Tests for Comparing Means (t tests and their non-parametric cousin)

Vera Demberg

Saarland University

Nov 23, 2021

Recap from last week

Last week, we were looking at tests for **categorial** data, i.e. situations where both DV and IV were nominal scale
(remember "categorial" = nominal scale)

Example:

dependent / response variable: death sentence (yes / no)

independent / predictor variable: white vs. nonwhite

Recap from last week

Last week, we were looking at tests for **categorial** data, i.e. situations where both DV and IV were nominal scale
(remember "categorial" = nominal scale)

Example:

dependent / response variable: death sentence (yes / no)

independent / predictor variable: white vs. nonwhite

But what if our dependent variable is **interval or ratio scale**?

Introduction

But what if our dependent variable is **interval or ratio scale**?
(You will probably encounter this situation very often.)

Introduction

But what if our dependent variable is **interval or ratio scale**?
(You will probably encounter this situation very often.)

- Are ratings for my user interface higher if I include new functionality?
- Is working memory capacity reduced by listening to music (relative to not listening to music)?
- Does a new drug increase or decrease blood pressure?

In all these situations, our outcome variable is an interval or ratio scale variable; and our predictor is a binary "grouping" variable. In other words, we want to compare the means of the two groups.

Closer to home...

Example

You built a dialog system that can converse with humans, and want to compare it against the standard system, for which it is well known that the average task completion time is 320 seconds.
You ask some people to interact with your system using the typical set of tasks, and record the completion times.
Now you wonder whether people can complete the task significantly faster when using your new dialog system.

We need to compare the mean completion time for the new system to the well-known and accepted time.

Introduction

Second Example

You built a dialog system that can converse with humans. There are however two alternative settings with the system, and you want to test whether it makes a difference to use the one or the other.
You ask some people to interact with the one or the other version of the system, and record the time it takes to complete a certain task.

We need to compare the mean completion time for each of the settings.

Introduction

Second Example

You built a dialog system that can converse with humans. There are however two alternative settings with the system, and you want to test whether it makes a difference to use the one or the other.
You ask some people to interact with the one or the other version of the system, and record the time it takes to complete a certain task.

We need to compare the mean completion time for each of the settings.

For all of these questions, we can use the t-test (and we'll look at some variants for how to use it in different situations), and see some tests similar to the t-test which we can use when certain assumptions of the t-test are violated.

Today: t-test and friends

1. One-sample z-test
2. One-sample t-test
3. Independent samples t-test
4. Independent samples t-test (Welch test)
5. Paired-samples t-test
6. Effect size
7. Checking normality of a sample
8. Wilcoxon test (for non-normal data)

Table of Contents

Simplest case

Simplest case:

- we somehow magically know *mean* μ and *sd* σ of the general population
- we have drawn a sample with some specific property
- we want to test whether this sample is plausible to come from the general population

$\rightarrow$ in this case, we can just use our normal distribution, i.e. the z-test that we already saw a few weeks ago.

Remember the central limit theorem?

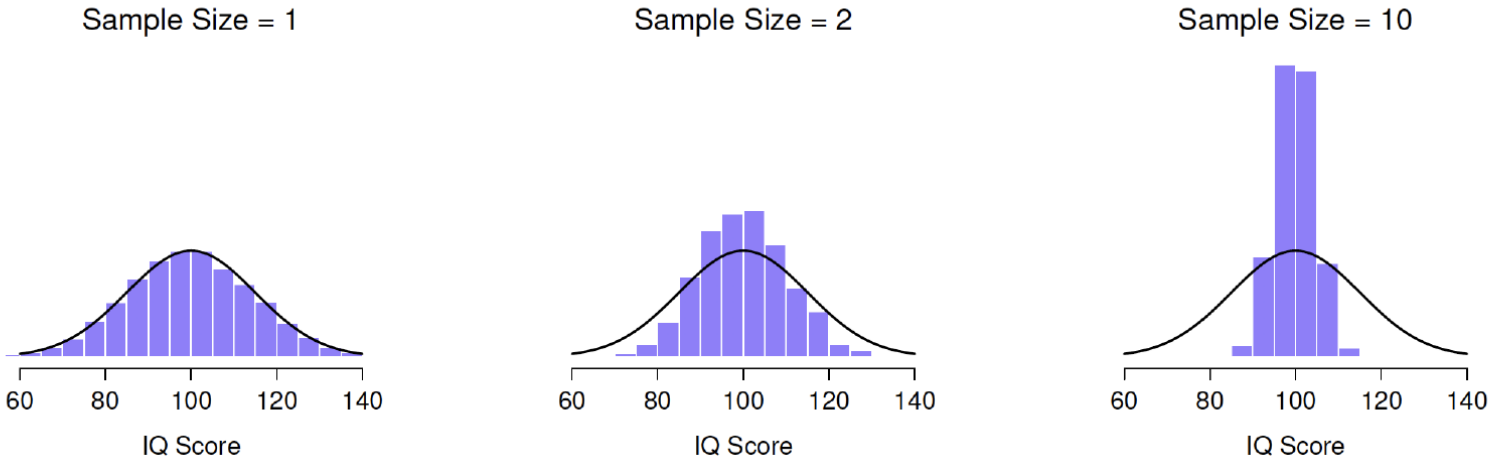
Central Limit Theorem

Among other things, the central limit theorem tells us that if the population distribution has mean μ and standard deviation σ, then the sampling distribution of the mean also has mean μ, and the standard error of the mean SEM is:

$$\mathrm{SEM} = \frac{\sigma}{\sqrt{N}}$$

Remember the central limit theorem?

We get less variance with larger samples.
The sampling distribution is always normal.



For each plot 10,000 samples of IQ data were generated. Then calculate the mean IQ observed within each of these data sets. The histograms in these plots show the distribution of these means (i.e., the sampling distribution of the mean).

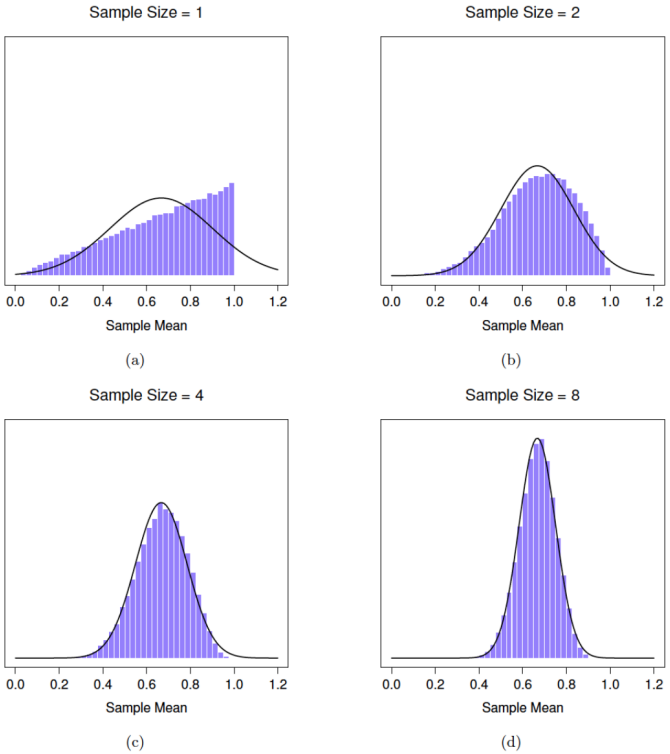
Central limit theorem

This allows us to use the tests we will be talking about if

- the sample is large enough that the sampling distribution is normal (argument via central limit theorem).

GOOD NEWS!

As long as your sample size isn't tiny, the sampling distribution of the mean will be approximately normal, no matter what your population distribution looks like!



- or the original distribution is normal

Example

Achenbach child behaviour checklist, which measures behaviour problems.

population mean: it is known that the population mean based on many thousands of individuals is 50.

population sd: it is known that the population standard deviation is 10.

sample mean: a sample of 15 children who was recently hospitalized with a severe illness, scored 56.0 on average

Do these children score significantly higher than normal children on the behavioural problems measure? Or is this just coincidence, considering that our sample is quite small?

second Example

(based on an actual study by Williamson 2008)

Achenbach Youth Self-Report Inventory checklist on anxiety levels.

population mean: it is known that the population mean based on many thousands of individuals is 50.

population sd: it is known that the population standard deviation is 10.

sample mean: a sample of 166 children with at least one depressed parent, these children showed a mean of 55.71 on the test.

sample sd: sample sd was 7.35

Do these children score significantly higher than average children on the anxiety level scale?

Calculating significance with sampling distributions

Study 1:

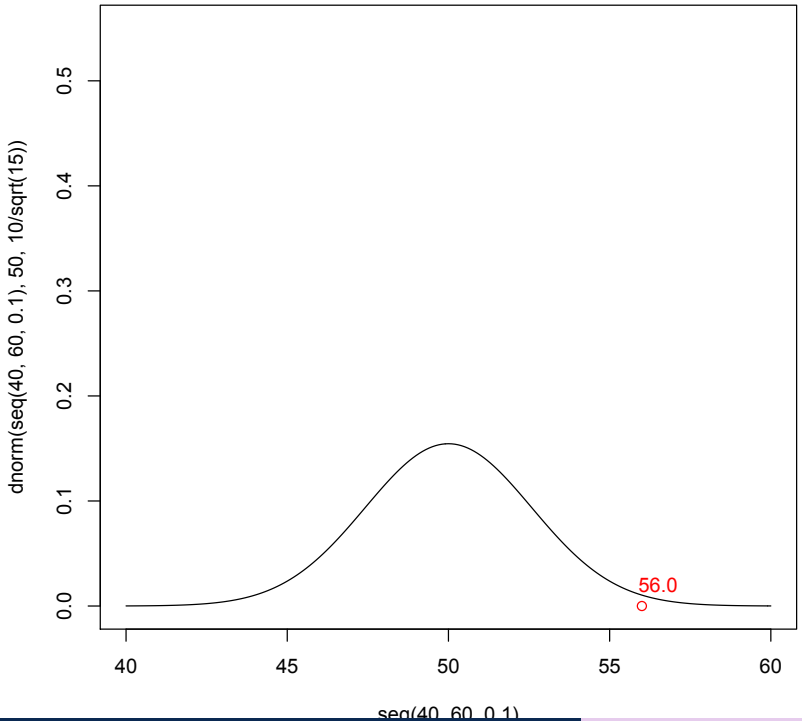
sample size: 15

SEM: $\frac{10}{\sqrt{15}}$

sample mean: 56.0

$z = \frac{56-50}{\frac{10}{\sqrt{15}}} = 2.32379$

p=0.01038



Study 2:

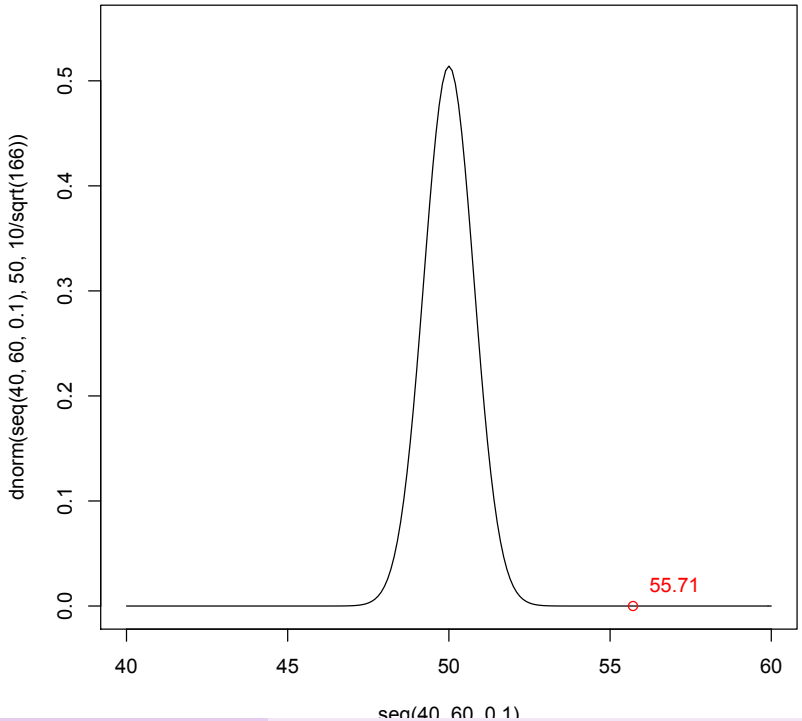
sample size: 166

SEM: $\frac{10}{\sqrt{166}}$

sample mean: 55.71

$z = \frac{55.71-50}{\frac{10}{\sqrt{166}}} = 7.35682$

p$<$ 0.0000001



Wait one more moment...

huh, why didn't we actually use the sample sd at all?

Achenbach Youth Self-Report Inventory checklist on anxiety levels.

population mean: it is known that the population mean based on many thousands of individuals is 50.

population sd: it is known that the population standard deviation is 10.

sample mean: a sample of 166 children with at least one depressed parent, these children showed a mean of 55.71 on the test.

sample sd: sample sd was 7.35

This is because we have the population standard deviation: for the purposes of our null hypothesis (testing whether these children are from the general population), **we assume that our sample sd is the same or smaller than the population sd**.
(this will be different for cases where we don't have the population sd.)

Assumptions of the z-test

Normality As usually described, the z-test assumes that the true population distribution is normal (or that the sampling distribution of the mean is normal). We can check this assumption if we feel worried about it.

Assumptions of the z-test

Normality As usually described, the z-test assumes that the true population distribution is normal (or that the sampling distribution of the mean is normal). We can check this assumption if we feel worried about it.

Independence The second assumption of the test is that the observations in your data set are not correlated with each other, or related to each other in some funny way. (e.g. you can't copy your data 10 times and say you got 10 times the sample size).

Assumptions of the z-test

Normality As usually described, the z-test assumes that the true population distribution is normal (or that the sampling distribution of the mean is normal). We can check this assumption if we feel worried about it.

Independence The second assumption of the test is that the observations in your data set are not correlated with each other, or related to each other in some funny way. (e.g. you can't copy your data 10 times and say you got 10 times the sample size).

Known standard deviation The third assumption of the z-test is that the true standard deviation of the population is known to the researcher. **This assumption is almost always wrong.**

Table of Contents

If the standard deviation of the population is not known...

Usually, we don't know the standard deviation of the population.

$\rightarrow$ We then need to *adjust for the uncertainty* of what the sd is.

Student's t distribution

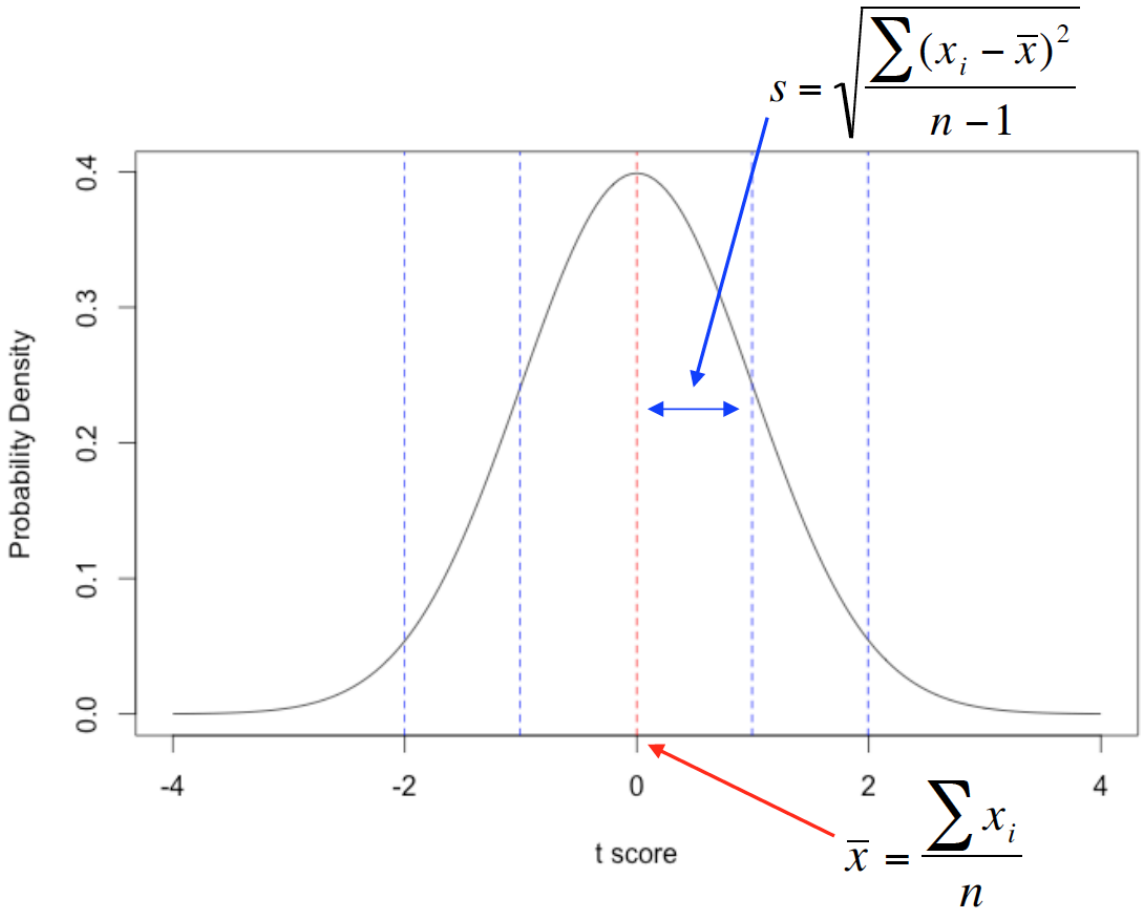
This problem luckily was solved by a William Sealy Gosset (Student, 1908).

The story behind why it's called "Student's t-test"

Gosset was working as a chemist for the Guinness brewery at the time (see J. F. Box, 1987). Because Guinness took a dim view of its employees publishing statistical analysis (apparently they felt it was a trade secret), he published the work under the pseudonym "A Student", and to this day, the full name of the t-test is actually Student's t-test.

t distribution
when σ is unknown

William S. Gosset
aka "Student"



$$t = \frac{x - \bar{x}}{s}$$

x – *individual score*

$\bar{x}$ – *mean in a sample*

s – *standard deviation in a sample*

n – *sample size*

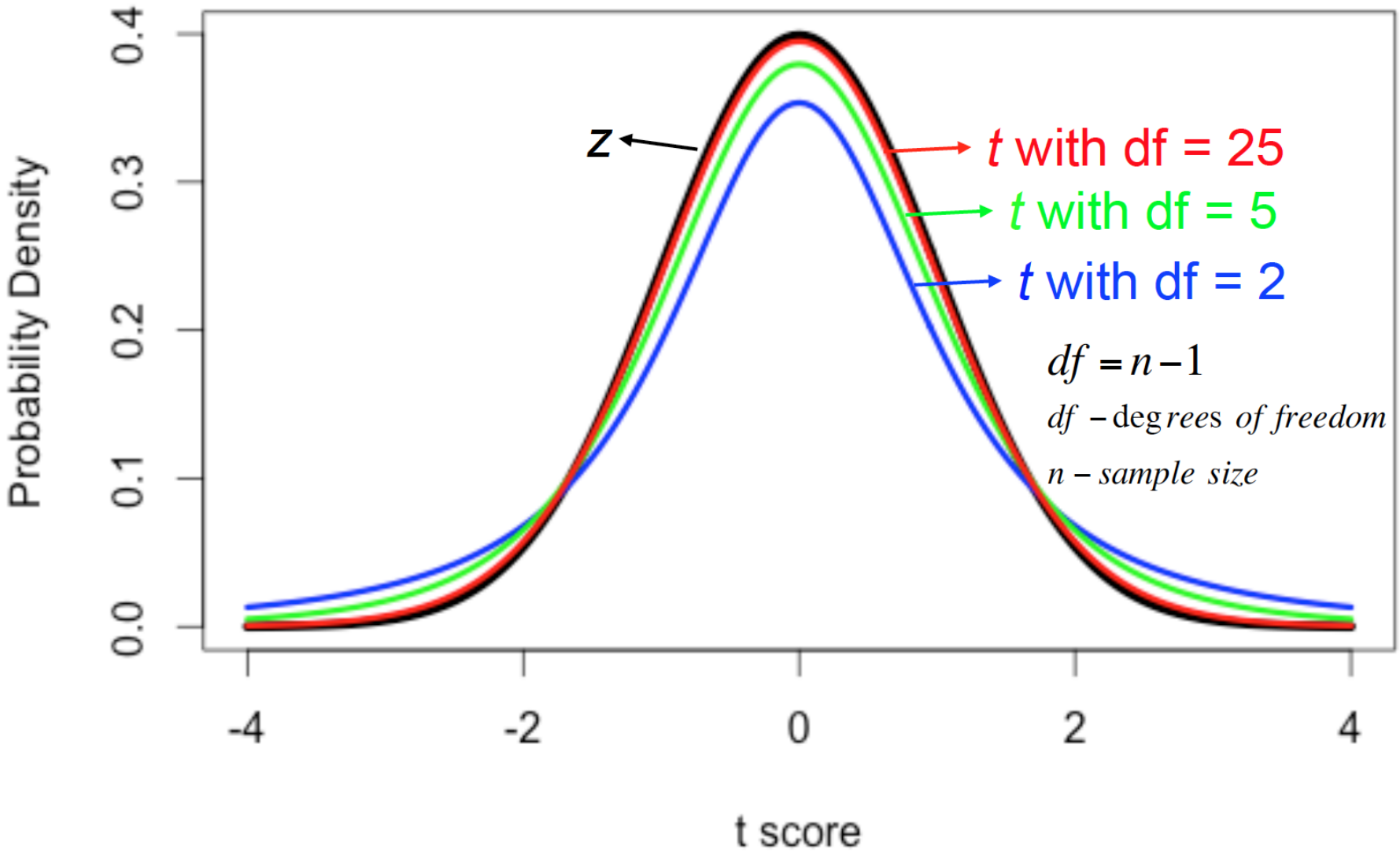
t value of a score represents how much that score is different from the mean in terms of standard deviations estimated from a sample

we can always estimate from a sample!!!

but.. the shape of the t distribution depends on the size of the sample

t distribution

shape and sample size



the t distribution

The t-distribution looks similar to the z distribution, but has heavier tails. This reflects the fact that we're uncertain about the standard deviation (i.e. that it's an estimate instead of the truly correct standard deviation).

As we have more degrees of freedom, the t distribution starts to look identical to the z distribution (and that's exactly how it should be – if we have 100 000 observations, we will have a very very good estimate of the population mean and population standard deviation).

The Student t Distribution

Description

Density, distribution function, quantile function and random generation for the t distribution with `df` degrees of freedom (and optional non-centrality parameter `ncp`).

Usage

```
dt(x, df, ncp, log = FALSE)
pt(q, df, ncp, lower.tail = TRUE, log.p = FALSE)
qt(p, df, ncp, lower.tail = TRUE, log.p = FALSE)
rt(n, df, ncp)
```

Arguments

`x, q`
: vector of quantiles.

`p`
: vector of probabilities.

`n`
: number of observations. If `length(n) > 1`, the length is taken to be the number required.

`df`
: degrees of freedom (> 0, maybe non-integer). `df = Inf` is allowed.

Calculating an example in R

Example from Navarro book

Grading of statistics introduction class, with grades generally being normally distributed, with population mean of 67.5.
The average grade of 20 psychology students from one year is 72.3, and the standard deviation from this sample is 9.52.
Were these students doing significantly better?

Calculating an example in R

population mean: 67.5
sample mean: 72.3
sample sd: 9.52
number of students: 20
We use the sample sd instead of the population sd.

$$t = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{N}}$$

```
> grades<-c(50, 60, 60, 64, 66, 66, 67, 69 ,70 ,74, 76, 76, 77, 79, 79, 79, 81, 82, 82, 89)
> t<-(mean(grades)-67.5)/(sd(grades)/sqrt(20))
> t
[1] 2.254713
> pt(t, 19)
[1] 0.9819274
> 1-pt(t, 19)
[1] 0.01807261
> (1-pt(t, 19))*2
[1] 0.03614522
```

Luckily, this can be done a lot more easily in R...

```
> t.test(grades,mu=67.5)

    One Sample t-test

data:  grades
t = 2.2547, df = 19, p-value = 0.03615
alternative hypothesis: true mean is not equal to 67.5
95 percent confidence interval:
 67.84422 76.75578
sample estimates:
mean of x
     72.3
```

even easier with Navarro's function

```
> oneSampleTTest( x=grades, mu=67.5 )

   One sample t-test

Data variable:   grades

Descriptive statistics:
            grades
   mean     72.300
   std dev.  9.521

Hypotheses:
   null:        population mean equals 67.5
   alternative: population mean not equal to 67.5

Test results:
   t-statistic:  2.255
   degrees of freedom:  19
   p-value:  0.036

Other information:
   two-sided 95% confidence interval:  [67.844, 76.756]
   estimated effect size (Cohen's d):  0.504
```

Correctly reporting this in a paper

"With a mean grade of 72.3, the psychology students scored slightly higher than the average grade of 67.5 ($t(19) = 2.25, p < .05$); the 95% confidence interval is [67.8, 76.8]."

Might also be reported as "$t(19) = 2.25, p < .05, CI_{95}$ =[67.8, 76.8]"

Correctly reporting this in a paper

"With a mean grade of 72.3, the psychology students scored slightly higher than the average grade of 67.5 ($t(19) = 2.25, p < .05$); the 95% confidence interval is [67.8, 76.8]."

Might also be reported as "$t(19) = 2.25, p < .05, CI_{95}$ =[67.8, 76.8]"

Being able to formulate the result of a test in this form is relevant for the exam!

Assumptions of the t-test

Normality the t-test assumes that the true population distribution is normal (or that the sampling distribution of the mean is normal).

Independence The t-test also assumes that the observations in your data set are not correlated with each other, or related to each other.

Table of Contents

The independent samples t-test

Most of the time, we actually have two samples that we want to compare (as opposed to comparing one sample with the population).

Example

Example from earlier

You built a dialog system that can converse with humans. There are however two alternative settings with the system, and you want to test whether it makes a difference to use the one or the other.
You ask some people to interact with the one or the other version of the system, and record the time it takes to complete a certain task.

We need to compare the mean completion times for each of the settings.

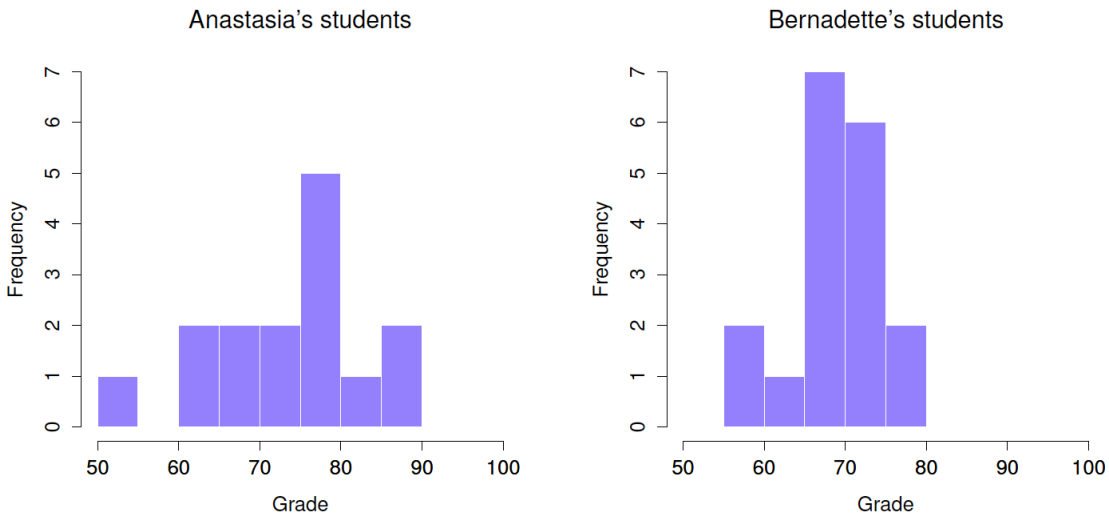
Another grading example

More grading...

A class has two tutors, and we want to find out which tutor is better by comparing the performance of the students in the final exam by tutor group.

The data

```
> head( harpo )
  grade      tutor
1    65  Anastasia
2    72 Bernadette
3    66 Bernadette
4    74  Anastasia
5    73  Anastasia
6    71 Bernadette
```



(a) (b)

Figure 13.6: Histograms showing the overall distribution of grades for students in Anastasia's class (panel a) and in Bernadette's class (panel b). Inspection of these histograms suggests that the students in Anastasia's class may be getting slightly better grades on average, though they also seem a little more variable.

	mean	std dev	N
Anastasia's students	74.53	9.00	15
Bernadette's students	69.06	5.77	18

t statistic for independent samples

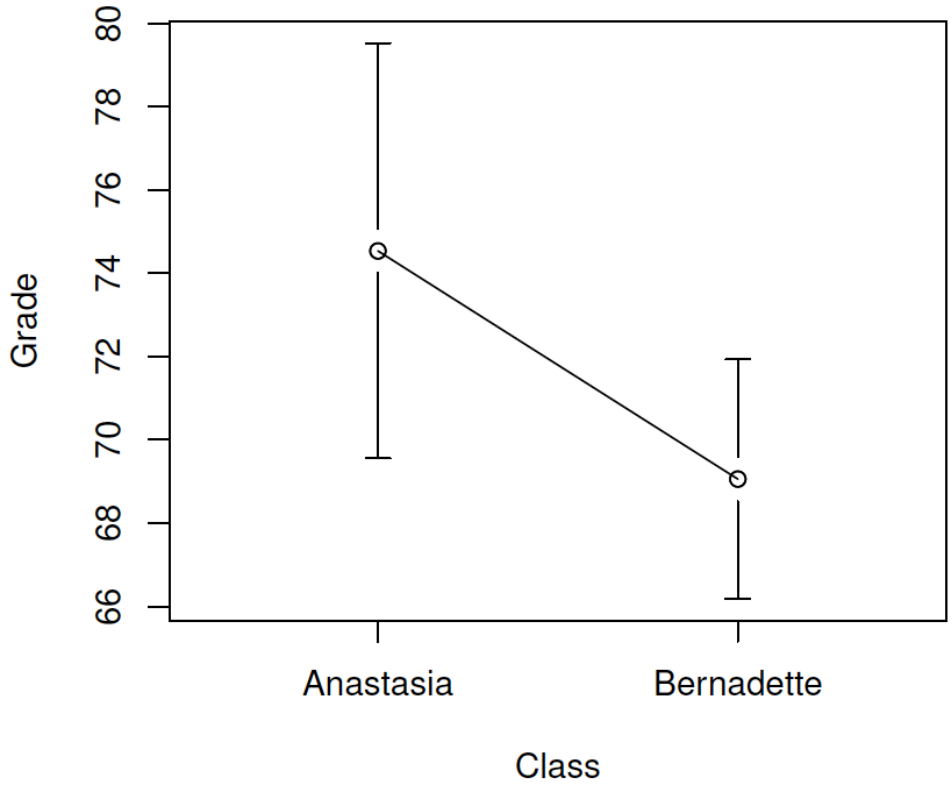


Figure 13.7: Plots showing the mean grade for the students in Anastasia's and Bernadette's tutorials. Error bars depict 95% confidence intervals around the mean. On the basis of visual inspection, it does look like there's a real difference between the groups, though it's hard to say for sure.

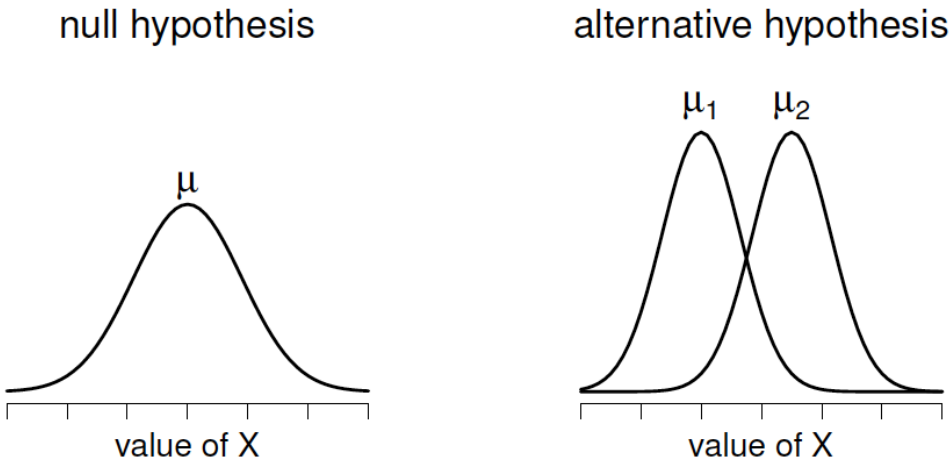


Figure 13.8: Graphical illustration of the null and alternative hypotheses assumed by the Student t-test. The null hypothesis assumes that both groups have the same mean μ, whereas the alternative assumes that they have different means μ_1 and μ_2. Notice that it is assumed that the population distributions are normal, and that, although the alternative hypothesis allows the group to have different means, it assumes they have the same standard deviation.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{SE}$$

How do we estimate the SE now?

Before:

- used standard error of the population (for the z-test)
- used standard error of the sample (for one sample t-test)

Independent samples (Student's t): With two samples, we need to take both standard deviations into account.

How do we estimate the SE now?

Before:

- used standard error of the population (for the z-test)
- used standard error of the sample (for one sample t-test)

Independent samples (Student's t): With two samples, we need to take both standard deviations into account. We do this by averaging variances (weighted average taking into account the number of observations in each sample).

$$\textit{weights} \quad w_1 = N_1 - 1; \qquad w_2 = N_2 - 1$$

here:

$$w_1 = 15 - 1; \qquad w_2 = 18 - 1$$

Pooled standard deviation and standard error

Now that we've assigned weights to each sample, we calculate the pooled estimate of the variance by taking the weighted average of the two variance estimates, $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$

$$\hat{\sigma}_p^2 = \frac{w_1\hat{\sigma}_1^2 + w_2\hat{\sigma}_2^2}{w_1 + w_2}$$

Finally, we convert the pooled variance estimate to a pooled standard deviation estimate, by taking the square root. This gives us the following formula for $\hat{\sigma}_p$,

$$\hat{\sigma}_p = \sqrt{\frac{w_1\hat{\sigma}_1^2 + w_2\hat{\sigma}_2^2}{w_1 + w_2}}$$

$$\text{SE}(\bar{X}_1 - \bar{X}_2) = \hat{\sigma}\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\text{SE}(\bar{X}_1 - \bar{X}_2)}$$

Degrees of freedom are now N-2 (one df lost for each sample).

Calculating independent samples t-test

Now we again have all ingredients to calculate the t-test by hand, this time for independent samples.

Calculating independent samples t-test

Now we again have all ingredients to calculate the t-test by hand, this time for independent samples.

One important point: Student's original independent samples t-test assumed that both samples have the same variation!

Calculating independent samples t-test

Now we again have all ingredients to calculate the t-test by hand, this time for independent samples.

One important point: Student's original independent samples t-test assumed that both samples have the same variation!

We'll now take a look at how to do this in R (and how to do it without same variations for the samples).

Doing it in R by hand

	mean	std dev	N
Anastasia's students	74.53	9.00	15
Bernadette's students	69.06	5.77	18

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} = \frac{74.53 - 69.06}{\sqrt{\frac{14*9.00^2+17*5.77^2}{14+17}} * \sqrt{\frac{1}{15} + \frac{1}{18}}} = 2.113$$

```
(1-pt(2.113, 31))*2
> 0.04275259
```

Doing it in R

```
> independentSamplesTTest(
      formula = grade ~ tutor,  # formula specifying outcome and group variables
      data = harpo,             # data frame that contains the variables
      var.equal = TRUE          # assume that the two groups have the same variance
  )

   Student's independent samples t-test

Outcome variable:   grade
Grouping variable:  tutor

Descriptive statistics:
            Anastasia Bernadette
   mean        74.533     69.056
   std dev.     8.999      5.775

Hypotheses:
   null:        population means equal for both groups
   alternative: different population means in each group

Test results:
   t-statistic:  2.115
   degrees of freedom:  31
   p-value:  0.043

Other information:
   two-sided 95% confidence interval:  [0.197, 10.759]
   estimated effect size (Cohen's d):  0.74
```

Formulating the findings

The mean grade in Anastasia's class was 74.5% (std dev = 9.0), whereas the mean in Bernadette's class was 69.1% (std dev = 5.8). A Student's independent samples t-test showed that this 5.4% difference was significant ($t(31) = 2.1$, $p < .05$, $CI_{95} = [0.2, 10.8]$, $d = .74$), suggesting that a genuine difference in learning outcomes has occurred.

```
   Student's independent samples t-test

Outcome variable:   grade
Grouping variable:  tutor

Descriptive statistics:
            Anastasia Bernadette
   mean        74.533     69.056
   std dev.     8.999      5.775

Hypotheses:
   null:        population means equal for both groups
   alternative: different population means in each group

Test results:
   t-statistic:  2.115
   degrees of freedom:  31
   p-value:  0.043

Other information:
   two-sided 95% confidence interval:  [0.197, 10.759]
   estimated effect size (Cohen's d):  0.74
```

Assumptions of independent-samples t-test

Normality the t-test assumes that the true population distribution is normal (or that the sampling distribution of the mean is normal).

Independence The t-test also assumes that the observations in your data set are not correlated with each other, or related to each other.

Homogeneity of variance (also called "homoscedasticity"). The third assumption is that the standard deviation is the same in both groups. You can test this assumption using the Levene test, which I'll talk about later (lecture 8). However, there's a very simple remedy for this assumption, which we'll get to next.

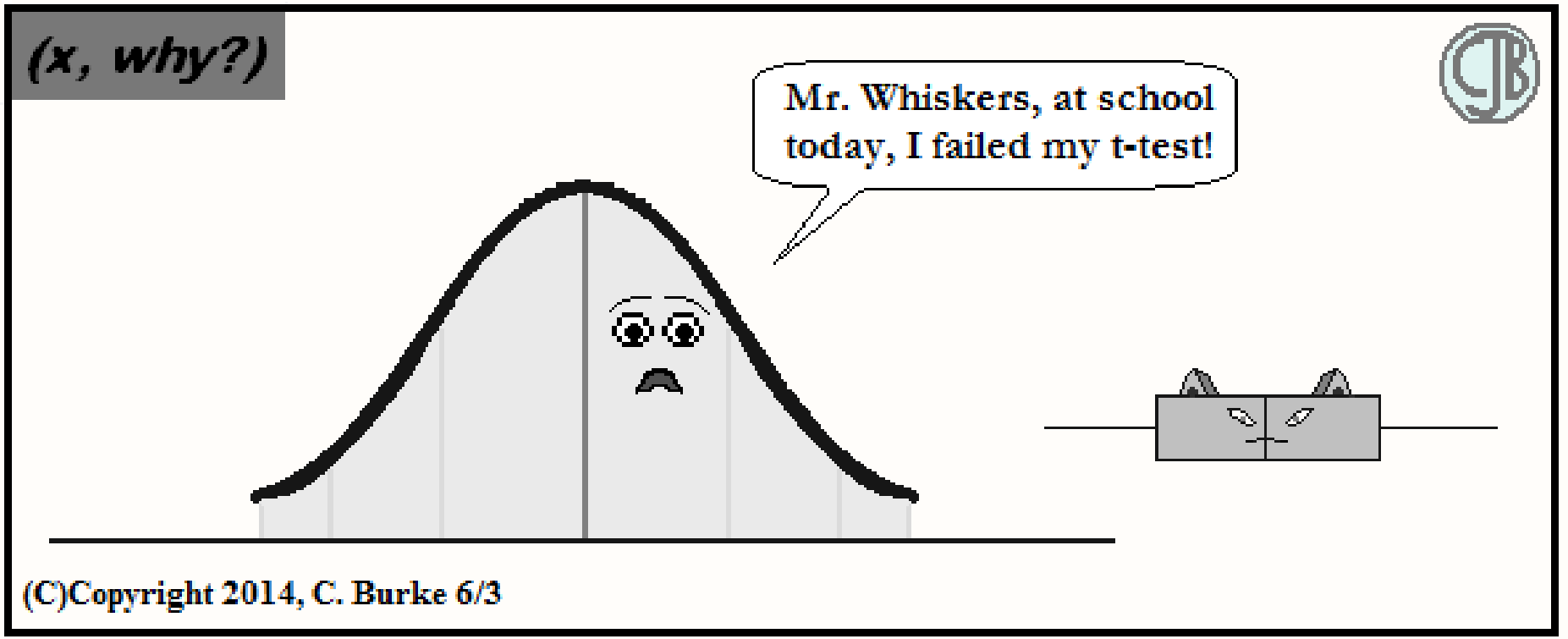

(x, why?)
Mr. Whiskers, at school today, I failed my t-test!
(C)Copyright 2014, C. Burke 6/3

Table of Contents

Why another t-test ?

The biggest problem with Student's independent samples t-test is the third assumption (same variance).
Why would two samples that have different means have the same variance?

Why another t-test ?

The biggest problem with Student's independent samples t-test is the third assumption (same variance).
Why would two samples that have different means have the same variance?

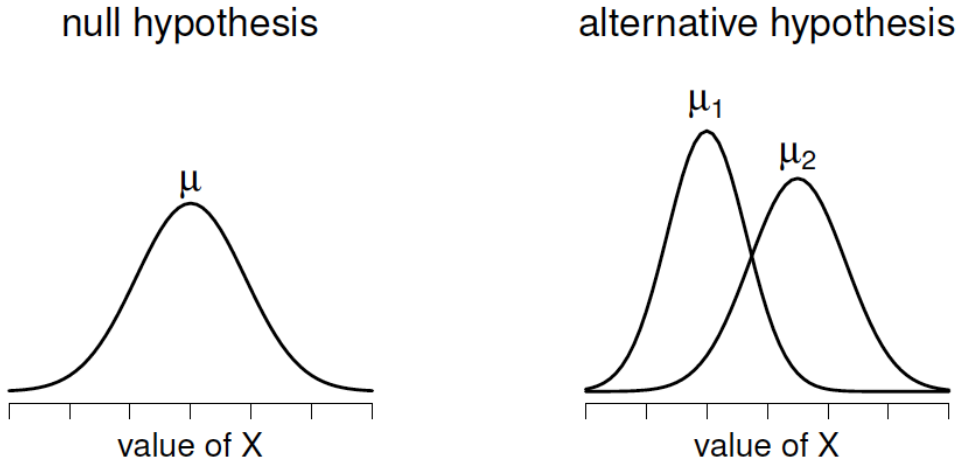


Figure 13.9: Graphical illustration of the null and alternative hypotheses assumed by the Welch t-test. Like the Student test (Figure 13.8) we assume that both samples are drawn from a normal population; but the alternative hypothesis no longer requires the two populations to have equal variance.

Welch vs. Student's t-test

Student	Welch
$t = \frac{\bar{X}_1 - \bar{X}_2}{SE(\bar{X}_1 - \bar{X}_2)}$	$t = \frac{\bar{X}_1 - \bar{X}_2}{SE(\bar{X}_1 - \bar{X}_2)}$
$SE(\bar{X}_1 - \bar{X}_2) = \hat{\sigma}\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$	$SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}$
$df = N - 2$	$\mathrm{df} = \frac{(\hat{\sigma}_1^2/N_1 + \hat{\sigma}_2^2/N_2)^2}{(\hat{\sigma}_1^2/N_1)^2/(N_1 - 1) + (\hat{\sigma}_2^2/N_2)^2/(N_2 - 1)}$

```
> independentSamplesTTest(
      formula = grade ~ tutor,  # formula specifying outcome and group variables
      data = harpo              # data frame that contains the variables
  )

   Welch's independent samples t-test

Outcome variable:   grade
Grouping variable:  tutor

Descriptive statistics:
            Anastasia Bernadette
   mean        74.533     69.056
   std dev.     8.999      5.775

Hypotheses:
   null:        population means equal for both groups
   alternative: different population means in each group

Test results:
   t-statistic:  2.034

   degrees of freedom:  23.025
   p-value:  0.054

Other information:
   two-sided 95% confidence interval:  [-0.092, 11.048]
   estimated effect size (Cohen's d):  0.724
```

Welch test assumptions

Same as for student's t-test , except we don't have to worry about homoscedasticity.

Table of Contents

The independence assumption

What if we have a set of data for which the independence assumption doesn't hold?

The independence assumption

What if we have a set of data for which the independence assumption doesn't hold?

A common research design (repeated measures) means that this assumption is violated.

Examples

Does listening to music reduce people's working memory capacity?
We could measure each person's working memory capacity in two conditions: with music, and without music.

We have a specific set of tasks or set of data that we want to predict, and have two different models for doing that.
In this case we're measuring each task / data point twice (once with each model).

Idea

If we have two measurements from one person or two measurements for the same task, this is important information to take into account, because the fact that the measurements do come from the exact same person or same task allow us to observe more directly the effect of the manipulation.

Solution:
the *paired t-test*

Example from Navarro book

Course with two exams.
Teacher claims that students work harder due to difficult first exam.
Does the grade improve from first exam to second exam?

	var	n	mean	sd	median
id*	1	20	10.50	5.92	10.5
grade_test1	2	20	56.98	6.62	57.7
grade_test2	3	20	58.38	6.41	59.7

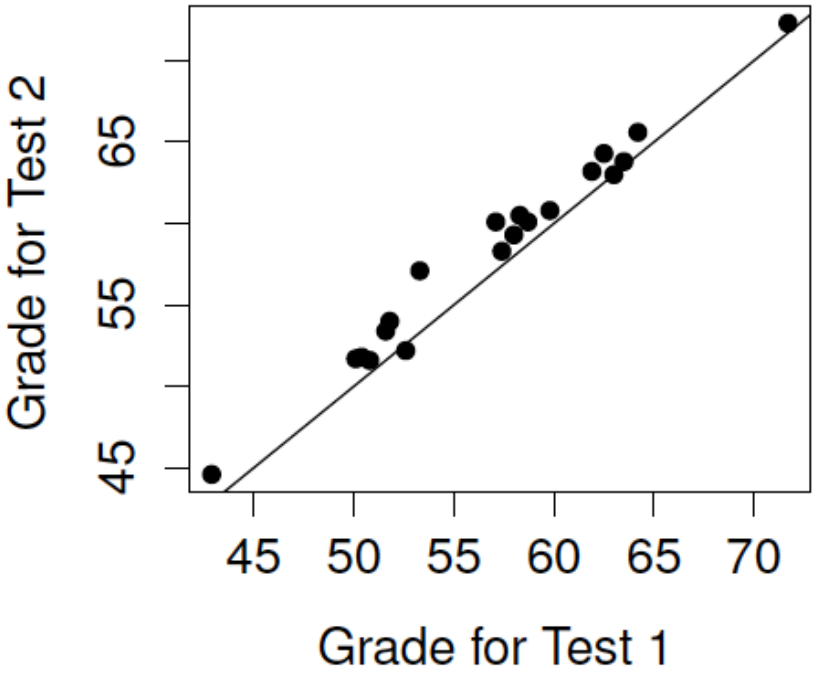
Example from Navarro book

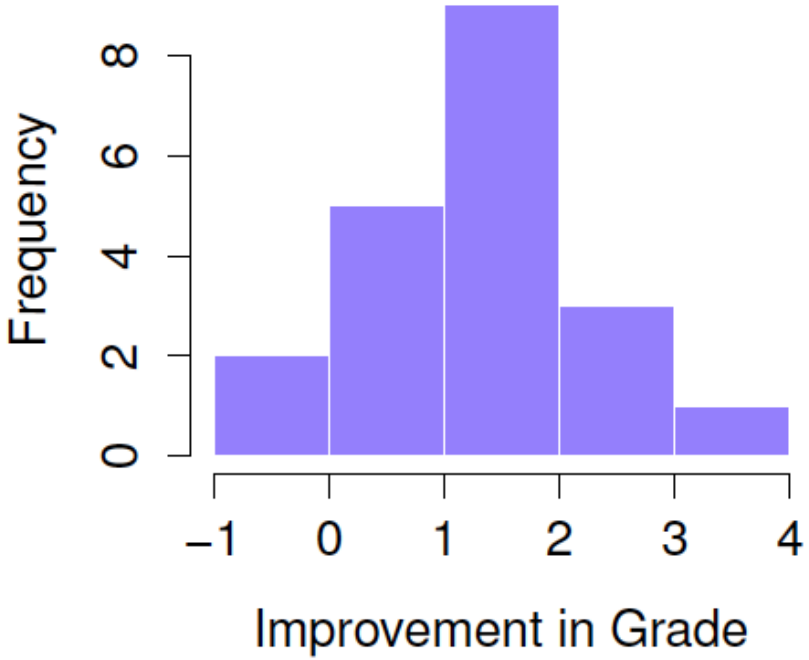
Course with two exams.
Teacher claims that students work harder due to difficult first exam.
Does the grade improve from first exam to second exam?

	var	n	mean	sd	median
id*	1	20	10.50	5.92	10.5
grade_test1	2	20	56.98	6.62	57.7
grade_test2	3	20	58.38	6.41	59.7

It kind of looks like no – basically same grades.

Paired data





But if we take a closer look, we can see that there's a real improvement here – most students scored higher in the second test than in the first test.

Each of the students improved a little bit.

paired vs. independent t-test

if we were to run an independent samples t-test

we'd be conflating between-subject variability (which we're not interested in) with the *within-subject variability* (which we are interested in).

Paired t-test

Core idea

Instead of running an independent sample *t*-test on the two data series, let's calculate the difference between first and second test

```
> head( chico )
       id grade_test1 grade_test2 improvement
1 student1        42.9        44.6         1.7
2 student2        51.8        54.0         2.2
3 student3        71.7        72.3         0.6
4 student4        51.6        53.4         1.8
5 student5        63.5        63.8         0.3
6 student6        58.0        59.3         1.3
```

Then, we can run a one-sample *t*-test on the *improvement*.

Paired t-test

Core idea

Instead of running an independent sample *t*-test on the two data series, let's calculate the difference between first and second test

```
> head( chico )
       id grade_test1 grade_test2 improvement
1 student1        42.9        44.6         1.7
2 student2        51.8        54.0         2.2
3 student3        71.7        72.3         0.6
4 student4        51.6        53.4         1.8
5 student5        63.5        63.8         0.3
6 student6        58.0        59.3         1.3
```

Then, we can run a one-sample *t*-test on the *improvement*.

(Calculating the improvement takes out the variation between subjects.)

```
> t.test(chico$grade_test2, chico$grade_test1, paired=TRUE)

    Paired t-test

data:  chico$grade_test2 and chico$grade_test1
t = 6.4754, df = 19, p-value = 3.321e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.9508686 1.8591314
sample estimates:
mean of the differences
                  1.405


> chico$improvement <- chico$grade_test2 - chico$grade_test1
> t.test(chico$improvement)

    One Sample t-test

data:  chico$improvement
t = 6.4754, df = 19, p-value = 3.321e-06
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.9508686 1.8591314
sample estimates:
mean of x
    1.405
```

```
> t.test(chico$grade_test2, chico$grade_test1, paired=TRUE)

    Paired t-test

data:  chico$grade_test2 and chico$grade_test1
t = 6.4754, df = 19, p-value = 3.321e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.9508686 1.8591314
sample estimates:
mean of the differences
                  1.405
```

Pairing can make a large difference in the result!

```
> t.test(chico$grade_test1, chico$grade_test2)

    Welch Two Sample t-test

data:  chico$grade_test1 and chico$grade_test2
t = -0.68231, df = 37.96, p-value = 0.4992
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.573752  2.763752
sample estimates:
mean of x mean of y
   56.980    58.385
```

Table of Contents

Effect size

A significant effect is not necessarily a large or an important effect.
How can we compare effect sizes across different experiments?
Most used for t-test : *Cohen's D*.

Cohen's D – Simple idea

Calculate the difference between the means of the two populations, and divide by the standard deviation. $d = \frac{mean1 - mean2}{sd}$

```
> pt<-t.test(chico$grade_test2, chico$grade_test1, paired=TRUE)
> pt

    Paired t-test

data:  chico$grade_test2 and chico$grade_test1
t = 6.4754, df = 19, p-value = 3.321e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.9508686 1.8591314
sample estimates:
mean of the differences
                  1.405

> names(pt)
[1] "statistic"   "parameter"   "p.value"     "conf.int"    "estimate"
[6] "null.value"  "alternative" "method"      "data.name"
> pt$estimate
mean of the differences
                  1.405
> pt$estimate[[1]]/sd(chico$improvement)
[1] 1.447952
```

But be careful with interpretation

For our paired *t*-test , the effect size is now with respect to the improvement in grades. If we want to calculate Cohen's D with respect to the original scale, we need to select that option in R.

```
> cohensD(chico$grade_test2, chico$grade_test1, method="paired")
[1] 1.447952
> cohensD(chico$grade_test2, chico$grade_test1)
[1] 0.2157646
```

Cohen's D for Welch's *t*-test

When calculating effect size for Welch's *t*-test , we of course need to take into account again the different standard deviations of the two samples:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}{2}}}$$

```
> cohensD( formula = grade ~ tutor,
+          data = harpo,
+          method = "unequal"
+ )
[1] 0.7244995
```

Effect size

Table 13.1: A (very) rough guide to interpreting Cohen's d. My personal recommendation is to not use these blindly. The d statistic has a natural interpretation in and of itself: it redescribes the different in means as the number of standard deviations that separates those means. So it's generally a good idea to think about what that means in practical terms. In some contexts a "small" effect could be of big practical importance. In other situations a "large" effect may not be all that interesting.

d-value	rough interpretation
about 0.2	"small" effect
about 0.5	"moderate" effect
about 0.8	"large" effect

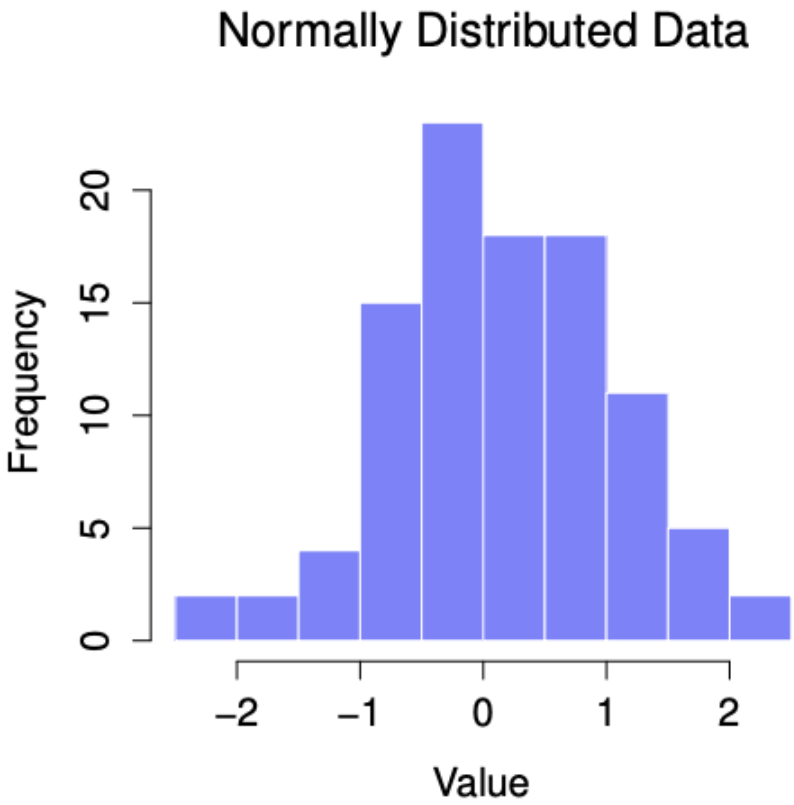
Table of Contents

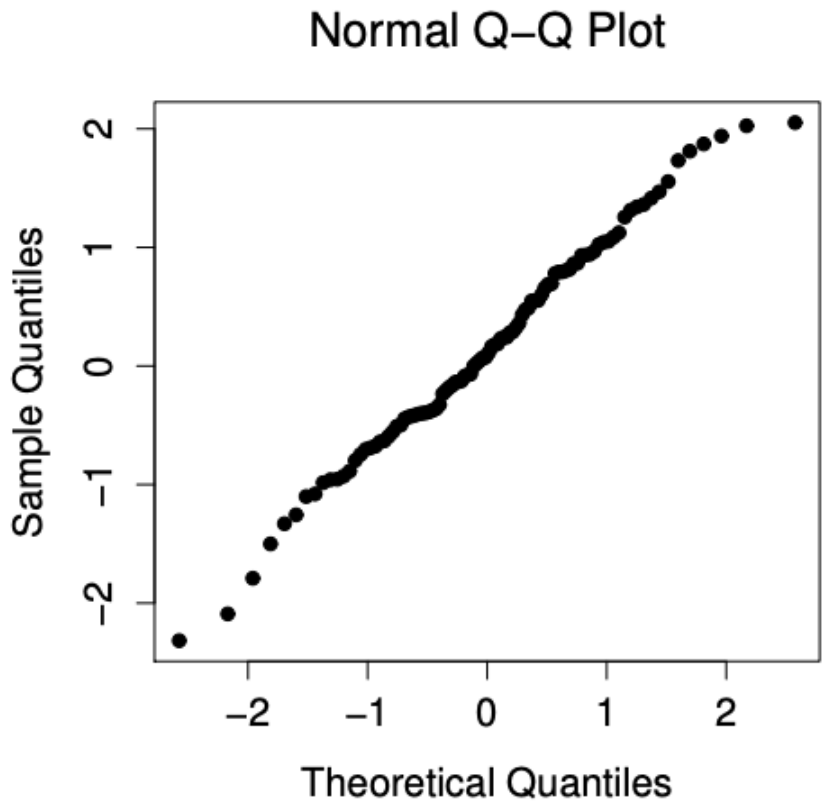
Assumptions

we have already talked a lot about assumptions and what to do if they don't hold...

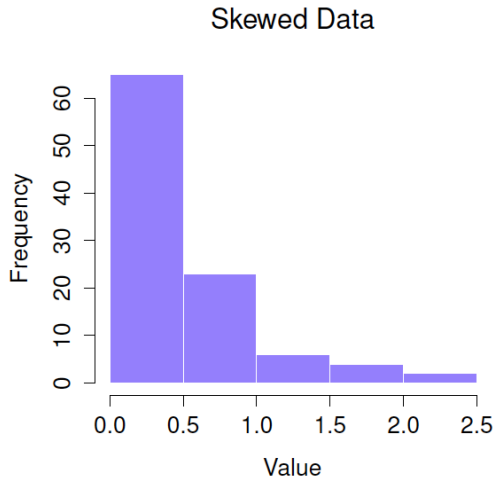
- ... so what about the **normality** assumption?
- How can we test it?
- And what can we do if it doesn't hold?

Checking for normality: QQ plot

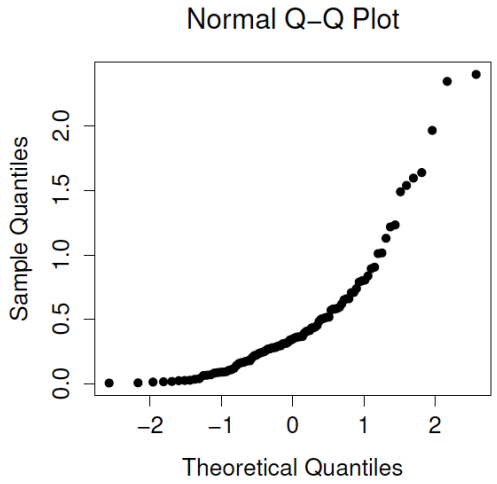




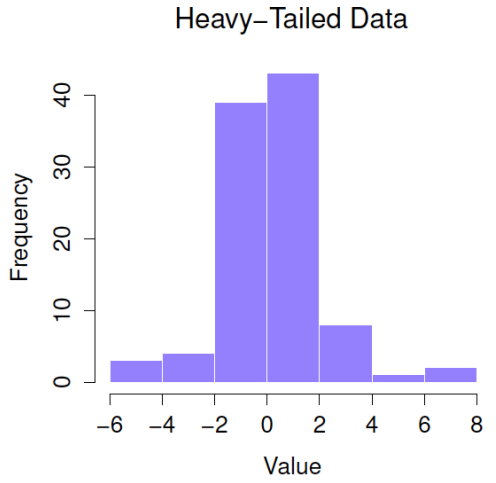
Inspecting QQ plots



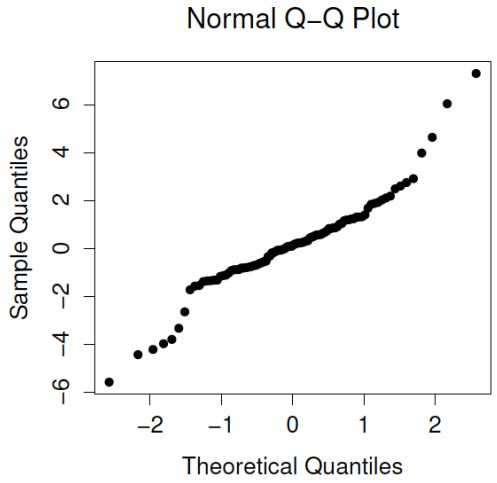
(a)



(b)



(c)



(d)

Plots are nice....

Well, plots look nice...
but can't I calculate some number
that I can then cite in my paper?

Yes, you can.
Use the Shapiro-Wilk normality test.

Plots are nice....

Well, plots look nice...
but can't I calculate some number that I can then cite in my paper?

Yes, you can.
Use the Shapiro-Wilk normality test.

```
> shapiro.test(chico$improvement)

    Shapiro-Wilk normality test

data:  chico$improvement
W = 0.9664, p-value = 0.6778

> shapiro.test(beaver1$temp)

    Shapiro-Wilk normality test

data:  beaver1$temp
W = 0.97031, p-value = 0.01226

> shapiro.test(beaver2$temp)

    Shapiro-Wilk normality test

data:  beaver2$temp
W = 0.93336, p-value = 7.764e-05
```

:-(

... so our beavers aren't normal...

What do we do now?

:-(

... so our beavers aren't normal...

What do we do now?

$\rightarrow$ Wilcoxon test :-)

Table of Contents

Wilcoxon test

- doesn't assume normality
- actually, no assumptions about shape
- it's a *non-parametric* test
- drawback: the Wilcoxon test is usually less powerful than the t-test (i.e., higher Type II error rate)

Two-sample Wilcoxon test aka "Mann-Whitney test"

Awesome data:

we have two groups and their awesomeness scores:

A 6.4, 10.7, 11.9, 7.3, 10.0

B 14.5, 10.4, 12.9, 11.7, 13.0

Do they differ in awesomeness?

Wilcoxon test

We compare all scores with all scores. Whenever group A scores higher, we place a checkmark.

		group B				
		14.5	10.4	12.4	11.7	13.0
group A	6.4	.	.	.	.	.
	10.7	.	✓	.	.	.
	11.9	.	✓	.	✓	.
	7.3	.	.	.	.	.
	10.0	.	.	.	.	.

Then we count the number of checkmarks, and use the W test statistic.

```
> wilcox.test(awesome$scores ~ awesome$group)

    Wilcoxon rank sum test

data:  awesome$scores by awesome$group
W = 3, p-value = 0.05556
alternative hypothesis: true location shift is not equal to 0

> t.test(awesome$scores ~ awesome$group)

    Welch Two Sample t-test

data:  awesome$scores by awesome$group
t = -2.5994, df = 6.9387, p-value = 0.03573
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.1926364 -0.2873636
sample estimates:
mean in group A mean in group B
           9.26           12.50
```

What is the W statistic?

The W statistic is similar to the normal distribution for large samples ($n > 15$), you can then calculate a z value from the W value and look it up in the standard normal distribution.
For smaller samples, there are separate W tables (see e.g. wikipedia or pages 441/ 442 in "Probability and statistics for computer scientists", by Michael Baron.)
In the Baron book, you can also find a more detailed explanation of the Wilcoxon test and the W distribution (pages 328 - 333); also, check out the explanation in the course book (by Dan Navarro) in Wilcoxon rank sum test (section 13.10.1) for more details.

One-sample Wilcoxon test

For one sample, or two paired samples.

```
> happiness
   before after change
1      30     6    -24
2      43    29    -14
3      21    11    -10
4      24    31      7
5      23    17     -6
6      40     2    -38
7      29    31      2
8      56    21    -35
9      38     8    -30
10     16    21      5
```

```
> wilcox.test(happiness$change)

    Wilcoxon signed rank test

data:  happiness$change
V = 7, p-value = 0.03711
alternative hypothesis: true location is not equal to 0
```

		all differences									
		-24	-14	-10	7	-6	-38	2	-35	-30	5
positive differences	7	.	.	.	✓	✓	.	✓	.	.	✓
	2	.	.	.	.	.	.	✓	.	.	.
	5	.	.	.	.	.	.	✓	.	.	✓

The V test statistic just counts binary outcomes and therefore uses the binomial distribution.

Summary

- A **one sample t-test** is used to compare a single sample mean against a hypothesised value for the population mean.
- An **independent samples t-test** is used to compare the means of two groups, and tests the null hypothesis that they have the same mean.
  - the **Student test** assumes that groups have same standard deviation
  - the **Welch test** does not.
- A **paired samples t-test** is used when you have two scores from each person, and you want to test the null hypothesis that the two scores have the same mean. (equivalent to one-sample t-test on difference between the two scores for each person)
- **Effect size** calculations for the difference between means can be calculated via the *Cohen's d* statistic.
- You can check the **normality** of a sample using QQ plots and the Shapiro-Wilk test.
- If your data are non-normal, you can use **Wilcoxon tests**.