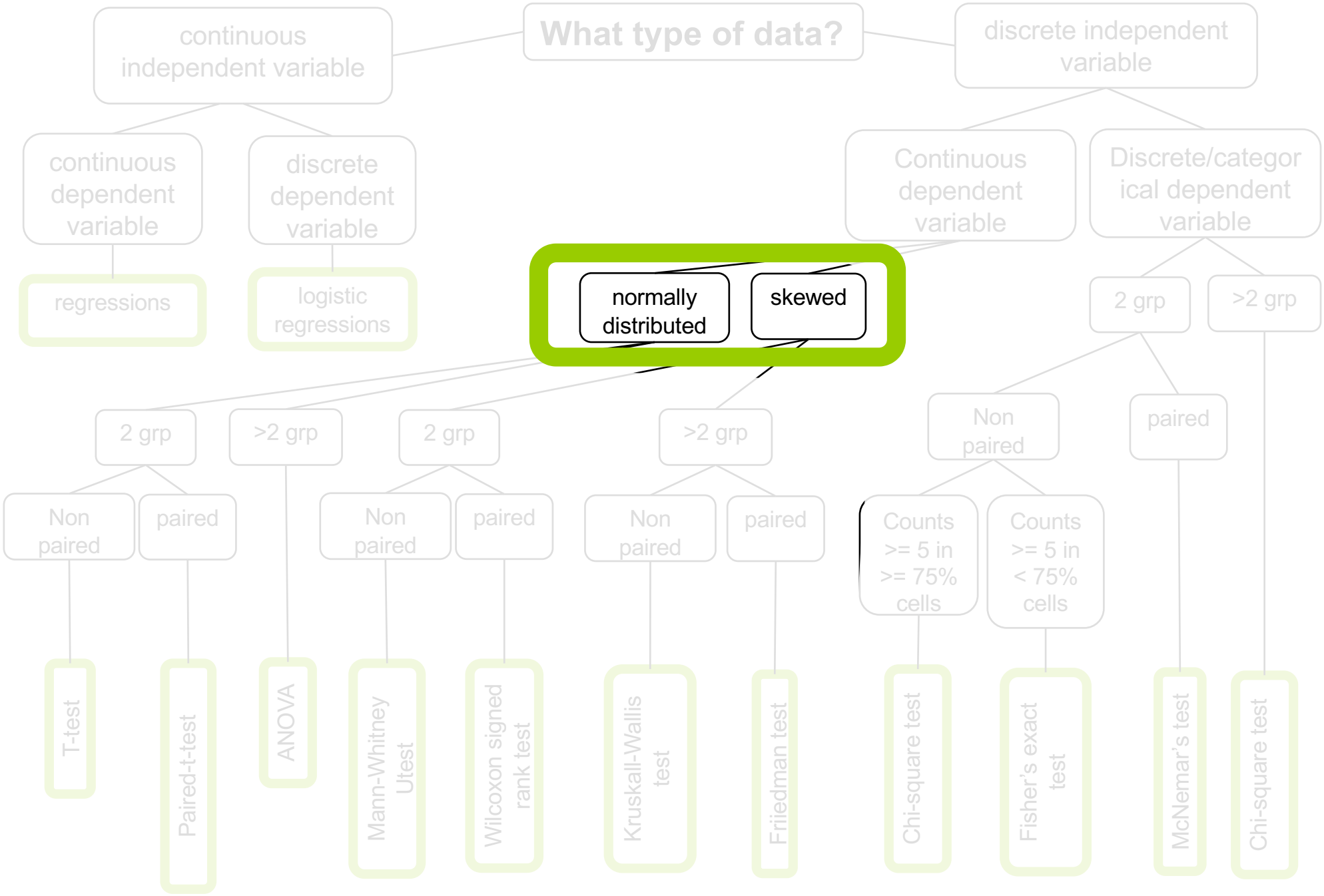
15 Normality tests and Homogeneity tests

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
Continuous dependent variable
Discrete/categorical 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
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::

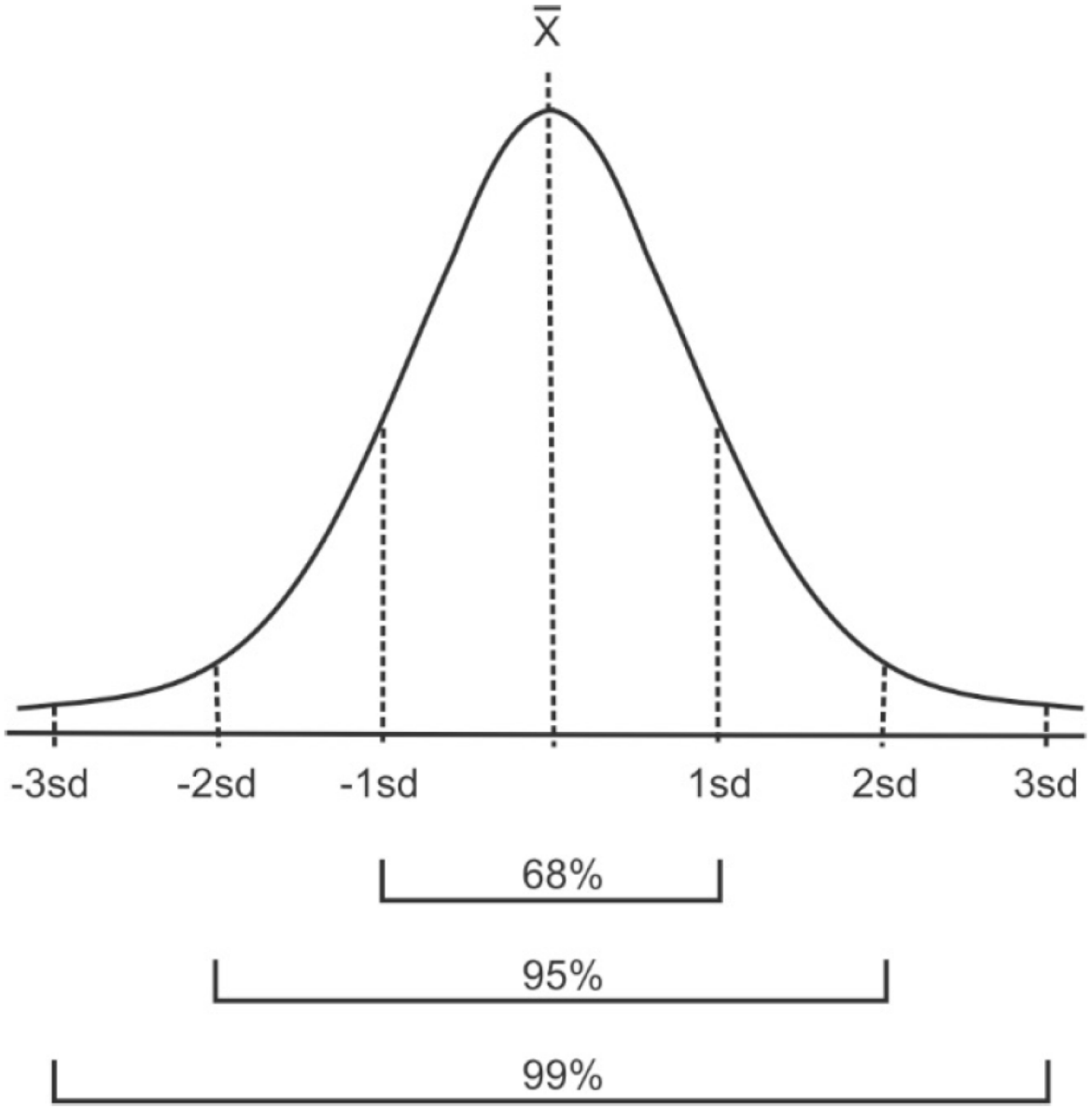
test if data is **normally distributed**

use **alternatives to non-parametric tests** if data not normally distributed

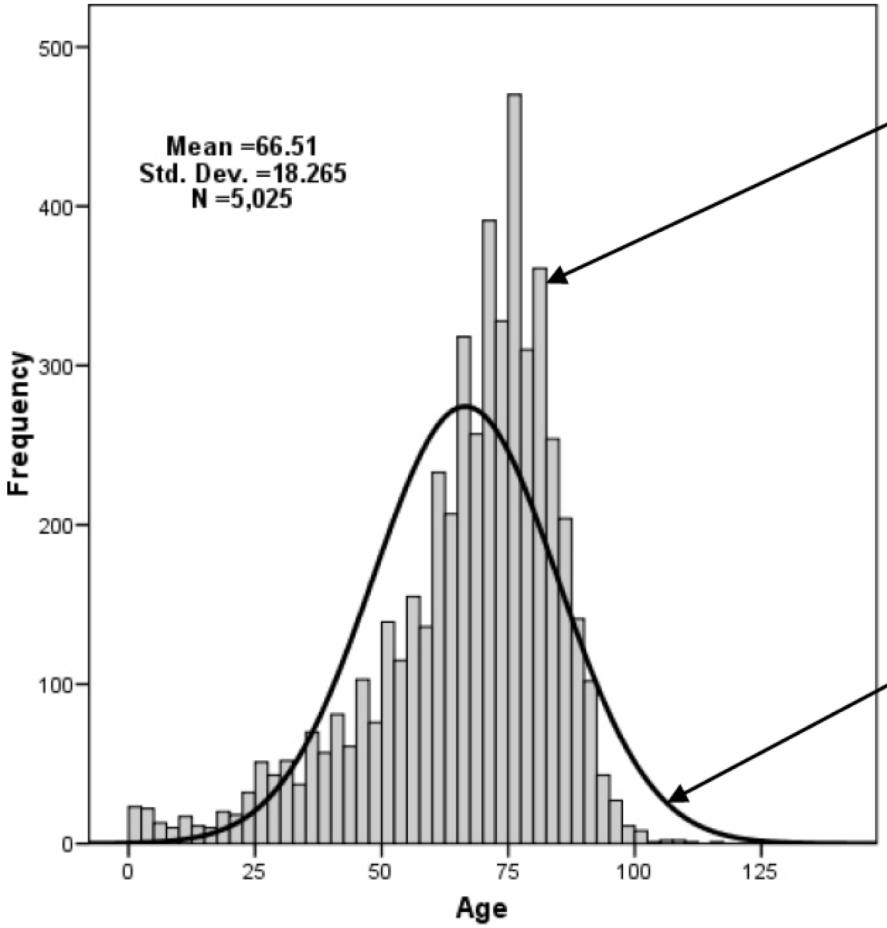
other verifications (assumptions) we need to do on data prior to doing certain statistical tests

assumption of normality

given the **mean** and **standard deviation** of a dataset = a theoretical normal distribution has those proportions



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.

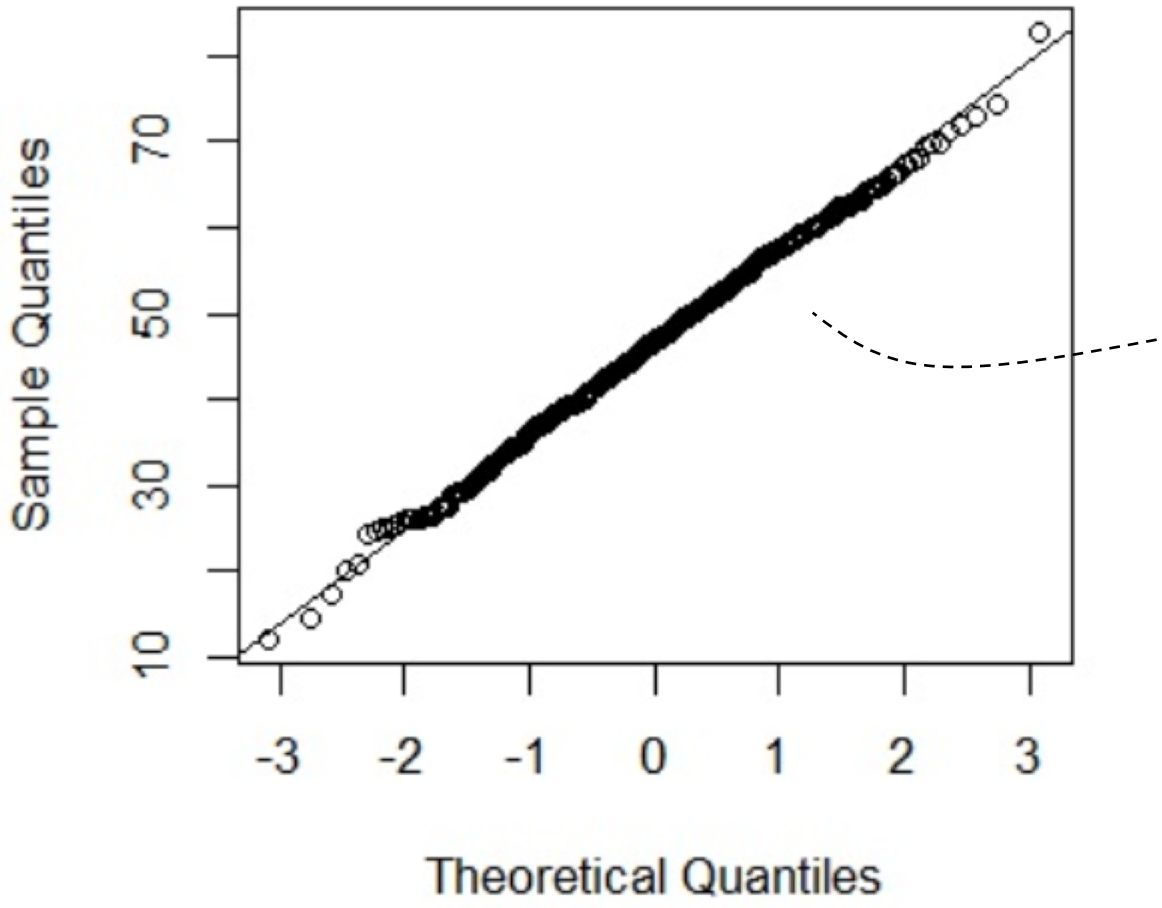
Theoretical normal distribution ***calculated*** from a mean of 66.51 and a standard deviation of 18.265.

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

several methods to check that, we are only going to look at some of them: **Q-Q probability plots**, **Kolmogorov-Smirnov test** and **Shapiro-Wilks test**

(some others: W/S test, Jarque-Bera test, D'Agostino test)

Quantile Quantile Probability Plots



if the two sets come from a population with the same distribution, the points should fall along a line

it is a plot of the quantiles of the first data set against the quantiles of the second data set

3.89 4.75 6.33 4.75 7.21 5.78 5.80 5.20 7.90

does the following sample comes from a normality distributed population?

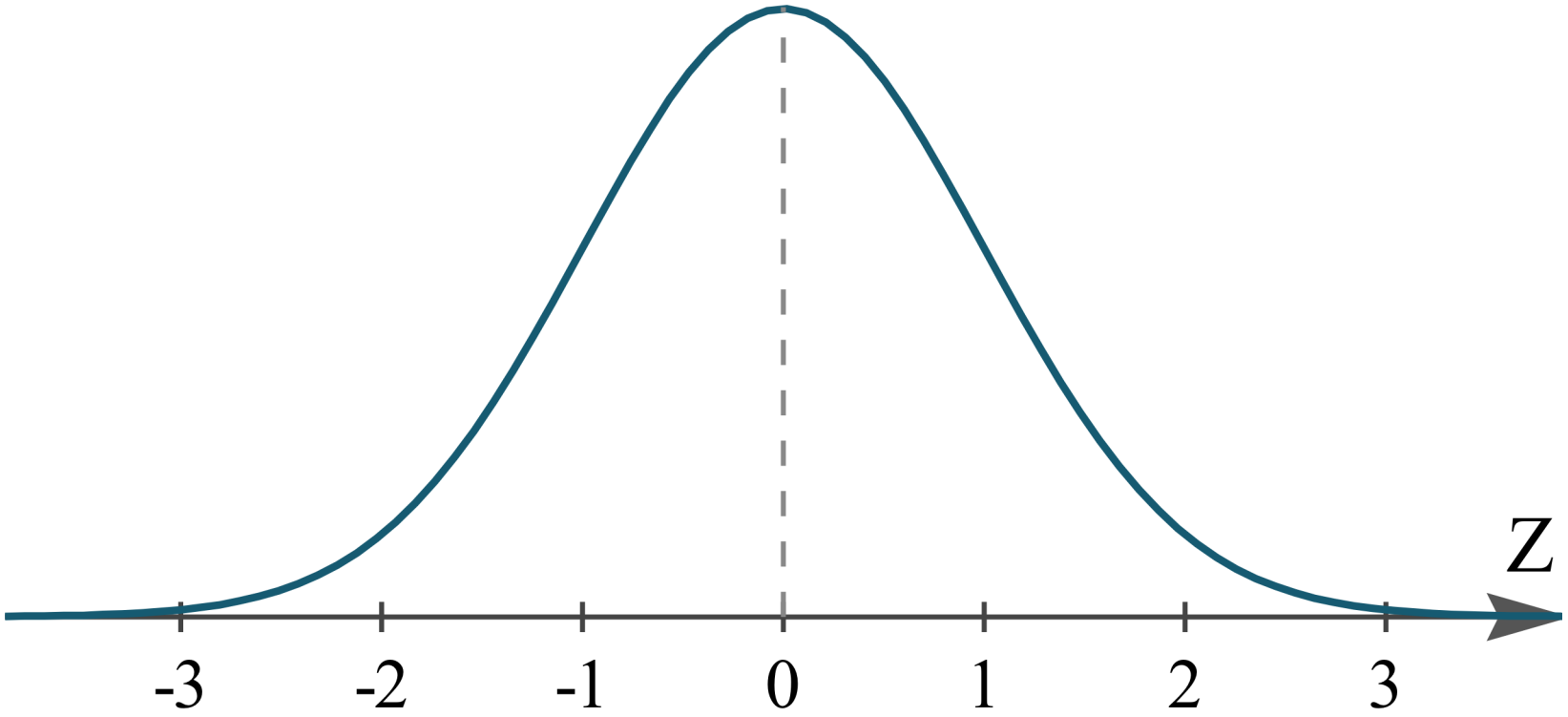
1. order the data:

3.89 4.75 4.75 5.20 5.78 5.80 6.33 7.21 7.90

2. plot these against appropriate quantile from the standard normal distribution

… let's look at this in detail

here a standard normal distribution truncated at -3 and 3

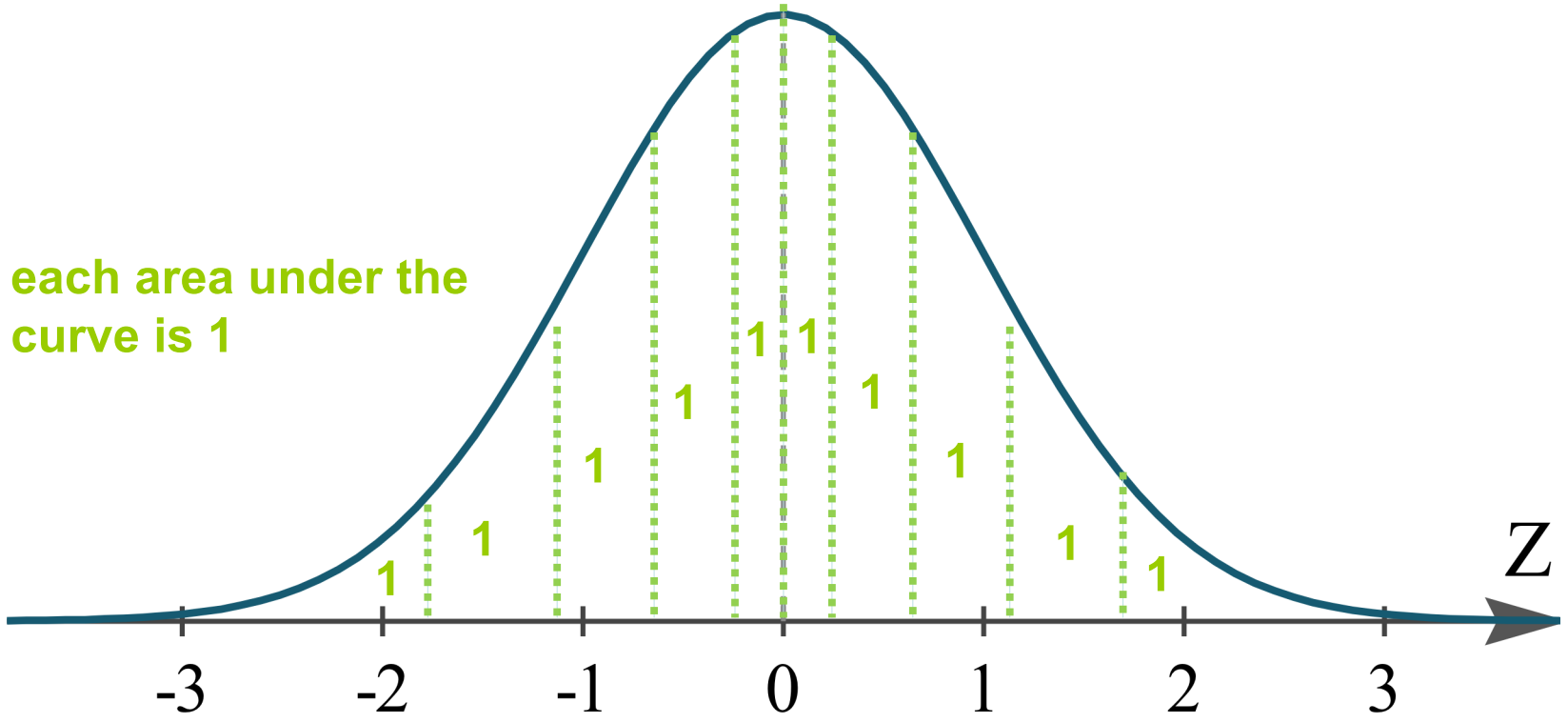


3. search for nine values on the normal distribution = split it in 10 areas

and here our **nine** sample values

3.89 4.75 4.75 5.20 5.78 5.80 6.33 7.21 7.90

here a standard normal distribution truncated at -3 and 3



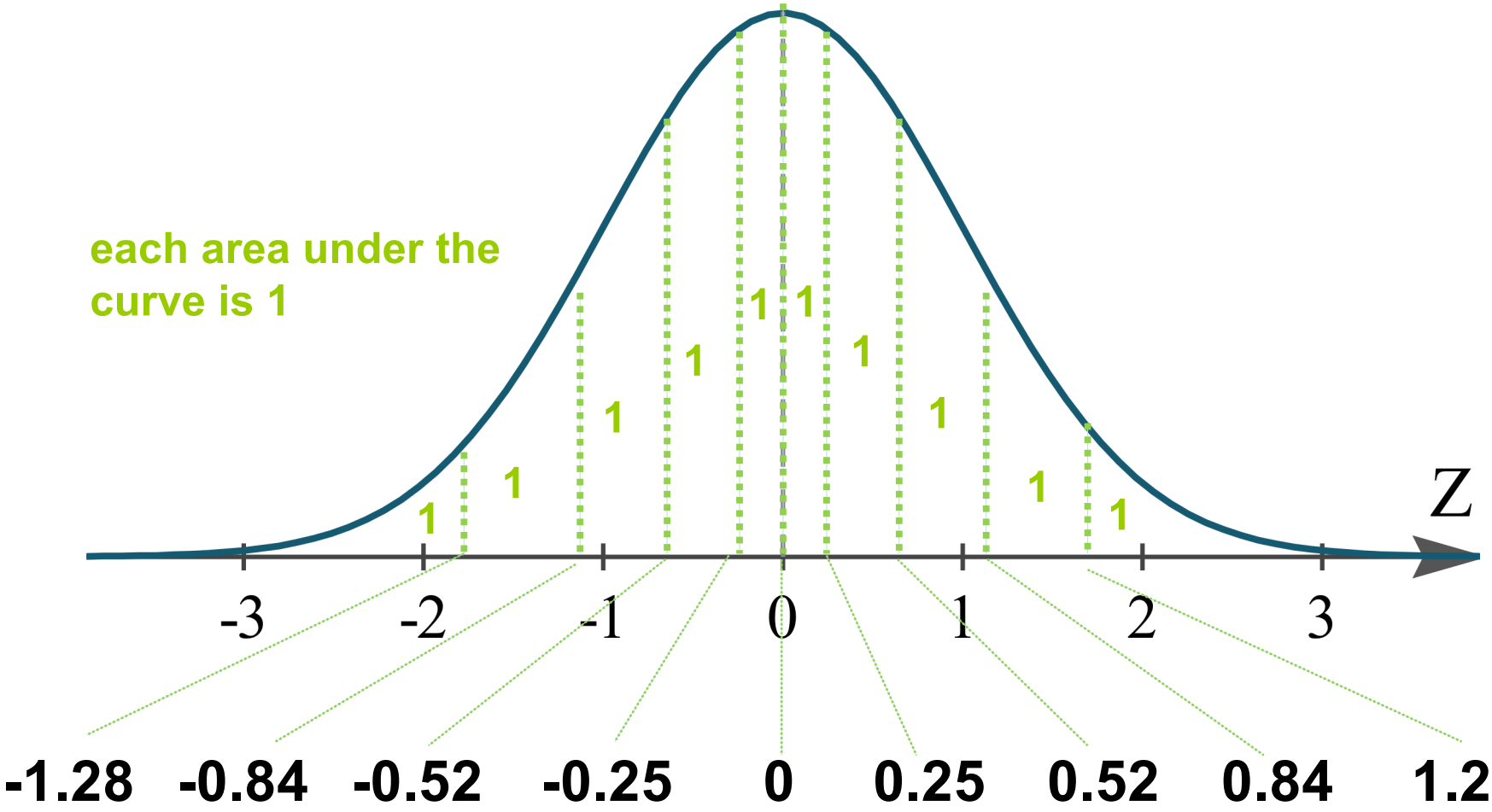
3. search for nine values on the normal distribution = split it in 10 areas

4. find the values that makes that happen

and here our **nine** sample values

3.89 4.75 4.75 5.20 5.78 5.80 6.33 7.21 7.90

here a standard normal distribution truncated at -3 and 3



and here our **nine** sample values

3.89 4.75 4.75 5.20 5.78 5.80 6.33 7.21 7.90

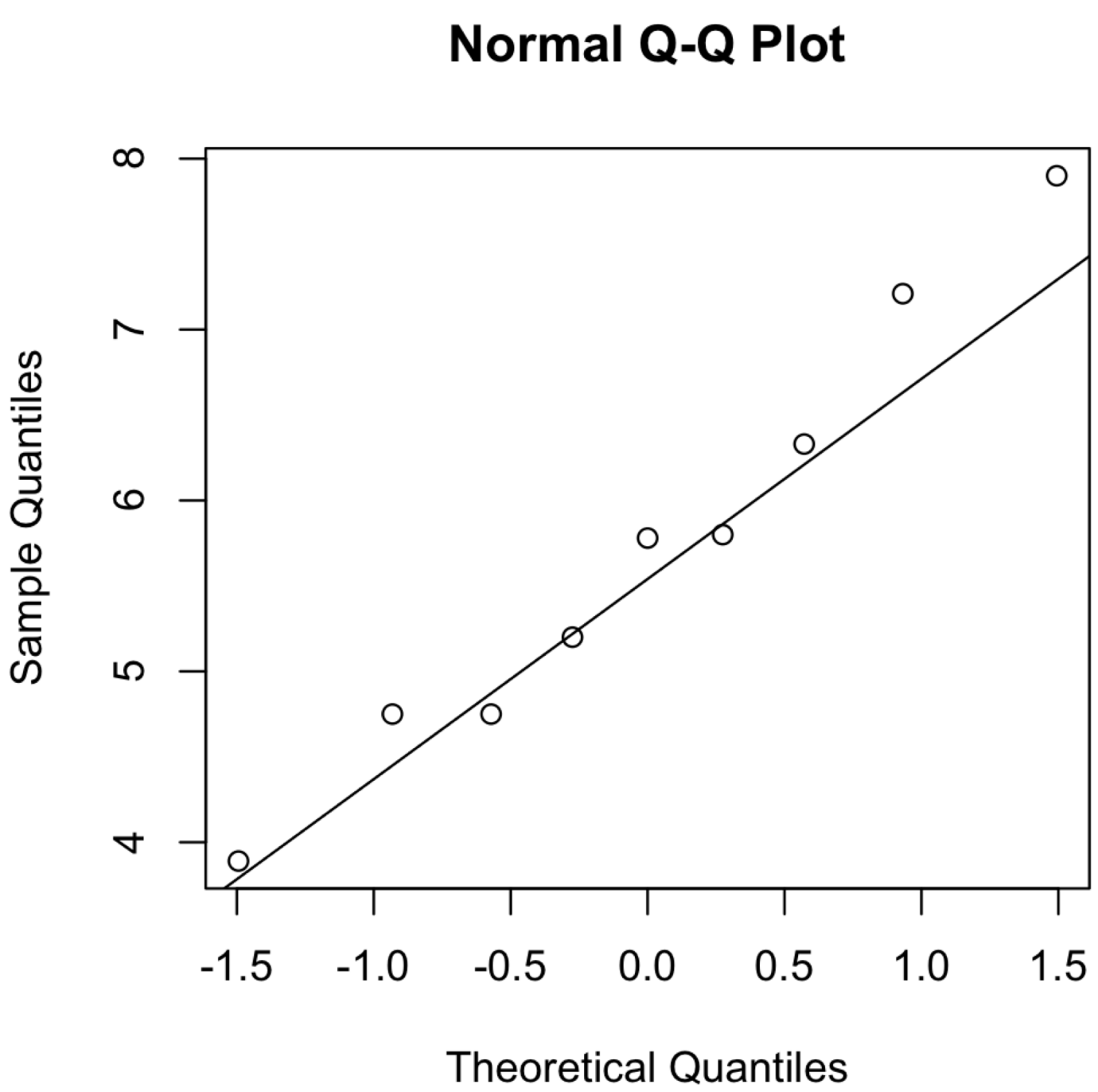
5. plot smallest value in our sample of size nine against what we expect to get as the smaller value in a sample of the same size from the standard normal distribution … we do that for all pairs

here what we would expect to get

-1.28 -0.84 -0.52 -0.25 0 0.25 0.52 0.84 1.28

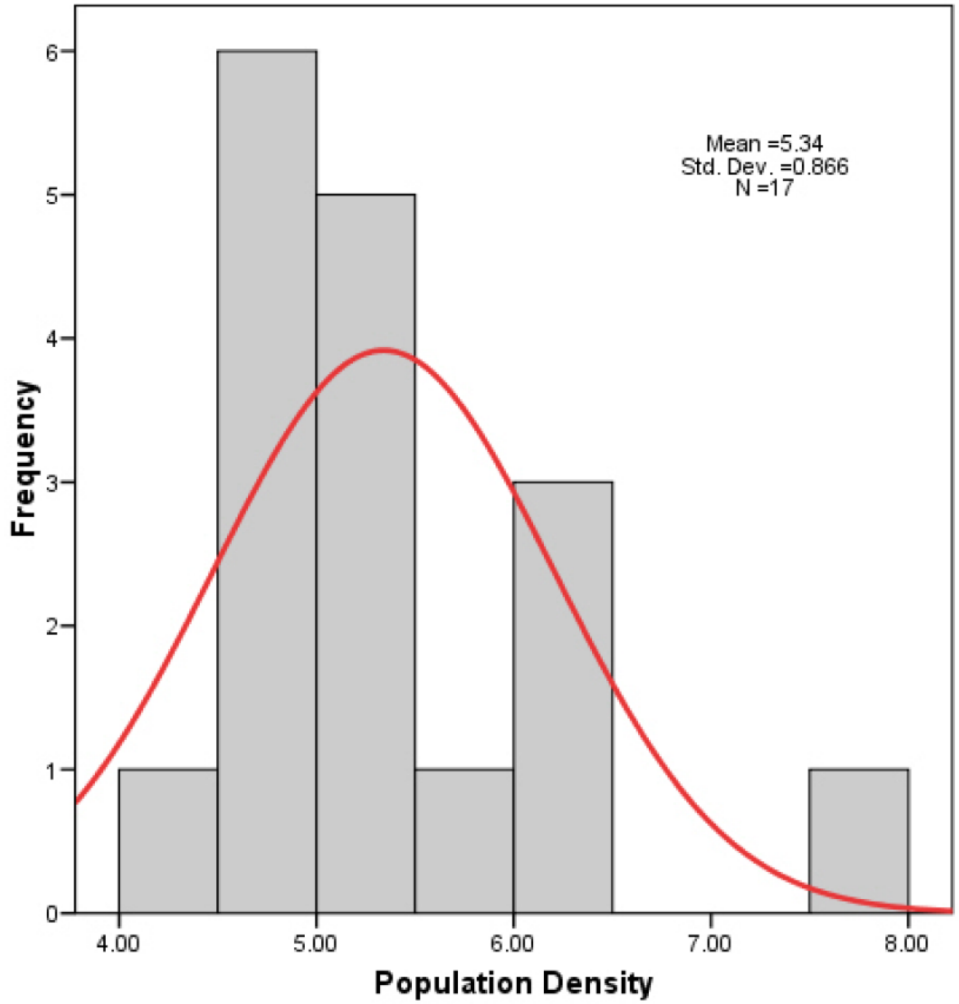
and here our **nine** sample values

3.89 4.75 4.75 5.20 5.78 5.80 6.33 7.21 7.90


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
4
5
6
7
8
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

```
data <- c(3.89, 4.75, 4.75, 5.20, 5.78, 5.80,
6.33, 7.21, 7.90)

# generate expected from normal distribution
u=seq(0.1,0.9,by=0.1)
expected <- qnorm(u)
print(expected)

[1] -1.2815516 -0.8416212 -0.5244005 -
0.2533471  0.0000000  0.2533471  0.5244005
[8]  0.8416212  1.281551

plot(expected,data)

#or simpler:
qqnorm(data)
qqline(data)
```

Q-Q plots are great **graphical tools for large sample** but not **very useful for small sample size**, e.g. this is the histogram of the last example: data do not ‘look’ normal, but they are not statistically different than normal.



statistical tests for normality are more precise since actual probabilities are calculated

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

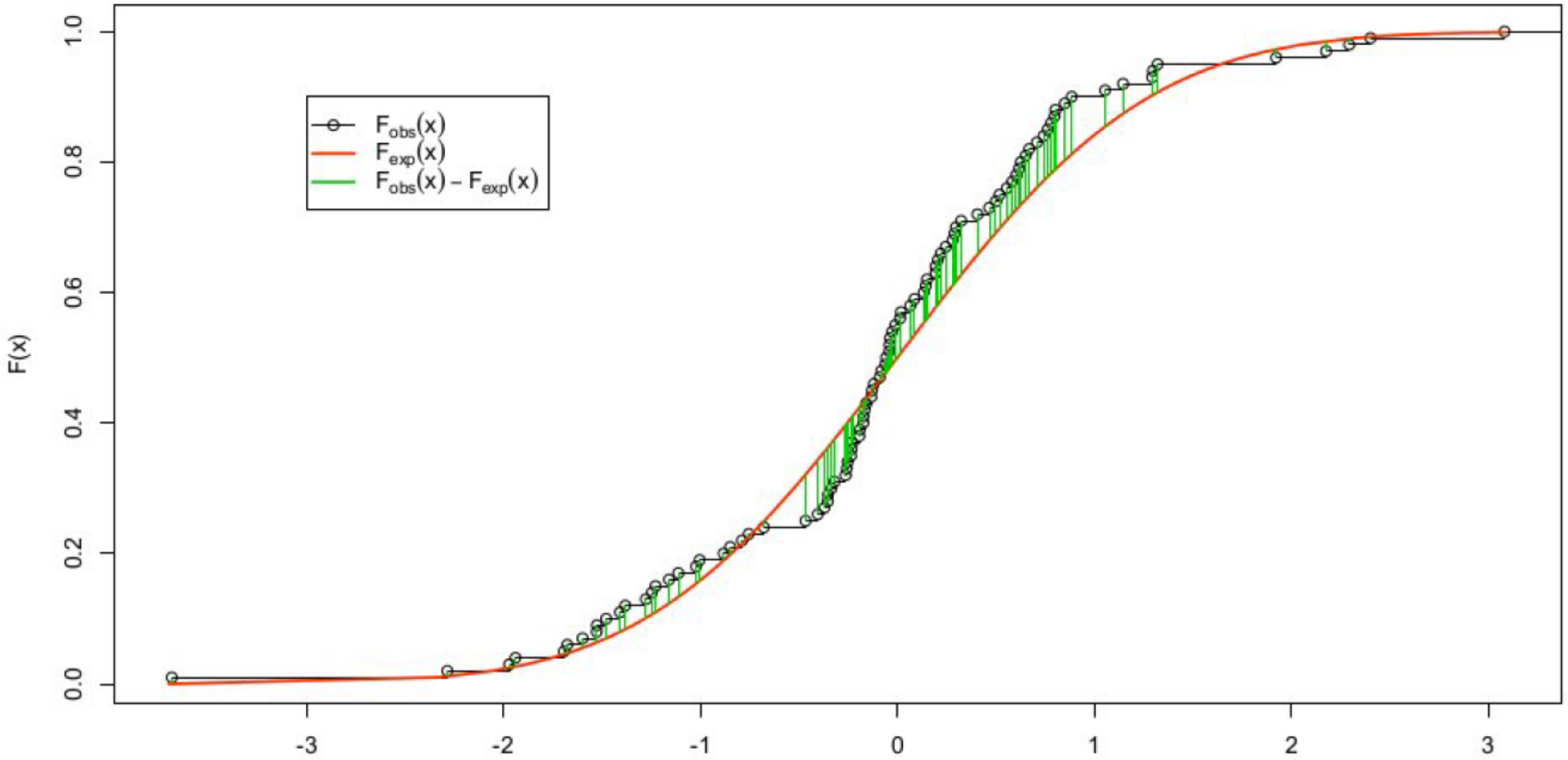
cumulative distribution function expected

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
```

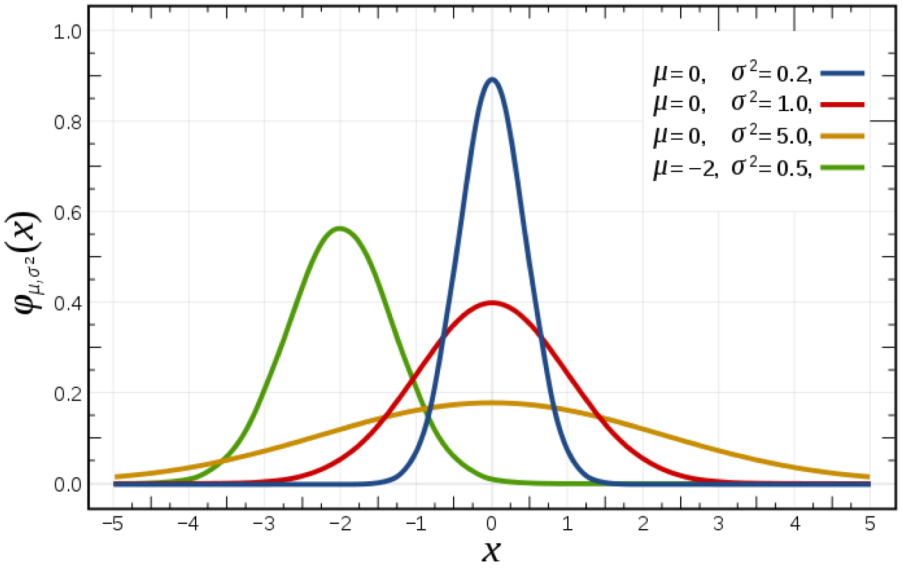
does the following sample of n=100 comes from a normality distributed population?

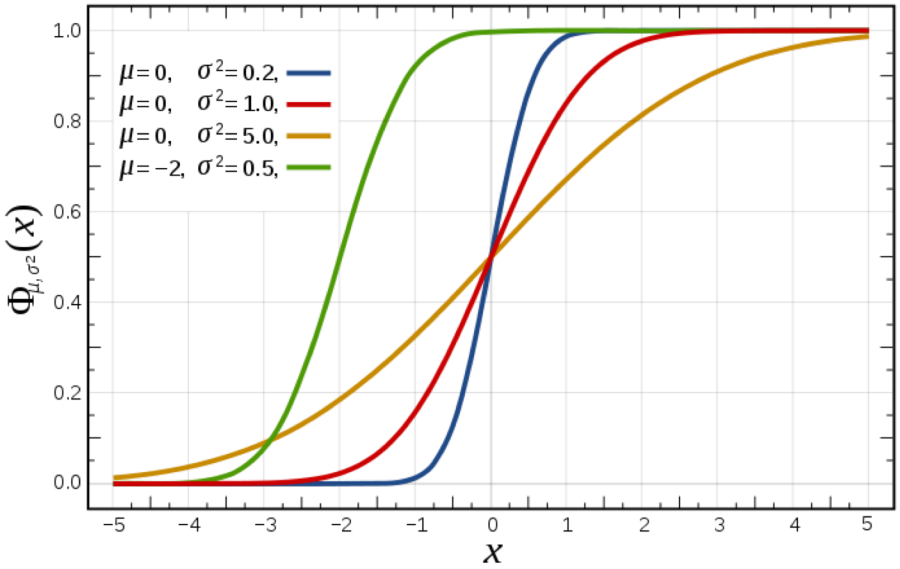
intuitively, we search for the maximum absolute distance between our data cumulative distribution function and the normal cumulative distribution function



so far we looked at **probability density function:** represents probability that the variate has the value x

another way to look at this is the **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 -0.13 -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  0.16  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}

0.01	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

now we have two tables Fobs and Fexp …

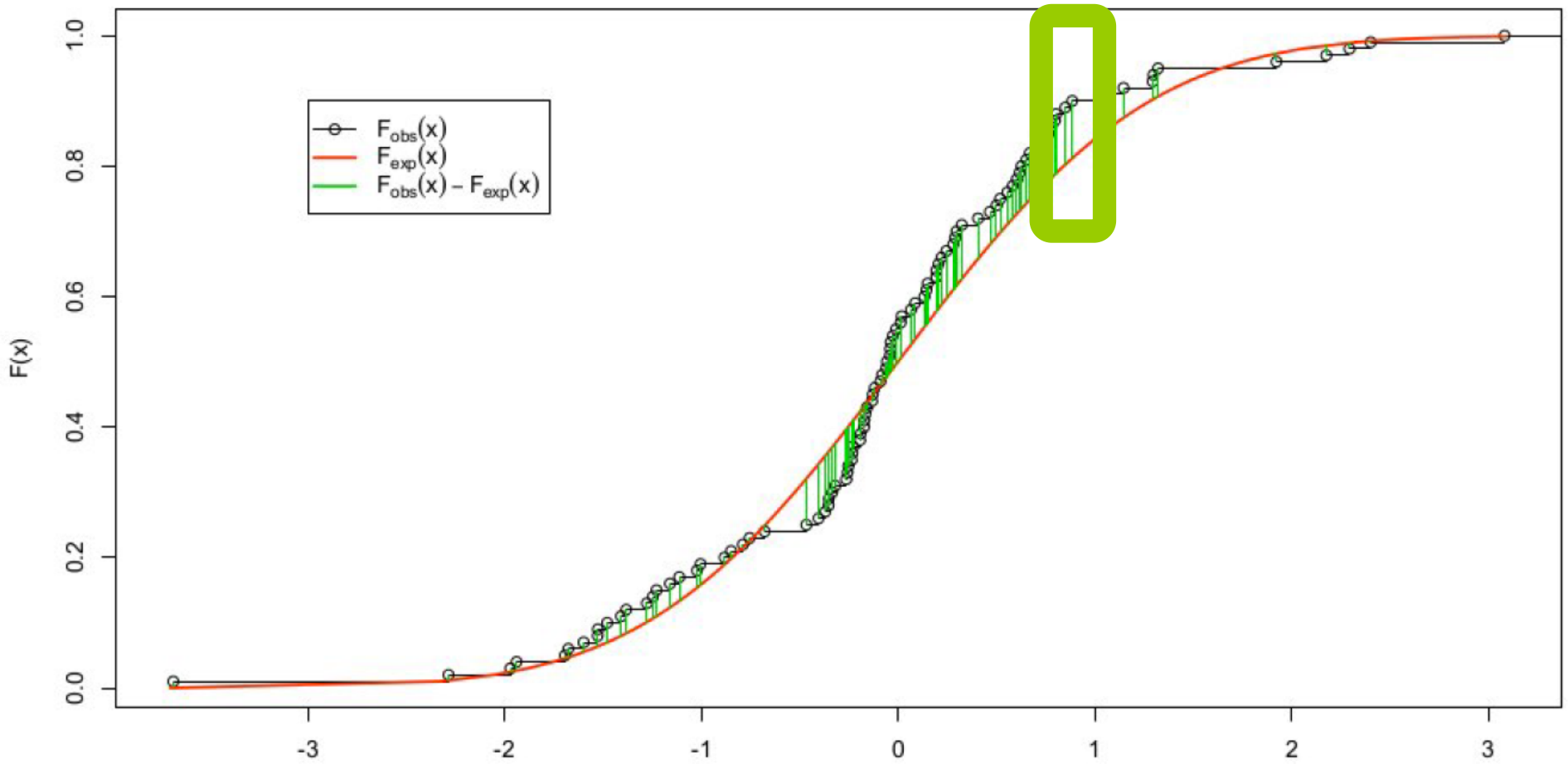
4. lets compute the absolute difference between the two and find the highest value

0.01	0.02	0.01	0.01	0.02	0.02	0.02	0.03	0.04	0.04
0.04	0.05	0.04	0.04	0.05	0.06	0.07	0.07	0.08	0.08
0.09	0.09	0.09	0.06	−0.04	−0.05	−0.05	−0.04	−0.03	−0.04
−0.03	−0.04	−0.03	−0.02	−0.01	0.00	0.01	0.01	0.02	0.03
0.04	0.05	0.03	0.04	0.05	0.06	0.06	0.06	0.07	0.08
0.09	0.10	0.11	0.12	0.11	0.12	0.12	0.13		0.11
0.12	0.12	0.13	0.14	0.15	0.16	0.17	0.18	0.19	0.19
0.18	0.12	0.11	0.10	0.11	0.10	0.08	0.09		0.09
0.09	0.09	0.10	0.10	0.10	0.11	0.11	0.12	0.13	0.11
0.06	0.06	0.04	0.05	0.06	−0.02	−0.03	−0.02	−0.01	0.00

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

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 0.19 > **0.136**

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

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 0.19 > **0.136** so null hypothesis rejected

H0: the samples come from a normal distribution

conclusion: the data do not come from a normal distribution

note KS is different than other tests we saw where we looked for a value below a critical level to reject the null, here it is the opposite (the larger the results the less likely is H0 so we reject it)

what if Dn < Dcrit?

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**

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

```
#sorting a table
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)
ysorted <- sort(y)

X <- rnorm(100)
p = ecdf(X) #cumulative distribution function
X = sort(X) # trick to get fexp
fexp <- p(X)
fobs <- p(ysorted)
KS = max(abs(fexp-fobs))
```

```
#or easier
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**

a bit more beefy but let’s go steps by steps …

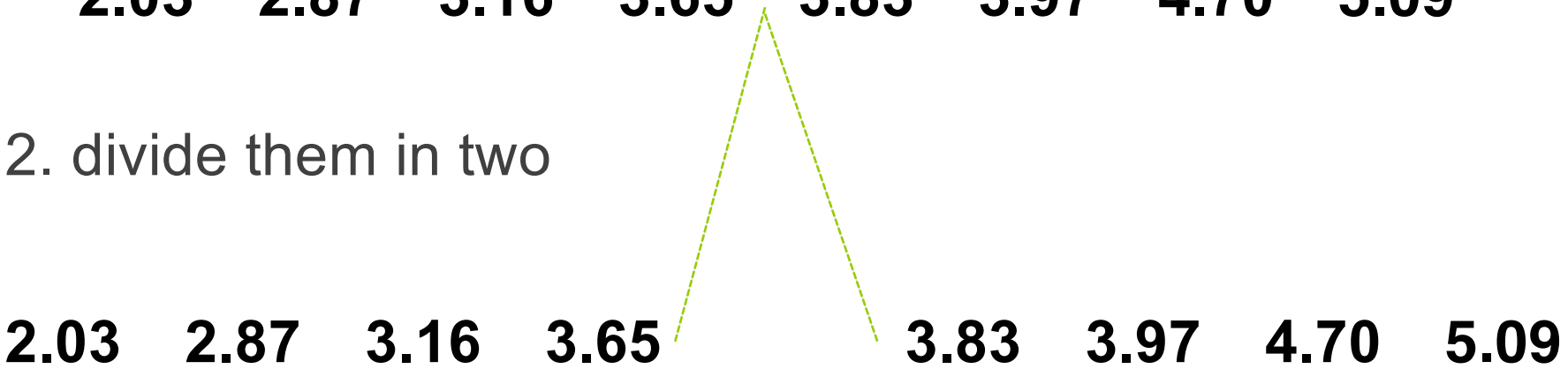
3.83 3.16 4.70 3.97 2.03 2.87 3.65 5.09

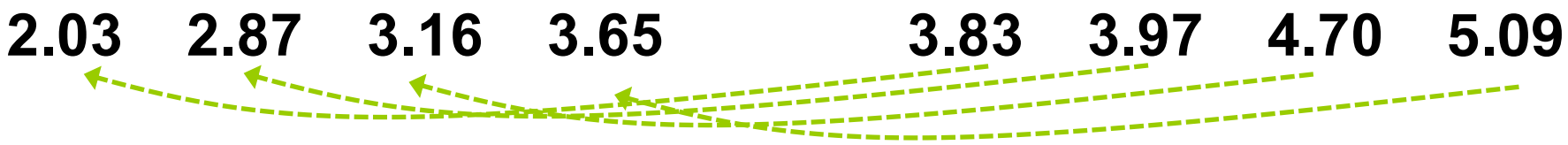
does the following sample comes from a normality distributed population?

1. order the data:

2.03 2.87 3.16 3.65 3.83 3.97 4.70 5.09

2. divide them in two





3. compute di the differences between both

3.06
1.83
0.81
0.18

4. multiply each of these by ai

good new we have shapiro-wilk table

n =	2	3	4	5	6	7	8	9	10	11	12	13	14
a1	0.7071	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.5888	0.5739	0.5601	0.5475	0.5359	0.5251
a2			0.1677	0.2413	0.2806	0.3031	0.3164	0.3244	0.3291	0.3315	0.3325	0.3325	0.3318
a3					0.0875	0.1401	0.1743	0.1976	0.2141	0.2260	0.2347	0.2412	0.2460
a4							0.0561	0.0947	0.1224	0.1429	0.1586	0.1707	0.1802
a5									0.0399	0.0695	0.0922	0.1099	0.1240
a6											0.0303	0.0539	0.0727
a7													0.0240

...

di		ai		
3.06	*	0.6052	=	1.851912
1.83	*	0.3164	=	0.579012
0.81	*	0.1743	=	0.141183
0.18	*	0.0561	=	0.010098

↓

total: **2.582205**

5. Divide the total by SS

$$W = \frac{\left(\sum_{i=1}^{[n/2]} a_i \left(X_{(n+1-i)} - X_{(i)}\right)\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{(2.582205)^2}{6.782549963} = 0.98307903$$

6. from the reference table of W (another table yeah!),
Wcrit(n=8 at 0.05)=0.818

and 0.983>Wcrit

0.983>Wcrit, so we cannot reject null hypothesis, so we **assume the data follows a normal distribution**

otherwise (if <) we could rejected the null hypothesis and conclude with 95% confidence that that the data are not normally distributed

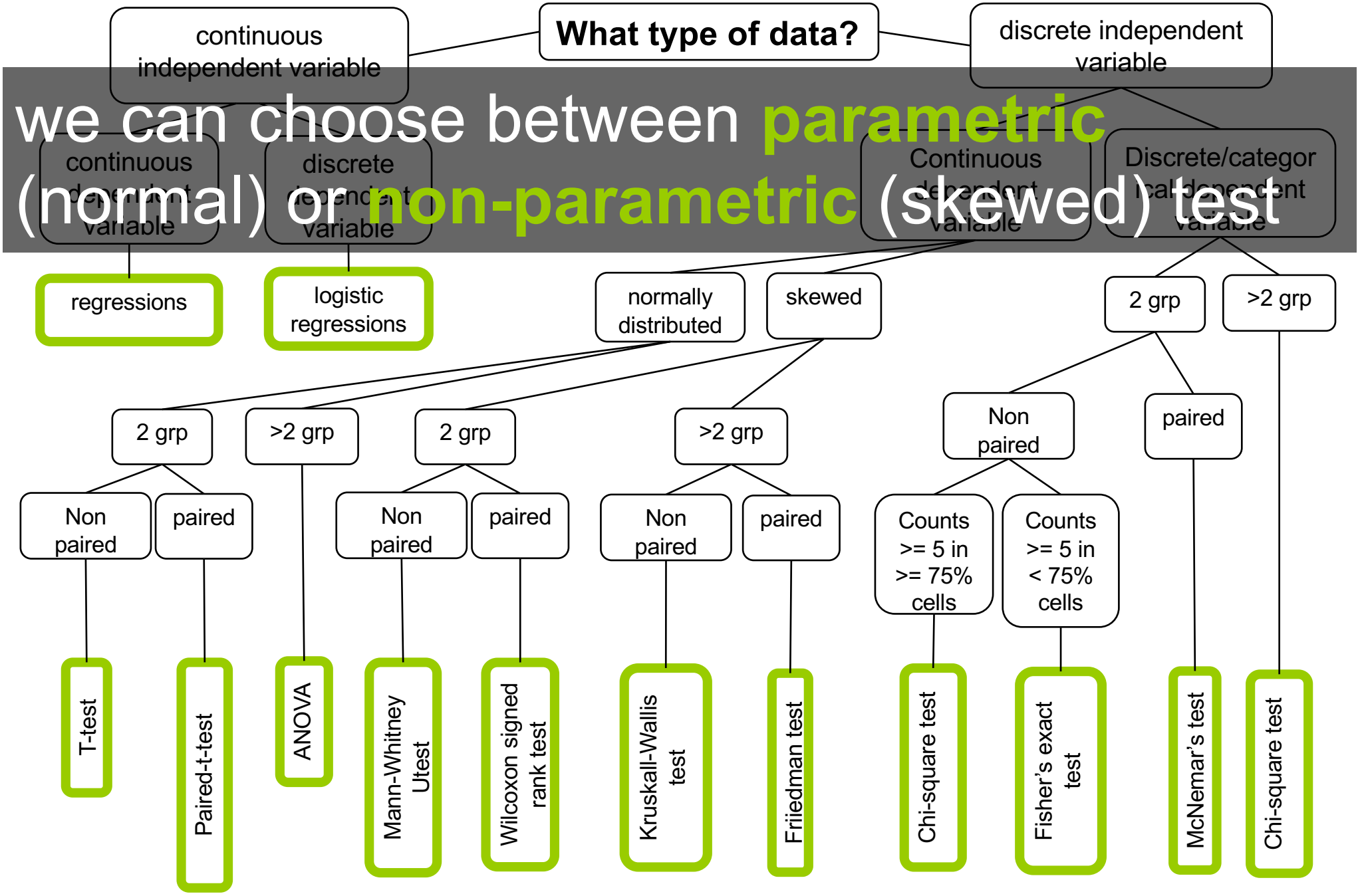
note we search for value below a critical level to reject the null, this is quite different from the results using the Kolmogorov-Smirnov test where this is the opposite

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

```
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
```

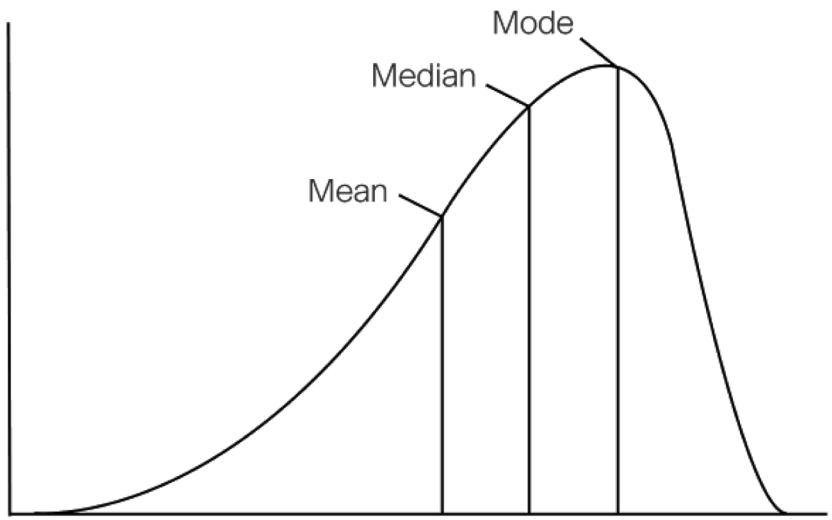
ok so we know how to check our data, now what?


What type of data?
continuous independent variable
discrete independent variable
we can choose between parametric (normal) or non-parametric (skewed) test
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

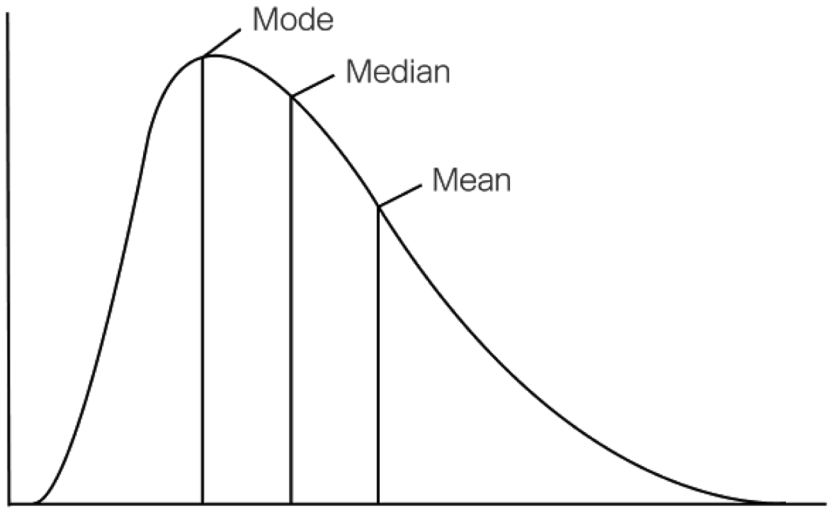
but if your data is not normally distributed you could also try to make it normal using **transformations**

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

transformations

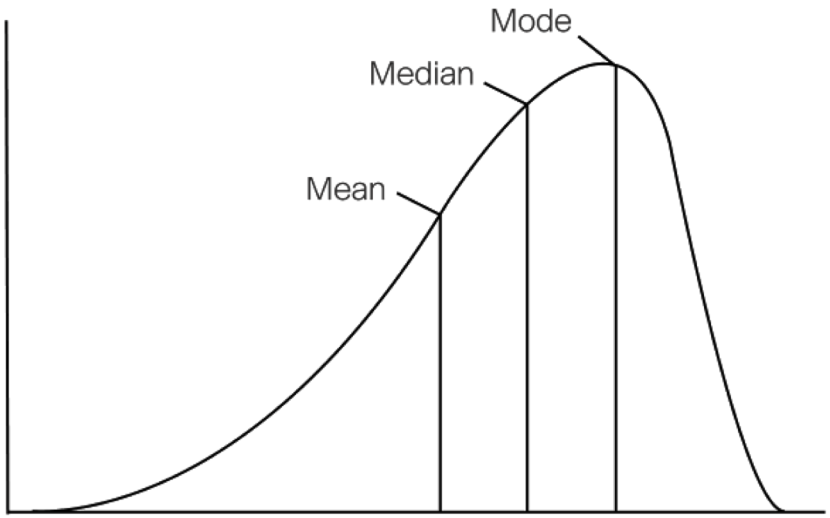


Left-Skewed (Negative Skewness)

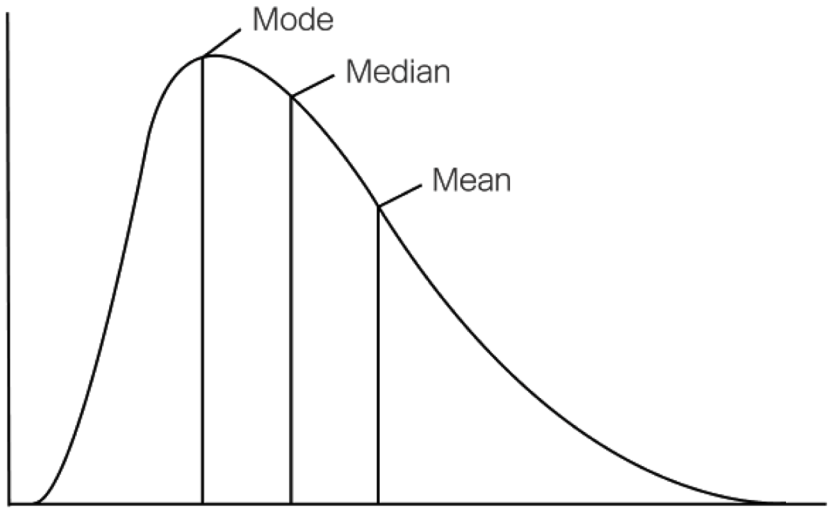


Right-Skewed (Negative Skewness)

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



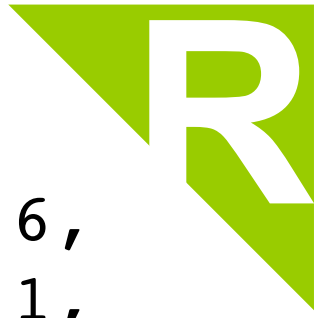
Left-Skewed (Negative Skewness)



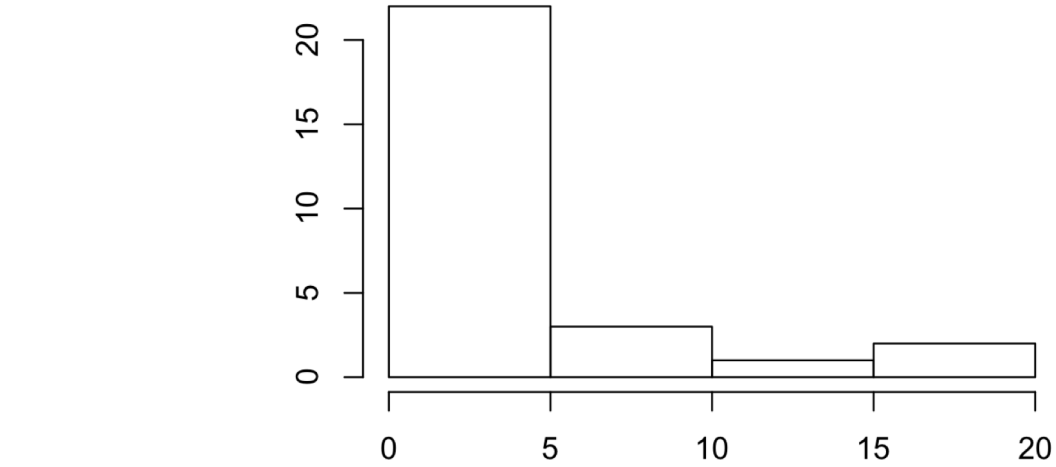
Right-Skewed (Negative Skewness)

common transformations for right skewed::

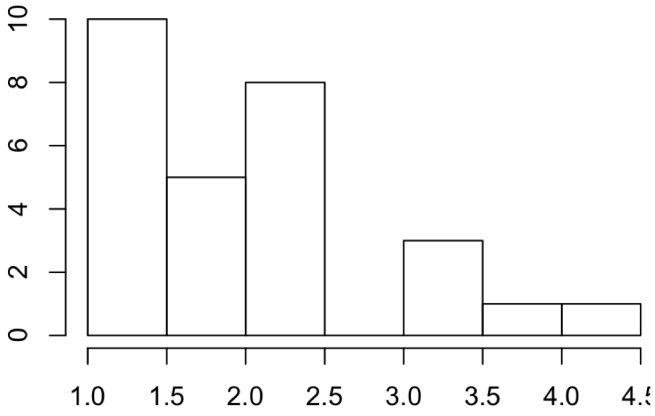
square, cube root and logarithmic



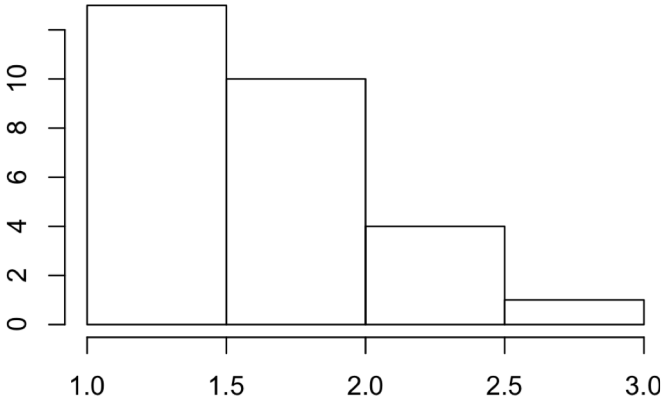
```
y <-c(1.0, 1.2, 1.1, 1.1, 2.4, 2.2, 2.6,
4.1, 5.0, 10.0, 4.0, 4.1, 4.2, 4.1, 5.1,
4.5, 5.0, 15.2, 10.0, 20.0, 1.1, 1.1, 1.2,
1.6, 2.2, 3.0, 4.0, 10.5)
hist(y)
qqnorm(y)
qqline(y)

y_sqrt = sqrt(y) #cube root
y_cub = sign(y) * abs(y)^(1/3) #square root
y_log = log(y) #logarithm

# you can now try
qqnorm(y_log)
qqline(y_log)
```



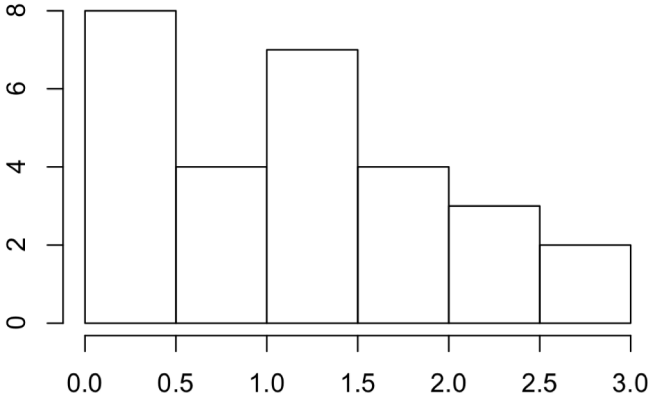
square root



cube root



logarithmic



sometimes it does not work and rather than trying a complex transformations it id better to use non parametric tests

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

other assumptions

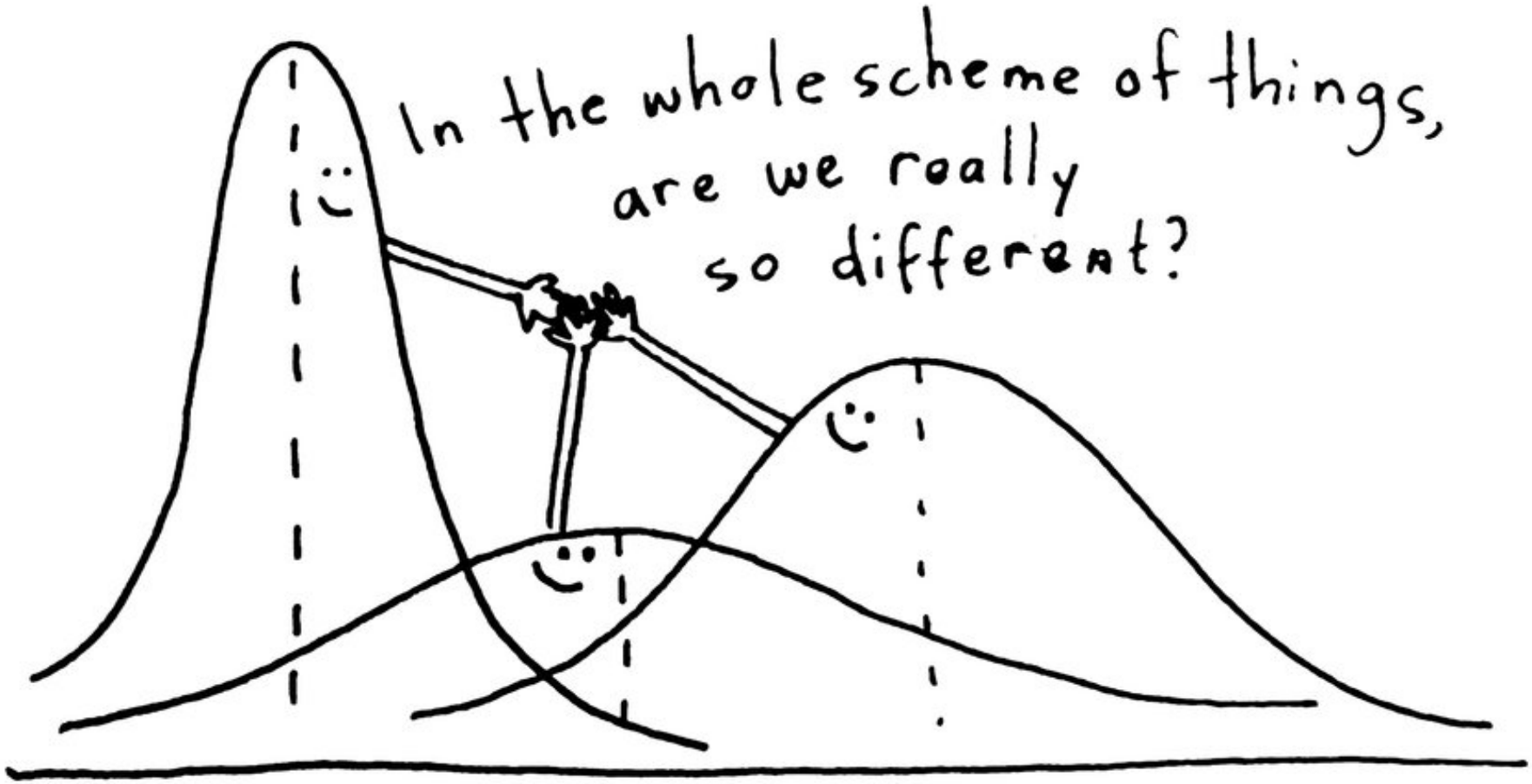
let’s have a look at the ANOVA we did when we tried to check if chocolate improves memorization

```
# first we run 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)

we won't go in detail with this test but the most important is this:

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

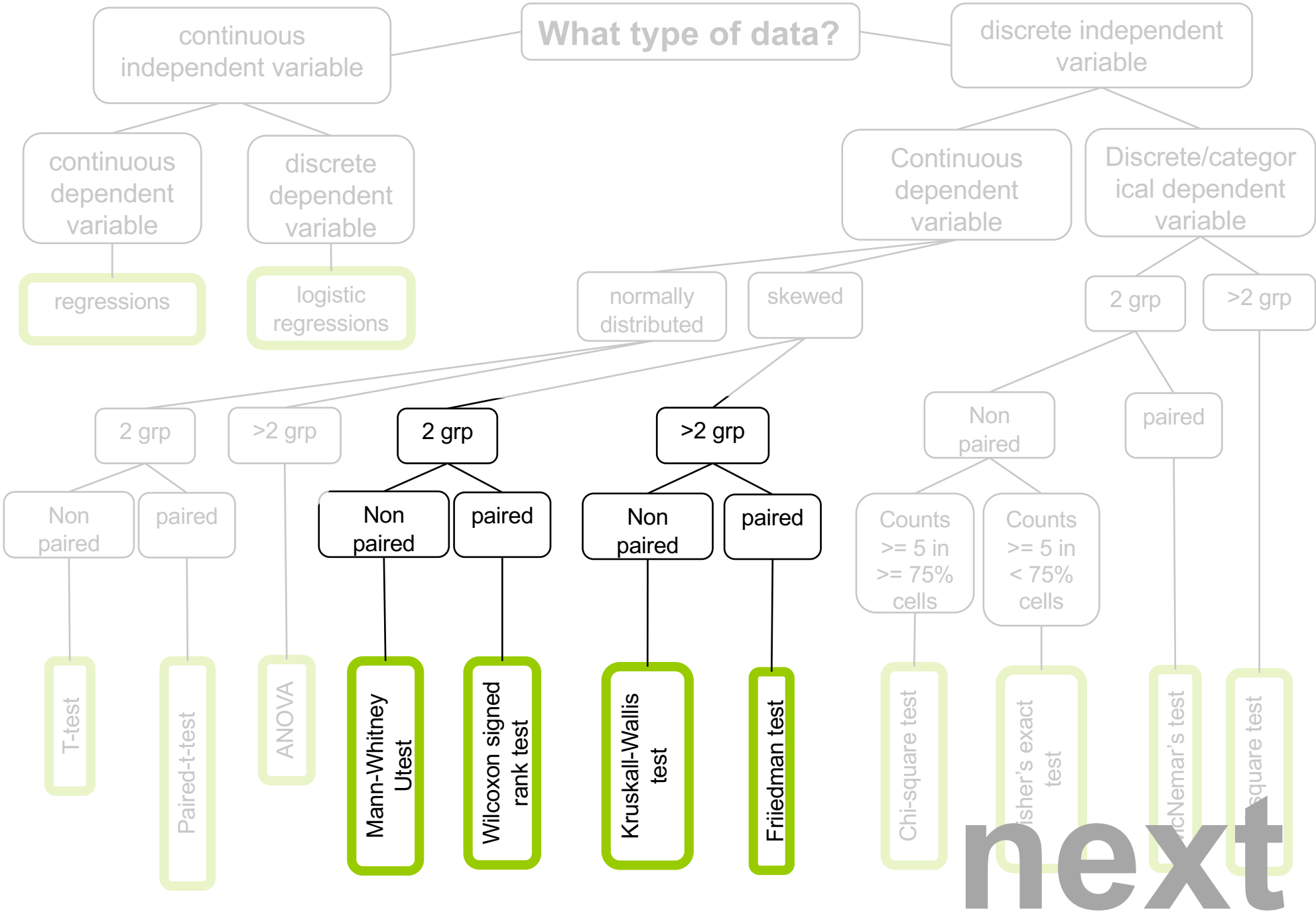


unfortunately **ANOVAs** do not work well in this case

summary

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

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
Non paired
paired
Non paired
paired
Non paired
paired
Non paired
paired
T-test
Paired-t-test
ANOVA
Mann-Whitney Utest
Wilcoxon signed rank test
Kruskall-Wallis test
Friiedman test
Counts >= 5 in >= 75% cells
Counts >= 5 in < 75% cells
Chi-square test
Fisher's exact test
McNemar's test
square test
next

end