Statistiek I

t-tests

John Nerbonne

CLCG, Rijks*universiteit* Groningen

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

university of groningen

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.

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
  - paired data: compares pairs of values; example: two measurements on each of 20 patients
  - single sample: (estimate mean)
- population statistics (μ, σ) unnecessary
  of course, sample statistics need
- appropriate with numeric data "normally distributed"
  see Mann-Whitney U-Test, Wilcoxon rank-sum test for non-parametric fall-backs

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 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). We have to check on $P(t(\text{df}))$, where dF is the degrees of freedom, normally $n - 1$.

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 M&M, Tabel E. 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.

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 dF $=$ Min$\{(n_1 - 1), (n_2 - 1)\}$ (using a smaller number makes showing significance harder, and therefore more reliable). **Notate bene:** S+ (and other statistical packages) often deviate from this conservative recommendation, using (something close to) dF $= (n_1 - 1) + (n_2 - 1) = n - 2$ (legitimate).
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

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

If three or more samples, use ANOVA (later in course).
If distribution unsuitable, using Mann-Whitney (later in course).

Independent Sample *t*-Test: Example

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

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

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

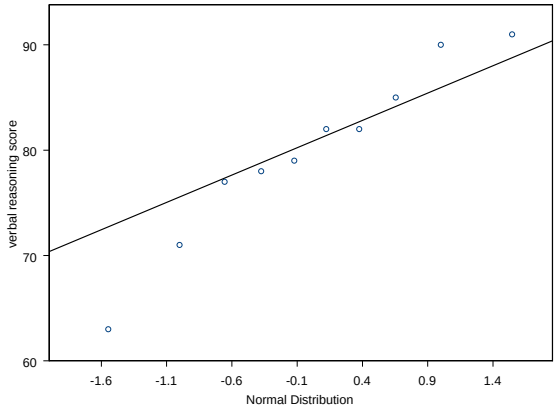
Hypotheses: $H_0 : \mu_m = \mu_f$
$H_a : \mu_m \neq \mu_f$

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

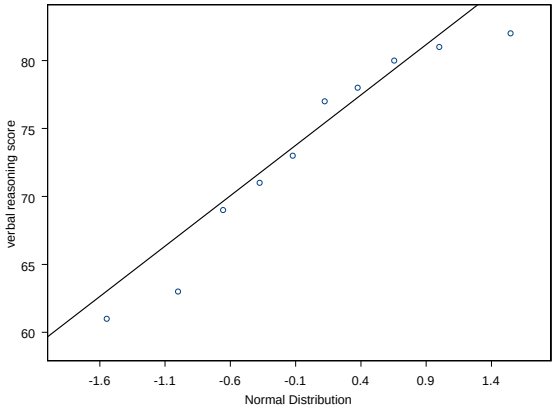
Independent Sample *t*-Test: Normality Test

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

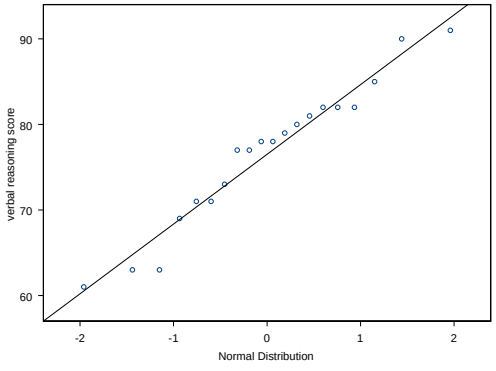
Boys



Girls

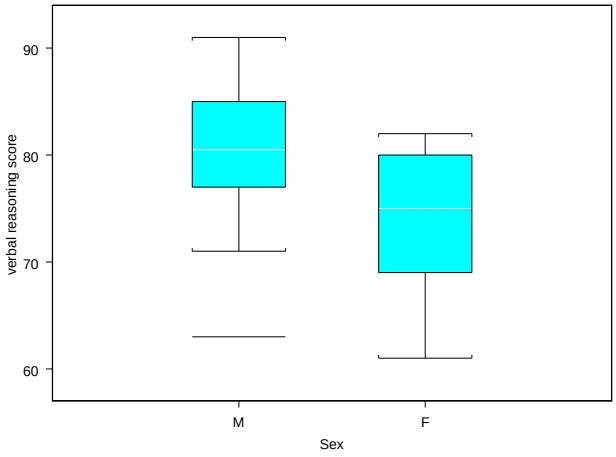


Normal Quantile Plots



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 technique: check KOLMOGOROV-SMIRNOV GOODNESS OF FIT (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 → Two Samples → *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

```
        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
```

Note that Table E in M&M, p.634 is for one-tailed test. Thus values $\approx$ 1/2 those of S+.
`Degrees of Freedom` $\approx (n_1 - 1) + (n_2 - 1)$ (less conservative than book)

university of groningen

Interpreting Results

We tested whether boys and girls differ in verbal reasoning ability. 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 verbal reasoning test, we applied a two-sided t-test after checking that the distributions were roughly normal. We discuss one outlier below.
The samples showed a 6-pt difference in average scores, yielding $p = 0.09$. Thus we did **not** reject the null hypothesis at the level $p \leq 0.05$. We retain H_0.
Discussion: Given the small sample, and low outlier in the higher scoring group, we might confirm the hypothesis in a larger study or by recalculating, eliminating this individual. Should we?

Reporting Results

1. State the issue in terms of the populations (not merely the samples). Formulate H_0 and H_a.
2. State how your hypothesis is to be tested, how samples were obtained, what procedures (test materials) were used to obtain measurements.
3. Identify the statistical test to be used, why.
4. Illustrate your research question graphically, if possible. —For example, with box plots, as above.
5. Present the results of the study on the sample, their significance level.
6. State conclusions about the hypotheses.
7. Discuss and interpret your results.

Practice this in laboratory exercises!

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 using single sample (not focus here, but useful below)
t-statistic:

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

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.
$H_0 : \mu = \mu_0$ and $H_a : \mu \neq \mu_0$
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 want to 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 ?

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

One-sided chance of $t = -1.75$ (M&M Table E, p.634, $df = 8$) is 0.05.

Interpreting Single Sample *t*-Test

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

Nota bene M&M, Table E gives *one-tailed chance* $P(t(8) < -1.86) = 0.05)$ Since this is a two-sided test, we note $P(|t(8)| > 1.86) = 0.10)$. 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

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

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).
Then we examine the difference between scores. We can check the hypothesis that the scores are from the same populations by check whether the average differences tends to be zero.
$H_0 : \mu_{(x_i - y_i)} = 0$ and $H_a : \mu_{(x_i - y_i)} \neq 0$
This can be regarded as a **single sample** of differences. We get the calculations (in some statistics packages), by calculating a set of differences, then applying a single-sample t-test to check the hypothesis that the mean is zero.
Some packages have built-in paired t-tests, e.g., SPSS.

Paired *t*-Test: Example

Example: Suppose you suspect that the test for verbal reasoning which prove boys better is flawed. You find a second. Now you look at results of *both tests* on 15 subjects.

subj	test1	test2	δ
1	7.6	7.3	0.3
2	10.2	9.1	1.1
⋮	⋮	⋮	⋮
15	8.4	7.5	0.9
m	6.85	6.23	0.6
sd	3.2	2.94	

Suppose we applied a *t*-test for independent samples, with H_0, H_a:

$$H_0 : \mu_{\text{test1}} = \mu_{\text{test2}}$$
$$H_a : \mu_{\text{test1}} \neq \mu_{\text{test2}}$$

Then $t = 0.546, p = 0.0546$, retain H_0 at $\alpha = 0.05$. No proof that tests are different.

Paired *t*-Tests

Paired *t*-test appropriate when there are two related samples (e.g., two measurements x, y of the same people). This is called PAIRED DATA. Then we can examine differences between the first and second elements of the pairs: $H_0 : \mu_{(x_i - y_i)} = 0$ and $H_a : \mu_{(x_i - y_i)} \neq 0$

1. This is **inappropriate** when scores differ in scale, e.g., when one score is %-percentage, and other in $[0, 600]$. Consider then REGRESSION.
2. In small samples, both sets of scores must be roughly normally distributed.

Paired *t*-Test: Application & Results

We assume two sets of data, X, Y. Let δ_i be $x_i - y_i$.

$$t = \frac{\bar{\delta}}{s_\delta / \sqrt{n}}$$

For data in last slide (p.24), $s_\delta \approx 0.4$

$$\begin{aligned} t &= \frac{0.6}{0.4/\sqrt{15}} \\ &= \frac{0.6}{0.4/3.9} \\ &= \frac{0.6}{0.1} \\ &= 6 \end{aligned}$$

For $dF = 14$, $p \approx 0.000$ (Table, p.643)

Paired *t*-Test: Interpretation

We reject the null hypothesis at $p \leq 0.001$-level. The tests do not yield the same results.
Discussion: The second test yields slightly, but consistently higher scores.
Note: We were **not** able to prove the two tests different using the *t* test for independent samples. But since we have two sets of test results for the same people, we can examine the data in pairs.
General lesson: More sophisticated statistics allow more sensitivity to data. Using the paired data, we could prove that the tests were different.
Using only unpaired data commits *error of second sort*: null hypothesis false, but not rejected

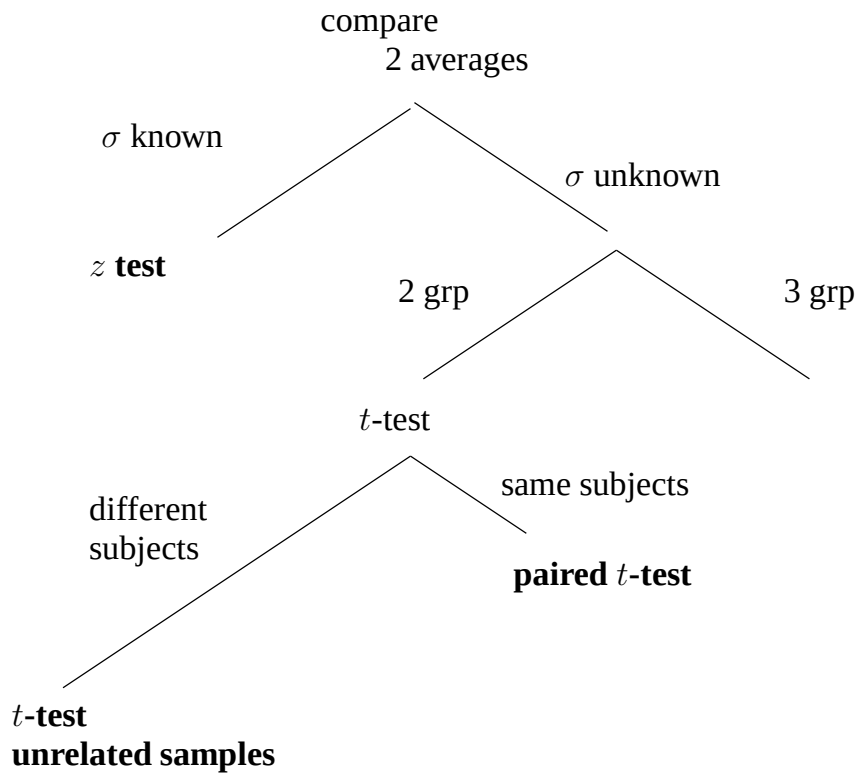
Nonparametric Alternatives to paired *t*-Tests

If distribution nonnormal, recommended alternative is the WILCOXON SIGN RANK TEST (treated later in this course).
If distribution nonnormal and asymmetric, we can note the sign of the differences and use those in the SIGN TEST (also later in the course). This is an application of the BINOMIAL DISTRIBUTION: note chance of sign of δ, either all positive or 14 positive, one negative:

$$\left(\binom{15}{1} + \binom{15}{0}\right)(0.5)^{15} \approx 0.00045$$

Sign test (M&M, p.409) — sometimes necessary when small samples too skewed for t test

z- vs. *t*-tests, including paired data



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 (μ, 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

university of groningen

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 $n \geq 15$

Nonparametric fallbacks:

- Independent samples $\rightarrow$ Mann-Whitney
- Paired t-test $\rightarrow$ Wilcoxon rank-sum 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: $0.05/3 = 0.17$ (set α at 0.1)

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

university of groningen

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

The recommendation for sample sizes of "about 30" stems from the the idea that small effect sizes (under 0.3σ) are uninteresting, at least until you are quite advanced.

Next Topic

Analysing Proportions