MAT20306 - Advanced Statistics

Lecture 2: Sample size calculations

Wilcoxon tests

WAGENINGEN UR
For quality of life

Biometris
Quantitative Methods brought to Life

A confidence interval (CI)

- A confidence interval is a range of “likely” values for a population parameter, confidence level is often 0.95.
- The width of the interval reflects the accuracy :
  narrow interval $\rightarrow$ accurate estimate,
  wide interval $\rightarrow$ inaccurate estimate
- The bounds of an (1-α) CI are random (depend on the sample), the parameter is a fixed (unknown) number.

- A (1-α)-CI for a parameter consists of all H_0-values V for which H_0: parameter = V is not rejected in two sided t-test with significance α.
- CI ≠ RR

Structure of a confidence interval

Limits of a two-sided 1-α confidence interval for a **parameter**:

estimate ± $t_{df}(\alpha/2)$ * standard error (estimate)

With $t_{df}(\alpha/2)$ from table 2, (or PQRS, or ...)

For one sample:

with Normality of *y* assumed:

$$\bar{y} \pm t_{n-1}(\alpha/2) \times s/\sqrt{n}$$

Example A, n=20:

give 0.95 CI for μ if $\bar{y} = 26.5$ and $s_y = 5.67$

NB. sometimes confidence intervals limits are calculated with a z-value:

estimate ± $z_{\alpha/2}$ * standard error (estimate) , with $z_{\alpha/2}$ from N(0,1)

WAGENINGEN UR For quality of life

The after-party: Analysing your data

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Age	20	26,5000	5,67079	1,26803

One-Sample Test

	Test Value = 25.8					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Age	,552	19	,587	,70000	-1,9540	3,3540

WAGENINGEN UR
For quality of life

The 4 elements in t-procedures

1. Confidence interval calculation
2. t-test (8 steps)

t-procedures

In t-procedure, **4 elements** are central:

A. **Parameter of interest**

B. **Estimator** (how do we estimate the parameter)

The Estimate (the outcome of the estimator in the sample)

C. **Standard error** (se) of the estimator / estimate, a measure of how certain we can be about the estimate

D. **Degrees of freedom** (df) for the t-distribution.

# samples & # variables	We have a research question about:	H_0:	Note:	TS:	Distribution when H_0 is true	1-α c.i.
1 sample 1 variable	Population expected value	$\mu=\mu_0$	σ is known	$z=\frac{\bar{y}-\mu_0}{\sigma/\sqrt{n}}$	$z \sim z(0, 1)$	$\bar{y} \pm z_{\alpha/2} * \sigma/\sqrt{n}$
1 sample 1 variable	Population expected valued	$\mu=\mu_0$	σ is unknown	$t=\frac{\bar{y}-\mu_0}{s/\sqrt{n}}$	$t \sim t(n-1)$	$\bar{y} \pm t_{\alpha/2} * s/\sqrt{n}$
2 samples 1 variable	Difference between two population expected values	$\mu_1 - \mu_2 = D_0$	$\sigma_1 = \sigma_2$ **OR** $\sigma_1 \neq \sigma_1$	$t=\frac{\bar{y}_1-\bar{y}_2-0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ $t'=\frac{\bar{y}_1-\bar{y}_2-0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$	$t \sim t(n_1+n_2-2)$ $t' \sim t(df)$ from SPSS output	$\bar{y}_1-\bar{y}_2 \pm t_{\alpha/2} s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ $(\bar{y}_1-\bar{y}_2) \pm t_{\alpha/2} * \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$
1 sample 2 variable	Population expected difference	$\mu_d = D_0$	Observations are paired	$t=\frac{\bar{d}-\mu_d}{s_d/\sqrt{n}}$	$t \sim t(n-1)$	$\bar{d} \pm t_{\alpha/2} * s_d/\sqrt{n}$

WAGENINGEN UR
For quality of life

Error probabilities and power in testing

https://www.youtube.com/watch?v=Dsa9ly4OSBk

four possibilities and probabilities		Reality	
		H_0 true	H_a true
Decision	H_0 rejected	**Type I error** $\boldsymbol{\alpha}$	correct, **power =** 1 - $\boldsymbol{\beta}$
	H_0 not rejected	correct **1 - $\boldsymbol{\alpha}$**	**Type II error** $\boldsymbol{\beta}$

- Type I error: P (Type I error) = α (continuous),
  P (Type I error) ≤ α (discrete),
  typically $\alpha = 0.05$ (or maybe smaller like 0.01).
- Type I error is under control: α is **chosen**.
- Type II error: P (Type II error) = β ,
  depends on $\boldsymbol{\alpha}$, on $\boldsymbol{\sigma}$, on **true parameter value**, and on sample size $\boldsymbol{n}$.

How large should *n be* to achieve small β, and large power (1- β), if the true parameter deviates by Δ from the hypothesized one?

Sample size calculations / power calculations

Question: How “large” should my experiment be?

Answer: that depends on

1) what you want (build up a **CI**, or perform a **test**)

2) How **precise** you want it. Specify precision criteria.

Information / some guess for the variance (or standard deviation), is also needed.

Illustration: simulation power 2 samples.xlsx – on blackboard / Practical

Case study 1A: cereal manufacturer (p.245 O&L6 / p. 256 O&L7)

A cereal manufacturer produces cereals in boxes with weight W. A machine is set to deliver a mean weight of 16.37 ounces. Standard deviation is 0.225 ounces.

- Manufacturer can be fined if the true mean is 16.27 (or less).
- Concern of the manufacturer: one machine is under filling
- Manufacturer takes a sample to determine whether the expected weight μ is less than 16.37.
- Precision requirement: If the real mean would be 16.27, the manufacturer wants the test to reject H_0: $\mu = 16.37$ (with large probability).
- How many boxes should the manufacturer take to see if the machine is OK (not to get a fine) or not?

Cereal machine: hypothesis test

- test H_0: $\mu = 16.37$ versus H_a: $\mu < 16.37$

Δ = smallest relevant difference between true μ and value 16.37

- risk of civil penalty when true mean weight is less than16.27
  $\rightarrow$ take $\Delta = 16.37 - 16.27 = 0.10$

$$n = \frac{\sigma^2 (z_\alpha + z_\beta)^2}{\Delta^2}$$

Means: 99% chance to reject H_0 if it is not true

- power = *0.99*, so $\beta = 0.01$ → $z_\beta = z_{0.01} = 2.33$
- test of size $\alpha = 0.05$ → $z_\alpha = z_{0.05} = 1.645$
- assumed standard deviation $\sigma = 0.225$

$$n = \frac{\sigma^2 (z_\alpha + z_\beta)^2}{\Delta^2} = \frac{0.225^2 (1.645 + 2.33)^2}{0.10^2} = 79.99$$

Case study 1A: cereal manufacturer (p.245 O&L6 / p. 256 O&L7)

Cereal manufacturer produces cereals in boxes with weight W. A machine is set to deliver a mean weight of 16.37 ounces. Standard deviation is 0.225 ounces.

The cereal manufacturer wants to check the machine and construct a 0.95 confidence interval for the mean population weight with an error margin of at most 0.1.

- How many boxes should a random sample contain?
- $\bar{y} \pm Error\ Margin\ (EM)$ with $EM = t_{df}(\alpha/2)*s/\sqrt{n}$
- → $n = t_{df}(\alpha/2)^2*s^2/(EM)^2$, so if $EM \leq 0.1$, then $n \geq t^2*s^2/0.1^2$.
- we use z in stead of t; $z_{0.025}=1.96$
- We use an estimation for s, in this case 0.225.
- $n \geq 0.225^2 * 1.96^2/ 0.1^2 = 19.44$ → n should be 20 at least.

Required sample size, one sample

Two possible aims :

1. Construct a $(1-\alpha)$ confidence interval for $\boldsymbol{\mu}$.
   Requirement: error margin ≤ E
   or interval width ≤ W (with W=2E)
   (E or W and α should be specified)

$$n = \frac{\sigma^2 (z_{\alpha/2})^2}{E^2}$$

2. Testing H_0: $\mu = V_0$, at size α (often $\alpha = 0.05$), if we want to reject H_0 with probability ($\pi = 1-\beta$) (the power) when in reality $\mu = V_1$:

For $\Delta = V_1 - V_0$ we usually choose the minimum relevant difference between μ_d and μ_0.

$$n = \frac{\sigma^2 (z_\alpha + z_\beta)^2}{\Delta^2} \quad (\text{one sided } \mathrm{H_a})$$

$$n = \frac{\sigma^2 (z_{\alpha/2} + z_\beta)^2}{\Delta^2} \quad (\text{two sided } \mathrm{H_a})$$

Required sample size, paired observations

Two possible aims:

1. Construct a (1-α) confidence interval for μ_d:
With requirement: error margin ≤ E
or interval width ≤ W (with W = 2E)

$$n = \frac{(z_{\alpha/2})^2 \times \sigma_d^2}{E^2}$$

2. Testing H_0: $\mu_d = V_0$, at size α (often $\alpha = 0.05$), if we want to reject H_0 with probability (1-β) (the power) when in reality $\mu_d = V_1$:

For $\Delta = V_1 - V_0$ we usually choose the minimum relevant difference between μ_d and μ_0.

$$n = \frac{\sigma_d^2 (z_\alpha + z_\beta)^2}{\Delta^2} \quad \text{(one sided } H_a\text{)}$$

$$n = \frac{\sigma_d^2 (z_{\alpha/2} + z_\beta)^2}{\Delta^2} \quad \text{(two sided } H_a\text{)}$$

Sample size: two sample t-test, O&L6 p.323 / O&L7 p.334

Two possible aims:

1. Construct a (**1-α**) confidence interval for $\mu_1 - \mu_2$:
With requirement: error margin ≤ **E**
or interval width ≤ **W** (with W = 2E)

$$n_1 = n_2 = 2\frac{(z_{\alpha/2})^2 \times \sigma^2}{E^2}$$

2. Testing H_0: $\mu_1 - \mu_2 = V_0$, at size α
(often 0.05), if we want to reject H_0
with probability **π = 1-β** (the power)
when in reality $\mu_1 - \mu_2 = V_1$.
So if the relevant difference is

Δ = V_1-V_0

$$n_1 = n_2 = 2\frac{\sigma^2(z_\alpha + z_\beta)^2}{\Delta^2} \quad \text{(one sided H}a\text{)}$$

$$n_1 = n_2 = 2\frac{\sigma^2(z_{\alpha/2} + z_\beta)^2}{\Delta^2} \quad \text{(two sided H}a\text{)}$$

Note the extra factor 2 in the expression for n (because a difference between two independent sample means is involved).

In all formula's there is also σ for which you need to have at least an estimate

Normality ...

Two non-parametric tests

Frank Wilcoxon
(1882 -1965)

INDIVIDUAL COMPARISONS BY RANKING METHODS

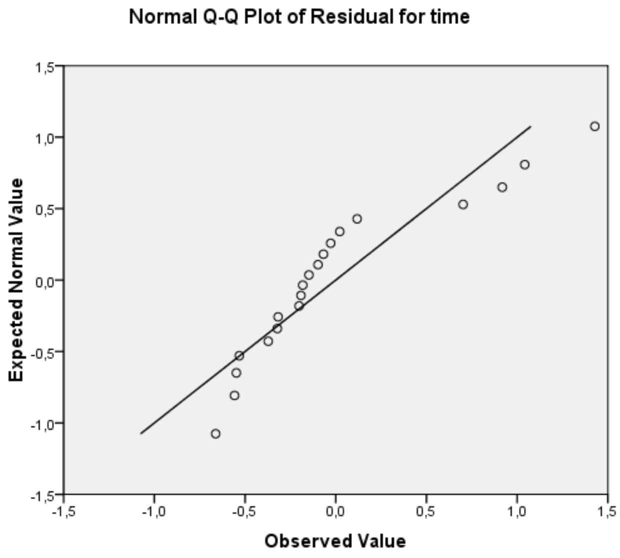
FRANK WILCOXON
American Cyanamid Co.

Biometrics Bulletin 1: 80–83, (1945)

In 1945, Frank Wilcoxon presented the *rank-sum test* and the *signed-rank test* that are named after him.

No normality

- So far, for inference a Normal distribution of the response was assumed.
- What if Normality cannot be assumed and samples are ‘small’ (for large samples we may rely on the central limit theorem).



- A possible solution: nonparametric methods.
- Nonparametric (or distribution free) methods: no specific distribution of the response variable is assumed.
- Ranks will be used instead of the original data
- Data have to be continuous or at the least ordinal.

- Other possible solutions for non-normality: transform the response variable or use other distributions (binomial, Poisson, gamma, …)

What is the idea: ranks!

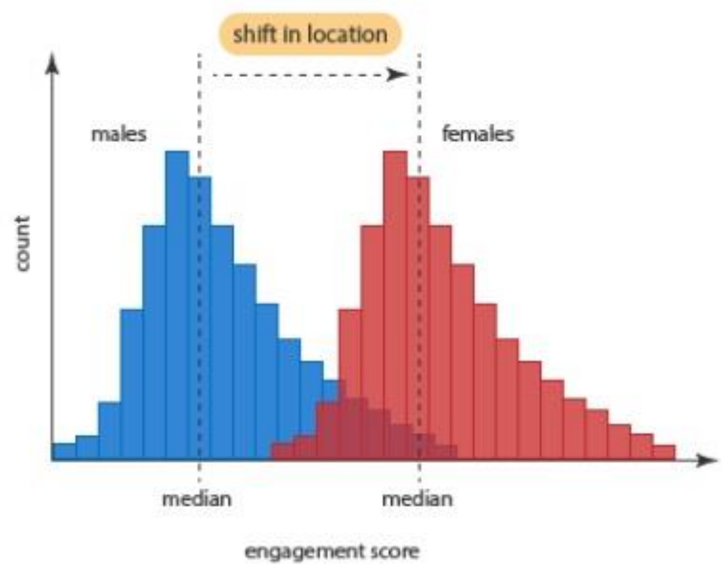
- We will replace the data by rank numbers:

  1 for the lowest observation, 2 for the lowest but one, … etc.
- two-sample t-test → **Wilcoxon rank sum test or Mann-Whitney U test**
- paired t-test → **Wilcoxon's signed rank test**
- Use of ranks usually does not extend in any useful way to more complicated problems.

SPSS uses Mann-Whitney test as name for Wilcoxon rank sum test.
Wilcoxon proposed the test for equal sample sizes, in 1945.
Mann & Whitney extended it towards unequal sample sizes, in 1947.

Situation 3a. Two samples, non-Normal observations

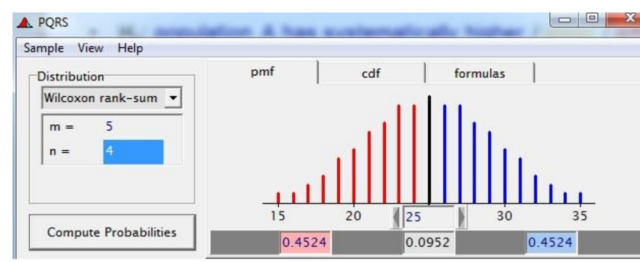
- Same setup as the two-sample t-test: two independent random samples from two populations or a comparison of two treatments.
- Independent samples, *say* $x_1,..,x_{n1}$ *and* $y_1,..,y_{n2}$, *of size* n_1 *and* n_2.
- For H_a, we think in terms of the shift alternative: two distributions of the same form that are shifted relative to each other.
- `Wilcoxon's rank-sum test (Mann Whitney U test)
- Note: under Normality, if the two standard deviations are equal the alternative is also a shift alternative.



Before the party

1. H_0 and H_a
2. Definition of the test statistic (TS)
3. Distribution of the TS if H_0 is true
4. Behaviour of TS, expected under H_a (larger / smaller / larger or smaller)
5. Type of p-value: L, R or 2-sided.

Dogs randomly receive treatment feed type 1 or 2, with n_1=5 and n_2=4 . Test if the 2 treatments lead to systematically different distributions of body weight gain.



Before the party

- H_0: the distribution of the observations in each population is the same.
- H_a: population 1 has systematically higher / lower / different values than population 2.

Observations (of both samples together) are replaced by ranks:
rank 1 for the lowest observation, …, rank (n_1 +n_2) for the highest.

When there are equal observations, these are averaged (mid-ranks), e.g if two equal observations should get ranks 5 and 6, each receives rank 5.5.

The test statistic: **W_1 = sum of ranks in sample 1**, or

W_2 = sum of ranks in sample 2

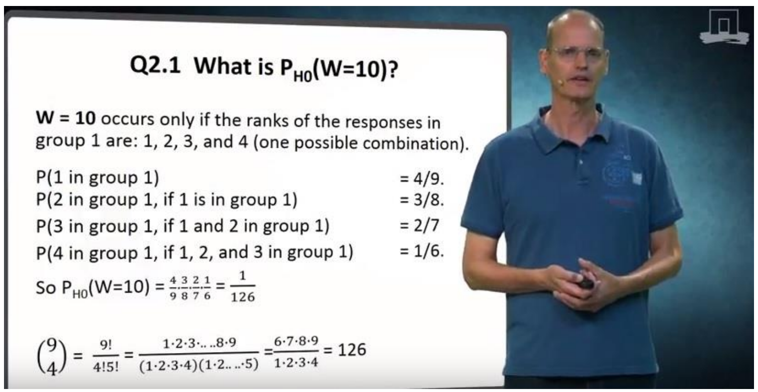
If SPSS output is available, choose the one indicated by SPSS.

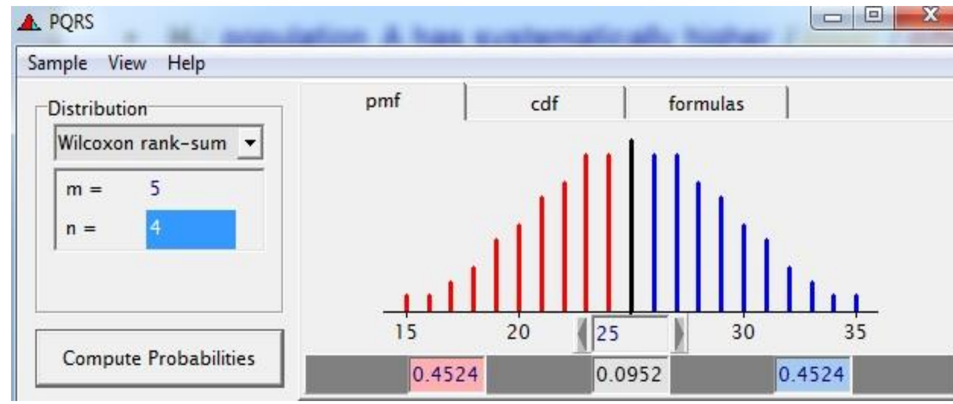
Test statistic is often denoted by *W*, but O&L call it *T*.

Before the party

3) Under H_0, W ~ Wilcoxon rank sum distribution (n_1, n_2).

- The exact distribution is shown by PQRS for the case of no ties.





4) Under H_a W tends to larger / smaller / larger or smaller values

5) → use RPV / LPV / 2-tailed PV.

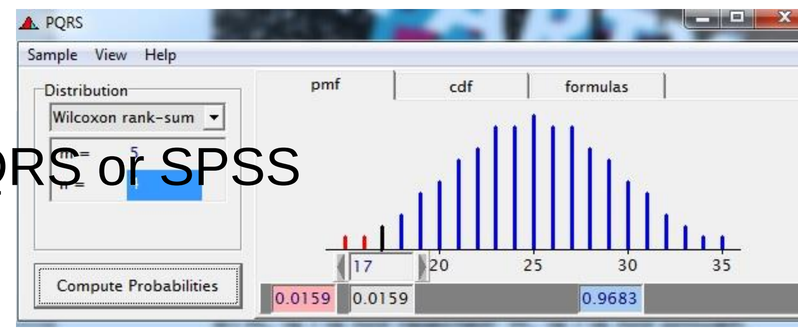
During the party: Descriptive (Sample) Statistics

n_1=5 and n_2=4

Weigh gain observations
A: 12, 25, 17, 11, 15
B: 18, 100, 20, 27

The after-party: Analysing your data

6) outcome W_1 : calculate it, or get it from SPSS



7) appropriate PV: get it from PQRS or SPSS

8) H_0 is / is not rejected; H_a is / is not proven.

8a)It is / is not shown that population 1 has systematically diferent values than population 2.

The after-party: Analysing your data

We will always use the P-value method.
Table 5 in O&L has critical values for W_1; <u>we do not use it.</u>

You should know how to calculate W for small samples, and how to use output from SPSS / PQRS / R to draw the right conclusion

SPSS uses smallest sum of ranks !!!
R uses the Mann-Whitnay test statistic !!!

Ranks

	Group	N	Mean Rank	Sum of Ranks
BodyWeightGain	1,00	5	3,40	17,00
	2,00	4	7,00	28,00
	Total	9		

Test Statistics[a]

	BodyWeightGain
Mann-Whitney U	2,000
Wilcoxon W	17,000
Z	-1,960
Asymp. Sig. (2-tailed)	,050
Exact Sig. [2*(1-tailed Sig.)]	,063[b]
Exact Sig. (2-tailed)	,063
Exact Sig. (1-tailed)	,032
Point Probability	,016

a. Grouping Variable: Group

b. Not corrected for ties.

```
> d1<-c(12, 25, 17, 11, 15)
> d1
[1] 12 25 17 11 15
> d2<-c(18, 100, 20, 27)
> d2
[1]  18 100  20  27
> wilcox.test(d1,d2,paired = FALSE)

        Wilcoxon rank sum test

data:  d1 and d2
W = 2, p-value = 0.06349
alternative hypothesis: true location shift is not equal to 0
```

Situation 2a. Paired observations, non-Normal differences.

Example:

The 1st and 2nd born twin of identical twins did a psychological test.

For each pair of twins we have a pair of test results (x, y).

We are interested whether the 1st born scores higher than the 2nd born.

The test results are actually scores, and there is some doubt about the normality assumption of the paired t-test.

Paired observations, *n* pairs (*x*, *y*), with *d*=*x*-*y*.

Experimental units: pair of twins.

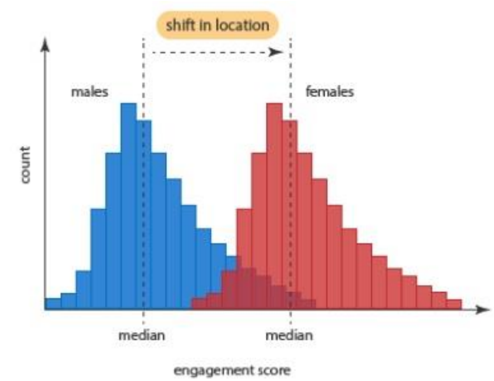
Measurement units: individual child

Before the party

Normality of differences is doubtful.

H_0: distribution of differences d is symmetrical around D_0,
H_a: differences d tend to be smaller than / larger than / unequal D_0.

D_0 is the H_0-value of the median of d.
Often D_0 is 0.



Use Wilcoxon's signed - rank test or the sign-test

Before the party

Normality of differences is doubtful.

H_0: distribution of differences d is symmetrical around D_0,
H_a: differences d tend to be smaller than / larger than / unequal D_0.

D_0 is the H_0-value of the median of d.
Often D_0 is 0.

H_0: no systematic difference in score among twins, or

H_0: distribution of differences between twins is symmetrical around 0

H_a: 1st twin tends to score higher than 2nd twin, or

H_a: differences between scores of 1st and 2nd born twin tend to be positive

Before the party

Calculate differences $d_i = (x_i - y_i) - D_0$.

Differences d_i that are zero, are left out.

Assign rank numbers to the absolute values of the remaining d_i .

When there are equal absolute differences, use mid-ranks.

Mark d_i*'s* with positive sign to get $T+$: sum of ranks of positive differences

Or

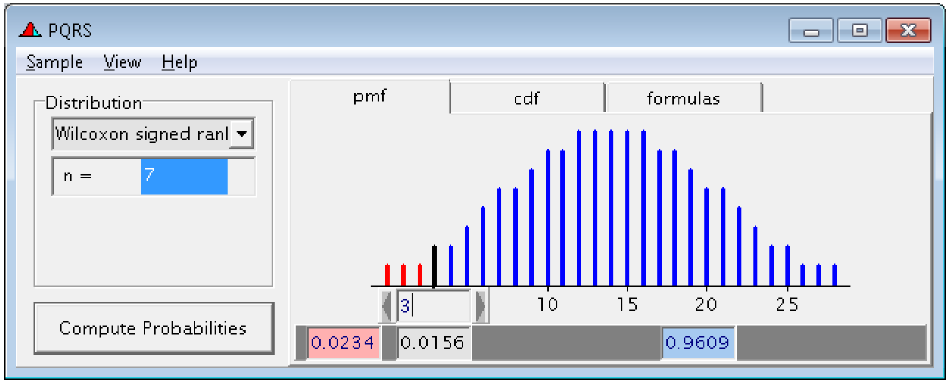
Mark d_i*'s* with negative sign to get $T-$: sum of ranks of negative differences

2) The test statistic: T_+ or T_-

Before the party

3) Under H_0: T- (or T+) ~ Wilcoxon signed rank (n) distribution

The exact distribution is shown by PQRS for the case of no ties.



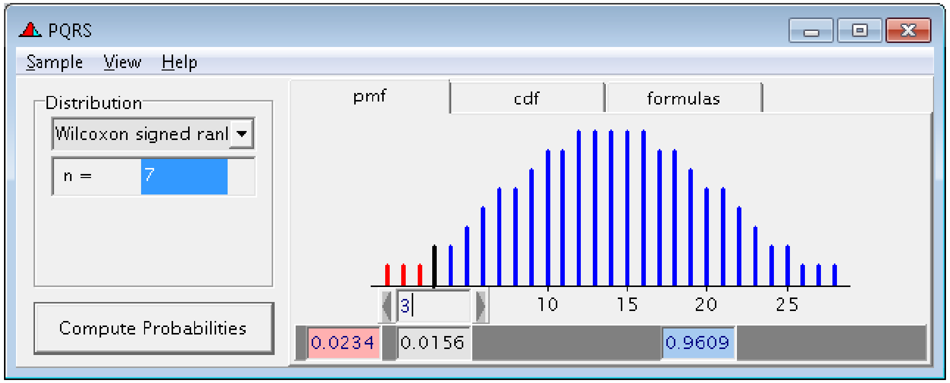
4) Under H_a T_+ or T_- tend to larger / smaller / larger or smaller values than under H_0

5) → use RPV / LPV / 2-tailed PV.

Before the party

3) Under H_0 T- (or T+) ~ Wilcoxon signed rank (n) distribution

The exact distribution is shown by PQRS for the case of no ties.



4) Under H_a T_+ or T_- tend to larger / smaller / larger or smaller values than under H_0

Under H_a, we expect positive differences $\rightarrow$ T- tends to be smaller than under H_0

5) → use RPV / LPV / 2-tailed PV.

We reject if the left sided P-value is smaller than 0.05 (= α)

During the party: Descriptive (Sample) Statistics

Pair of twins	1	2	3	4	5	6	7
1st born (x)	3.3	3.7	3.4	3.0	3.8	3.7	3.3
2nd born (y)	2.9	3.5	3.7	2.5	2.9	3.0	3.2

The after-party: Analysing your data

Pair of twins	1	2	3	4	5	6	7
1^{st} born (x)	3.3	3.7	3.4	3.0	3.8	3.7	3.3
2^{nd} born (y)	2.9	3.5	3.7	2.5	2.9	3.0	3.2
d	0.4	0.2	-0.3	0.5	0.9	0.7	0.1
|d|	0.4	0.2	0.3	0.5	0.9	0.7	0.1
Rank |d|	4	2	3	5	7	6	1

6) Outcome Test statistic:

Sum of ranks of positive d's : $T_+ = 4 + 2 + 5 + 7 + 6 + 1 = 25$

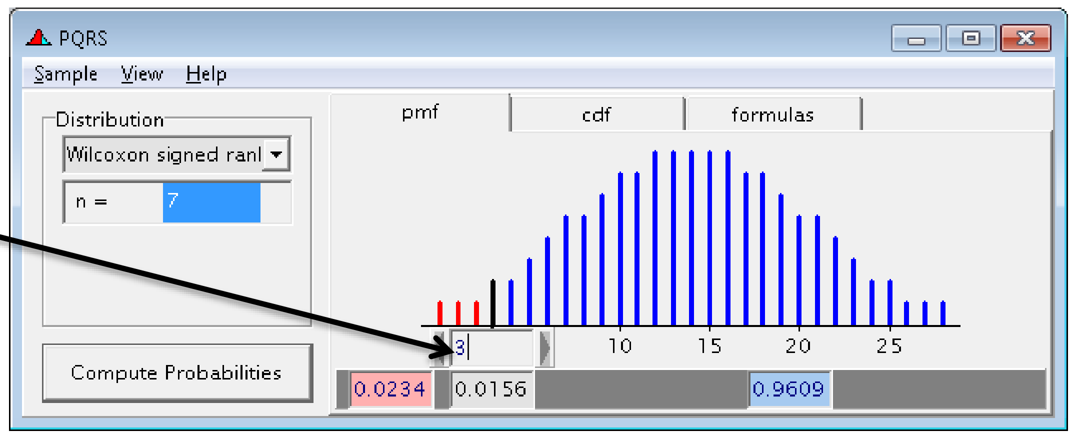
Sum of ranks of negative d's : $T_- = 3$

Now we need PQRS or SPSS output to give us the P-value.
Table 6 in O&L allows to find RR: **but we do not use it.**

The after party with PQRS

- Using $\boldsymbol{T}_{-} = 3$

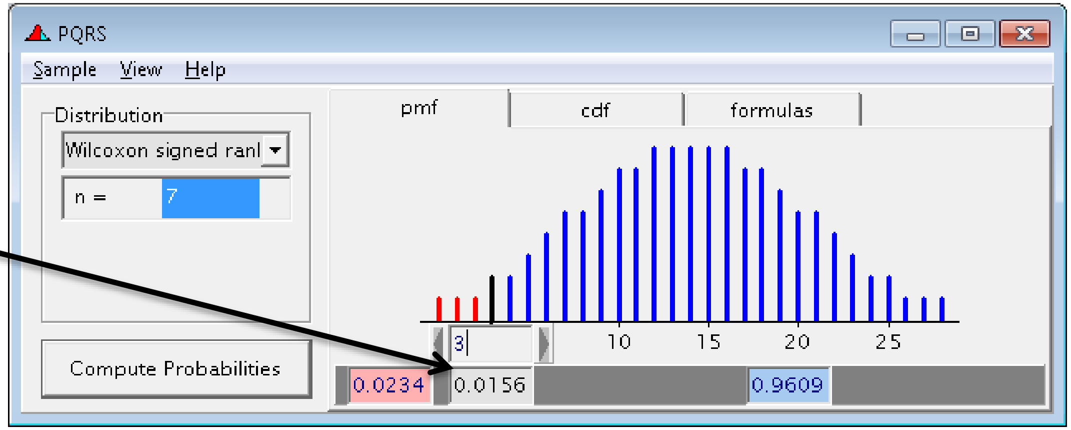
Look $\boldsymbol{T}_{-}$ up in left tail



- 7) P-value = 0.0234 + 0.0156 = 0.039 < 0.05,
- 8) H_0 is rejected, H_a is accepted. It is shown (α=0.05) that first borns systematically score higher on the test than 2nd borns of twins.

The after party with PQRS

- Using $T_- = 3$

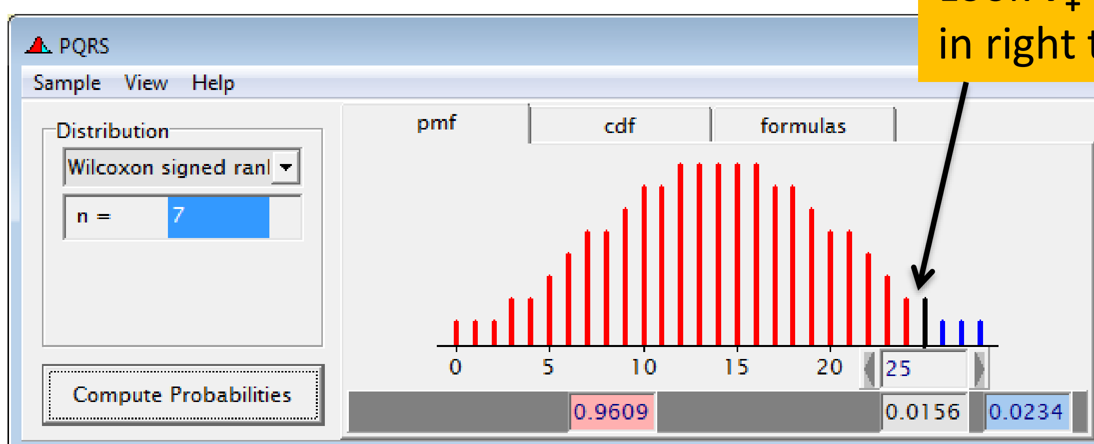


Look T_- up in left tail

- 7) P-value = 0.0234 + 0.0156 = 0.039 < 0.05,
- 8) H_0 is rejected, H_a is accepted. It is shown (α=0.05) that first borns systematically score higher on the test than 2nd borns of twins.

- If TS is T_+, then
- Under H_0 T_+ has the same distribution; Under H_a T+ tends to larger values, so use RPV.
- **7)** RPV = 0.039 <0.05 → same conclusion

Look T_+ up in right tail



The after party with SPSS / R output

Ranks

		N	Mean Rank	Sum of Ranks
firstborn - secondborn	Negative Ranks	1[a]	3.00	3.00
	Positive Ranks	6[b]	4.17	25.00
	Ties	0[c]		
	Total	7		

a. firstborn < secondborn
b. firstborn > secondborn
c. firstborn = secondborn

outcomes of T_- and T_+

Test Statistics[b]

	firstborn - secondborn
Z	-1.859[a]
Asymp. Sig. (2-tailed)	.063
Exact Sig. (2-tailed)	.078
Exact Sig. (1-tailed)	.039
Point Probability	.016

a. Based on negative ranks.
b. Wilcoxon Signed Ranks Test

One sided p-value

```
> b1<-c(3.3, 3.7, 3.4, 3, 3.8, 3.7, 3.3)
> b2<-c(2.9, 3.5, 3.7, 2.5, 2.9, 3.0, 3.2)
> wilcox.test(b1,b2,paired = TRUE)

        Wilcoxon signed rank test

data:  b1 and b2
V = 25, p-value = 0.07813
alternative hypothesis: true location shift is not equal to 0
```

Normal approximations

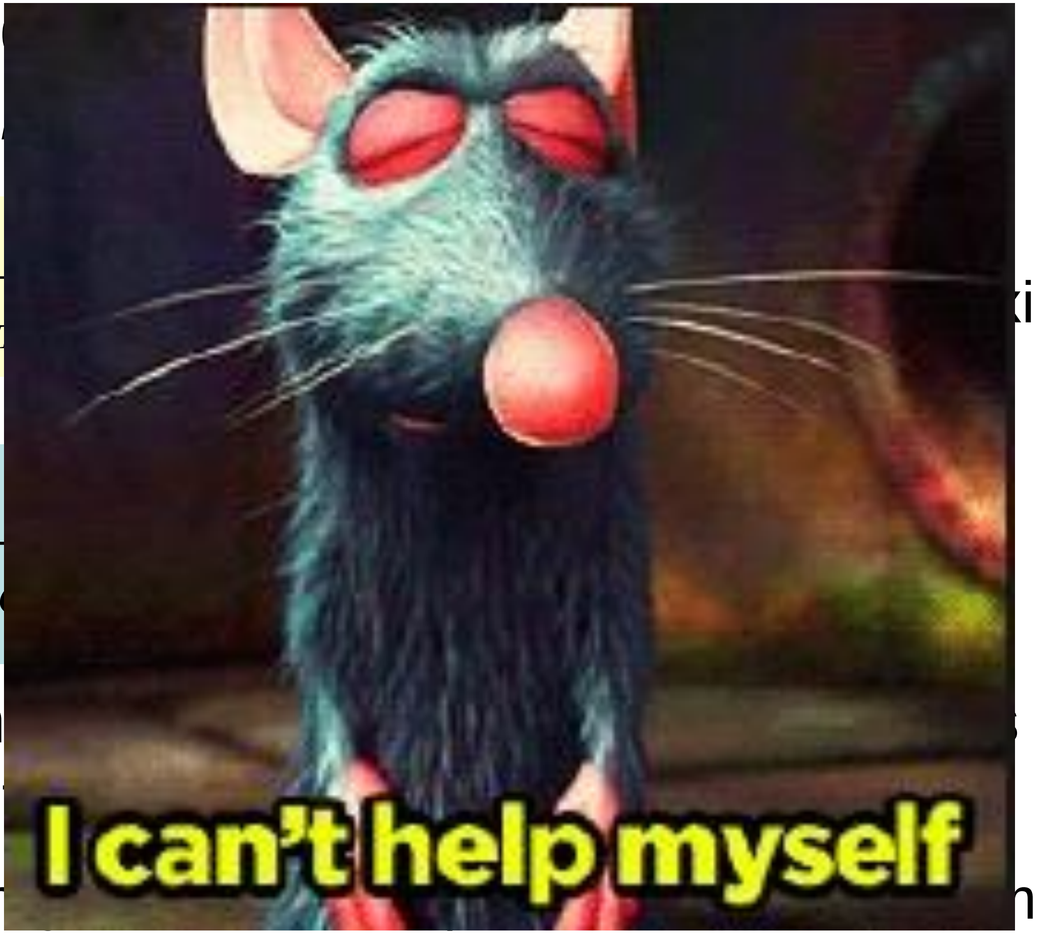
- We can also use a z-test based on a normal approximation (O&L p32

- For large

$$z = \frac{T -}{\sigma_T}$$

imately.

$$z = \frac{W}{}$$



- You do n for σ_T. You can skip

- SPSS ca nal approximations.

Normal approximations

- We can also use a z-test based on a normal approximation (O&L p320). **You can skip this.**
- For large n ($n > 50$) we can use the z-test:

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{T - \frac{1}{4}n(n+1)}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}}$$ ~ N(0,1) approximately.

$$z = \frac{W - \mu_W}{\sigma_W} = \frac{W - \frac{1}{2}n_1(n_1 + n_2 + 1)}{\sqrt{\frac{1}{12}n_1 n_2(n_1 + n_2 + 1)}}$$

- You do not have to know or use the formulas for σ_T. You can skip them.
- SPSS can give both exact probabilities, and normal approximations.