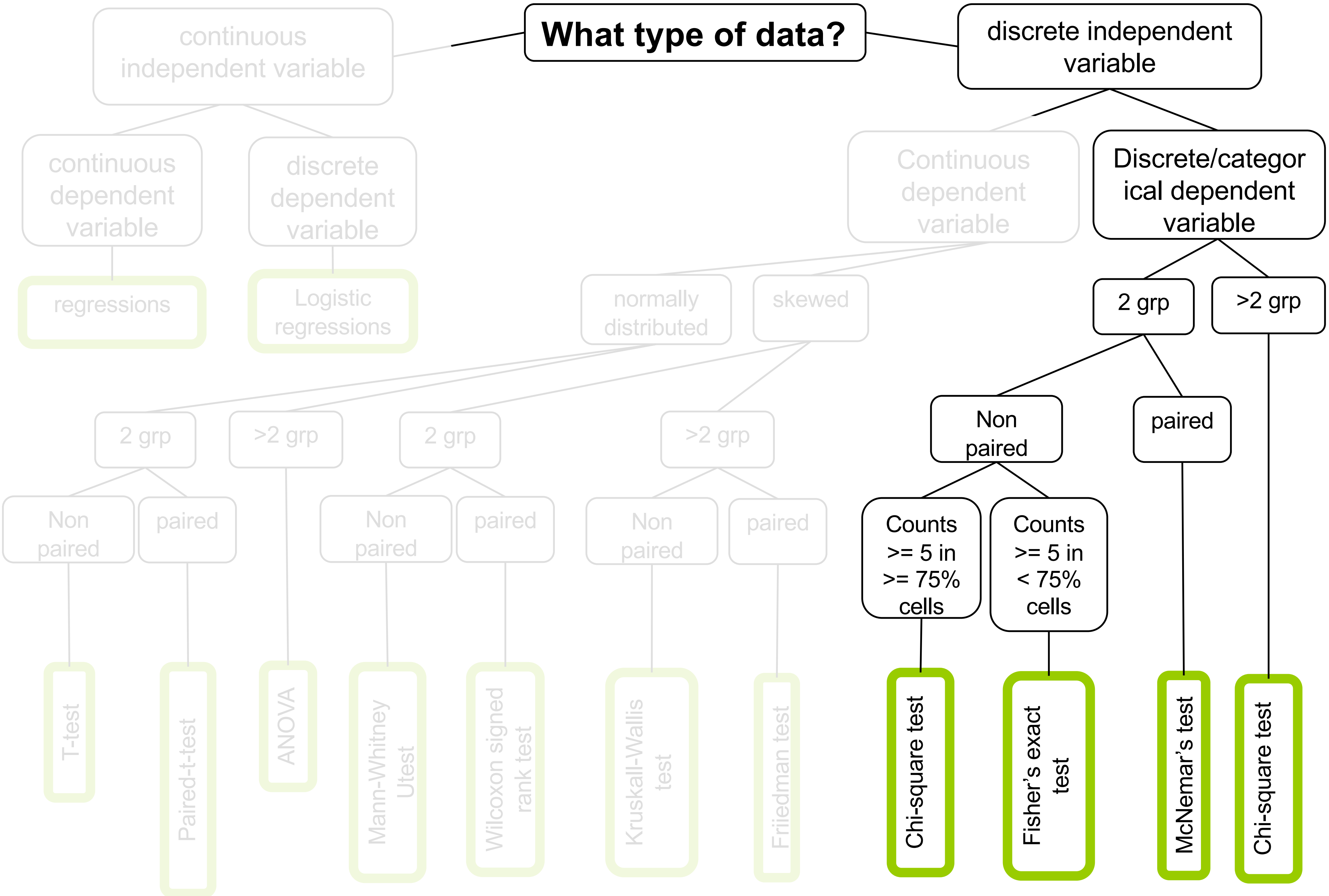
Chi-square
and Anova ++

17

Probability and Statistics

COMS10011
Dr. Anne Roudaut
csxar@bristol.ac.uk
https://github.com/coms10011



What type of data?
continuous independent variable
discrete independent variable
continuous dependent variable
discrete dependent variable
regressions
Logistic regressions
Continuous dependent variable
Discrete/categorical 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

until now, we did statistical test using means or medians but the assumptions for means have eliminated certain types of variables (e.g. gender)

mean not appropriate measure of central tendency for nominal (categorical type) data

Chi-square can do do!

there are two types of Chi-square tests:

goodness of fit test (for one variable only)
contingency table test (for two variables at a time)

goodness of fit

looks to **see if a single variable fits some hypothesised probability distribution**

e.g. in a population of students, there would be an equal number of students who like or dislike brussels sprouts

in fact we don’t even have to go 50/50, we may theorize that only 1/4 (25%) will like them (because they are disgusting!)

	# of persons
like BP	11
Dislike BP	139
	150 (total)

	# of persons	%expected	# expected
like BP	11	25%	37.5 (25% of 150)
Dislike BP	139	75%	112.5 (75% of 150)
	150 (total)	100%	150 (total)

observed cases

expected cases

$$X^2 = \sum \frac{(o-e)^2}{e}$$

$$= \frac{(11-37.5)^2}{37.5} + \frac{(139-112.5)^2}{112.5}$$

$$= 24.96$$

now, like with all the test we have seen, we look into a table, here the Chi-square table)

Critical values of the Chi-square distribution with *d* degrees of freedom

	Probability of exceeding the critical value						
d	0.05	0.01	0.001	*d*	0.05	0.01	0.001
1	$\boxed{3.841}$	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

degree of freedom DF
= number of group - 1

24.96 > 3.841 so we **reject the null hypothesis**

our theory of 25% Brussel Sprout lovers does not hold

	# of persons	%expected	# expected
like BP	25	25%	25 (25% of 100)
Dislike BP	75	75%	75 (75% of 100)
	100 (total)	100%	100 (total)

observed cases

expected cases

$$X^2 = \sum \frac{(o-e)^2}{e}$$

$$= \frac{(25-25)^2}{25} + \frac{(75-75)^2}{75} = 0$$

if data was perfect fit (pvalue would be = 1)
… cannot reject null (thus cannot conclude)

let’s see if our theory holds with a raise of hand

who like Brussel sprout?

who dislike Brussel sprout?

```
table = c(11,139)

        male female Sum
sport     26      3  29
family    24     22  46
Sum       50     25  75

chisq.test(tulip, p = c(1/4, 3/4))

Chi-squared test for given probabilities


data:  tulip
X-squared = 24.969, df = 1, p-value = 5.826e-07
```

this example is fairly simple but Chi-square also work with more data, e.g. 30% prefer eating chicken for Christmas dinner, 50% prefer turkey, 10% prefer vegetarian option, 10% prefer other types of meat

… problem sheet 5 will be about that

contingency tables

public opinion surveys tend to show there is a relationship between gender and *something*, e.g. preference in sport car vs. family car (public opinion surveys are very stereotypical!)

so here we have **two variables/groups**: gender (female or male) and car preference (sport or family)

	male	female
sport	26	3
family	24	22

we do a **Chi-square contingency table test** to prove preference of car is related to or dependant upon gender but we don’t have ”expected value” here so we first need to calculate them for each cell

$$E_{ij} = \frac{R_i C_j}{N}$$

where R = row, C= column, N = total,

for ith row and jth column

	male	female	Sum
sport	26	3	29
family	24	22	46
Sum	50	25	75

1. we compute the sums in all direction

2. for each cell, multiplying that cells row and column totals and dividing by our total sample size

e.g. case (sport, male)= (29 * 50) / 75

e.g. case (family, male= (46 * 50) / 75

…

Obs.	male	female
sport	26	3
family	24	22

Exp.	male	female
sport	19.3	0.9
family	30.6	15.3

4. use same Chi-square formula than before

5. Calculate degree of freedom as DF =(number of rows−1)∗(number of columns−1) (here = 1)

6. Use the Chi-square table to conclude!

```
table = matrix(c(26, 24, 3, 22), ncol=2)
colnames(table) = c('male', 'female')
rownames(table) = c('sport','family’)
addmargins(table)

       male female Sum
sport    26      3  29
family   24     22  46
Sum      50     25  75

chisq.test(table,correct=FALSE) #must use correct=FALSE
for a 2 by 2 table otherwise = TRUE


Pearson's Chi-squared test


data:  table
X-squared = 11.244, df = 1, p-value = 0.0007986
```

we reject the null and conclude that car type preference is dependant of gender

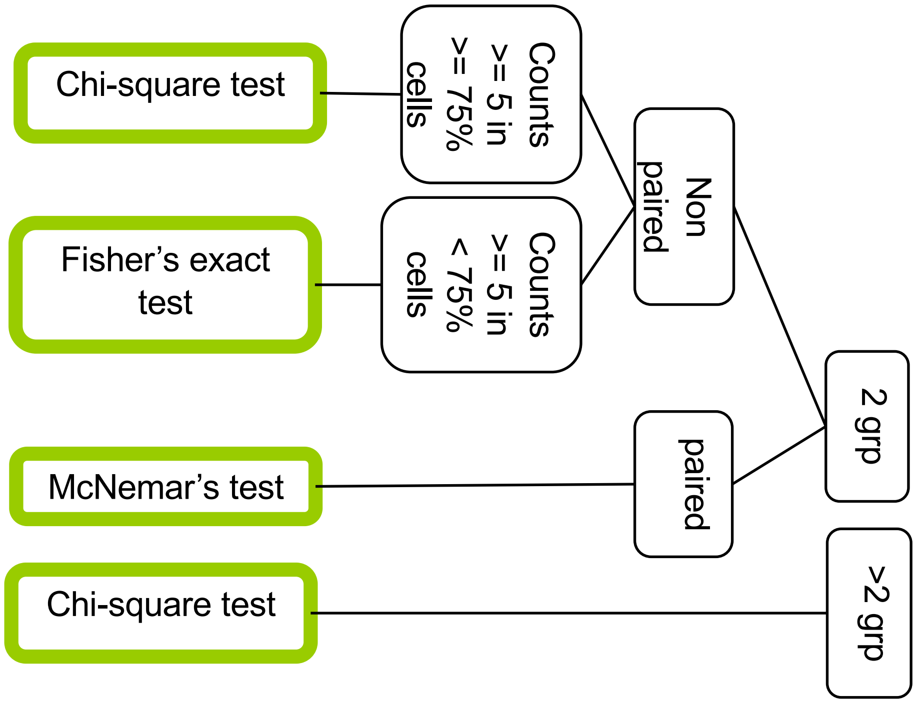
Chi-square contingency test::

if you have two categorical variables, and you'd like to determine whether the variables are independent

(sometimes called a test of independence)

H0: the 2 categorical variables are independent (no relationship between the variables)

in R


>2 grp
2 grp
paired
Non paired
Counts >= 5 in >= 75% cells
Counts >= 5 in < 75% cells
Chi-square test
Fisher's exact test
McNemar's test
Chi-square test

```
> library(MASS)         # load the MASS package
> tbl = table(survey$Smoke, survey$Exer)
> tbl                   # the contingency table

        Freq None Some
  Heavy    7    1    3
  Never   87   18   84
  Occas   12    3    4
  Regul    9    1    7

>chisq.test(tbl) # or fisher.test(tbl) if Counts >= 5 in <
75% cells


        Pearson's Chi-squared test

data:  table(survey$Smoke, survey$Exer)
X-squared = 5.4885, df = 6, p-value = 0.4828
```

```
# mcnemar example on presidential Approval Ratings:
Approval of the President's performance in office in two
surveys, one month apart, for a random sample of 1600
voting-age Americans.

Performance <- matrix(c(794, 86, 150, 570), nrow = 2,
dimnames = list("1st Survey" = c("Approve", "Disapprove"),
"2nd Survey" = c("Approve", "Disapprove")))
Performance

            2nd Survey
1st Survey   Approve Disapprove
  Approve        794        150
  Disapprove      86        570


mcnemar.test(Performance)
```

that was easy and we are done with Chi-square!

now I would like to briefly come back to ANOVA for a bit
(oh not again!)

statistic tests on multiple variables

until now, we did statistical test on one independent variable (with multiple conditions or groups) and one dependant variable

e.g. effect of **chocolate, baseline, punishment (IV)** on **memorization score (DV)**

now it is possible to do tests for multiple IVs and multiple DVs (e.g. with CHI-square contingency table we look at two IVs).

however doing so decrease the power of your experiment (because you run more tests)

so it only works with powerful tests based on ANOVA (i.e. continuous variable and assumption of normality and homogeneity assumed)

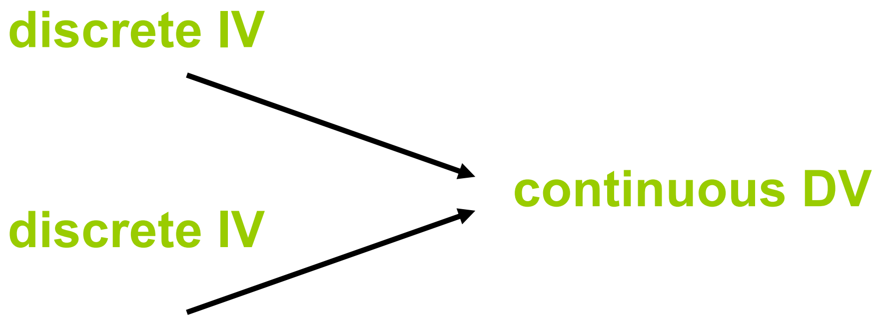
e.g. **two-ways ANOVA, MANOVA, ANVOVA**

discrete IV $\longrightarrow$ **continuous DV**

one-way ANOVA::

compare the effect of **one discrete independent variables**, having 2 or more levels on **one dependant variable**

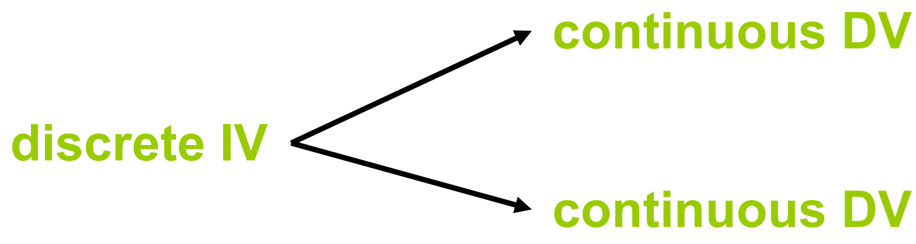
e.g. effect of alcohol consumption (none, 2-pints, 4-pints) on attractiveness ratings



two-way ANOVA::

compare the effect of **two discrete independent variables**, each of them having 2 or more levels on **one dependant variable**

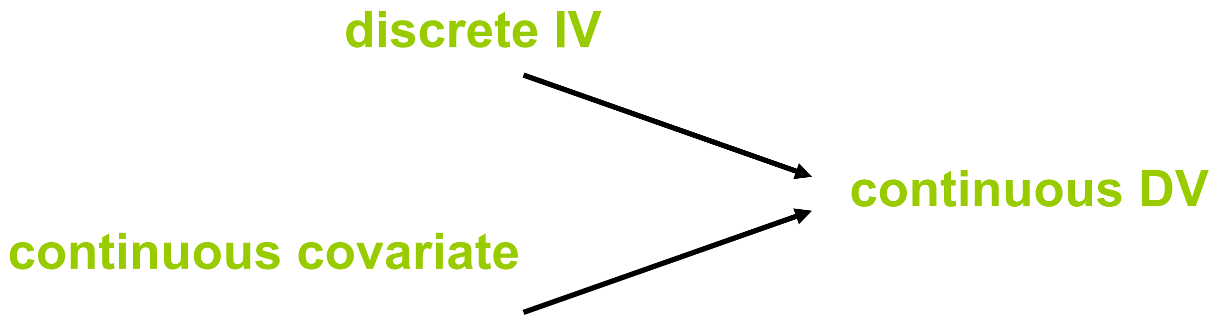
e.g. effect of gender (female, male) and alcohol consumption (none, 2-pints, 4-pints) on attractiveness ratings



one-way MANOVA::

(multivariate analysis of variance) compare the effect of **one independent variable**, having 2 or more levels on **two dependant variables**

e.g. effect of different memorization enhancing drugs (placebo, drug A, drug B) on memorization skills and emotional ratings (to find the sweet spot for a drug that enhance skills without depressing people!)



one-way ANCOVA::

(analysis of **co**variance) compare the effect of **one independent variable**, having 2 or more levels and **one continuous covariate** on **one dependant variable**

e.g. effect of phone sizes (iphone 4, iphone 5, iphone 6, iphone 7) on the amplitude of phone movements made when texting **given the measure of the participants hand width (covariate)** discrete IV → continuous DV

continuous covariate → continuous DV

one-way MANCOVA::

(multivariate analysis of covariance) compare the effect of **one independent variable**, having 2 or more levels and **one continuous covariate** on **two dependant variable**

… you can even two a two-way MANCOVA (but your experience might not have much statistical power because there are too many tests to perform) although these tests exist, my advice would be to keep the experimental design as simple as possible as you can, analysis will be easier and more powerful

… we will look at power next time with a guest lecturer: Luluah Al-Barrack

two ways ANOVA practically

```
# two-ways anova in R (I added a gender column in our
chocolate vs. reward vs. punishment file)
library(ez)
dat = read.csv("HCI2018resultsTwoWays.csv", header = TRUE)
ezANOVA(dat,id,between=.(group,gender),dv=score)
```

```
        Effect DFn DFd       F            p p<.05        ges
1        group   2  54 72.7776 4.709005e-16     * 0.72939818
2       gender   1  54  1.4112 2.400561e-01       0.02546778
3 group:gender   2  54  0.8064 4.517654e-01       0.02900052
```

Like with one way with have the effect for each IV

As well as the **interaction** between both

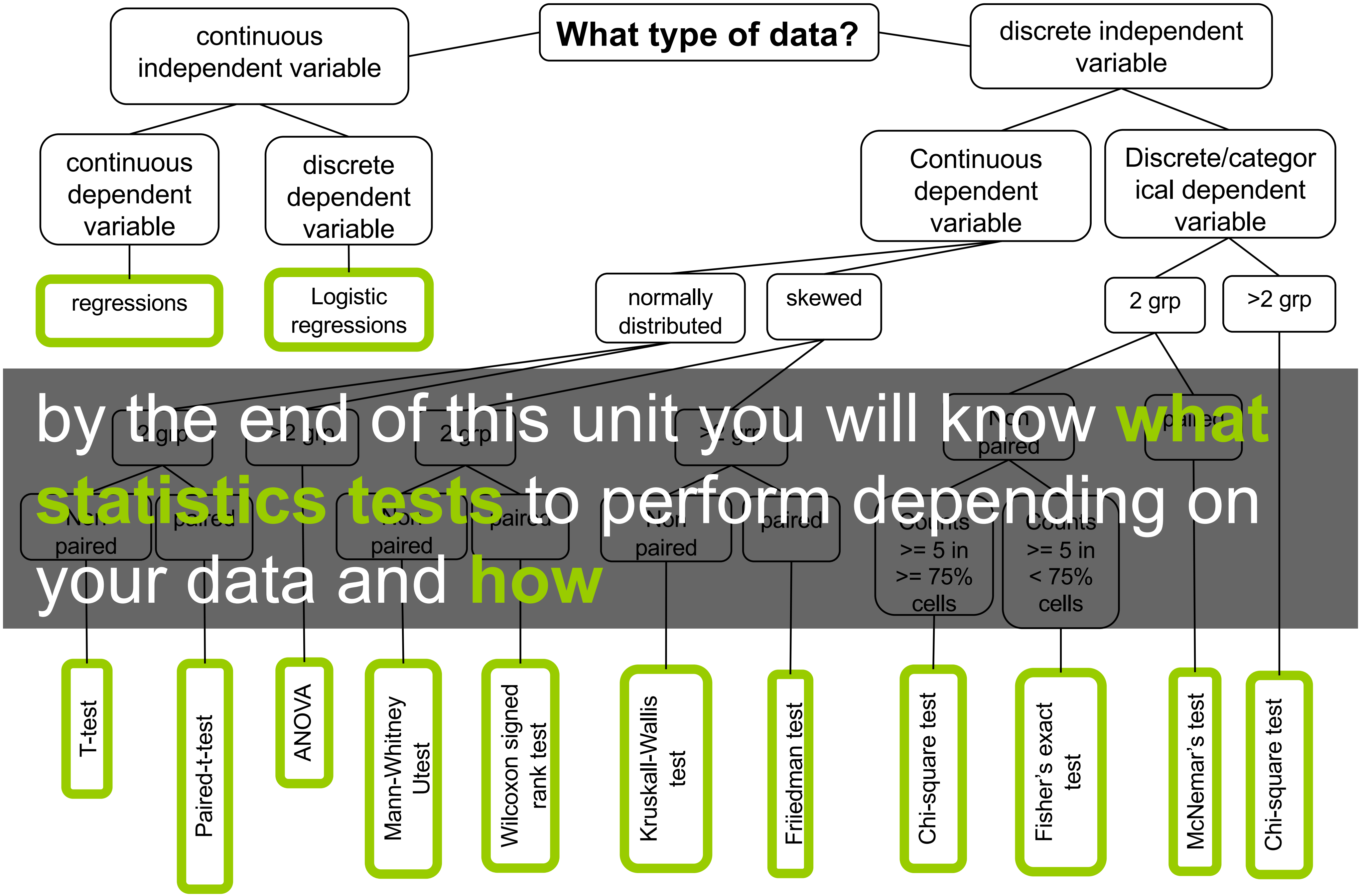
```
# note here our two IV are between but we could have them
both within or a combination of within and between we
would write:
ezANOVA(dat,id,within=group,between=gender,dv=score)
```

we can write (only the significant one)

A two-way ANOVA showed a significant effect on IV1 (Fdf = F_value, p<0.05), on IV2 (Fdf = F_value, p<0.05) and on the interaction IV1 x IV2 (Fdf = F_value, p<0.05)

and then you can do your post-hoc comparison tests (although there are more to do!) and conclude

summary



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
paired
Non paired
>= 5 in >= 75% cells
>= 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
by the end of this unit you will know what statistics tests to perform depending on your data and how

1. Linear regression
2. Hypothesis testing, comparing things
3. Experimental design a: T-test
4. Experimental design b: ANOVA
5. How T-test and ANOVA work
6. Non-parametric tests a, normality tests
7. Non-parametric tests b
8. Categorical data: Chi-square
9. **Sample size, power and effect size (luluah)**
10. **Alternatives to p value testing**
11. **Questions before exam**

unit menu

1. Be able to give the CHI-square formula (goodness of fit and contingency table)
2. Calculate a CHI-square by hand on an example with a single variable and conclude
3. Explain what is the different between goodness of fit and contingency table methods
4. Explain what is a two-way ANOVA, MANOVA and ANCOVA and be able to explain the differences between them

take away

end