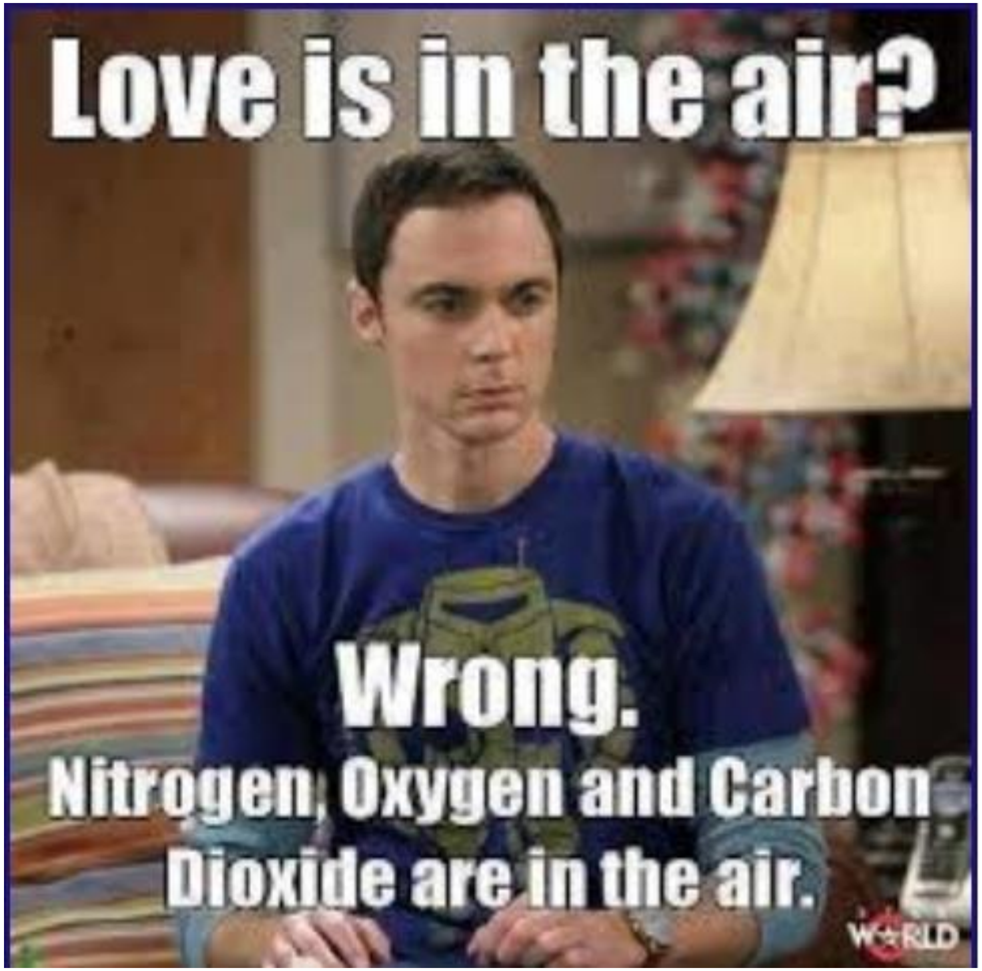
MAT20306 - Advanced Statistics

Lecture 1: Inference: t-procedures

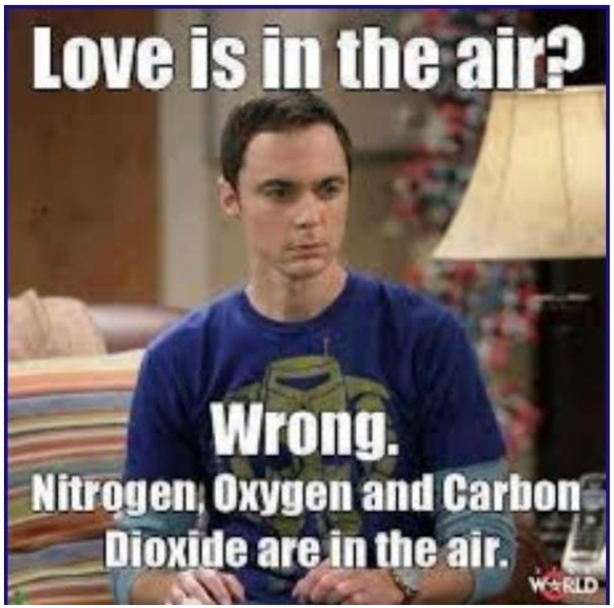
WAGENINGEN UR
For quality of life

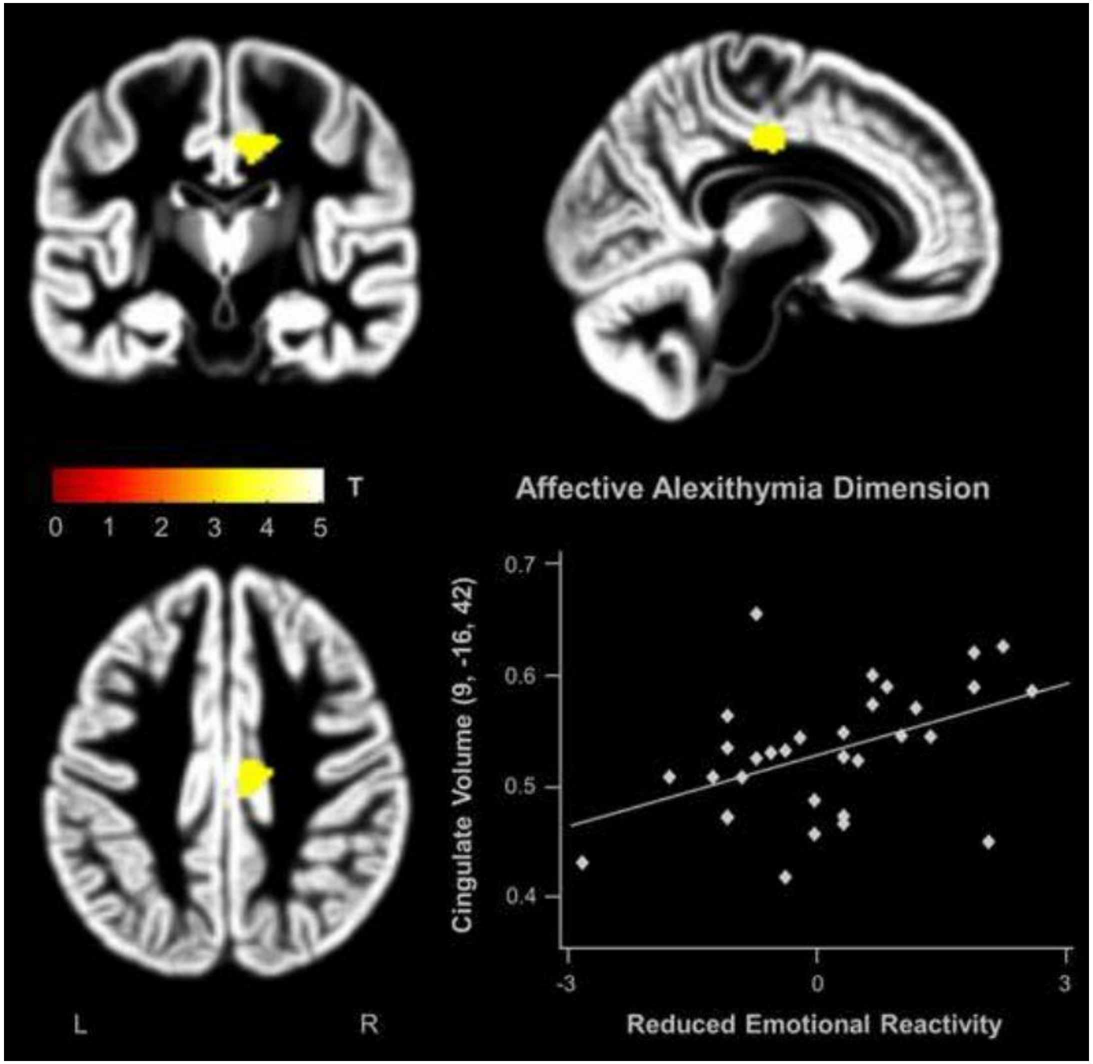
Biometris
Quantitative Methods brought to Life

Why statistics?



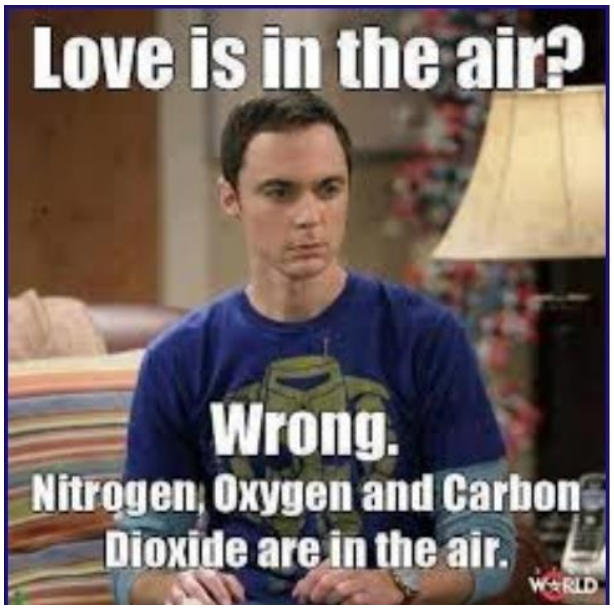
Simple Linear Regression: Example

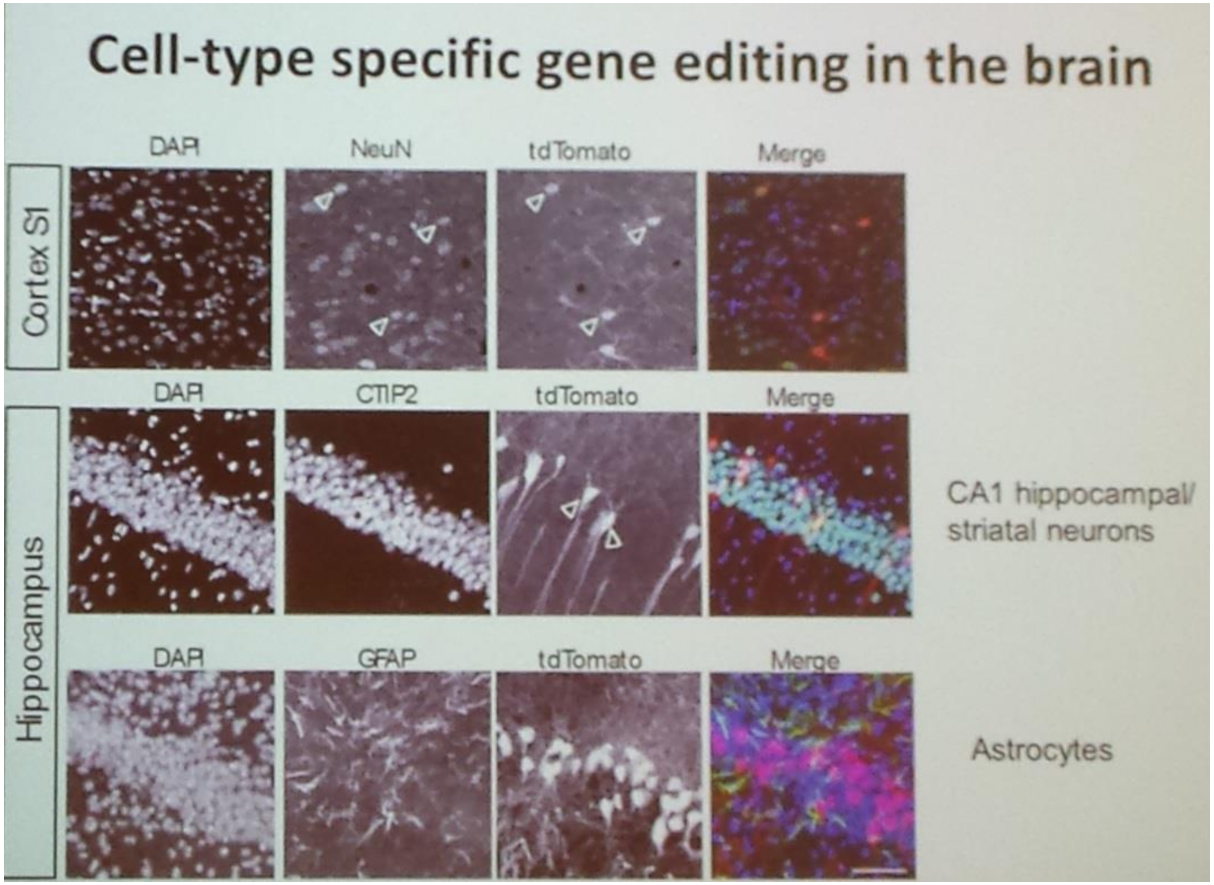




Simple Linear Regression: Implications

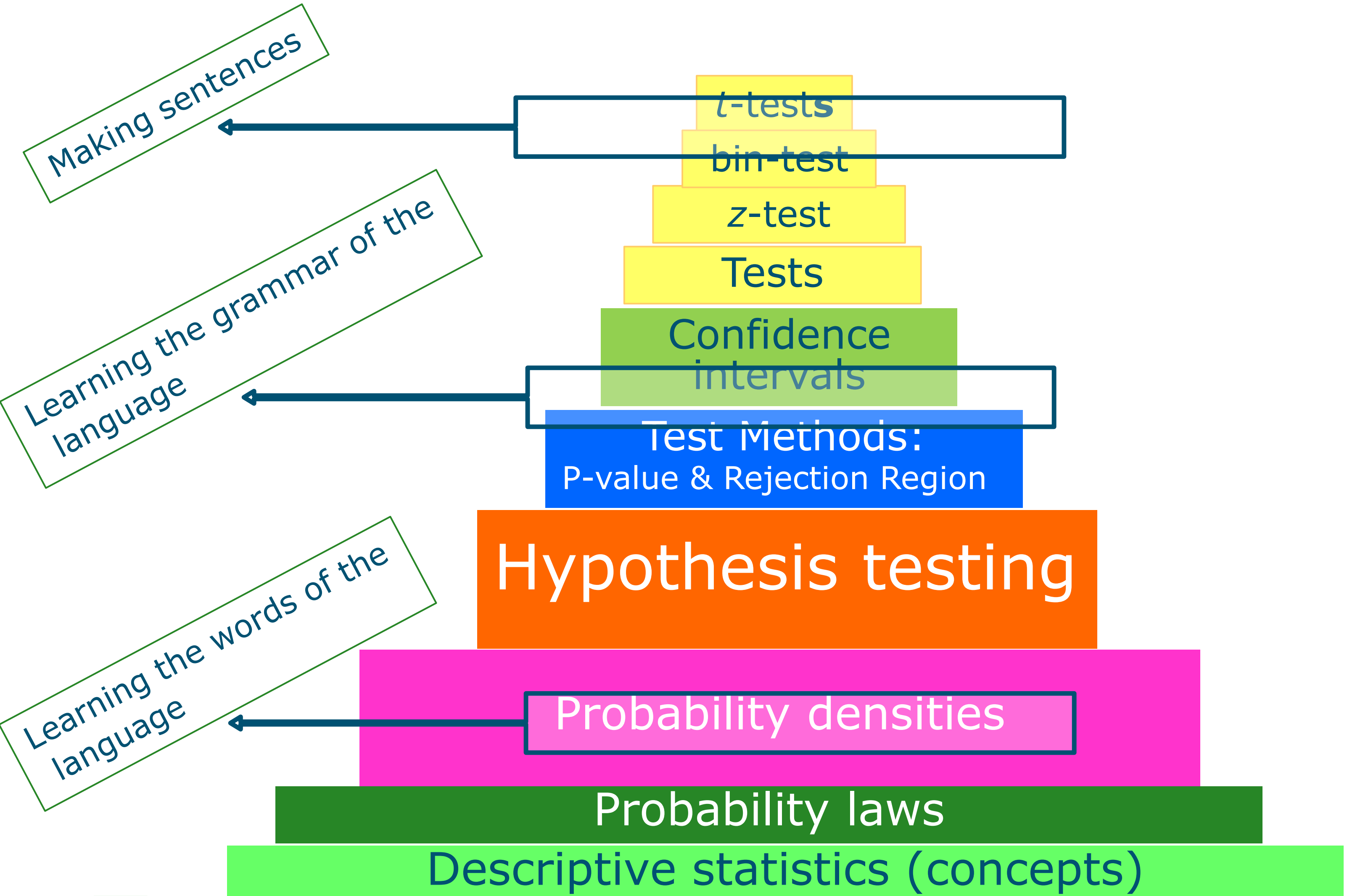
Jennifer Doudna, professor of Biomedical Sciences at UC Berkeley





WAGENINGEN UR
For quality of life

Topics @ Basic Level – Lecture 1 outline



WAGENINGEN UR
For quality of life

Example A:


FRAT PARTY | BACK TO THE ZEROS
pagina laatst gewijzigd op vrijdag 5 februari 2016 om 04:57
Line-up
Statistieken
Brooxx
Djennah
Emilé Laurent
Jeremy Herkul
NGHTSHD
MC: Maga Ranx
5 bezoekers · herkomst
3 geïnteresseerd · herkomst
305 bezoekers
287 geïnteresseerd
20 / 80 · mannen / vrouwen
25.8 · gemiddelde leeftijd
27.0 · leeftijd mediaan
150 × bekeken sinds 11 januari 2016

Example A:

FRAT PARTY | BACK TO THE ZEROS

pagina laatst gewijzigd op vrijdag 5 februari 2016 om 04:57

Line-up	Statistieken	
Brooxx	5	bezoekers · herkomst
Djennah	3	geïnteresseerd · herkomst
Emilé Laurent	305	bezoekers
Jeremy Herkul	287	geïnteresseerd
NGHTSHD	20 / 80	mannen / vrouwen
MC: Maga Ranx	25.8	gemiddelde leeftijd
	27.0	leeftijd mediaan
	150 ×	bekeken sinds 11 januari 2016

Research question?

Is the mean age of the people present at the party higher than 25.8 ?

Setup: one sample t-test

- We want to compare a characteristic in a population to some fixed number (25,8 years).
- The people "interviewed" at the party are the sampling units.
- The response is their age, measured per person (so the person is also the observed or observational unit).
- The scientist makes a guess about the population mean (age) based on one random sample.

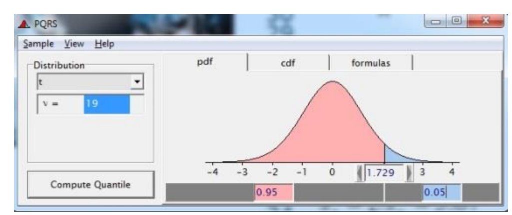
- The population is a **physical population**.

  The type of research is **observational.**

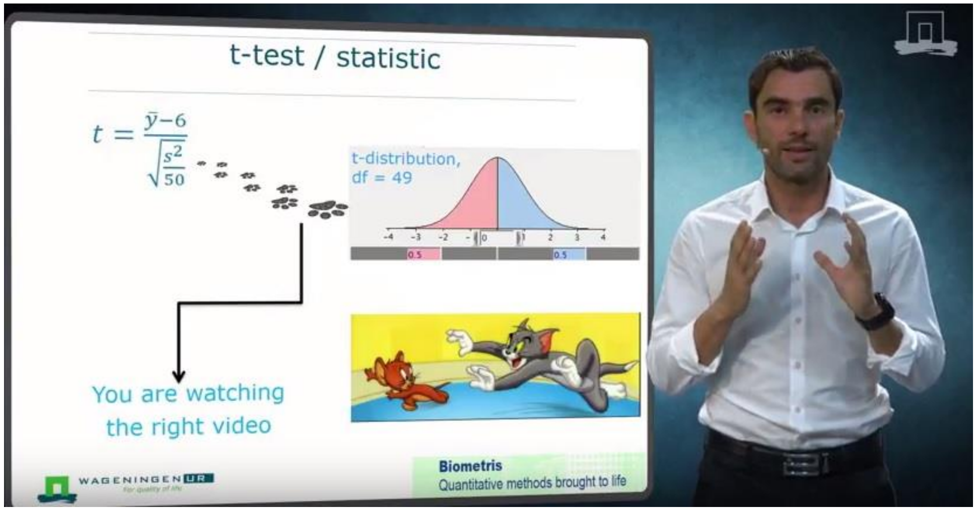
Before the party (observational study)

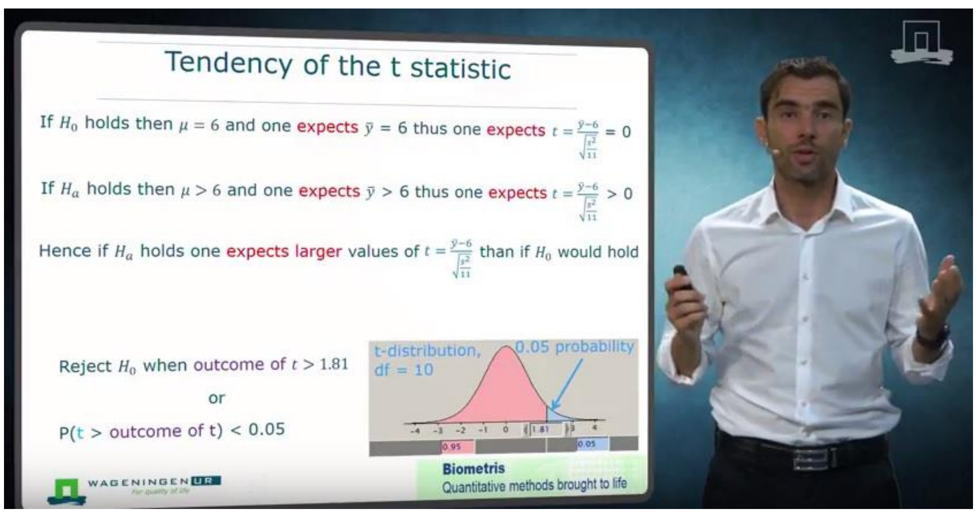
Before the party (observational study)

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 rejection region (RR): L, R or 2-sided. Specify the RR for given α

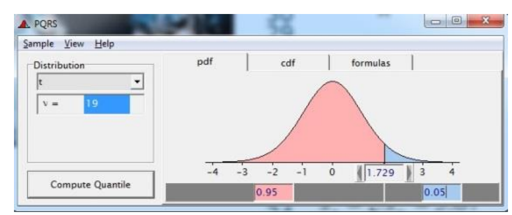


Before the party (observational study)





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 rejection region (RR): L, R or 2-sided. Specify the RR for given α



During the party: Descriptive (Sample) Statistics

18
25
28
23
30
32
22
30
29
19
33
21
22
34
19
36
22
24
34
29

n=20

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = 26.5$$

$$s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1};$$

$$s_y = \sqrt{s_y^2} = 5.67$$

The after-party: Analysing your data

6. Outcome of the test statistic
7. Is the outcome in the critical region?
8. Conclude whether H_0 is rejected or not

8b Formulate the conclusion in words:
H_0 is (not) rejected, H_a is (not) proven,
it is (not) shown that … (H_a in words)

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

Testing: the 8 steps

1. H_0 and H_a
2. Definition of the test statistic (TS)
3. Distribution of the test statistic if H_0 is true
4. Behaviour of TS, expected under H_A (larger / smaller / larger or smaller)
5. Type of rejection region (RR): L, R or 2-sided. Specify the RR for given α

DATA come in

6. Outcome of the test statistic
7. Is outcome in the critical region?
8. Conclude whether H_0 is rejected or not

8b Formulate the conclusion in words:

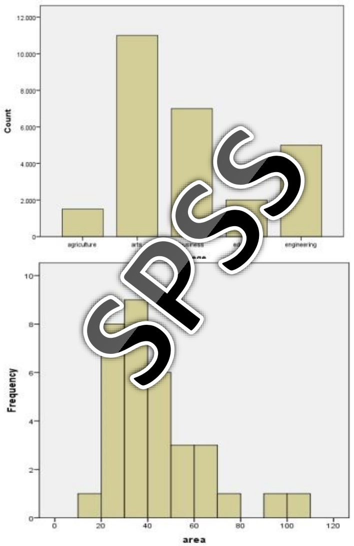
H_0 is (not) rejected, H_a is (not) proven,
it is (not) shown that … (H_a in words)

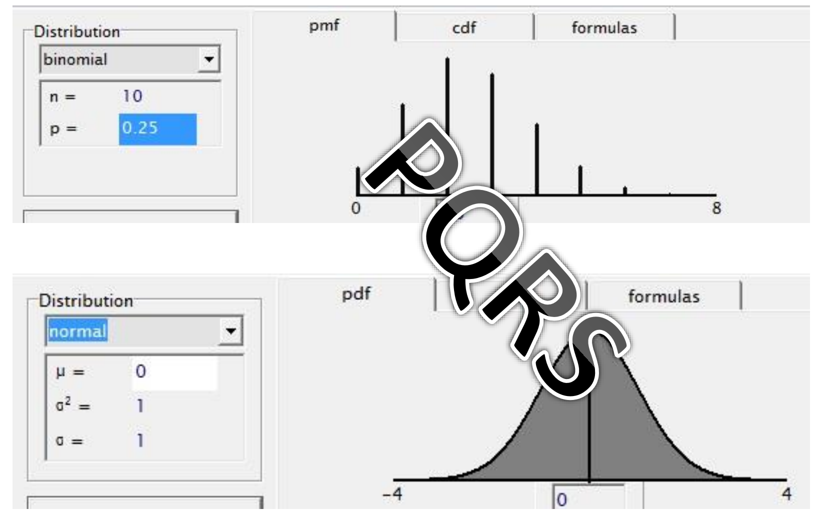
5A. type of P-value.
(L, R or 2-tailed PV)

7A. Determine P-value
Is P-value < α?

alternative steps when the P-value is used, rather than the rejection region

Random sample VS Population





Actual outcomes of the variable	Possible outcomes of the variable
Relative frequency	Probability
Bar chart / Histogram	Probability distribution
$\bar{y}$ (sample mean), s (sample st. dev.)	μ (expected value), σ (population st. dev.)

Testing: Quantiles (RR) and Probabilities (p-values)

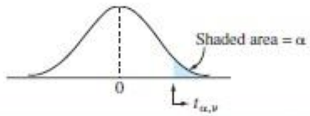
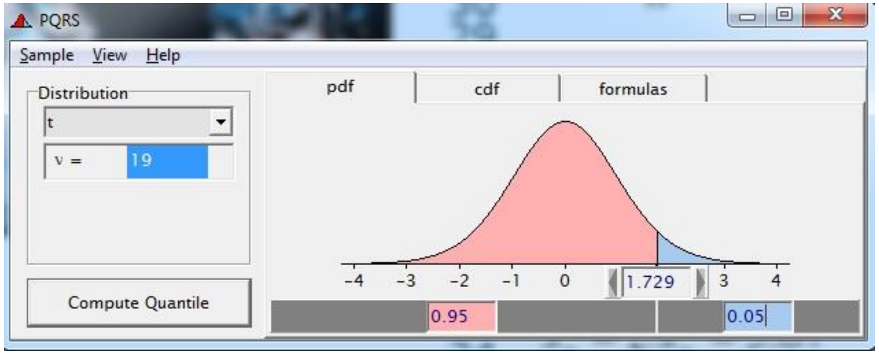
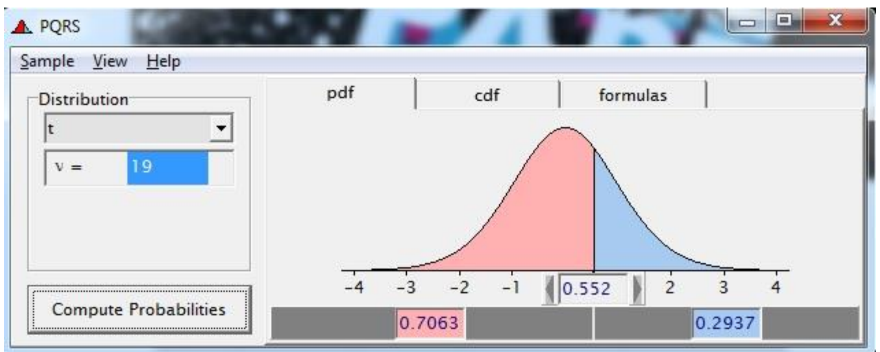


TABLE 2
Percentage points of Student's t distribution

	Right-Tail Probability (α)								
df	**.40**	**.25**	**.10**	**.05**	**.025**	**.01**	**.005**	**.001**	**.0005**
1	.325	1.000	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	.289	.816	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	.277	.765	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	.271	.741	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	.267	.727	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	.265	.718	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	.262	.706	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	.261	.703	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	.260	.700	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	.260	.697	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	.259	.695	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	.258	.691	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	.257	.688	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	.257	.687	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	.256	.685	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	.256	.684	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	.256	.684	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	.256	.684	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	.256	.683	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	.256	.683	1.310	1.697	2.042	2.457	2.750	3.385	3.646
35	.255	.682	1.306	1.690	2.030	2.438	2.724	3.340	3.591
40	.255	.681	1.303	1.684	2.021	2.423	2.704	3.307	3.551
50	.255	.679	1.299	1.676	2.009	2.403	2.678	3.261	3.496
60	.254	.679	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	.254	.677	1.289	1.658	1.980	2.358	2.617	3.160	3.373
inf.	.253	.674	1.282	1.645	1.960	2.326	2.576	3.090	3.291





WAGENINGEN UR
For quality of life

Example B: Tape worms in sheep

Researchers investigate the effectiveness of a new drug D for tape worms in the stomach of sheep.

24 sheep, infected with worms, are randomly divided into 2 groups.

Sheep in one group receive the new drug, sheep in the other group receive no treatment (NT).

After 6 months the worms are counted.

Setup: 2-independent samples t-test

- We compare the mean number of worms in two "populations": one of sheep receiving the new drug, one of sheep receiving no treatment. The two populations are not physical, but hypothetical.
- The 24 sheep are not randomly selected. Randomization is introduced by a lottery deciding which 12 sheep get the new treatment. Any 12-12 result of the randomization is OK,
- The sheep are the experimental units.
- The response is the number of tape worms in the sheep, so the sheep (or the stomach of the sheep) is the measured unit.

Observational study vs. Experimental study

	Observational	Experimental
Level of control of conditions:	low	high
Randomization	sampling	random treatment allocation to exp. units
Risk of confounders	high	low
Cause-effect conclusion?	No	yes

Sampling unit = unit that is actually sampled

Experimental unit = unit to which treatment is randomly assigned

For both types

Observed / measured unit = unit upon which a response is measured.

This can be a sub-unit of the experimental units or sampling unit.

WAGENINGEN UR For quality of life

Statistical model (step 0 in the analysis)

- The samples are assumed to be from two normal populations:

  $y_1 \sim N(\mu_1, \sigma^2)$, for sheep receiving the new drug

  $y_2 \sim N(\mu_2, \sigma^2)$, for receiving no treatment

- Responses y_1 ... y_{24} from the 24 sheep are independent (random sample of sheep , randomly assigned to the groups).
- Note that equal variance σ^2 for the two distributions is assumed.
- So, we assume: - normality of the observations

  - equal variance in the two populations

  - independence of the observations

Before the party (experimental study)

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 rejection region (RR): L, R or 2-sided. Specify the RR for given α

WAGENINGEN UR
For quality of life

H_0, H_a and a statistical test – sheep example

- We call the population means to be compared e.g.: μ_1 and μ_2
  - μ_1 for the new drug μ_2 for no treatment
- The research hypothesis is that μ_1 is less than μ_2. This goes into the alternative hypothesis, so
- **1)** $H_a: \mu_1 - \mu_2 < 0.$
- **1)** $H_0: \mu_1 - \mu_2 = 0.$ (the null hypothesis, always with "=")

WAGENINGEN UR
For quality of life

The test statistic

- The test statistic measures how much the data deviate from H_0 value of the parameters

2) $$t = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{se(\bar{y}_1 - \bar{y}_2)} = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- Where $s_p = \sqrt{{s_p}^2}$, with s_p^2 the pooled estimate of the variance

$$ {s_p}^2 = \frac{(n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2}{n_1 + n_2 - 2}$$

What to expect for t under H_0 is specified in step 3.

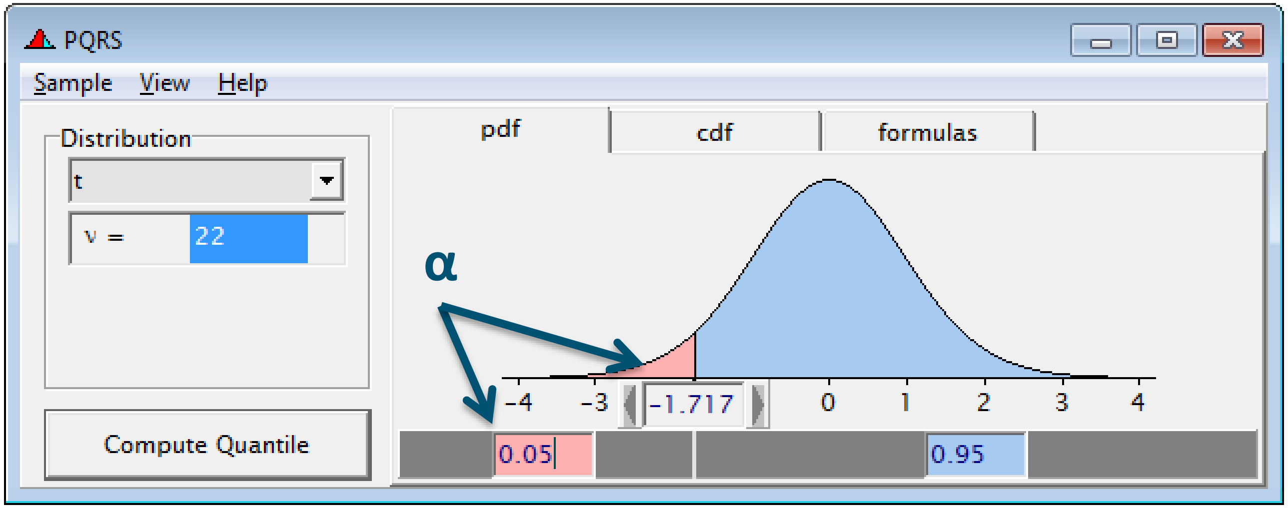
Under H_0 *t* follows a t-distribution with *df* =(12-1)+(12-1)=22

3) $$t \overset{H_0}{\sim} t_{22}$$

WAGENINGEN UR
For quality of life

The t-distribution and the rejection region

4) Under H_a t tends to smaller values → so Rejection Region is left-sided



5) Rejection region: $t < - t_{22}(0.05) = -1.717$ (see also Table 2)

During the party: Descriptive (Sample) Statistics

<u>Summary statistics</u>	group 1	group 2
Sample sizes:	$n_D = 12$	$n_{NT} = 12$
Sample means:	$\bar{y}_D = 26.58$	$\bar{y}_{NT} = 39.67$
Sample standard deviations:	$s_D = 14.36$	$s_{NT} = 13.86$

The after-party: Analysing your data

6. Outcome of the test statistic
7. Is the outcome in the critical region / p-value < α ?
8. Conclude whether H_0 is rejected or not

8b Formulate the conclusion in words:

H_0 is (not) rejected, H_a is (not) proven,

it is (not) shown that … (H_a in words)

Outcome of the test statistic in the sheep example

$\bar{y}_1 = 26.58, \bar{y}_2 = 39.67, s_1 = 14.36, s_2 = 13.86$

$$s_p{}^2 = \frac{(12-1)*s_1^2 + (12-1)*s_2^2}{24-2} = \frac{11*(14.36)^2 + 11*(13.86)^2}{22} = 199.2$$

$$s_p = \sqrt{199.2} = 14.11 \rightarrow \quad t = \frac{(26.58-39.67)-0}{14.11\sqrt{\frac{1}{12}+\frac{1}{12}}} = \mathbf{-2.272}$$

6) Outcome: t=-2.272

7) t < -1.717 , so t is in the RR

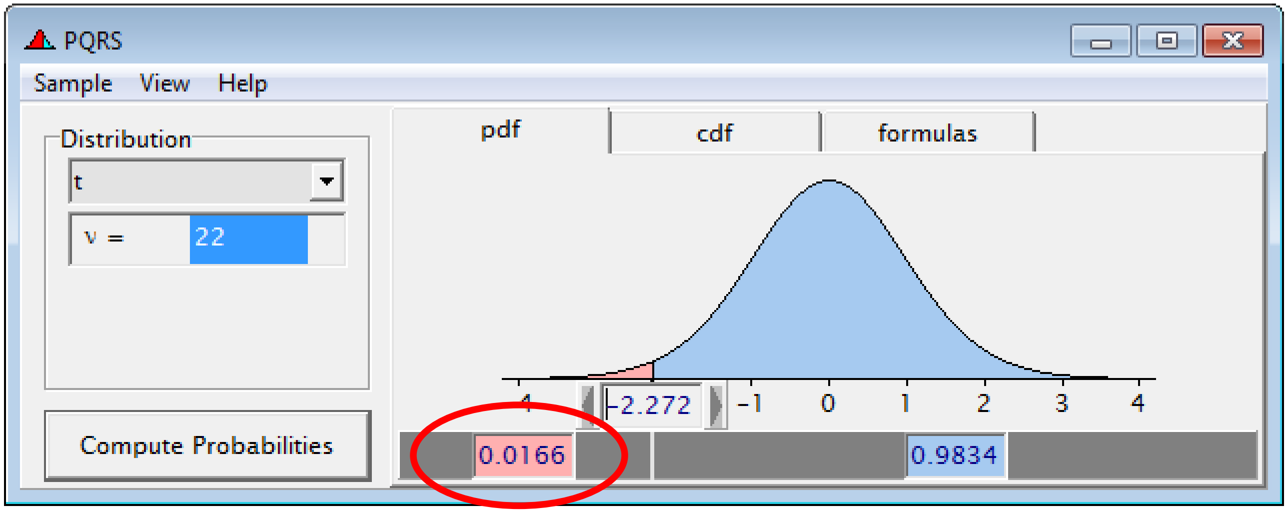
8) so H_0 is rejected, H_a is proven.

8b) It is shown that the **expected (mean)** number of tapeworms is lower in animals treated with the new drug.

The P-value -method

6) t= -2.272

7A) P-value = LPV = $P(t \leq -2.272) = 0.0166$



P-value < 0.05, so

8) we reject H_0, etc.

SPSS

$$t = \frac{(\bar{y}_1 - \bar{y}_2)}{s_p \sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}}$$

$$2P(|t| > |t_{obs}|$$

two sided P – value

$$s_p \sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}$$

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_A : \sigma_1^2 \neq \sigma_2^2$$

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Nworms	Equal variances assumed	.205	.655	-2.271	22	.033	-13.083	5.761	-25.032	-1.135
	Equal variances not assumed			-2.271	21.972	.033	-13.083	5.761	-25.033	-1.134

$$t' = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$\bar{y}_1 - \bar{y}_2$$

$$2P(|t'| > |t'_{obs}|)$$

two sided P – value

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

As t<0, LPV=0.033/2

RPV = 1- LPV = 1-0.033/2

Example C: Blood pressure change

A physician records the blood pressure before (x) and after 2 weeks (y) of medication use for 16 patients: d = x-y.
She regards them as a random sample from the population of all people with high blood pressure in the area.

<u>Summary statistics</u>

Sample size:	$n = 16$
Sample mean:	$\bar{d} = 6$
Sample st. deviation:	$s_d = 12$

Research question?

Setup: paired t-test

- We want to blood pressure before and after medication use. The population could be: all patients (in this practice or in the Netherlands) that have high blood pressure.
- The patients are the sampling units.
- The response is blood pressure, measured twice per patient, and the patient-moment combinations are the observed units

WAGENINGEN UR
For quality of 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 $\neq$ RR

WAGENINGEN UR
For quality of life

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 pop. Mean:

with Normality 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, with $z_{\alpha/2}$ from N(0,1) distribution.

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

WAGENINGEN UR
For quality of life

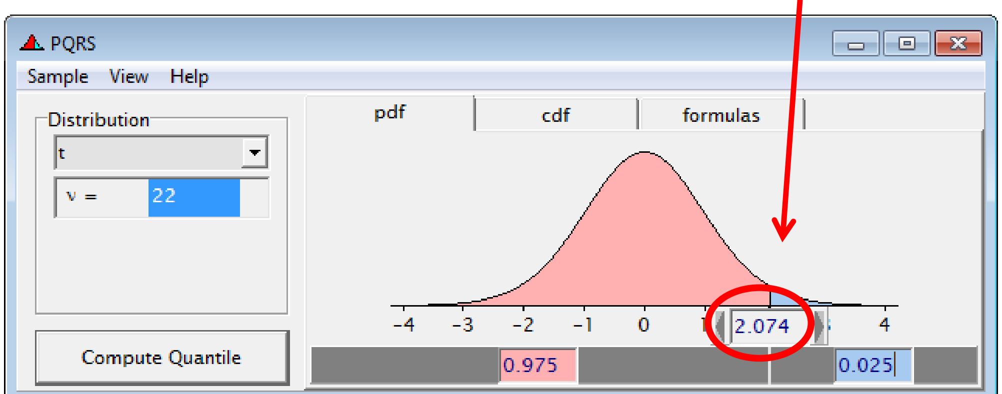
Confidence interval for the sheep example

estimate ± constant * standard error (estimate)

$\bar{y}_1 = 26.58$
$\bar{y}_2 = 39.67$
estimate = 26.58 - 39.67 = -13.03

t-distr.
df = 22
constant = 2.074

$se = 14.11\sqrt{\frac{2}{12}} = 5.76$



0.95-confidence interval:

$$(-13.03 \pm 2.074 * 5.76)$$
$$=$$
$$(\mathbf{-25.0}\,,\ \mathbf{-1.1})$$

For quality of life

Are we there yet ? ...

WAGENINGEN UR
For quality of life

One sided vs two sided

- We expect the drug to be better that no treatment at all, so: $H_0: \mu_1 - \mu_2 = 0$

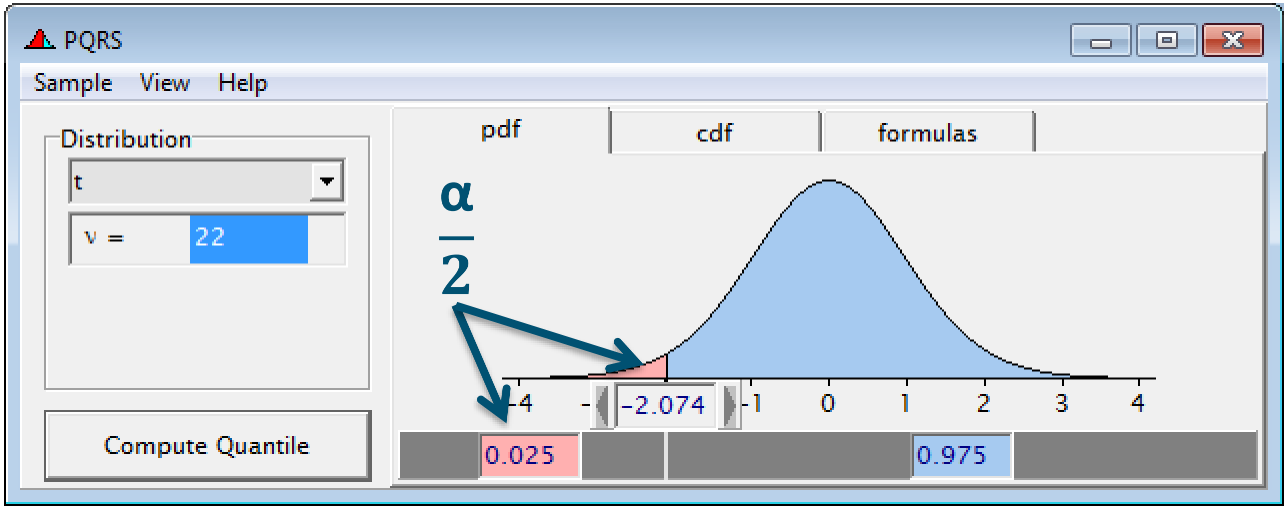
$$H_A: \mu_1 - \mu_2 < \ 0 \qquad \text{(one sided } H_A\text{)}$$

- What if we had no expectation?

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_A: \mu_1 - \mu_2 \ \neq 0 \qquad \text{(two sided } H_A\text{)}$$

two sided H_A: critical region



Rejection region:

all outcomes of t **smaller** than −2.074, and

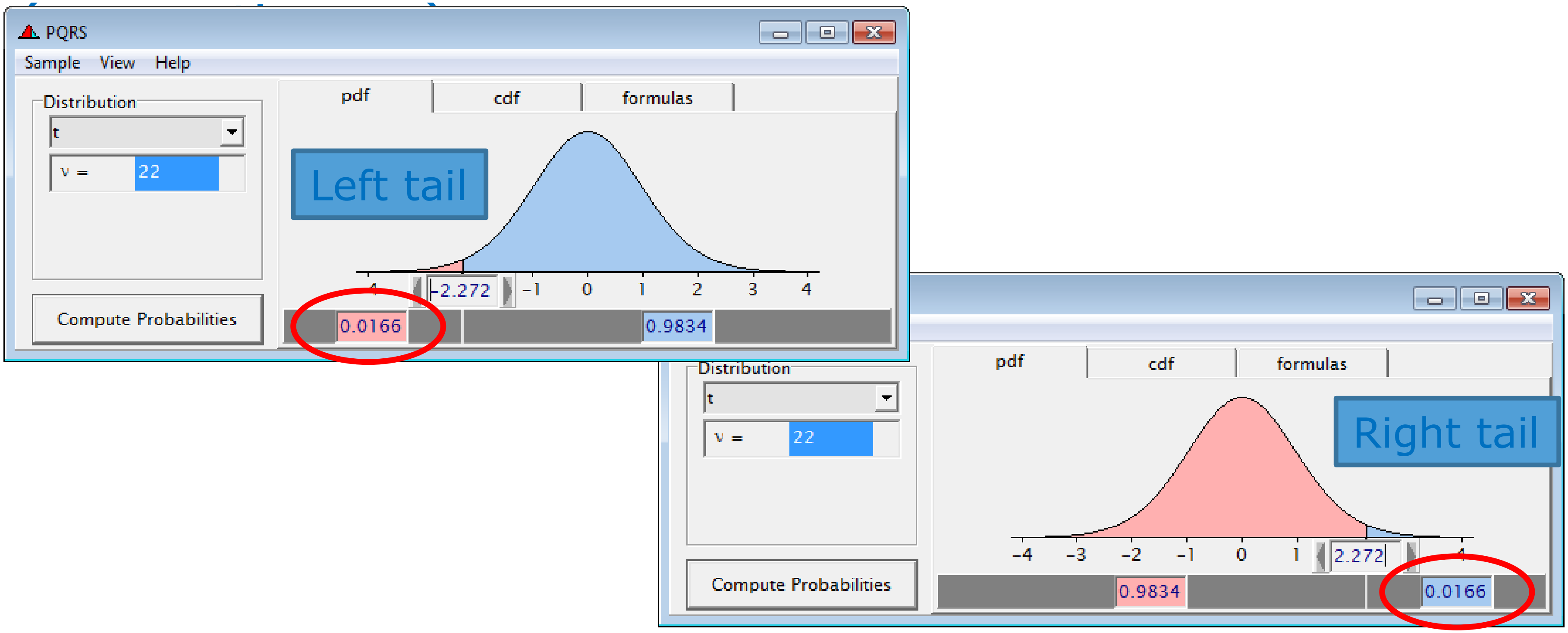
all outcomes of t **larger** than +2.074.

RR: (-∞, −2.074) and (2.074, ∞)

WAGENINGEN UR
For quality of life

two sided H_A: p-value

P-value = probability under H_0 for the outcome of test statistic t and anything more extreme



P-value = $P(t < -2.272) + P(t > 2.272) = 2 * 0.0166 = 0.0332$

Equal variances or not?

What if variances of both samples are not equal?

We cannot pool the variances in S_p.

Use different expression for the standard error.

$$equal\ variances:\ t = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{se(\bar{y}_1 - \bar{y}_2)} = \frac{(\bar{y}_1 - \bar{y}_2)\ - 0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$unequal\ variances:\ t = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{se(\bar{y}_1 - \bar{y}_2)} = \frac{(\bar{y}_1 - \bar{y}_2)\ - 0}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

SPSS

$$t=\frac{(\bar{y}_1-\bar{y}_2)}{s_p\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})}}$$

$$2P(|\,t\,|>|\,t_{obs}\,|$$
$$\text{two sided P} - \text{value}$$

$$s_p\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})}$$

$$H_0:\sigma_1^2=\sigma_2^2$$
$$H_A:\sigma_1^2\neq\sigma_2^2$$

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Nworms	Equal variances assumed	.205	.655	-2.271	22	.033	-13.083	5.761	-25.032	-1.135
	Equal variances not assumed			-2.271	21.972	.033	-13.083	5.761	-25.033	-1.134

$$t'=\frac{\bar{y}_1-\bar{y}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

$$\bar{y}_1-\bar{y}_2$$

$$2P(|\,t'|>|\,t'_{obs}\,|)$$
$$\text{two sided P} - \text{value}$$

$$\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

As t<0, LPV=0.033/2

RPV = 1- LPV = 1-0.033/2

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.

WAGENINGEN UR
For quality of life

# 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	$z=\dfrac{\bar{y}-\mu_0}{\sigma/\sqrt{n}}$	$z \sim z(0, 1)$	$\bar{y} \pm z_{\alpha/2} * \sigma/\sqrt{n}$
1 sample 1 variable	Population expected valued	$\mu=\mu_0$	σ is unknown	$t=\dfrac{\bar{y}-\mu_0}{s/\sqrt{n}}$	$t \sim t(n-1)$	$\bar{y} \pm t_{\alpha/2} * s/\sqrt{n}$
2 samples 1 variable	Difference between two population expected values	$\mu_1-\mu_2=D_0$	$\sigma_1 = \sigma_1$ **OR** $\sigma_1 \neq \sigma_1$	$t=\dfrac{\bar{y}_1-\bar{y}_2-0}{s_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$ $t'=\dfrac{\bar{y}_1-\bar{y}_2-0}{\sqrt{\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}}}$	$t \sim t(n_1+n_2-2)$ $t' \sim t(df)$ from SPSS output	$\bar{y}_1-\bar{y}_2 \pm t_{\alpha/2} s_p \sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}$ $(\bar{y}_1-\bar{y}_2) \pm t_{\alpha/2} * \sqrt{\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}}$
1 sample 2 variable	Population expected difference	$\mu_d = D_0$	Observations are paired	$t=\dfrac{\bar{d}-\mu_d}{s_d/\sqrt{n}}$	$t \sim t(n-1)$	$\bar{d} \pm t_{\alpha/2} * s_d/\sqrt{n}$