Lecture 1: t tests and CLT

http://www.stats.ox.ac.uk/∼winkel/phs.html

Dr Matthias Winkel

Outline

I. z test for unknown population mean - review

II. Limitations of the z test

III. t test for unknown population mean

IV. t test for comparing two matched samples

V. t test for comparing two independent samples

VI. Non-Normal data and the Central Limit Theorem

I. z test for unknown population mean

The Achenbach Child Behaviour Checklist is designed so that scores from normal children are Normal with mean $\mu_0 = 50$ and standard deviation $\sigma = 10$, $N(50, 10^2)$.

We are given a sample of $n = 5$ children under stress with an average score of $\bar{X} = 56.0$.

Question: Is there evidence that children under stress show an abnormal behaviour?

Test hypotheses and test statistic

Null hypothesis: $H_0 : \mu = \mu_0$

Research hypothesis: $H_1 : \mu \neq \mu_0$ (two-sided).

Level of significance: $\alpha = 5\%$.

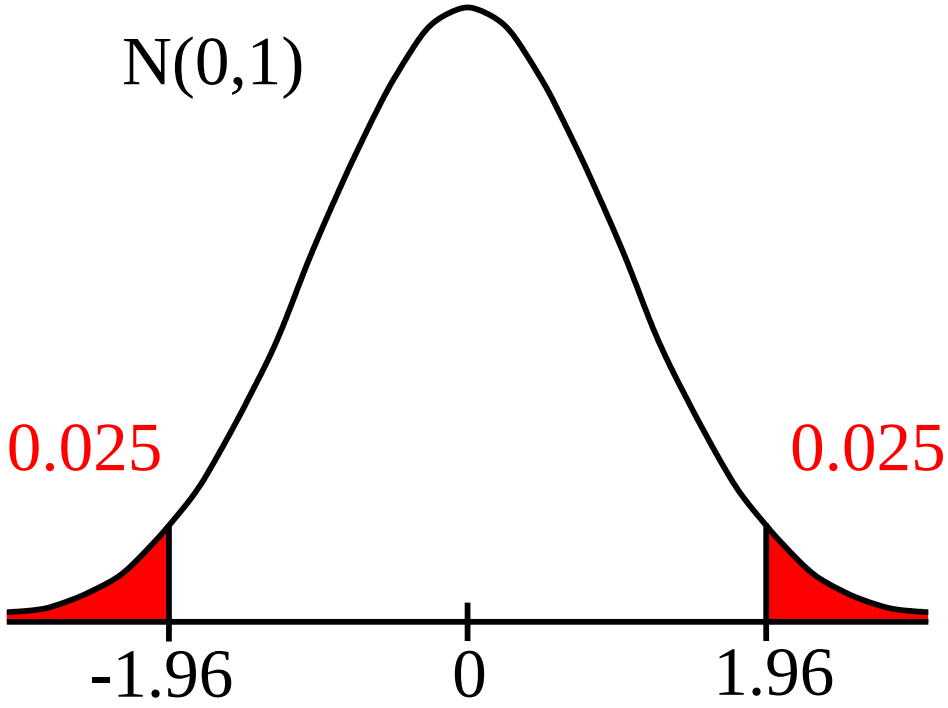
Under the Null hypothesis

$$X_1, \ldots, X_n \sim N(\mu_0, \sigma^2)$$

$$\Rightarrow \quad Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

The **data** given yield $z = \dfrac{\bar{x} - \mu_0}{\sigma\sqrt{n}} = \dfrac{56 - 50}{10\sqrt{5}} = 1.34.$

Critical region and conclusion



Test procedure, based on z table $P(Z > 1.96) = 0.025$:

If $|z| > 1.96$ then reject H_0, accept H_1.

If $|z| \leq 1.96$ then accept H_0, reject H_1.

Conclusion: Since $|z| = 1.34 \leq 1.96$, we cannot reject H_0, i.e. there is no significant evidence of abnormal behaviour.

II. Limitations of the (exact) z test

1. **Standard deviation** must be **known** (under the null hypothesis).

If not, estimate standard deviation and perform t test.

2. Data so far had to come from a **Normal population**.

If not, the Central Limit Theorem might allow us to still perform approximate z and t tests.

The rest of this lecture deals with these two issues.

III. t test for unknown population mean

The z test does NOT apply if σ is unknown, or more precisely, it is not exact even for large Normal populations.

If it was not known that $\sigma = 10$ for a population of normal children, we would estimate σ^2 by the sample variance

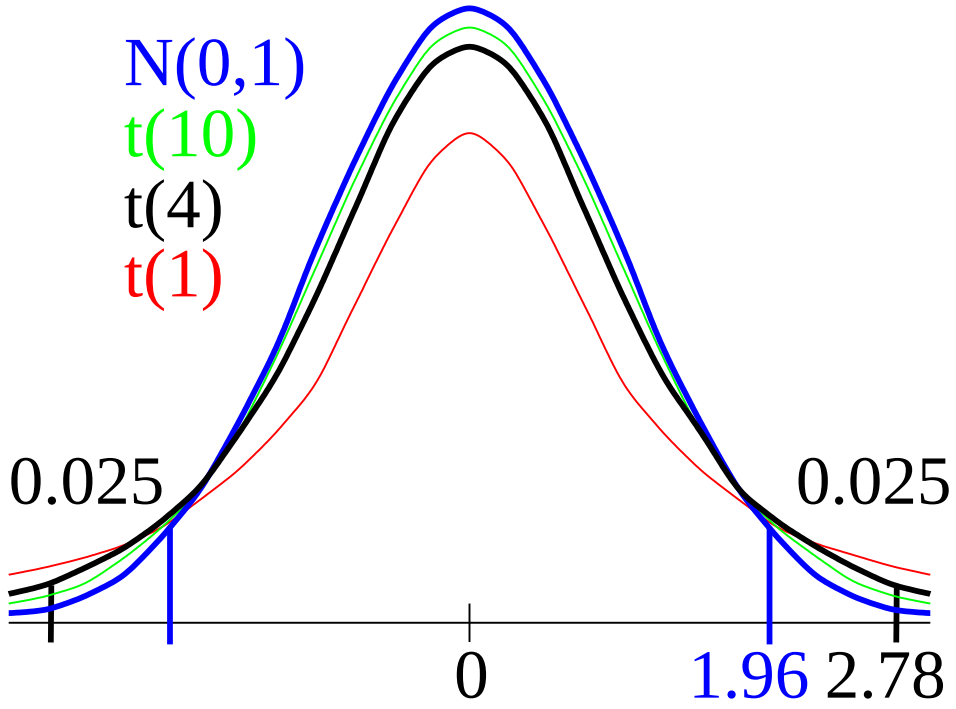
$$S^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \bar{X})^2.$$

We then replace σ in the test statistic by S:

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{\bar{X} - \mu_0}{\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2}/\sqrt{n}}.$$

$T \not\sim N(0,1)$ due to the X's in the denominator, $T \sim t_{n-1}$.

pdf's of t distributions



t distributions have **thicker tails** than the Normal distribution. Therefore the t **critical value is higher** than the z critical value.

The parameter: degrees of freedom (d.f.)

The statistic

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \qquad \left[\text{under } H_0 \text{ where } X_k \sim N(\mu, \sigma^2)\right]$$

has a parameter $n-1$ that measures how good or bad the estimate S of σ is:

If n is small (i.e. only **few observations**), the estimate is bad and T is **far from Normal**

If n is large (i.e. **many observations**), the estimate is good and T is **close to Normal**

Since $S^2/\sigma^2 \sim \chi^2_{n-1}$ produces the parameter $n-1$, it is called a degrees of freedom parameter as for the Chi-squared distribution.

The example step by step

Null hypothesis: $H_0 : \mu = 50, \quad [\sigma \text{ not specified}]$

Research hypothesis: $H_1 : \mu \neq 50$

Level of significance: $\alpha = 5\%$

Assumption: Normal observations

Data: $X_1, \ldots, X_5$ such that $\bar{X} = 56$, $S = 8.5$

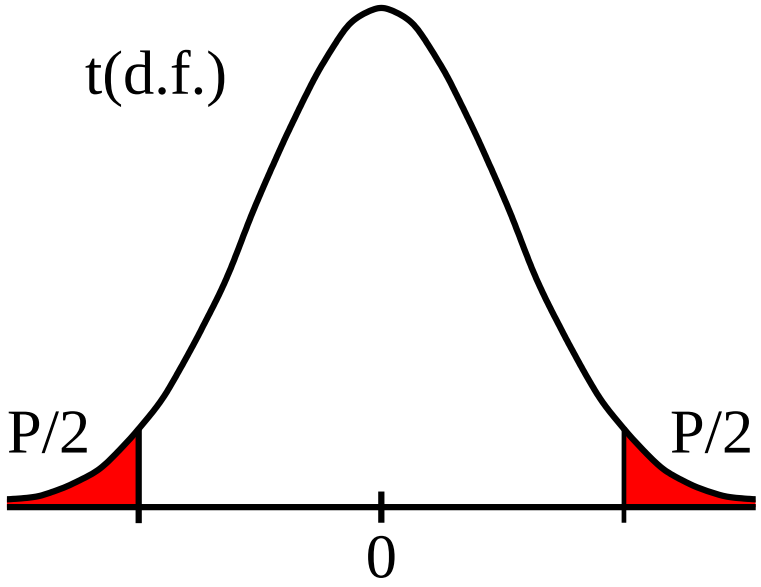
Under the Null hypothesis: $T = \dfrac{\bar{X} - 50}{S/\sqrt{5}} \sim t_4$

t **critical value** (from table) $2.78 > 1.58 = |T|$ implies (again): No evidence of abnormality for stressed children.

t table

The t table looks similar to the chi-squared table:

d.f.	P=0.10	P=0.05	P=0.01
1	6.31	12.71	63.7
2	2.92	4.30	9.93
3	2.35	3.18	5.84
4	2.13	2.78	4.60
5	2.02	2.57	4.03
⋮	⋮	⋮	⋮
∞	1.65	1.96	2.58



$P = 0.05$ corresponds to $\alpha = 5\%$ in the two-sided case.

$P = 0.10$ can be used for $\alpha = 5\%$ in the one-sided case.

In our example we applied: $P(t(4) \notin [-2.78, 2.78]) = 0.05$.

IV.-V. t tests for comparing two samples

Often tests are not between a known and an unknown population, but between **two unknown populations**.

Example: Two methods I and II are to be compared. Suppose data collected represent the quality of the method.

1. Methods I and II are applied to the same subjects/objects or ones of essentially identical type (matched pairs: IV.).
2. Methods I and II are applied to different subject/object groups. The group size may differ, but the distribution to groups must not be systematic in order that differences are due to the methods and not to the distribution process (independent design: V.).

IV. t test for comparing two matched samples

The PEFR (Peak Expiratory Flow Rate) is the maximum rate of airflow (litre/min) that can be achieved during a sudden forced expiration from a position of full inspiration (GPnotebook).

We are given two instruments for measuring PEFR, a Wright Peak Flow Meter (I) and a Mini Peak Flow Meter (II).

Question: Is there a bias between instruments I and II?

The data

10 subjects produced PEFRs on both instruments I and II:

Subject k	I	II	Difference D_k
1	490	525	-35
2	397	415	-18
3	512	508	4
4	401	444	-43
5	470	500	-30
6	415	460	-45
7	431	390	41
8	429	432	-3
9	420	420	0
10	421	443	-22

Sample mean $\bar{D} = -15.1$, sample std deviation $S = 26.3$

The test

Null hypothesis: $\mu_I = \mu_{II}$

Research hypothesis: $\mu_I \neq \mu_{II}$

Significance level: 5%

Data: relevant are only the differences $D_1, \ldots, D_{10}$

Assumption: $D_1, \ldots, D_{10}$ are from a Normal population

Test statistic: $T = \dfrac{\bar{D}}{S/\sqrt{10}} \sim t_9$ [one-sample test on D]

Critical value from t table: 2.26

Observed value: $T = -15.1/(26.3/\sqrt{10}) = -1.86$

Conclusion: no significant evidence of a bias ($|T| \leq 2.26$)

General remarks on the design

Matching should **always** be done, whenever possible, for various reasons.

1. It often doubles the data, since every subject provides two results. More data represents the populations better.
2. It eliminates variability between subjects (some can produce higher scores than others).
3. Differences may be Normally distributed even when individual scores are not.

However, matching is **not always possible** as we shall see.

A lot of care is required for conclusions to be justified.

1. For each subject, a coin toss determined the **order** (I,II) or (II,I), to spread exhaustion and familiarisation effects
2. For each subject, only the **second score** on each instrument was recorded, to rule out usage problem effects
3. Subjects should have been **chosen at random** from the population, to rule out effects specific to subpopulations.
4. One should not rely on single instruments but expect **variability between individual instruments** of each type.

Most of these design remarks can be summarised as 'randomisation'. Randomisation reduces all systematic errors.

V. t test for comparing two independent samples

In the PEFR instrument comparison, assume that every subject only produces a score with **one** of the instruments, say, the first five I, the last five II.

This is indeed a waste of resources here, but if our subjects were e.g. oranges and the instruments squeeze them to orange juice, we would not be able to reuse any orange for another instrument and just record the amount of juice (in ml) for one of them.

Question: Is there a bias between instruments I and II?

The test

Null hypothesis: $\mu_I = \mu_{II}$

Research hypothesis: $\mu_I \neq \mu_{II}$

Significance level: 5%

Data: two samples $X_{I1}, \ldots, X_{I5}$ and $X_{II1}, \ldots, X_{II5}$

Assumption: each sample is from a Normal population

Test statistic: $T = \dfrac{\bar{X}_I - \bar{X}_{II}}{sd} \sim t_8$ [why? and what is sd?]

Critical value from t table: 2.31

Observed value: $T = 0.95$ [once we know what sd is]

Conclusion: no significant evidence of a bias ($|T| \leq 2.31$)

What is sd?

The denominator sd must be an estimate for the standard deviation of the numerator $\bar{X}_I - \bar{X}_{II}$. We know

$$Var(\bar{X}_I - \bar{X}_{II}) = Var(\bar{X}_I) + Var(\bar{X}_{II}) = \frac{\sigma_I^2}{n_I} + \frac{\sigma_{II}^2}{n_{II}}$$

and can estimate σ_I by S_I and σ_{II} by S_{II}, but we need

Assumption: $\sigma_I^2 = \sigma_{II}^2 =: \sigma^2$

to identify the distribution of T. If $n_I = n_{II}$, we are done.

If $n_I \neq n_{II}$, we use a 'fairer' weighted estimate of σ^2

$$S^2 = \frac{(n_I - 1)S_I^2 + (n_{II} - 1)S_{II}^2}{n_1 + n_2 - 2}$$

This indicates the degrees of freedom to be $n_1 + n_2 - 2$.

VI. Non-Normal data and the Central Limit Theorem

Central Limit Theorem: For a sample $X_1, \ldots, X_n$ from a population with mean μ and variance σ^2, approximately

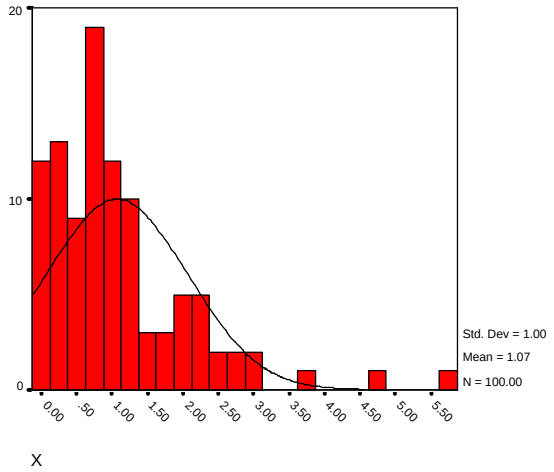
$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

The approximation becomes exact in the limit $n \to \infty$.

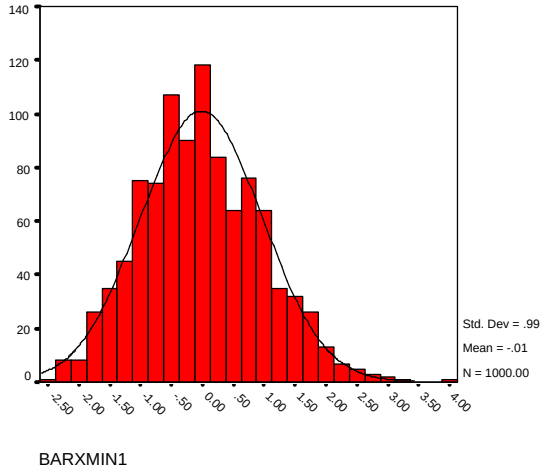
The approximation is bad for small sample sizes and must not be used. Some authors recommend sample sizes of at least 30 or 50 for practical use. The quality of approximation depends on the distribution. Skewed distributions e.g. take longer to converge than symmetric ones.

Illustration of the Central Limit Theorem

The lifetime of light bulbs is often assumed to be exponentially distributed (with mean 1 and variance 1, say). We have a sample $X_1, \ldots, X_n$ of size $n = 100$.



The Central Limit Theorem tells us that $\sqrt{n}(\bar{X} - 1)$ is approximately standard Normal.



Previous examples of applications of the CLT

1. Normal approximation of the Binomial distribution
2. Normal approximation of the Poisson distribution
3. One sample test for a proportion p (application of 1.)
4. Two sample test for a difference between proportions
5. Chi-squared Goodness-of-Fit Test and Association Test

It is **not** part of this course to work out exactly **why** these apply the CLT, **but** you should understand e.g. that the z test for a proportion p is based on discrete data and hence the continuous Normal distribution arises by approximation, the z test is hence not exact and you require large samples for the approximation to work.

Tests for unknown population means

1. The z test for one or two non-Normal samples can be applied in an approximate way, in practice usually for $n \geq 50$.

2. Instead of the t test, one can use the z test in the case of Normal samples of size $n \geq 50$ since the t distribution approaches the z distribution as the degrees of freedom increase.

3. It is NOT recommended to combine the two approximations and use an approximate t test when the data contain clear signs of non-Normality, except with sample sizes $n \geq 100$. There are weaker but more robust non-parametric tests for the small sample case.

Exam Paper Questions

Year	Human Sci TT	Psych MT	Psych HT	Psych TT
2001	(5)		4 (6)	(5)
2000	(7)	(6) (7)	6	(7)
1999	(5)	8	(9)	(5)
1998	(5)		4 (7)	(5)
1997	(6)		(6)	(6)

The question numbers in parentheses refer to comparison exercises between parametric and nonparametric test, so only the parametric part can be done at this stage.