Statistiek I

t-tests

John Nerbonne

CLCG, Rijks*universiteit* Groningen

http://www.let.rug.nl/nerbonne/teach/Statistiek-I/

Overview

t-Tests

To test an average or pair of averages when σ is known, we use z-tests

But often σ is unknown, e.g., in specially constructed psycholinguistics tests, in tests of reactions of readers or software users to new books or products.

In general, σ is known only for standard tests (IQ tests, CITO tests, ...).

***t*-Tests** incorporate an “estimation” of σ (based on the standard deviation SD of the sample in order to reason in the same way as in z-tests.

Different *t*-Tests

Student's *t*-Test ('Student' pseudonym of Guiness employee without publication rights)

- three versions:
  - independent samples compares two means
    determine whether difference is significant
  - single sample (estimate mean)
  - paired data compares pairs of values
    example: two measurements on each of 20 patients
- population statistic σ unnecessary
  of course, sample statistics needed
- appropriate with numeric data "roughly normally distributed"
  see Mann-Whitney U-Test, Wilcoxon signed rank test for non-parametric fall-backs (later in course)

Excursus: Degrees of Freedom

Most hypothesis-tests require that one specify DEGREES OF FREEDOM (*dF*) *vrijheidsgraden* — the number of ways in which data could vary (and still yield same result).

Example: 5 data points, mean
If mean & 4 data points known, fifth is determined
Mean 6, data is $4, 5, 7, 8$ and one unknown
☐ fifth $= 6$

There are **four** degrees of freedom in this set.

In general, with n numbers, $n - 1$ degrees of freedom (for the mean).

The *t* Statistic

t statistic:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \qquad z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Note that s is used (not σ). Of course, s should be good estimate. Cf. z test.

n is (usually) the number of items in the sample

Always used with respect to a number of **degrees of freedom**, normally $n - 1$ (below we discuss exceptions)

To know the probability of a t statistic we refer to the tables (e.g., M&M, Tabel E, Field, Table A2). We have to check on $P(t(\mathrm{df}))$, where dF is the degrees of freedom, normally $n - 1$.

t Tables in Field

A.2. Critical values of the *t*-distribution

	Two-Tailed Test		One-Tailed Test	
df	0.05	0.01	0.05	0.01
1	12.71	63.66	6.31	31.82
2	4.30	9.92	2.92	6.96
3	3.18	5.84	2.35	4.54
4	2.78	4.60	2.13	3.75
5	2.57	4.03	2.02	3.36
6	2.45	3.71	1.94	3.14
7	2.36	3.50	1.89	3.00
8	2.31	3.36	1.86	2.90
9	2.26	3.25	1.83	2.82
10	2.23	3.17	1.81	2.76
11	2.20	3.11	1.80	2.72
12	2.18	3.05	1.78	2.68
13	2.16	3.01	1.77	2.65
14	2.14	2.98	1.76	2.62
15	2.13	2.95	1.75	2.60
16	2.12	2.92	1.75	2.58

t Tables

Given degrees of freedom, dF, and chance of $t \leq p$, what t value is needed?

dF/p		0.05	0.01	0.001	
10	...	1.8	2.8	4.1	...
20	...	1.7	2.5	3.6	...
30	...	1.7	2.5	3.4	...
40	...	1.7	2.4	3.3	...
⋮	⋮	⋮	⋮	⋮	⋮
100	...	1.660	2.364	3.174	...
z	...	1.645	2.326	3.091	...

Note comparison to z. For $n \geq 100$, use z (differences neglible).

Compare t-tables. Be able to use table both to check $P(t)$ and, for a given p, to find min $t|P(t) \leq p$

z vs t

Suppose you mistakenly used s in place of σ in a z test with 10 or 20 elements in the sample, what would the effect be?

dF/p		0.05	0.01	0.001	
10	...	1.8	2.8	4.1	...
20	...	1.7	2.5	3.6	...
z	...	1.65	2.33	3.09	...

You would, e.g., treat a differences of $+2.33s$ as significant at the level 0.01 level (only 1% likely), when in fact you need to show differences of $+2.8$ or $+2.5$, respectively, to prove this level of significance.

Applying a z test using s instead of σ **overstates** the significance of the results (and increases the chance of a type I error, and a false positive decision on the H_a).

Independent Sample t-Tests

Two samples, unrelated data points (e.g., not before-after scores).

Compares sample means, $\bar{x}_1, \bar{x}_2$, wrt significant difference.

H_0 is always $\mu_1 = \mu_2$, i.e., that two populations have the same mean.
Two-sided alternative is $H_a : \mu_1 \neq \mu_2$ We use $\bar{x}_1, \bar{x}_2$ to estimate μ_1, μ_2:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s \cdot \sqrt{1/n_1 + 1/n_2}}$$

Degrees of freedom: use what SPSS or R provides, roughly

$$\mathrm{dF} = (n_1 - 1) + (n_2 - 1) = n - 2$$

Notate bene: Some books more conservative, $\mathrm{dF} \approx \mathrm{Min}\{(n_1 - 1), (n_2 - 1)\}$
— fewer *dF* makes showing significance harder, results more reliable

t increases with large diff. in means, or with small standard deviations (like z).

Independent Sample t-Test: Assumptions

Assumptions of Independent Sample t-tests (M&M, §7.1.5; Field, §9.3.2)

1. Exactly two samples **unrelated**
2. Distribution roughly normal.
   Field is vague on this, but rules of thumb are:
   - normality required if $n < 15$
   - no large skew or outliers if $n < 40$

Equal variances (sd's) often mentioned as condition, but easily corrected for (Welch test).

If three or more samples, use ANOVA (later in course, Stats II).

If distribution unsuitable, using Mann-Whitney (later in course).

Independent Sample *t*-Test: Example

Example: You wish to know whether there's a difference in verbal reasoning in boys vs. girls. There are tests, but no published σ. You test for a difference in average ability in these two samples.

Assume two variables, `VReason, Sex`

1	78	M
2	90	M
⋮	⋮	⋮
19	71	F
20	82	F

Two **independent** samples (no two scores from same person).

Example t-Test: One-Sided or Two-Sided?

Example: You wish to know whether there's a difference in verbal reasoning in boys vs. girls. There are tests, but no published σ. You test for a difference in average ability in the two samples.

No question of one being *better* than the other.

This is a **two-sided question.**

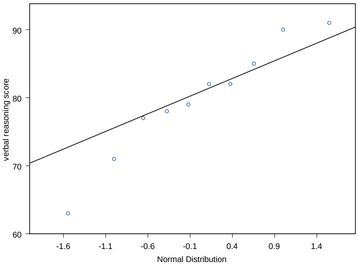
Hypotheses: $$\begin{array}{l} H_0: \ \mu_m = \mu_f \\ H_a: \ \mu_m \neq \mu_f \end{array}$$

What would hypotheses be if we asked whether boys are better than girls?

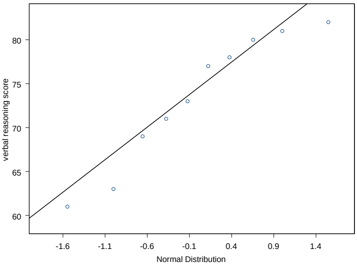
Independent Sample *t*-Test: Normality Test

$n_1 = n_2 = 10$, so for *t*-test, distributions must be roughly normal. Are they?

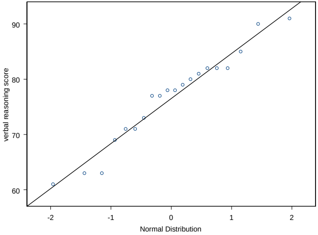
Boys



Girls



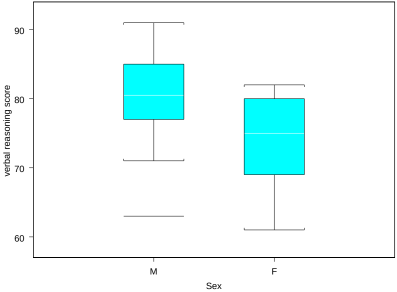
Normal Quantile Plots



Plot expected normal distribution quantiles (x axis) against quantiles in samples. If distribution is normal, the line is roughly straight. Here: distribution roughly normal.

More exact techniques: check KOLMOGOROV-SMIRNOV GOODNESS OF FIT or SHAPIRO-WILKS TEST (both uses H_0: distribution normal). If rejected, alternative tests are needed (e.g., Mann-Whitney).

Visualizing Comparison



Box plots show middle 50% in box, median in line, 1.5 interquartile range (up to most distant).

Independent Sample *t*-Test: Example

But is difference statistically significant?

Invoke (in S+)

1. Compare samples $\rightarrow$ Two Samples $\rightarrow$ *t*-Test
2. Specify variable `VReason`, groups `Sex`
3. Specify two-sided comparison, no assumption of equal variance (conservative).

Calculates *t*,
dF,
two-tailed probablity (*p*-value)

Results (in S+)

```
        Welch Modified Two-Sample t-Test

data:  x: VReasn with Sex = M , and y: VReasn with Sex = F
t = 1.7747, df = 17.726, p-value = 0.0931
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -1.166146  13.766146
sample estimates:
 mean of x mean of y
      79.8      73.5
```

Why “Welch”? — No assumption of equal variances.

Results (in S+)

```
        Welch Modified Two-Sample t-Test

data:  x: VReasn with Sex = M , and y: VReasn with Sex = F
t = 1.7747, df = 17.726, p-value = 0.0931
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -1.166146  13.766146
sample estimates:
 mean of x mean of y
      79.8      73.5
```

Why “Welch”? — No assumption of equal variances.

Results (in SPSS) (slightly diff. data)

Group Statistics

	Boys/Girls	N	Mean	Std. Dev.	Std. Err.
Verbal Reas.	Boys	10	79,8	7,96	2,52
	Girls	10	73,5	7,43	2,35

Independent Samples Test

	Levene's test		t-test for Equality of Means					95% C.I.	
	F	Sig.	t	df	Sig. (2-tail)	Mean Diff.	SE Diff.	Lower	Upper
Equal var.	,053	,821	1,83	18	,084	6,3	3,44	-,93	13,5
No Equal var.			1,83	17,9	,084	6,3	3,44	-,93	13,5

Results

	Levene's test		t-test for Equality of Means					95\% C.I.	
	F	Sig.	t	df	Sig. (2-tail)	Mean Diff.	SE Diff.	Lower	Upper
Equal var.	,053	,821	1,83	18	,084	6,3	3,44	-,93	13,5
No Equal var.			1,83	17,9	,084	6,3	3,44	-,93	13,5

Note that some tables give values for one-tailed tests (Moore & McCabe; Field gives both one-tailed and two-tailed probabilities). One-tailed values $\approx$ 1/2 those of SPSS.

`Degrees of Freedom` $\approx (n_1 - 1) + (n_2 - 1)$

`Equal Variance` vs. `No Equal Variance` usually unimportant — even when variances are very different.

How do you understand the 95% CI = (-0.9, 13.5)?

Effect Size

```
                    Group Statistics
            Boys/Girls   N    Mean   Std. Dev.    Std. Err.
Verbal Reas.    Boys    10    79,8     7,96         2,52
                Girls   10    73,5     7,43         2,35
```

COHEN'S d: $(m_1 - m_2)/sd_{\text{all}} = (79.8 - 73.5)/10.9 \approx 0.6\text{sd}$

Cohen's d: How many standard deviations apart are the two means?
See http://en.wikiversity.org/wiki/Cohen's d

For interpretable differences, effect size may be reported practically, e.g. in money spent, survival rates, time saved, etc.

Reporting Results

We tested whether boys and girls differ in verbal reasoning. We selected 10 healthy individuals of each group randomly, and obtained their scores on *<Named> Verbal Reasoning Assessment*. We identify the hypotheses:

$H_0 : \ \mu_m = \mu_f$ (male and female the same)
$H_a : \ \mu_m \neq \mu_f$ (male and female different)

Since σ is unknown for this test, we applied a two-sided t-test after checking that the distributions were roughly normal. We discuss one outlier below.

Although the difference was large (Cohen's d = 0.6 sd, see box-and-whiskers plots), it turned out not to reach significance, $t(18) = 1.83$, $p = 0.084$, so we did **not** reject the null hypothesis $p \leq 0.05$. We retain H_0.

Questions

1. Given the small sample, and low outlier in the higher group, we might confirm H_a by recalculating, eliminating this individual. Should we?
2. Have we reported effect size?

Reporting Results

We tested whether boys and girls differ in verbal reasoning. We selected 10 healthy individuals of each group randomly, and obtained their scores on *<Named> Verbal Reasoning Assessment*. We identify the hypotheses:

H_0 : $\mu_m = \mu_f$ (male and female the same)
H_a : $\mu_m \neq \mu_f$ (male and female different)

Since σ is unknown for this test, we applied a two-sided t-test after checking that the distributions were roughly normal. We discuss one outlier below.

Although the difference was large (Cohen's d = 0.6 sd, see box-and-whiskers plots), it turned out not to reach significance, $t(18) = 1.83$, $p = 0.084$, so we did **not** reject the null hypothesis $p \leq 0.05$. We retain H_0.

Questions

1. Given the small sample, and low outlier in the higher group, we might confirm H_a by recalculating, eliminating this individual. Should we?
2. Have we reported effect size?

Reporting Results — Guidelines

Guidelines next week!

One-Sided *t*-Test

If testing directional hypothesis, e.g., that boys are **better** than girls in 3-dim. thinking, then one can divide 2-tailed prob. obtained above by 2. (Since $0.09/2 < 0.05$, you could conclude immediately that the null hypothesis is rejected at the $p = 0.05$-level.)

But you can avoid even this level of calculation, by specifying the one-sided hypothesis in S+.

```
        Welch Modified Two-Sample t-Test

data:  x: VReasn with Sex = M , and y: VReasn with Sex = F
t = 1.7747, df = 17.726, p-value = 0.0466
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.1392472        NA
 ...
```

Single Sample *t*-Tests

Moore & McCabe introduce paired sample *t*-tests using single sample *t*-tests (not focus here, but useful below)

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \qquad\qquad z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

where $dF = n - 1$. Recall that t increases in magnitude with large diff. in means, or with small standard deviations.

Use e.g. to test whether μ has a particular value, μ_0.

$$H_0 : \mu = \mu_0 \qquad \text{and} \qquad H_a : \mu \neq \mu_0$$

Then $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$, where in general, scores of large magnitude (positive or negative) indicate large differences (reason to reject H_0).

Single Sample t-Tests

Example: Claim that test has been developed to determine "EQ" (Emotional IQ). Test shows that $\mu = 90$ (in general population), no info on σ. We test:

$$H_0 : \mu = 90$$
$$H_a : \mu \neq 90$$

Measure 9 randomly chosen Groningers ($df = n - 1 = 8$). Result: $\bar{x} = 87.2, s = 5$ (Could the restriction to Groningen be claimed to bias results? ;-)

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{87.2 - 90}{5/\sqrt{9}} = \frac{-2.8}{1.6} = -1.75$$

Two-tailed chance of $t = -1.75$ ($df = 8$, Field, Tbl A.2) $\gg 0.05$, Cohen's $d = -2.8/5 = -0.56$ (but note that we're interested mostly in the size, not the sign).

Interpreting Single Sample t-Test

$$H_0 : \mu_{\mathrm{EQ}} = 90$$
$$H_a : \mu_{\mathrm{EQ}} \neq 90$$

The value -1.75 falls within the central 90% of the distribution.

Result: insufficient reason for rejection. Retain H_0.

Discussion: The small sample gives insufficient reason to reject the claim (at $p = 0.05$ level of significance) that the test has a mean of 90.

Single Sample t-Tests

Example 2: Confidence Intervals

As above, check system for "EQ". Measure $\bar{x}$ and derive 95% confidence interval.

1. Measure 9 randomly chosen Groningers ($df = n - 1 = 8$). Result: $\bar{x} = 87.2, s = 5$
2. Find (in table) t^* such that $P(|t(8)| > t^*) = 0.05$, which means that $P(t(8) > t^*) = 0.025$. Result: $t^* = 2.3$
3. Derive 95%-Confidence Interval

$$= \quad \bar{x} \pm t^*(s/\sqrt{n})$$

Practice in SPSS Lab 2!

Calculating Confidence Intervals with t

$P(t(8) > 2.3) = 0.025$ The t-based 95%-confidence interval is

$$\begin{aligned} &= \bar{x} \pm t^*(s/\sqrt{n}) \\ & 87.2 \pm 2.3(5/\sqrt{9}) \\ & 87.2 \pm 3.7 \quad = \quad (83.5, 90.9) \end{aligned}$$

Subject to same qualifications as other t-tests.

- Sensitive to skewness and outliers (as is mean!) if $n < 40$. Look at data!
- Only use when distribution approximately normal when $n < 15$. Look at Normal-Quantile Plot.

Paired t-Tests

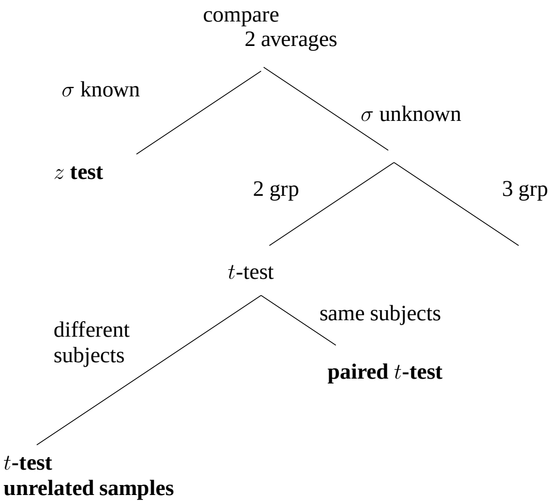
More powerful application of t-test possible if data comes in pairs. Then pairwise comparison of data points possible (instead of just mean).

Conceptually, this is very clever. All the sources of individual variation are effectively cancelled by examining data in pairs!

Naturally, many sources remain, e.g. measurement error, variation within a given individual.

Next week!

z- vs. t-tests, including paired data



Nonparametric alternatives to *t*-tests

If distribution nonnormal, recommended alternative is the MANN-WHITNEY U TEST (treated later in this course).

The Mann-Whitney test is equivalent to the Wilcoxon Rank Sum test.

Some journals in aphasiology require the use of the nonparametric test because they do not want results to depend on the normality of the distribution.

t-Tests: Summary

Simple *t* statistic:

$$t = \frac{m_1 - m_2}{s/\sqrt{n}}$$

- for numeric data, compares means of two groups, determines whether difference is significant
- population statistics (σ) unnecessary
  of course, sample statistics need
- three applications:
  - independent samples: compares two means
  - single sample (e.g., to estimate mean, or check hypothesis about mean)
  - paired: compares pairs of values
    example: two measurements on each of 20 patients

t-Tests: summary

Assumptions with all *t*-tests

1. Exactly two samples **unrelated measurements**
2. Distribution roughly normal if $n < 15$
   No large skew or outliers if $n < 40$

Additionally

1. Pay attention to whether variance is equal (reported in SPSS automatically via Levene's test)
2. Report effect size using Cohen's $d = (m_1 - m_2)/s$

Nonparametric fallbacks:

- Independent samples $\rightarrow$ Mann-Whitney
- Paired *t*-test $\rightarrow$ Wilcoxon signed rank test

Multiple Tests

Applying multiple tests risks finding apparent significance through sheer chance.
Example: Suppose you run three tests, always seeking a result significant at 0.05. The chance of finding this in one of the three is Bonferroni's **family-wise α-level**

$$\begin{aligned}
\alpha_{FW} &= 1 - (1 - \alpha)^n \\
&= 1 - (1 - .05)^3 \\
&= 1 - (.95)^3 \\
&= 1 - .857 = 0.143
\end{aligned}$$

To guarantee a family-wise alpha of 0.05, divide this by number of tests
Example: $0.05/3 = 0.017$ (set α at 0.1)

ANALYSIS OF VARIANCE is better in these cases (topic later).

Effect Size and Sample Size

Statistical significance obtains when an effect is unlikely to have arisen by chance. Very small differences may be significant when samples are large, i.e., these small differences are (probably) not due to chance.

As we saw in discussion of z, a difference of two standard errors or more ($z \geq 2$ or $z \leq -2$) is likely to arise in less than 5% of the time due to chance.

diff (in σ's)	n	p
0.01	40,000	0.05
0.1	400	0.05
0.25	64	0.05
0.37	30	0.05
0.5	16	0.05

We recommend samples of “about 30” because small effect sizes (under 0.4σ) are uninteresting, unless differences are important (e.g., health).

Watch out for effect size when reading research reports!

Proportions aka percentages

The relative frequency of nominal (categorical) data is sometimes an issue. We concentrate on the case where there are exactly two categories (buy/not-buy, click-through, not-, ... etc.)

Example: percentage of women customers for an e-business site. Sex is male vs. female, therefore categorical data.

Question: is the percentage of women significantly greater at one sort of site (e.g., films) as opposed to another (e.g., music)?

Two Approaches

1. Proportions may be viewed as numerical data. Use *t*-test.
   See `http://home.clara.net/sisa/`
2. Others, e.g. Moore & McCabe (5.1): *t*-tests do **not** apply since proportions reflect count data.
   Field is quiet on this.

t-Test on Proportions

If we are willing to regard counts as numerical scores, then we have a reason to use the *t*-test.

If the counts come from a yes-no (BINOMIAL) distribution, then distribution is roughly normal.
Prerequisite: Minimally 10 in each of the two categories.

We illustrate this technique on the question of whether there is a difference in the proportion (chance) of inflection errors made by Frisian children depending on whether they grow up in an exclusively Frisian setting or a mixed Frisian-Dutch setting.

This is a **two-sample** application.

t-test for Proportions in Two Samples

Nynke van den Bergh studies children acquiring Frisian. There are two groups:

- children who hear only Frisian at home and in child-care settings
- children who hear Frisian at home and Dutch in child-care settings.

The question is whether the mixed setting will lead to more interference errors—errors whether the child uses a Dutch pattern instead of a Frisian one.

Van den Bergh has studied patterns of the type:

Produced	Target	Translation
Gean mei boartsie	gean mei boartsj<u>en</u>	'go play along'
Ik kin swimmen	ik kin swimm<u>e</u>	'I can swim'

Frisian/Dutch Interference Hypothesis

Van den Bergh's null hypothesis is of course, that there is **no** difference in the proportion of incorrect inflections in the two populations (of expressions of inflection among children from the purely Frisian environment on the one hand as opposed to the children from the mixed environment on the other). Her alternative hypothesis is that the children from the mixed environment show more errors due to interference.

The hypothesis is therefore **one-sided**:

H_0 $p_F = p_M$, where p_F is the error percentage of children in the purely Frisian environment, p_M the error percentage of children in mixed environment.

H_a $p_F < p_M$

Frisian/Dutch Interference Data

Van den Bergh's data for kids 5 years, 11 months old:

Setting	**Correct**	**Incorrect**
Pure Frisian	85 (97.7%)	2 (2.3%)
Mixed	167 (89.8%)	19 (10.2%)

We wish to assume that there is **no** difference in the proportions in the two populations (the population of kids' expressions in the pure setting and their population in the mixed setting), and ask how likely these samples are given that assumption.

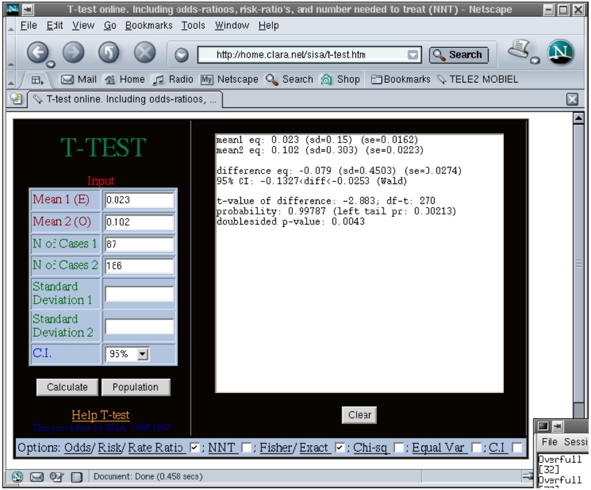
Frisian/Dutch Interference

We can test whether two proportions are significantly different at several online web-sites for statistics, e.g.

`http://home.clara.net/sisa/t-test.htm` Sisa

We only need to input the proportions, 0.023 error rate vs. 0.102 error, and the total number of elements, 87 and 186, respectively.

Sisa Calculations



Sisa Results

For this period, the Sisa web site calculated the following:

```
mean1 eq: 0.023 (sd=0.15) (se=0.0162)
mean2 eq: 0.102 (sd=0.303) (se=0.0223)
...
t-value of difference: -2.883; df-t: 270
probability: 0.99787 (left tail pr: 0.00213)
doublesided p-value: 0.0043
```

We're interested in the one-sided p value: whether the children in purely Frisian environments make **fewer** mistakes than those in mixed environments.

Indeed they make significantly fewer mistakes at this period ($p < 0.01$).

Two Samples Proportions

The data is usually nominal (categorical). **Example:** percentage of women customers for e-business sites. Sex is male vs. female, therefore categorical data. Question in that situation might be: does a significantly greater percentage of women visit one site as opposed to another.

View proportions as numerical data, recommending *t*-test
See `http://home.clara.net/sisa/`
Remember that minimally ten in each category must be present!

Next Topic

Paired t-tests, reporting results