16

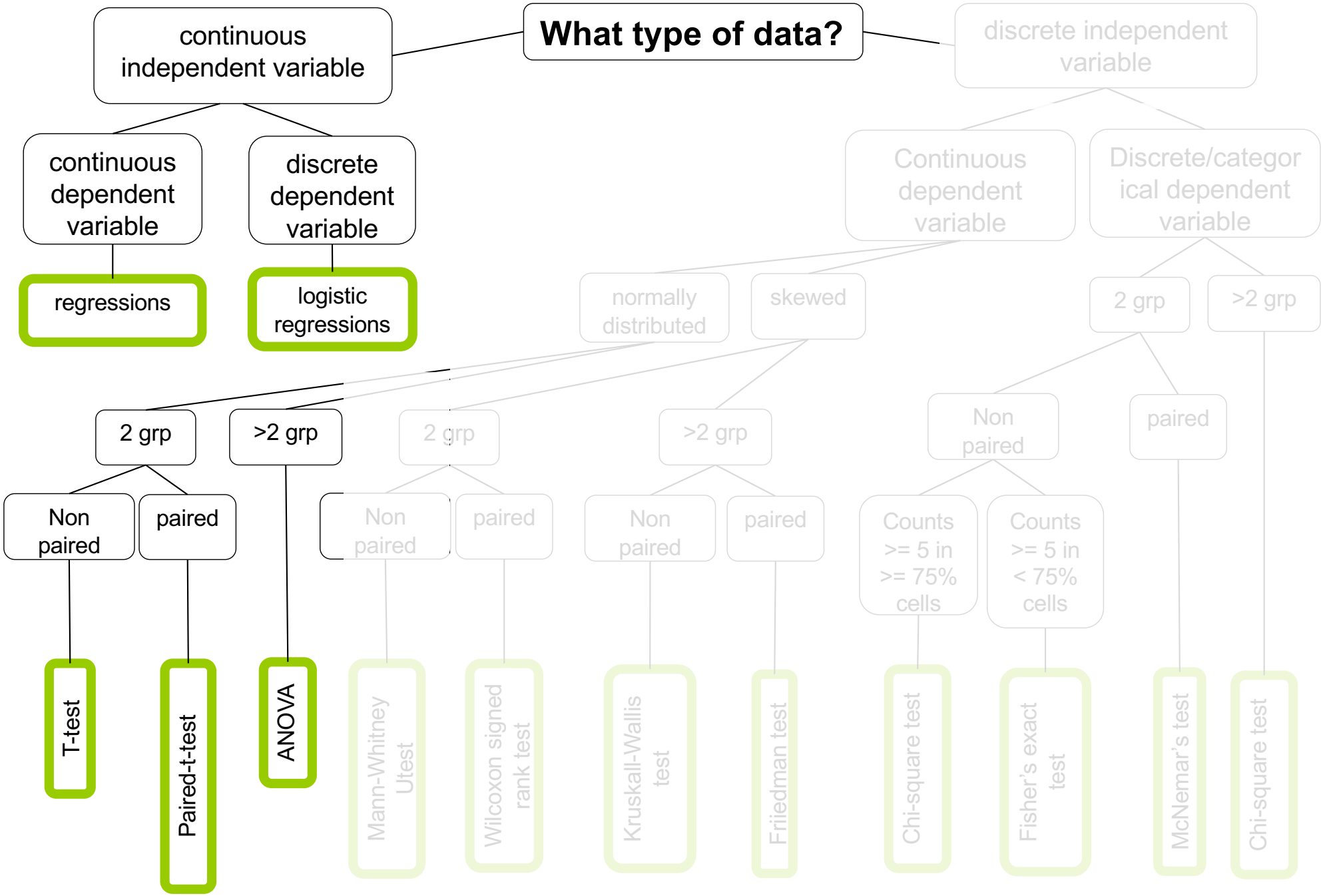
Normality and Homogeneity tests

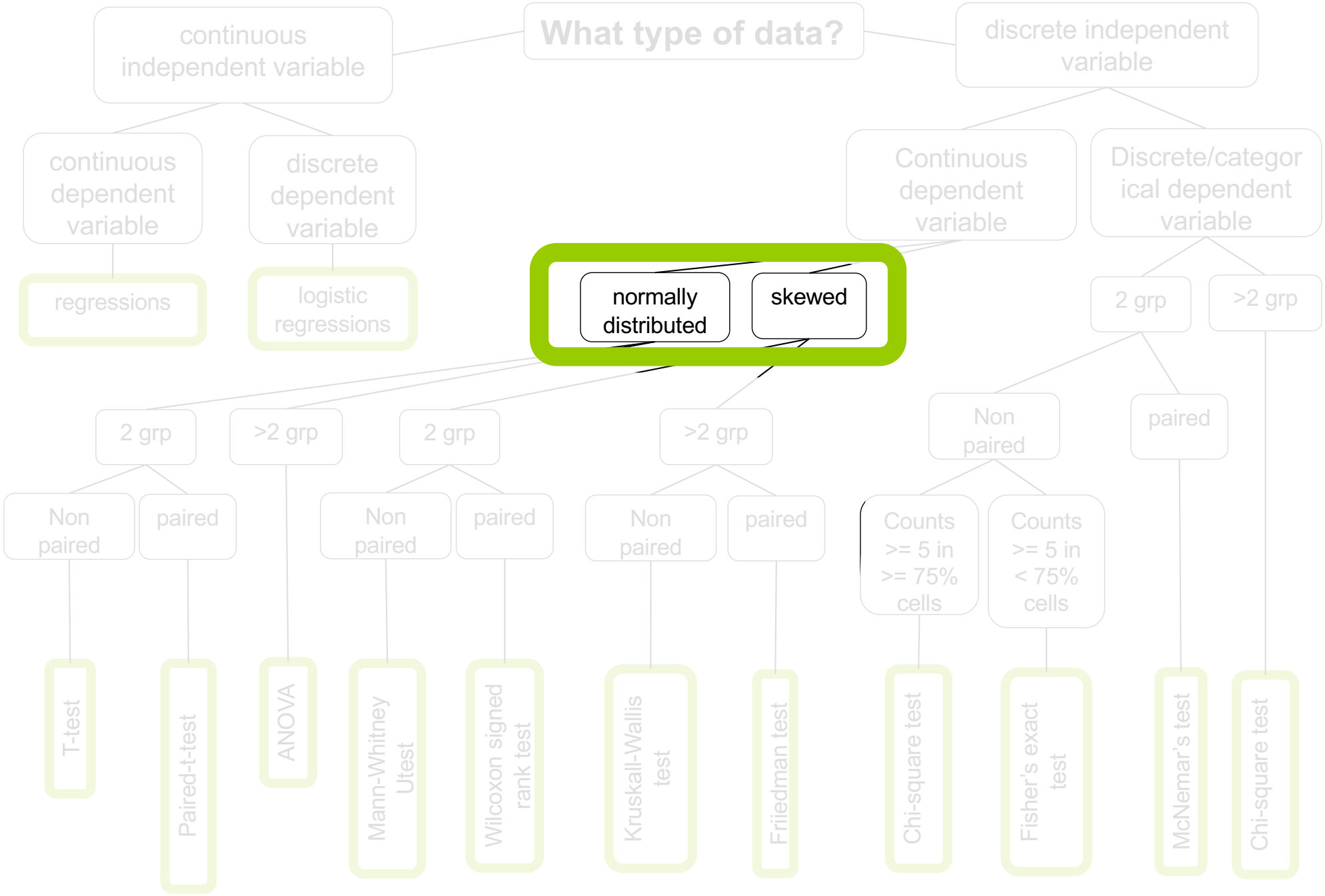
Probability and Statistics

COMS10011

Dr. Anne Roudaut
csxar@bristol.ac.uk

(Thanks S. Massa, Oxford)


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
Non paired
paired
2 grp
>2 grp
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


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

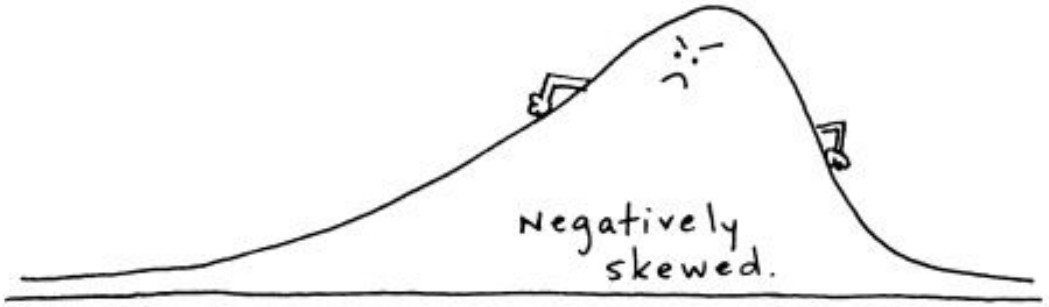
today::

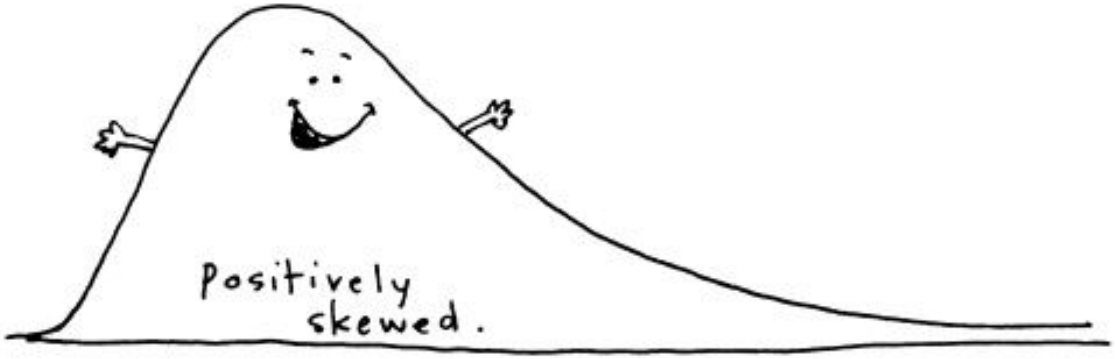
look into **assumption of normality** and **of homogeneity**

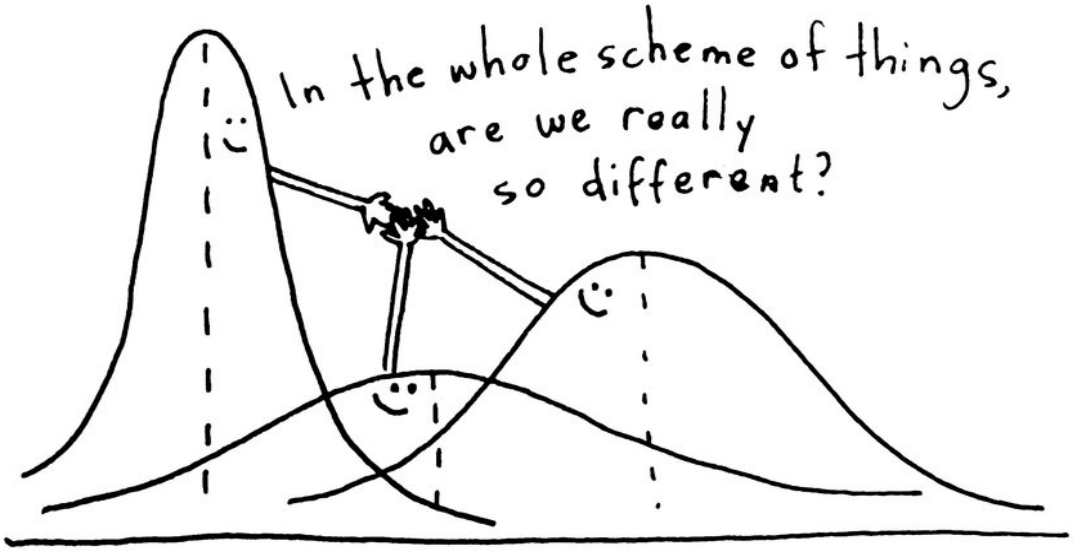
see what to do otherwise

tests we have seen so far (t-test, anova) assume that data **follow curve of normal distribution** and have **homogenous variance**

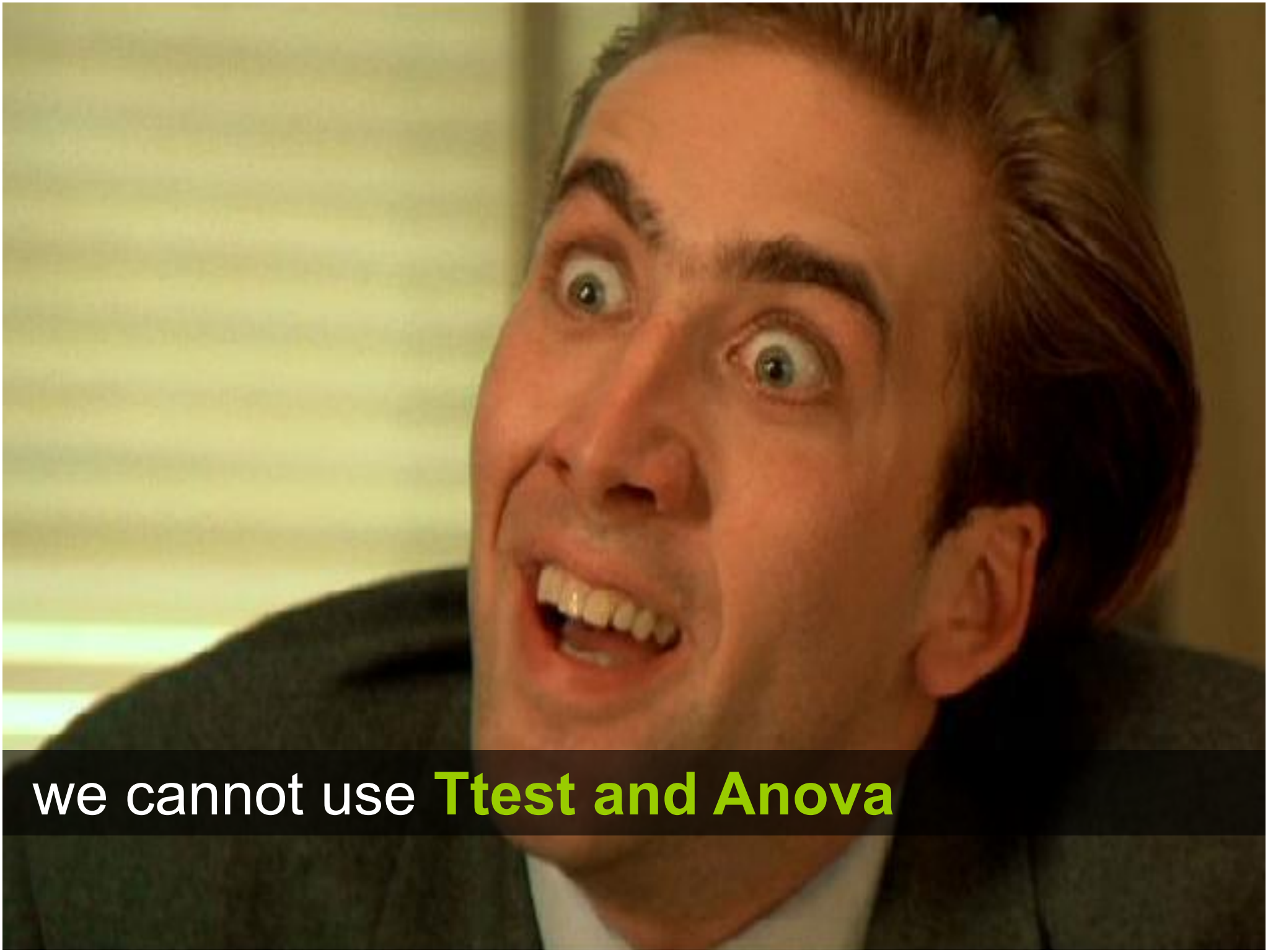
but if we have
distributions like this …







… or that

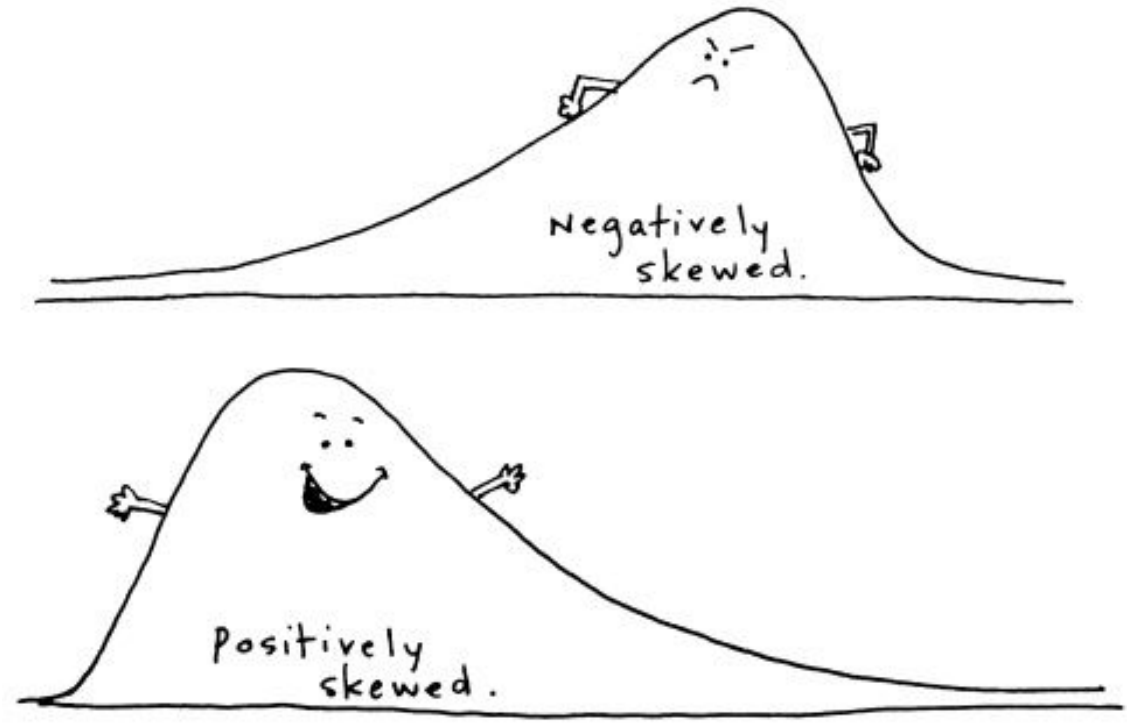

we cannot use Ttest and Anova

use parametric tests (ttest, anova)

if data **follow curve of normal distribution** with **homogeneous variance**

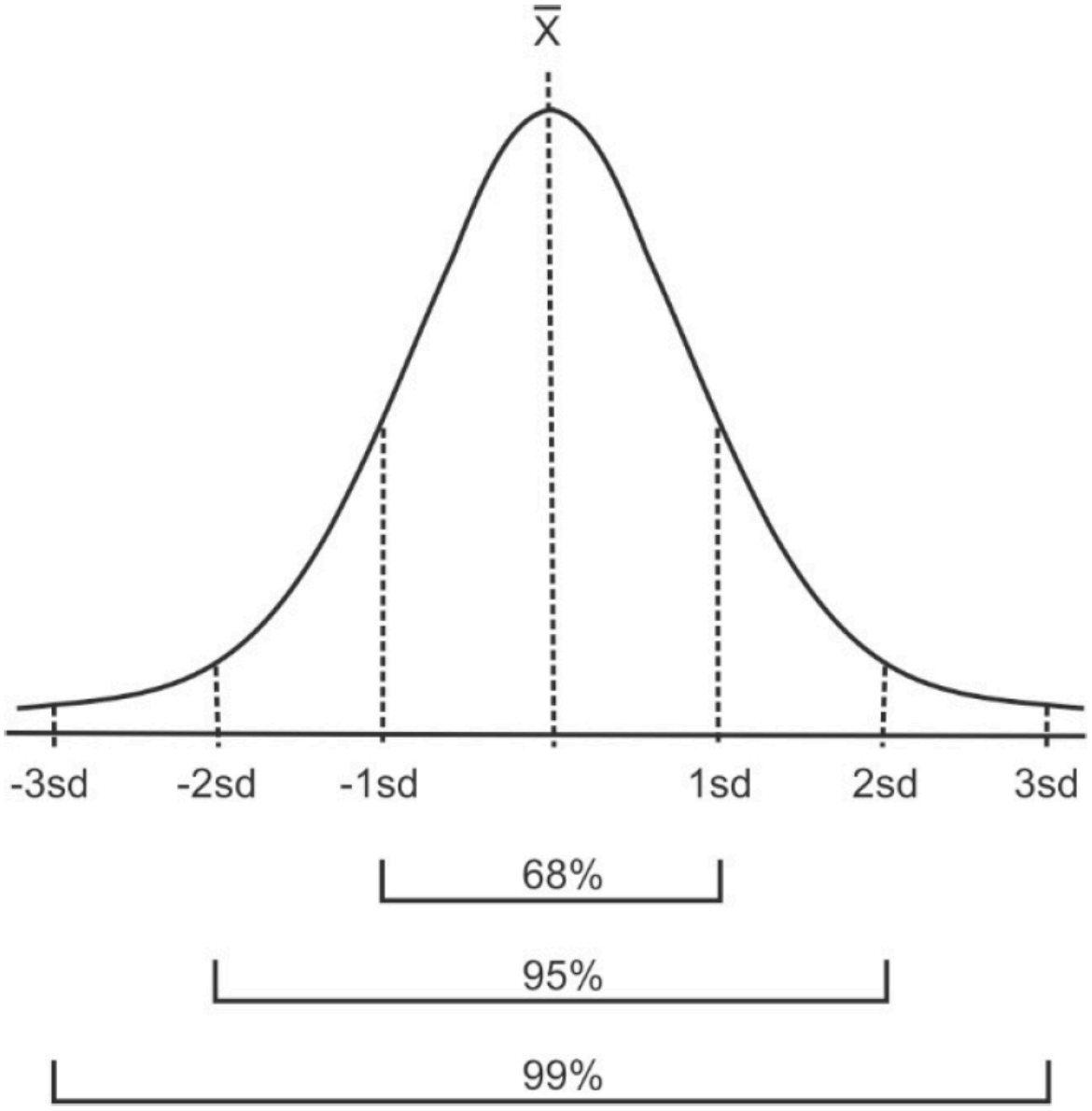
otherwise …

use non-parametric tests

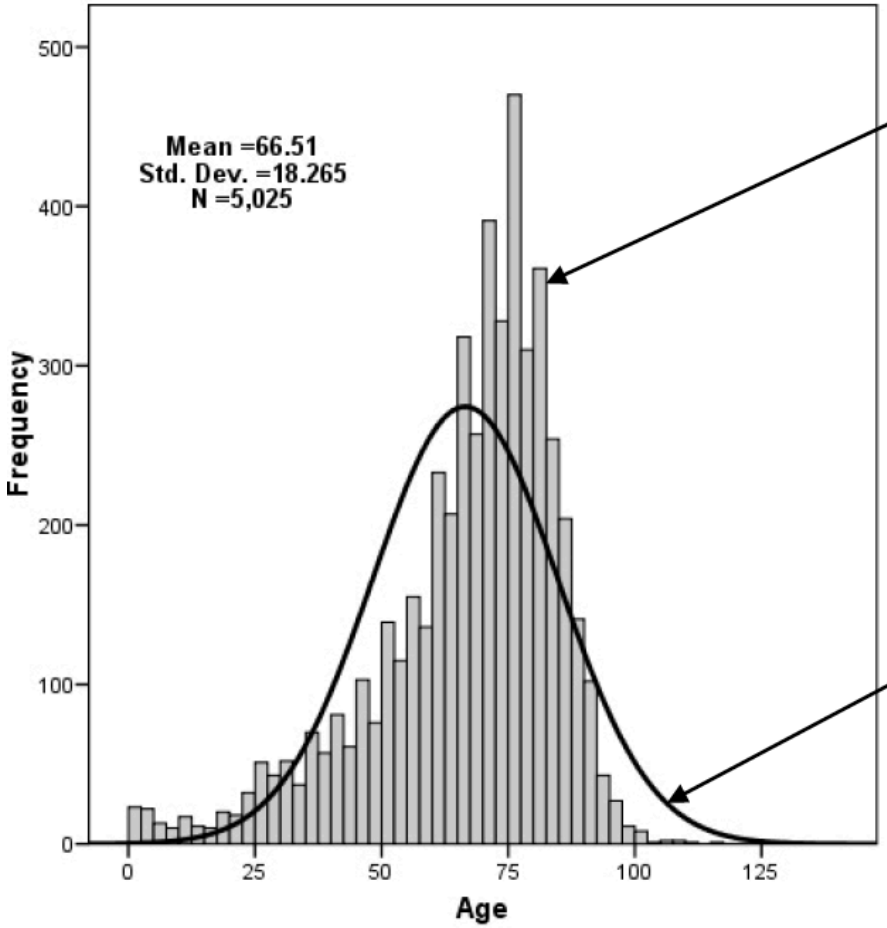


assumption of normality

given the **mean** and **standard deviation** of a dataset = a theoretical normal distribution has those proportions (Z-score)



this theoretical normal distribution can then be compared to the actual distribution of the data.



The ***actual*** data distribution that has a mean of 66.51 and a standard deviation of 18.265.

Theoretical normal distribution ***calculated*** from a mean of 66.51 and a standard deviation of 18.265.

<are the actual data statistically different than the computed normal curve? >

several methods to check that, we are only going to look at two: **Kolmogorov-Smirnov test** and **Shapiro-Wilks test**

Kolmogorov-Smirnov

- works best for data sets with n > 50
- not sensitive to problems in the tails

Shapiro-Wilks

- works best for data sets with n < 50
- doesn't work well if several values are same

Kolmogorov-Smirnov test

$$D_n = \max_x |F_{\text{exp}}(x) - F_{\text{obs}}(x)|$$

cumulative distribution function observed

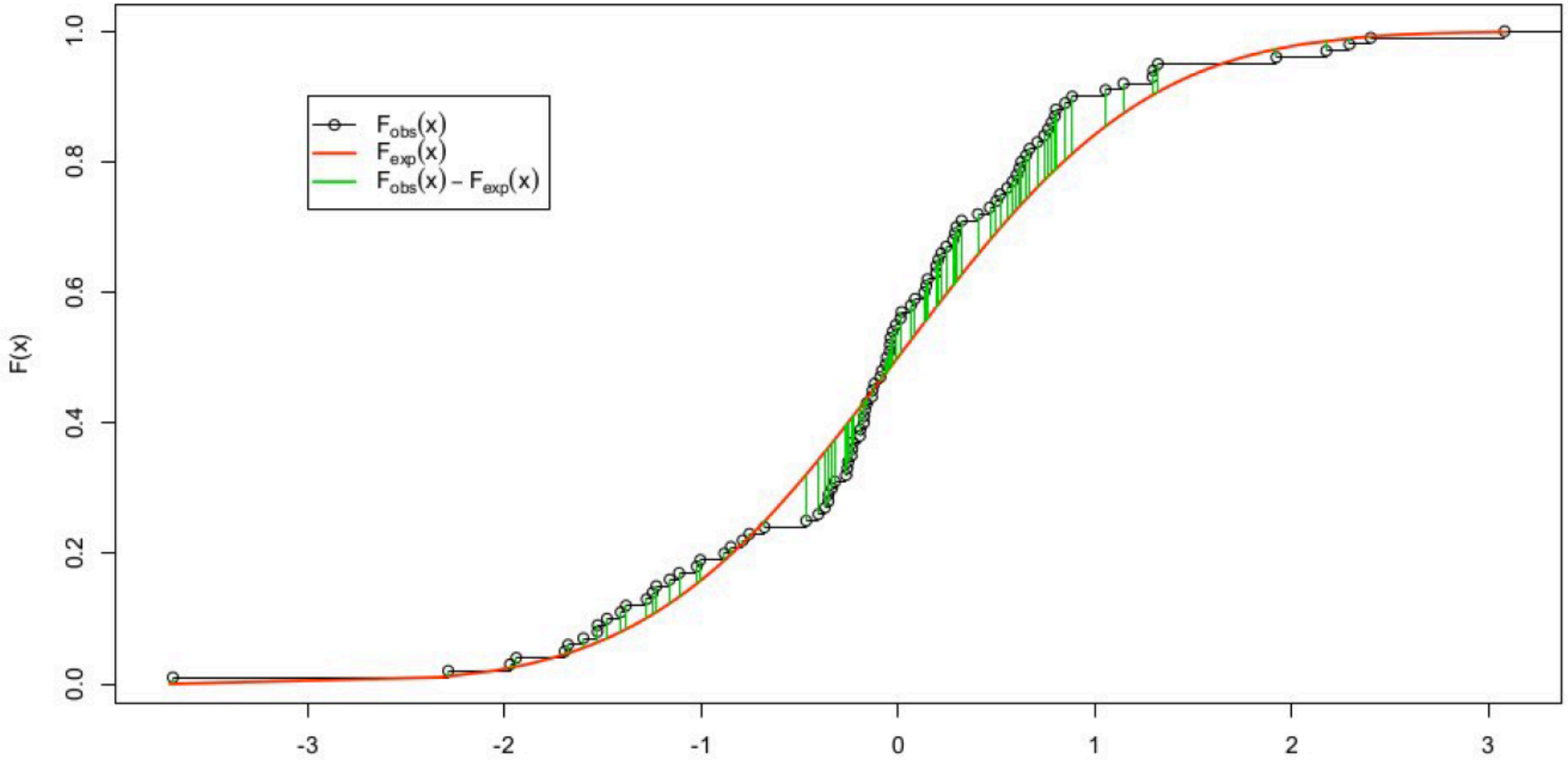
cumulative distribution function expected

can generate a **p-value**

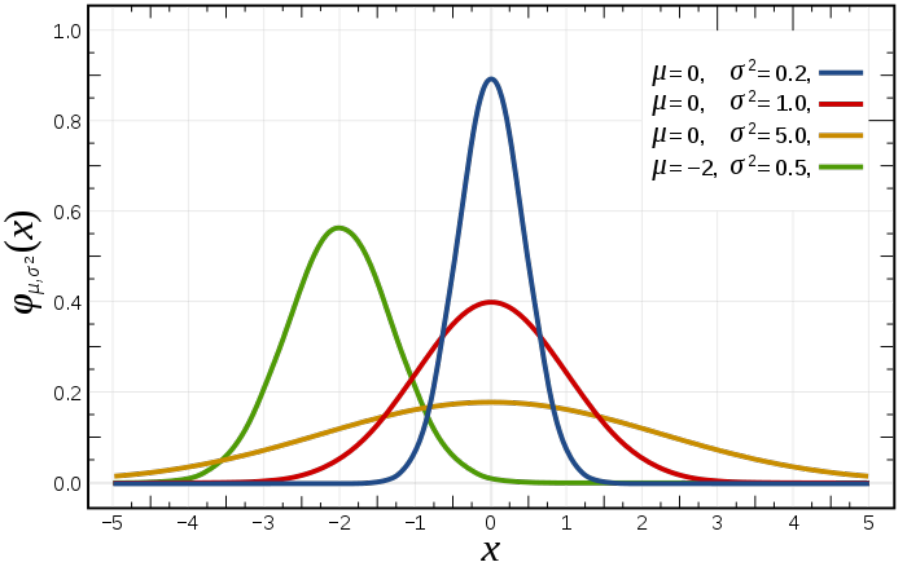
-0.16	-0.68	-0.32	-0.85	0.89	-2.28	0.63	0.41	0.15	0.74
1.30	-0.13	0.80	-0.75	0.28	-1.00	0.14	-1.38	-0.04	-0.25
-0.17	1.29	0.47	-1.23	0.21	-0.04	0.07	-0.08	0.32	-0.17
0.13	-1.94	0.78	0.19	-0.12	-0.19	0.76	-1.48	-0.01	0.20
-1.97	-0.37	3.08	-0.40	0.80	0.01	1.32	-0.47	2.29	-0.26
-1.52	-0.06	-1.02	1.06	0.60	1.15	1.92	-0.06	-0.19	0.67
0.29	0.58	0.02	2.18	-0.04	-0.13	-0.79	-1.28	-1.41	-0.23
0.65	-0.26	-0.17	-1.53	-1.69	-1.60	0.09	-1.11	0.30	0.71
-0.88	-0.03	0.56	-3.68	2.40	0.62	0.52	-1.25	0.85	-0.09
-0.23	-1.16	0.22	-1.68	0.50	-0.35	-0.35	-0.33	-0.24	0.25

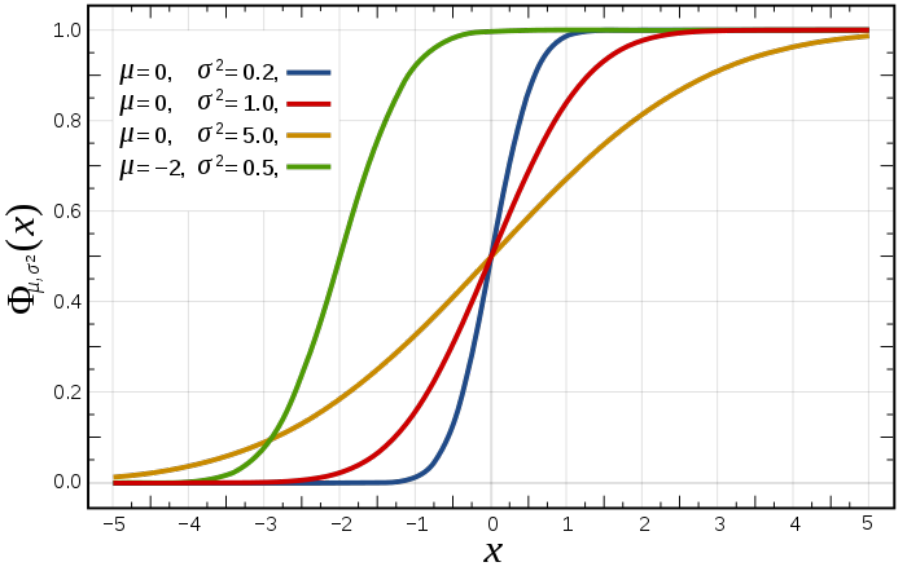
does the following sample of n=100 comes from a normality distributed population?

intuitively, we search for the maximum absolute distance between our data cumulative distribution function and the normal cumulative distribution function



another way to represent **probability density function** = **cumulative distribution function**, represents probability that the variable takes a value less than or equal to x





-0.16	-0.68	-0.32	-0.85	0.89	-2.28	0.63	0.41	0.15	0.74
1.30	-0.13	0.80	-0.75	0.28	-1.00	0.14	-1.38	-0.04	-0.25
-0.17	1.29	0.47	-1.23	0.21	-0.04	0.07	-0.08	0.32	-0.17
0.13	-1.94	0.78	0.19	-0.12	-0.19	0.76	-1.48	-0.01	0.20
-1.97	-0.37	3.08	-0.40	0.80	0.01	1.32	-0.47	2.29	-0.26
-1.52	-0.06	-1.02	1.06	0.60	1.15	1.92	-0.06	-0.19	0.67
0.29	0.58	0.02	2.18	-0.04	-0.13	-0.79	-1.28	-1.41	-0.23
0.65	-0.26	-0.17	-1.53	-1.69	-1.60	0.09	-1.11	0.30	0.71
-0.88	-0.03	0.56	-3.68	2.40	0.62	0.52	-1.25	0.85	-0.09
-0.23	-1.16	0.22	-1.68	0.50	-0.35	-0.35	-0.33	-0.24	0.25

does the following sample of n=100 comes from a normality distributed population?

1. order the data:

-3.68	-2.28	-1.97	-1.94	-1.69	-1.68	-1.60	-1.53	-1.52	-1.48
-1.41	-1.38	-1.28	-1.25	-1.23	-1.16	-1.11	-1.02	-1.00	-0.88
-0.85	-0.79	-0.75	-0.68	-0.47	-0.40	-0.37	-0.35	-0.35	-0.33
-0.32	-0.26	-0.26	-0.25	-0.24	-0.23	-0.23	-0.19	-0.19	-0.17
-0.17	-0.17	-0.16	-0.13	-0.13	-0.12	-0.09	-0.08	-0.06	-0.06
-0.04	-0.04	-0.04	-0.03	-0.01	0.01	0.02	0.07	0.09	0.13
0.14	0.15	0.19	0.20	0.21	0.22	0.25	0.28	0.29	0.30
0.32	0.41	0.47	0.50	0.52	0.56	0.58	0.60	0.62	0.63
0.65	0.67	0.71	0.74	0.76	0.78	0.80	0.80	0.85	0.89
1.06	1.15	1.29	1.30	1.32	1.92	2.18	2.29	2.40	3.08

2. compute the empirical distribution function

$$F_{\text{obs}}(-3.68) = \frac{1}{100}, \quad F_{\text{obs}}(-2.28) = \frac{2}{100}, \ldots, \quad F_{\text{obs}}(3.08) = 1$$

F_{obs}

0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18	0.19	0.20
0.21	0.22	0.23	0.24	0.25	0.26	0.27	0.28	0.29	0.30
0.31	0.32	0.33	0.34	0.35	0.36	0.37	0.38	0.39	0.40
0.41	0.42	0.43	0.44	0.45	0.46	0.47	0.48	0.49	0.50
0.51	0.52	0.53	0.54	0.55	0.56	0.57	0.58	0.59	0.60
0.61	0.62	0.63	0.64	0.65	0.66	0.67	0.68	0.69	0.70
0.71	0.72	0.73	0.74	0.75	0.76	0.77	0.78	0.79	0.80
0.81	0.82	0.83	0.84	0.85	0.86	0.87	0.88	0.89	0.90
0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99	1.00

3. for each observation xi from the data, compute:

$$F_{\text{exp}}(x_i) = P(Z \leq x_i)$$

(in this case, the expected distribution function is standard normal so use the normal table)

F_{exp}

0.000	0.011	0.024	0.026	0.045	0.047	0.055	0.064	0.064	0.070
0.080	0.084	0.101	0.107	0.110	0.123	0.133	0.154	0.158	0.189
0.198	0.215	0.226	0.249	0.321	0.343	0.356	0.362	0.363	0.369
0.375	0.396	0.399	0.400	0.407	0.409	0.410	0.423	0.425	0.432
0.432	0.434	0.437	0.447	0.449	0.453	0.464	0.468	0.476	0.477
0.484	0.484	0.485	0.490	0.496	0.505	0.508	0.526	0.535	0.553
0.557	0.560	0.577	0.577	0.582	0.588	0.597	0.610	0.614	0.617
0.627	0.658	0.680	0.692	0.698	0.711	0.720	0.727	0.732	0.735
0.743	0.748	0.761	0.771	0.777	0.783	0.788	0.789	0.803	0.812
0.854	0.874	0.902	0.903	0.907	0.973	0.985	0.989	0.992	0.999

now we have two tables Fobs and Fexp …

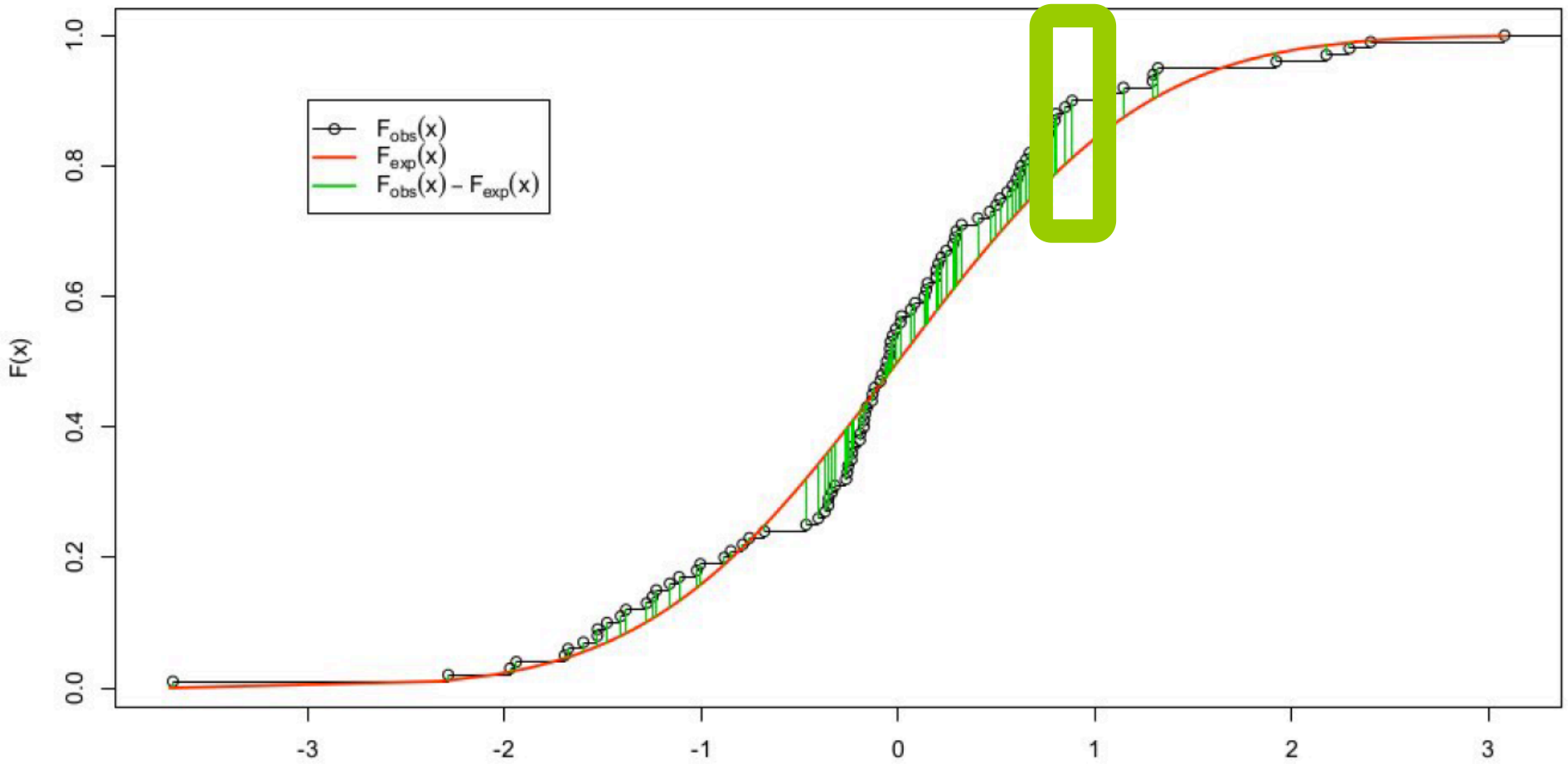
4. lets compute the absolute difference between the two and find the highest value

0.010	0.009	0.006	0.014	0.004	0.014	0.015	0.017	0.026	0.031
0.031	0.036	0.030	0.034	0.041	0.037	0.037	0.026	0.031	0.011
0.012	0.005	0.003	0.008	0.069	0.085	0.086	0.083	0.073	0.071
0.064	0.077	0.067	0.061	0.055	0.049	0.039	0.045	0.035	0.033
0.023	0.013	0.006	0.008	0.002	0.008	0.006	0.012	0.014	0.024
0.026	0.036	0.046	0.052	0.054	0.056	0.062	0.052	0.054	0.048
0.054	0.060	0.055	0.061	0.067	0.073	0.071	0.070	0.076	0.082
0.084	0.061	0.049	0.049	0.052	0.048	0.051		0.058	0.064
0.068	0.071	0.069	0.070	0.074	0.078	0.082	0.092	0.088	0.087
0.055	0.045	0.029	0.037	0.043	0.013	0.015		0.002	0.001

$$D_n = \max_x |F_{\text{exp}}(x) - F_{\text{obs}}(x)|$$

this is the D searched

we have calculated the maximum absolute distance between expected and observed distribution functions



5. at 95% level the critical value is approximately given by

$$D_{\text{crit},0.05} = \frac{1.36}{\sqrt{n}}$$

we have a sample size of n = 100 so Dcrit = 0.136

and 0.092 < **0.136** **= do not reject null hypothesis**

$$D_{\text{crit},0.05} = \frac{1.36}{\sqrt{n}}$$

n	α 0.01	α 0.05	α 0.1	α 0.15	α 0.2
1	0.995	0.975	0.950	0.925	0.900
2	0.929	0.842	0.776	0.726	0.684
3	0.828	0.708	0.642	0.597	0.565
4	0.733	0.624	0.564	0.525	0.494
5	0.669	0.565	0.510	0.474	0.446
6	0.618	0.521	0.470	0.436	0.410
7	0.577	0.486	0.438	0.405	0.381
8	0.543	0.457	0.411	0.381	0.358
9	0.514	0.432	0.388	0.360	0.339
10	0.490	0.410	0.368	0.342	0.322
11	0.468	0.391	0.352	0.326	0.307
12	0.450	0.375	0.338	0.313	0.295
13	0.433	0.361	0.325	0.302	0.284
14	0.418	0.349	0.314	0.292	0.274
15	0.404	0.338	0.304	0.283	0.266
16	0.392	0.328	0.295	0.274	0.258
17	0.381	0.318	0.286	0.266	0.250
18	0.371	0.309	0.278	0.259	0.244
19	0.363	0.301	0.272	0.252	0.237
20	0.356	0.294	0.264	0.246	0.231
25	0.320	0.270	0.240	0.220	0.210
30	0.290	0.240	0.220	0.200	0.190
35	0.270	0.230	0.210	0.190	0.180
40	0.250	0.210	0.190	0.180	0.170
45	0.240	0.200	0.180	0.170	0.160
50	0.230	0.190	0.170	0.160	0.150
OVER 50	$\frac{1.63}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.07}{\sqrt{n}}$

there is a plethora of **tables / sampling distributions** that are established and are the basis of all statistic tests

so 0.092 < **0.136** so null hypothesis not rejected

H0: the two distribution are similar

conclusion ?

here is a tricky bit … remember lecture on hypothesis testing, we cannot prove that two things are equal so we are going to **assume** that the normality is met

which is why we call this **assumption of normality**

we can assume data follow a normal distribution and proceed with parametric test such as anova

more example on GitHub repository or at
http://www.real-statistics.com/tests-normality-and-symmetry/statistical-tests-normality-symmetry/kolmogorov-smirnov-test/

```
y <-c(-0.16,-0.68,-0.32,-0.85,0.89,-2.28,0.63,0.41,0.15,0.74,1.30,-0.13,0.80,-0.75,0.28,-
1.00,0.14,-1.38,-0.04,-0.25,-0.17,1.29,0.47,-1.23,0.21,-0.04,0.07,-0.08,0.32,-0.17,0.13,-
1.94,0.78,0.19,-0.12,-0.19,0.76,-1.48,-0.01,0.20,-1.97,-0.37,3.08,-0.40,0.80,0.01,1.32,-0.47,2.29,-
0.26,-1.52,-0.06,-1.02,1.06,0.60,1.15,1.92,-0.06,-0.19,0.67,0.29,0.58,0.02,2.18,-0.04,-0.13,-0.79,-
1.28,-1.41,-0.23,0.65,-0.26,-0.17,-1.53,-1.69,-1.60,0.09,-1.11,0.30,0.71,-0.88,-0.03,0.56,-
3.68,2.40,0.62,0.52,-1.25,0.85,-0.09,-0.23,-1.16,0.22,-1.68,0.50,-0.35,-0.35,-0.33,-0.24,0.25)
X <- rnorm(100)
ks.test(X,y)

Two-sample Kolmogorov-Smirnov test


data:  X and y
D = 0.19, p-value = 0.05410262
alternative hypothesis: two-sided

#note that if you run the code you will have
different D (because of the random rnorm
generation) but likely that your pvalue will
always be above 0.05
```

Kolmogorov-Smirnov works well with **sample size > 50**
but when the sample is smaller Shapiro-Wilks works best

Shapiro-Wilks test

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

x(i) is the ith order statistic

SS (sum of squared difference)

$$(a_1, \cdots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m^T)^{1/2}}, \textit{where } m = (m_1, \cdots, m_n)^T$$

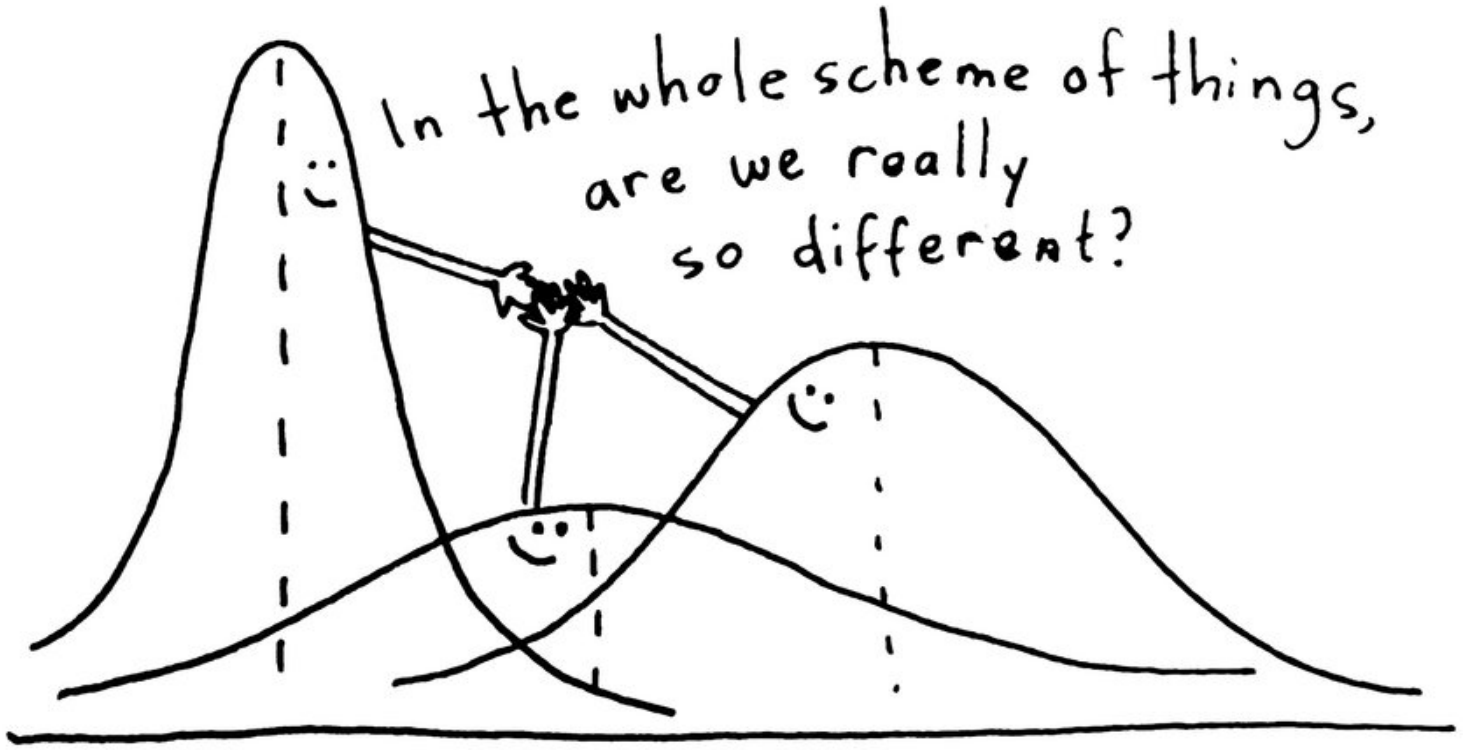
m1, ..., mn are the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution, and **V** is the covariance matrix of those order statistics.

can generate a **p-value**

```
y <-c(3.83, 3.16, 4.70, 3.97, 2.03, 2.87,
3.65, 5.09)
shapiro.test(y)
```

```
Shapiro-Wilk normality test

data:  y
W = 0.98317, p-value = 0.9769
```

= we cannot reject the null hypothesis and we assume the data is normally distributed



assumption of homogeneity

ANOVA we did when we tried to check if chocolate improves memorization

```
# we ran the one-way anova
dat = read.csv("HCIXP-anova.csv", header =
TRUE)
library(ez)
ezANOVA(dat,id,between=group,dv=score)

  Effect DFn DFd         F                 p
p<.05        ges
1  group   2  57 154.8886 9.056612e-24     *
0.8445923

$`Levene's Test for Homogeneity of Variance
  DFn DFd      SSn  SSd         F         p
p<.05
1   2  57 1.433333 29.3 1.394198 0.2563608
```

the levene's test checks for **homogeneity of variances** (null hypothesis is that all variances are equal)

if p-value < 0.05 means variances not equal and parametric tests such as ANOVA **are not suited** (need non-parametric tests)

if p-value > 0.05 we can **assume** that data have homogenous variance

we know how to check our data

… now what?

use parametric tests (ttest, anova)

if data **follow curve of normal distribution** with **homogeneous variance**

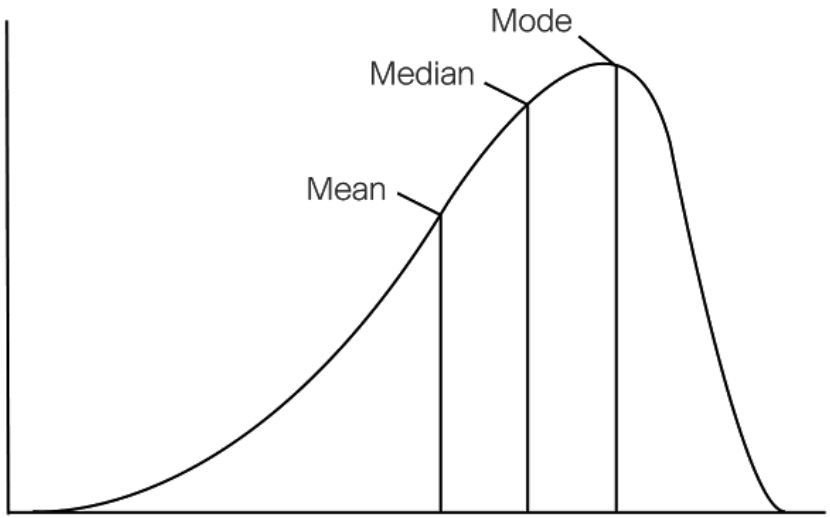
otherwise …

use non-parametric tests

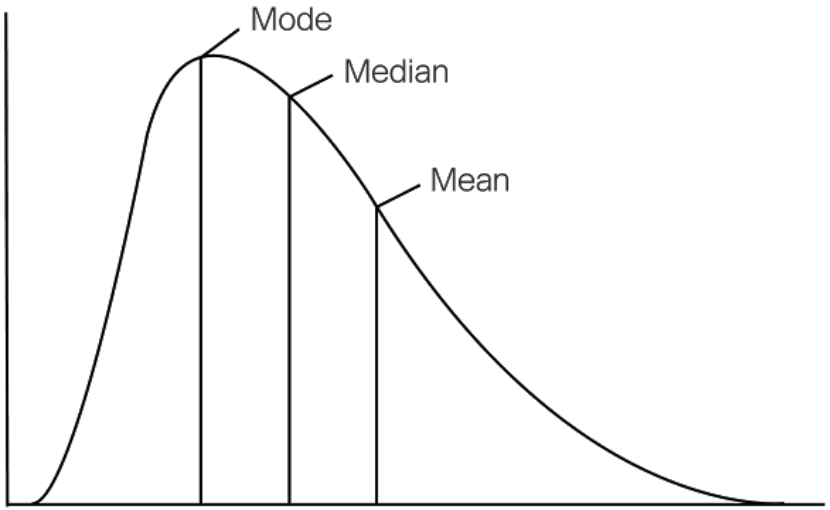
but if your data is not normally distributed you could also try to make it normal using **transformations**

... more generally because parametric tests are more robust than non-parametric ones

transformations

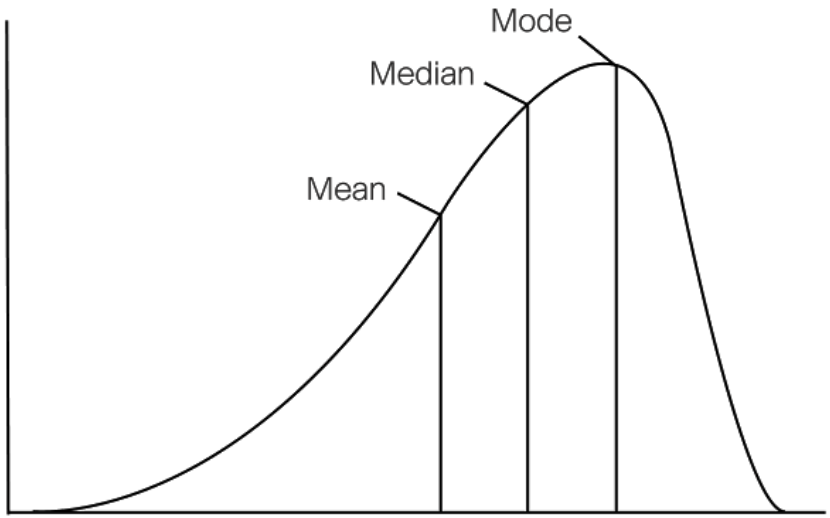


Left-Skewed (Negative Skewness)

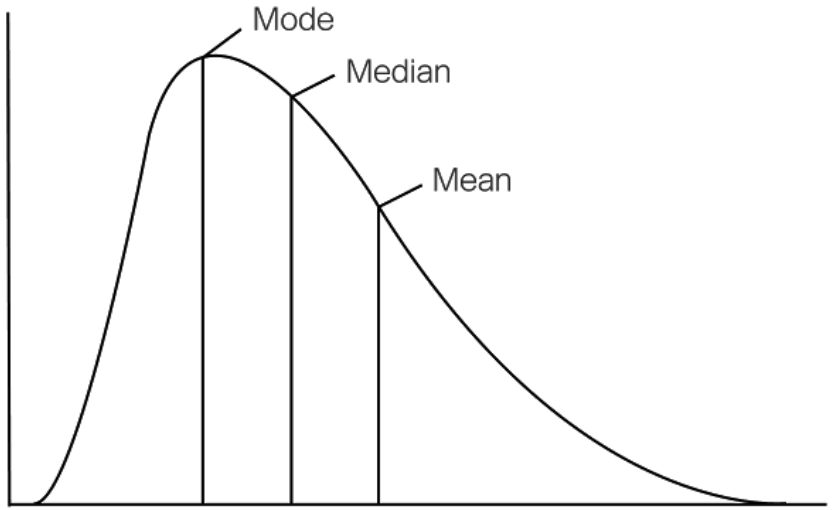


Right-Skewed (Negative Skewness)

common transformations for left skewed::
square root, cube root, log



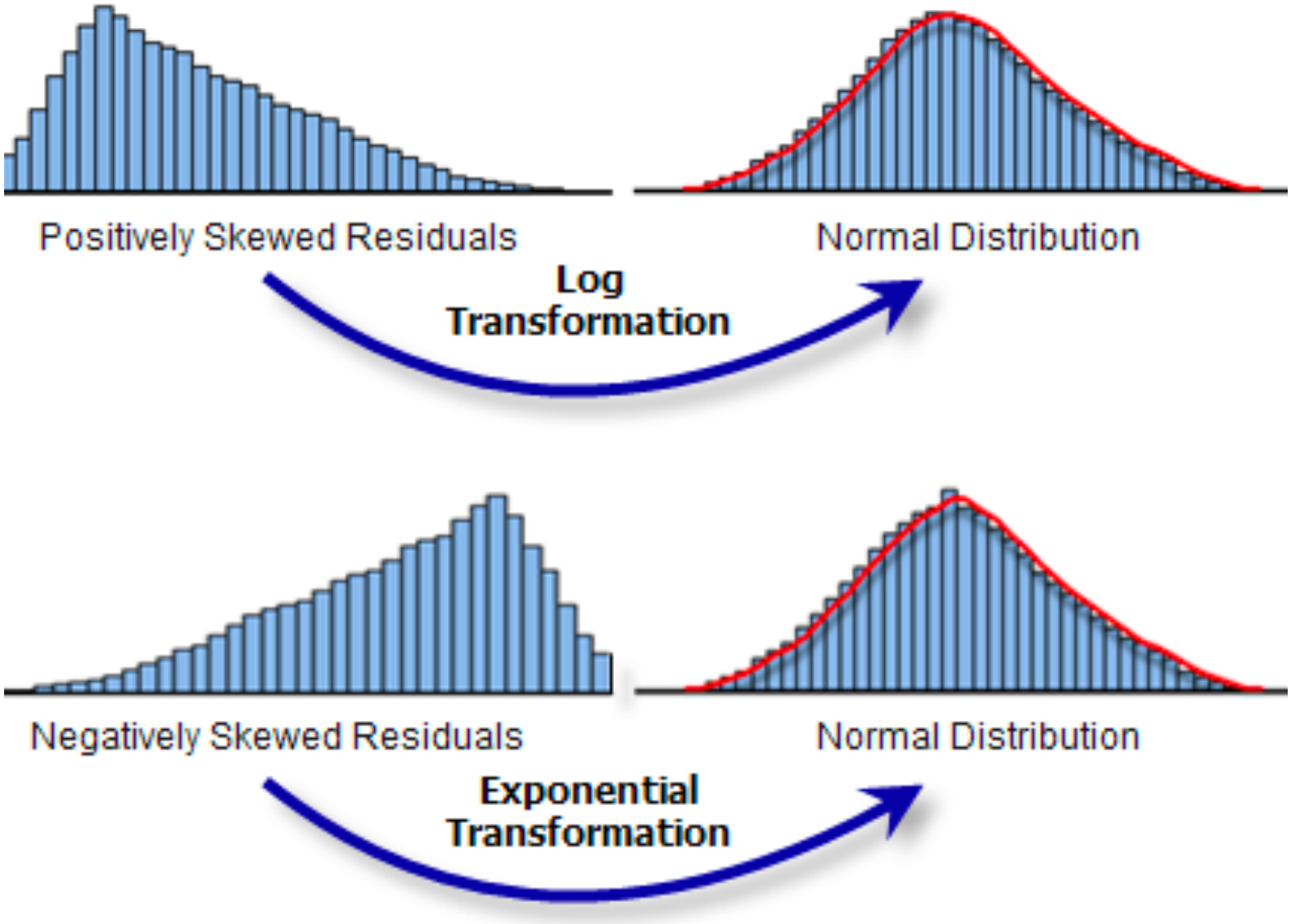
Left-Skewed (Negative Skewness)



Right-Skewed (Negative Skewness)

common transformations for right skewed::

square, cube root and logarithmic


Positively Skewed Residuals
Normal Distribution
Log
Transformation
Negatively Skewed Residuals
Normal Distribution
Exponential
Transformation

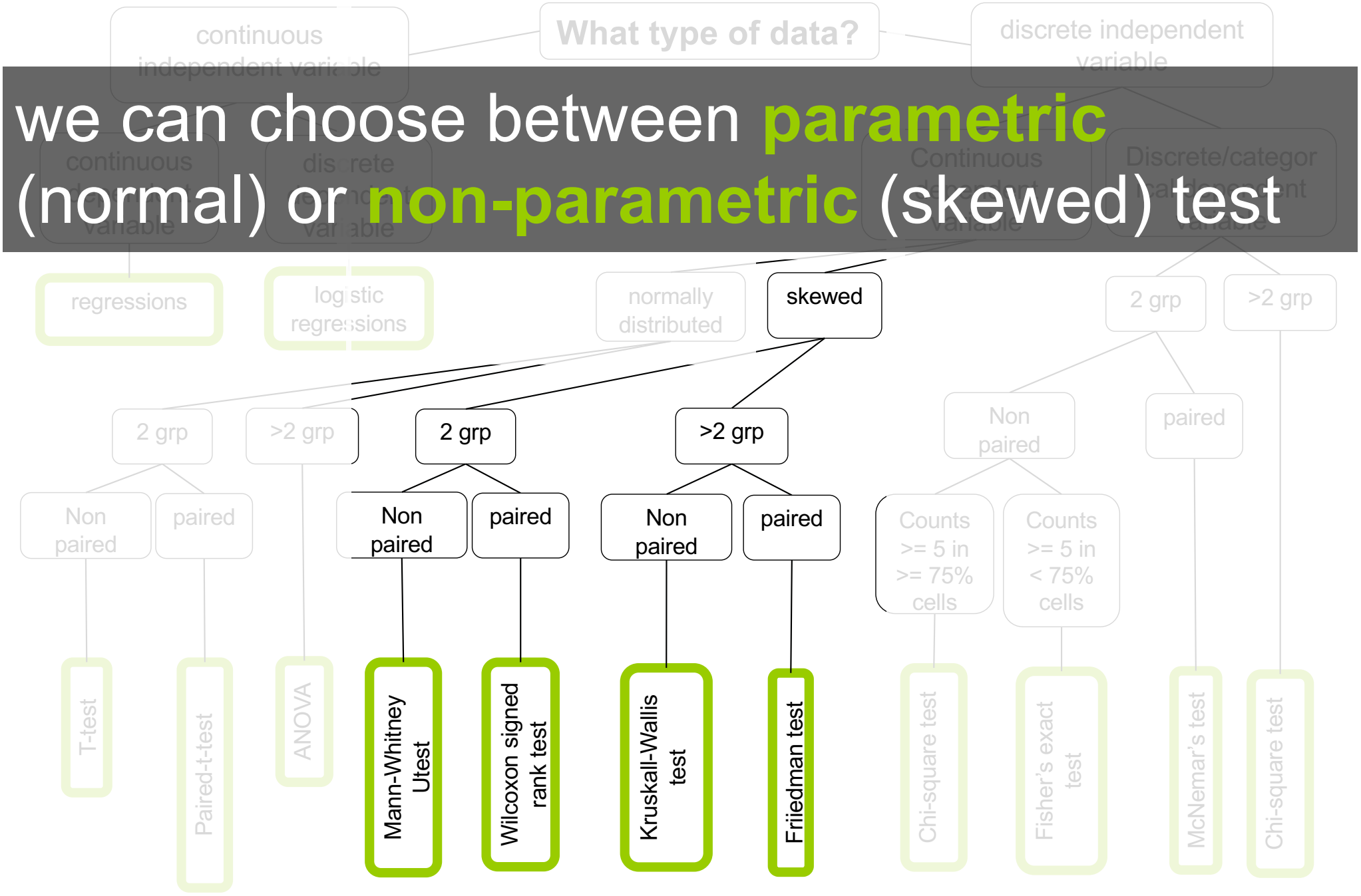
sometimes skewed distributions could come from **outliers**
so make sure to get rid of them!

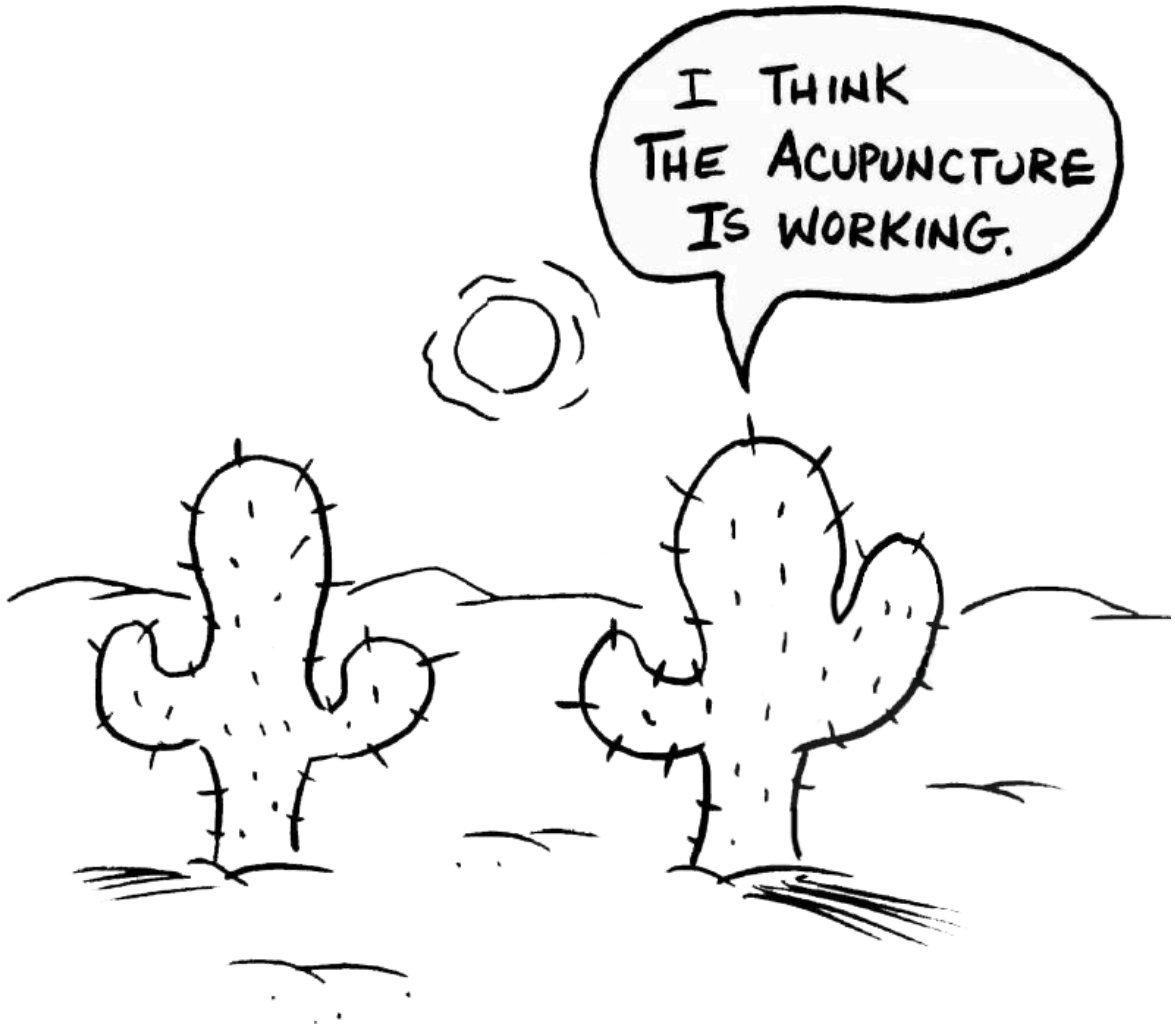
sometimes it does not work …

```
y <-c(-0.16,-0.68,-0.32,-0.85,0.89,-2.28,0.63,0.41,0.15,0.74,1.30,-0.13,0.80,-0.75,0.28,-
1.00,0.14,-1.38,-0.04,-0.25,-0.17,1.29,0.47,-1.23,0.21,-0.04,0.07,-0.08,0.32,-0.17,0.13,-
1.94,0.78,0.19,-0.12,-0.19,0.76,-1.48,-0.01,0.20,-1.97,-0.37,3.08,-0.40,0.80,0.01,1.32,-0.47,2.29,-
0.26,-1.52,-0.06,-1.02,1.06,0.60,1.15,1.92,-0.06,-0.19,0.67,0.29,0.58,0.02,2.18,-0.04,-0.13,-0.79,-
1.28,-1.41,-0.23,0.65,-0.26,-0.17,-1.53,-1.69,-1.60,0.09,-1.11,0.30,0.71,-0.88,-0.03,0.56,-
3.68,2.40,0.62,0.52,-1.25,0.85,-0.09,-0.23,-1.16,0.22,-1.68,0.50,-0.35,-0.35,-0.33,-0.24,0.25)
hist(y)


y_sqrt = sqrt(y) #cube root
hist(y_sqrt)


# or y_cub = sign(y) * abs(y)^(1/3) #square root
# or y_log = log(y) #logarithm
```

if your transformation makes your data follow a normal distribution then do all the subsequent statistical test with that new dependent variable (e.g. y_sqrt instead of y)

you have tried everything and still not good?


What type of data?
continuous independent variable
discrete independent variable
we can choose between parametric (normal) or non-parametric (skewed) test
regressions
logistic regressions
normally distributed
skewed
2 grp
>2 grp
Non paired
paired
T-test
Paired-t-test
ANOVA
Mann-Whitney Utest
Wilcoxon signed rank test
Kruskall-Wallis test
Friedman test
Counts >= 5 in >= 75% cells
Counts >= 5 in < 75% cells
Chi-square test
Fisher's exact test
McNemar's test
Chi-square test



acupuncture exercise

What is your hypothesis?

(a sentence that can derive a test)

What is your hypothesis?

(a sentence that can derive a test)

H = sticking needles in place XX of a participants will reduce their back pain

What are the dependent and independent variables?

1. What are our two conditions?

What are the dependent and independent variables?

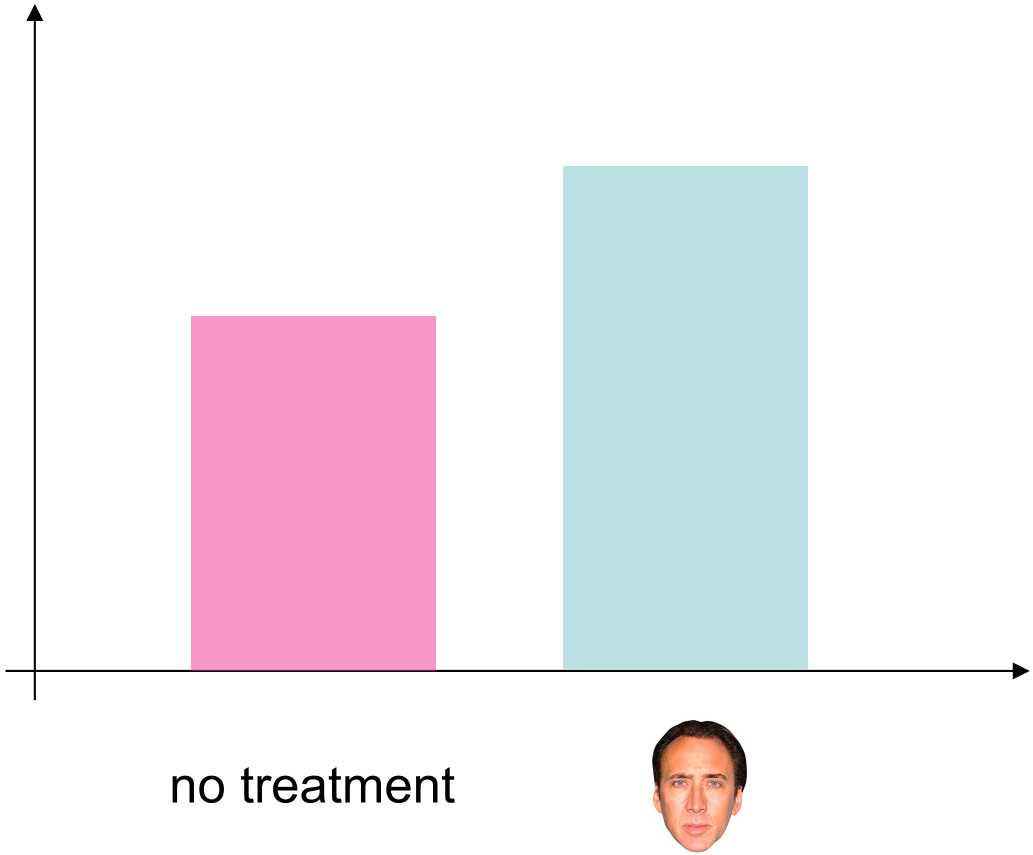
1. What are our two conditions?

IV = condition 1: receive acupuncture treatment
condition 2: do no receive acupuncture treatment

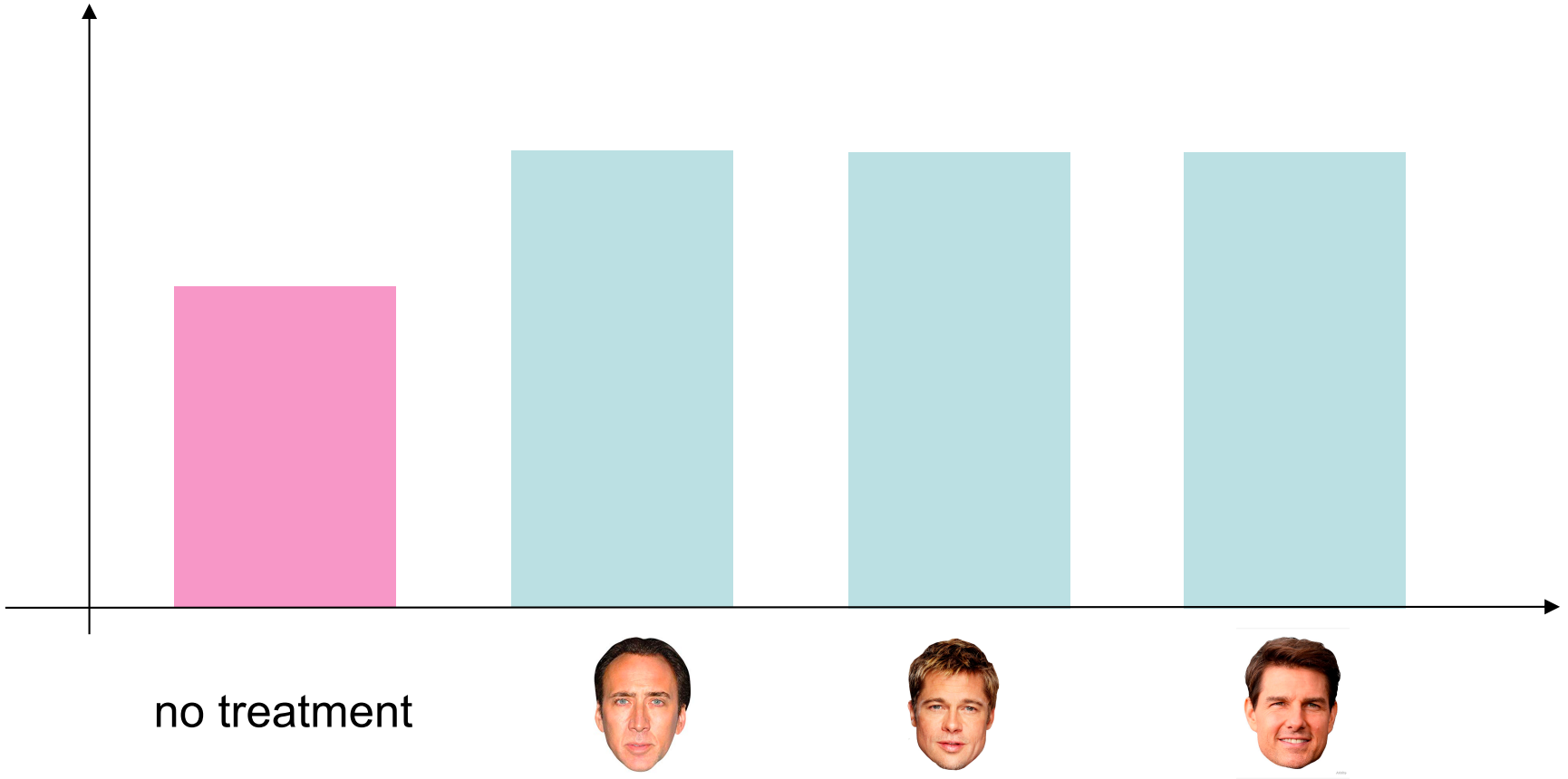
would that really work?

instead of acupuncture, imagine another treatment for back pain such as …

rubbing photos of Nicolas Cage on patients' body improve back pain



does it prove rubbing photos of NC is the miracle treatment?



and now?

does it prove rubbing photos of celebrities is the miracle treatment?

nope … may be the rubbing action does the deal

… or going to a “doctor” / practice to do a study that is supposed to improve your back does the deal

= placebo effect

What are the dependent and independent variables?

1. What are our two conditions?

IV = condition 1: acupuncture

condition 2: **fake acupuncture (needle not placed properly)**

What are the dependent and independent variables?

1. What are our two conditions?

IV = condition 1: acupuncture

condition 2: **fake acupuncture (needle not placed properly)**

2. What are you measuring and how?

What are the dependent and independent variables?

1. What are our two conditions?

IV = condition 1: acupuncture

condition 2: **fake acupuncture (needle not placed properly)**

2. What are you measuring and how?

DV = pain level

On a scale of 1 to 5 rate your back-pain level

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

true that back pain could be very different depending on participants,

you could make sure to recruit people with same problem

you could also ask the question before and after the treatment **= your DV is not the pain level but the difference in pain level from before to after**

Is this a within or between-subjects experiment?

Do they do each condition or only one?

Between subject (X participants do acupuncture, X others to fake acupuncture)

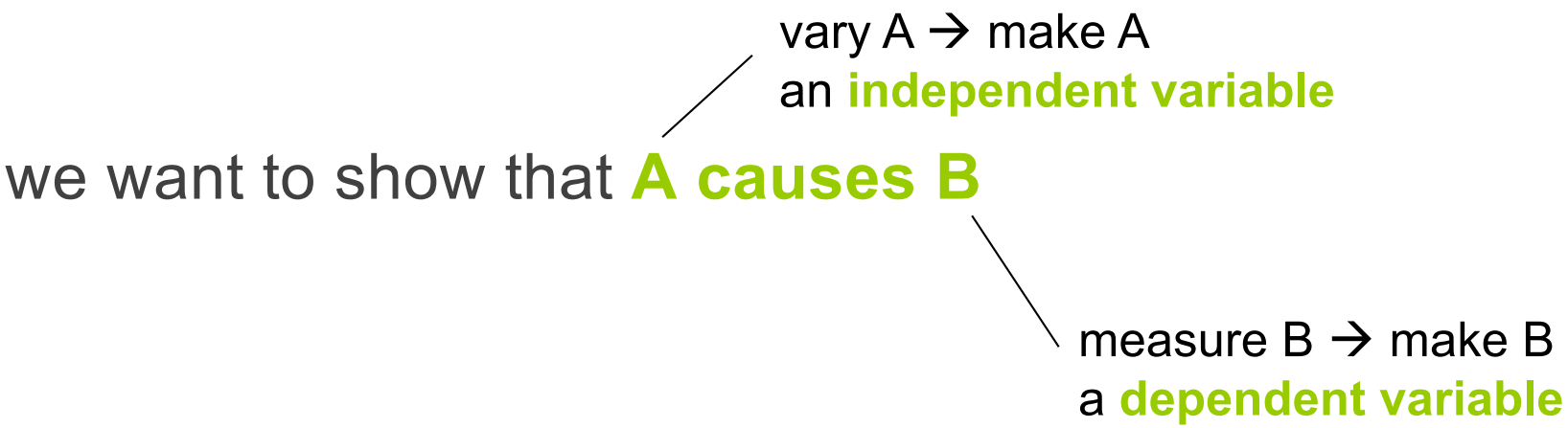
why is this a better choice?

imagine a within-participants

10 participants, invite them, ask for the pain level, make them do a treatment (won't know if acupuncture or fake)

Just after (or even a week after), ask again pain level, make them do the other treatment

arguable: what if the first test totally fix their back issue?


vary A → make A
an independent variable
we want to show that A causes B
measure B → make B
a dependent variable

summary

1. Give the names of tests we can use to check normality and explain their differences and when to use them
2. Explain what is the goal of a test of homogeneity of variance and what to do if the variances are not equal
3. I will **not ask** you to them by hand in the exam
4. Explain what to do if data not normal (transforming the data or using non-parametric tests)

take away

end