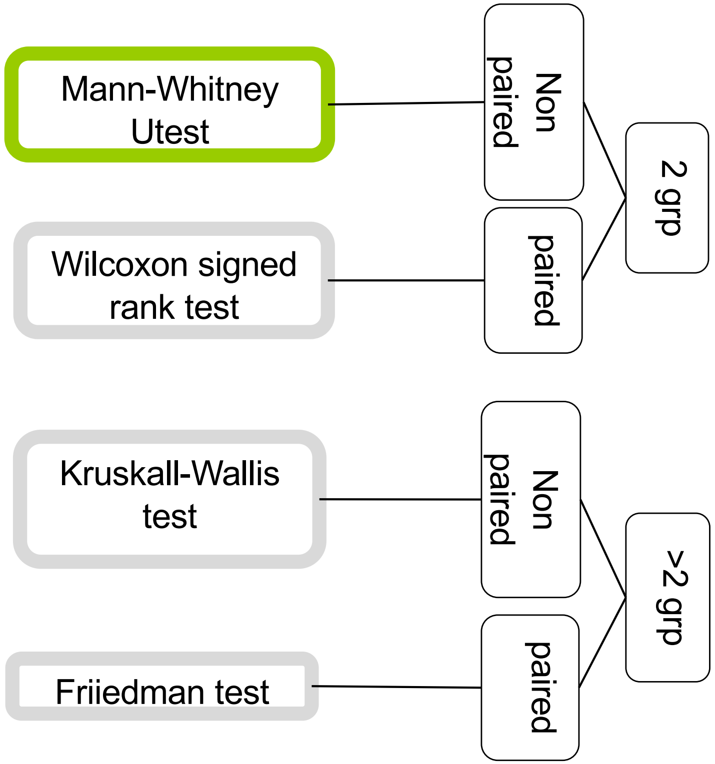
```r
# Overlaid histograms with means
library(ggplot2)
ggplot(dat, aes(x=score, fill=group)) +
geom_histogram(binwidth=1, alpha=.3, position="identity")
+ geom_vline(data=cdat, aes(xintercept=score.mean,
colour=group), linetype="dashed", size=1) +
expand_limits(x = 0, y = 0)
```



15
10
5
0
count
0
1
2
score
group
A
B

```
shapiro.test(dat$score)

Shapiro-Wilk normality test

data:  dat$score
W = 0.72514, p-value = 7.196e-07

= definitely not normal!
```


try this with a friend during reading weeks
https://tinyurl.com/statsBristol
tea
milk
water
don't tell them how you made the cup
tea
water
milk


Mann-Whitney Utest
Wilcoxon signed rank test
Kruskall-Wallis test
Friiedman test
Non paired
paired
2 grp
Non paired
paired
>2 grp

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

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


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

Wilcoxon rank sum test with continuity correction


W = 212, p-value = 0.07612
```


unfortunately our experience was inconclusive
(because pvalue > 0.05)

what can be the reasons that there is no different?

-> low sample size

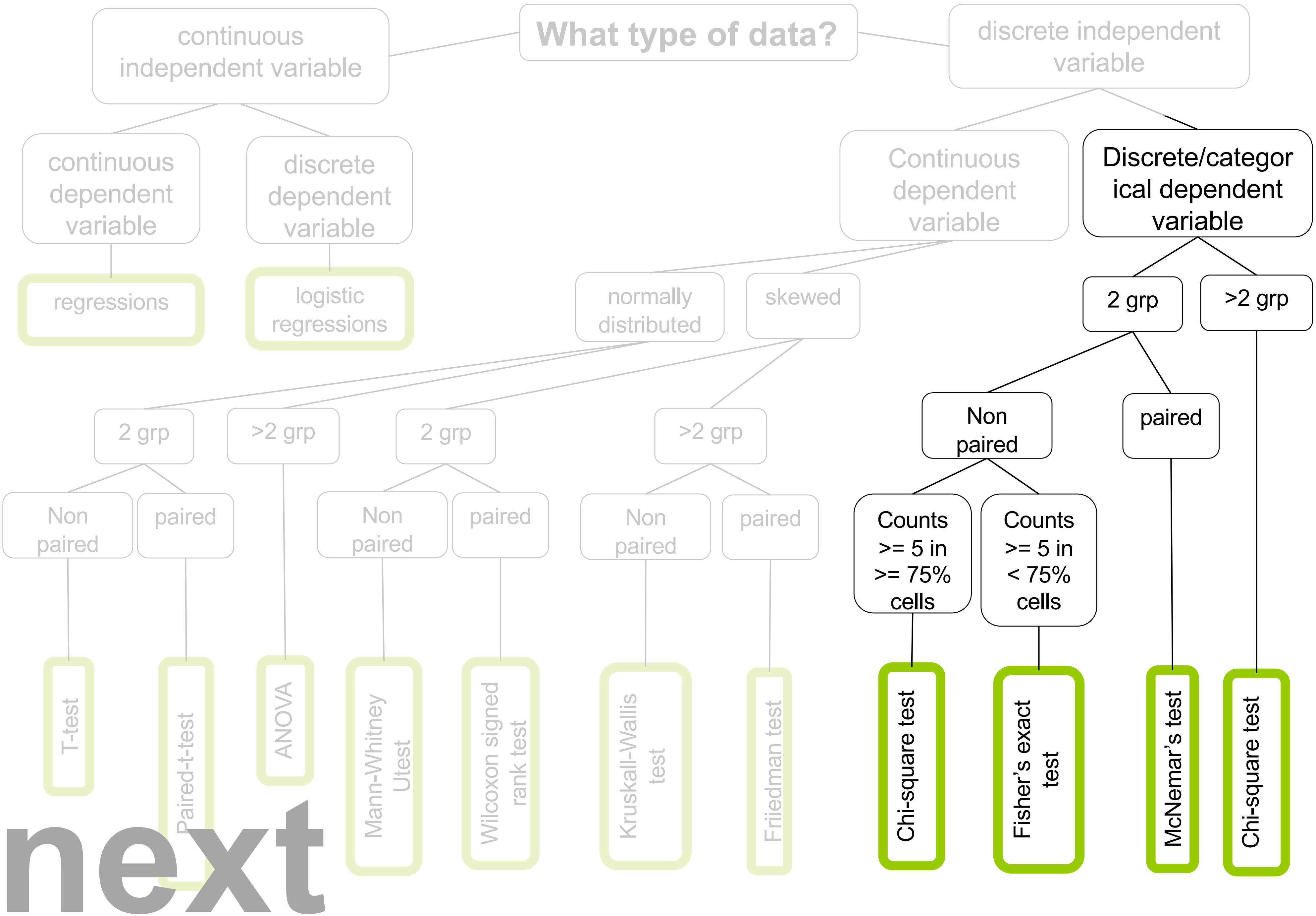
-> too much “noise” = did we control enough?

-> too weak signal = may be there is actually no difference in taste after all

summary

1. Give the name of the four non-parametric tests seen today and when to use them
2. Explain the basis of Mann Whitney and Wilcoxon test, aka that they use ranks rather than mean
3. I **will not ask** you to do it by hand in the exam

take away


What type of data?
continuous independent variable
discrete independent variable
continuous dependent variable
discrete dependent variable
Continuous dependent variable
Discrete/categorical 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
Friedman test
Chi-square test
Fisher's exact test
McNemar's test
Chi-square test
next

end