

T-Test: Paired Samples

Data Analysis for Psychology in R 1
Semester 2, Week 8

Dr Emma Waterston

Department of Psychology
The University of Edinburgh

Learning Objectives

- Understand when to use an paired sample t -test
- Understand the null hypothesis for an paired sample t -test
- Understand how to calculate the test statistic
- Know how to conduct the test in [R](#)

Topics for Today

- Conceptual background and introduction to our example
- Calculations and R-functions
- Assumptions and effect size

Paired T-Test Purpose & Data

- The paired sample t -test is used when we want to test the difference in mean scores for a sample comprising matched (or naturally related) pairs.
- Examples:
 - Pre-test and post-test score with an intervention administered between the time points
 - A participant experiences both experimental conditions (e.g., caffeine and placebo)
- Data Requirements
 - A continuously measured variable.
 - A binary variable denoting pairing.

t-statistic

$$t = \frac{\bar{d} - \mu_{d_0}}{SE_{\bar{d}}} \quad \text{where} \quad SE_{\bar{d}} = \frac{s_d}{\sqrt{n}}$$

- \bar{d} = mean of the individual difference scores (d_i) where $d_i = x_{i1} - x_{i2}$
- μ_{d_0} is the hypothesised population mean difference in the null hypothesis (which is usually assumed to be 0)
- $SE_{\bar{d}}$ = standard error of mean difference (d_i)
 - s_d = standard deviation of the difference scores (d_i)
 - n = sample size = number of matched pairs
- Sampling distribution is a t -distribution with $n - 1$ degrees of freedom
- Note, this is just essentially a one sample test on the difference scores

Hypotheses

- Two-tailed:

$$H_0 : \mu_d = \mu_{d_0}$$

$$H_1 : \mu_d \neq \mu_{d_0}$$

- One-tailed

$$H_0 : \mu_d = \mu_{d_0}$$

$$H_1 : \mu_d < \mu_{d_0}$$

$$H_1 : \mu_d > \mu_{d_0}$$

- Two-tailed:

$$H_0 : \mu_d - \mu_{d_0} = 0$$

$$H_1 : \mu_d - \mu_{d_0} \neq 0$$

- One-tailed

$$H_0 : \mu_d - \mu_{d_0} = 0$$

$$H_1 : \mu_d - \mu_{d_0} < 0$$

$$H_1 : \mu_d - \mu_{d_0} > 0$$

Questions?

Example

- I want to assess whether a time-management course influenced levels of exam stress in students.
- I ask 50 students to take a self-report stress measure during their winter exams.
- At the beginning of semester 2 they take a time management course.
- I then assess their self-report stress in the summer exam block.
 - Let's assume for the sake of this example that I have been able to control the volume and difficulty of the exams the students take in each block.

Data

```
## # A tibble: 6 × 3
##   ID      stress time
##   <chr>   <dbl> <fct>
## 1 ID1      14 t1
## 2 ID2       7 t1
## 3 ID3       8 t1
## 4 ID4       8 t1
## 5 ID5       7 t1
## 6 ID6       7 t1
```

Hypotheses

- I elect to use a two-tailed test with alpha (α) of .01
- I want to be quite sure the intervention has worked and stress levels are different.
- So my hypotheses are:

$$H_0 : \mu_d = \mu_{d_0}$$

$$H_1 : \mu_d \neq \mu_{d_0}$$

Questions?

Calculation

- Steps in my calculations:
 - Calculate the difference scores for individuals d_i
 - Calculate the mean of the difference scores \bar{d}
 - Calculate the s_d of the difference scores
 - Check I know my n
 - Calculate the standard error of mean difference ($SE_{\bar{d}}$)
- Use all this to calculate t

Data Organisation

- Our data is currently in what is referred to as long format.
 - All the scores are in one column, with two entries per participant.
- To calculate the d_i values, we will convert this to wide format.
 - Where there are two columns representing the score at time 1 and time 2
 - And a single row per person

Data Organisation

```
exam_wide <- exam %>%  
  pivot_wider(id_cols = ID,  
              names_from = time,  
              values_from = stress)  
head(exam_wide)
```

```
## # A tibble: 6 × 3  
##   ID      t1    t2  
##   <chr> <dbl> <dbl>  
## 1 ID1     14     7  
## 2 ID2      7     7  
## 3 ID3      8     9  
## 4 ID4      8    12  
## 5 ID5      7    10  
## 6 ID6      7     9
```

Calculation

```
exam_wide %>%  
  mutate(dif = t1 - t2) %>%  
  summarise(  
    dbar = mean(dif),  
    Sd = sd(dif),  
    mu_d0 = 0,  
    n = n()) %>%  
  mutate(  
    SEd = (Sd /sqrt(n)),  
    t = ((dbar-mu_d0)/SEd)  
  ) %>%  
  kable(digits = 2) %>%  
  kable_styling(full_width = FALSE)
```

dbar	Sd	mu_d0	n	SEd	t
2.1	3.55	0	50	0.5	4.19

Calculation

dbar	Sd	mu_d0	n	SEd	t
2.1	3.55	0	50	0.5	4.19

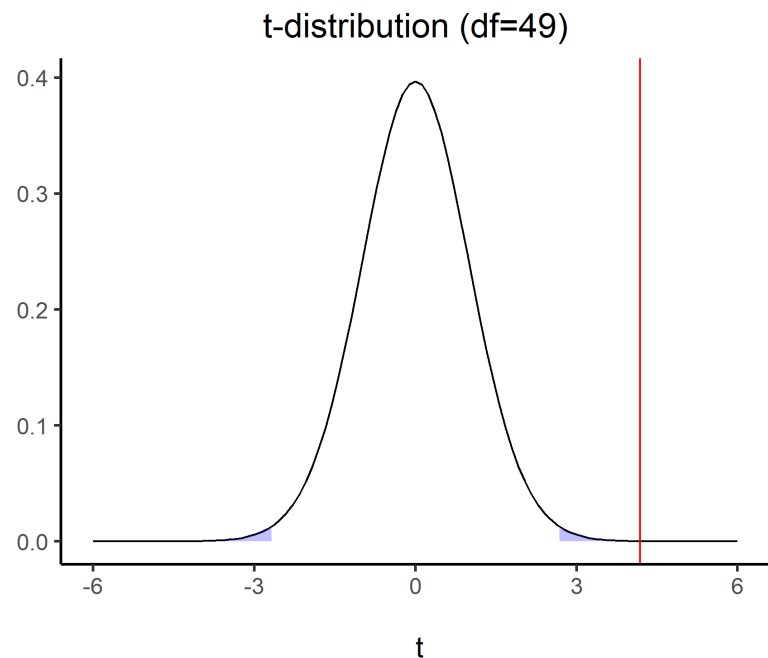
$$t = \frac{\bar{d} - \mu_{d_0}}{SE_{\bar{d}}} = \frac{2.1 - 0}{\frac{3.55}{\sqrt{50}}} = \frac{2.1}{0.5} = 4.20$$

- So in our example $t = 4.20$
- Note: When doing hand calculations there might be a small amount of rounding error when we compare to t calculated in [R](#).

Is my test significant?

- So we have all the pieces we need:
 - $t = 4.19$
 - $df = n - 1 = 50 - 1 = 49$
 - Hypothesis to test (two-tailed)
 - $\alpha = .01$
- So now all we need is the critical value from the associated t -distribution in order to make our decision.

Is my test significant?



```
tibble(  
  LowerCrit = round(qt(0.005, 49), 2),  
  UpperCrit = round(qt(0.995, 49), 2),  
  Exactp = round(2*(1-pt(calc[[6]]), 49)), 5)  
)
```

```
## # A tibble: 1 × 3  
##   LowerCrit UpperCrit Exactp  
##   <dbl>     <dbl>   <dbl>  
## 1     -2.68      2.68 0.00012
```

Is my test significant?

- So our critical value is 2.68
 - Our t -statistic (4.19) is larger than this
 - So we reject the null hypothesis

- Wide Format Data

```
# two numeric columns
res_wide <- t.test(exam_wide$t1, exam_wide$t2,
  paired = TRUE,
  mu = 0,
  alternative = "two.sided",
  conf.level = 0.99)
res_wide
```

```
##
##      Paired t-test
##
## data:  exam_wide$t1 and exam_wide$t2
## t = 4.1864, df = 49, p-value = 0.0001174
## alternative hypothesis: true mean difference is not equal to 0
## 99 percent confidence interval:
##  0.7556557 3.4443443
## sample estimates:
## mean difference
##           2.1
```

- Long Format Data

```
#one numeric column, one binary column
res_long <- t.test(exam$stress ~ exam$time,
  paired = TRUE,
  mu = 0,
  alternative = "two.sided",
  conf.level = 0.99)
res_long
```

Write-up

A paired-sample t -test was conducted in order to determine if a statistically significant ($\alpha = .01$) mean difference in self-report stress was present, pre- and post-time management intervention in a sample of 50 undergraduate students. The pre-intervention mean score was higher ($M = 9.72, SD = 2.19$) than the post intervention score ($M = 7.62, SD = 2.55$). The difference was statistically significant ($t(49) = 4.19, p < .001, two - tailed$). We are 99% confident that post-intervention scores were between 0.76 and 3.44 points lower than pre-intervention scores. Thus, we reject the null hypothesis of no difference.

Questions?

Assumption checks summary

	Description	One-Sample t-test	Independent Sample t-test	Paired Sample t-test
Normality	Continuous variable (and difference) is normally distributed.	Yes (Population)	Yes (Both groups/ Difference)	Yes (Both groups/ Difference)
Tests:	Descriptive Statistics; Shapiro-Wilks Test; QQ-plot			
Independence	Observations are sampled independently.	Yes	Yes (within and across groups)	Yes (within groups)
Tests:	None. Design issue.			
Homogeneity of variance	Population level standard deviation is the same in both groups.	NA	Yes	NA
Tests:	F-test			
Matched Pairs in data	For paired sample, each observation must have matched pair.	NA	NA	Yes
Tests:	None. Data structure issue.			

Assumptions

- Normality of the difference scores (d_i)
- Independence of observations **within** group/time
- Data are matched pairs (design)

Adding the difference scores

- Our assumptions concern the difference scores.
- We showed these earlier in our calculations.
- Here we will add them to `exam_wide` for ease.

```
exam_wide <- exam_wide %>%  
  mutate(  
    dif = t1 - t2)
```

Normality: Skew

Verbal label	Magnitude of skew in absolute value
Generally not problematic	$ \text{Skew} < 1$
Slight concern	$1 > \text{Skew} < 2$
Investigate impact	$ \text{Skew} > 2$

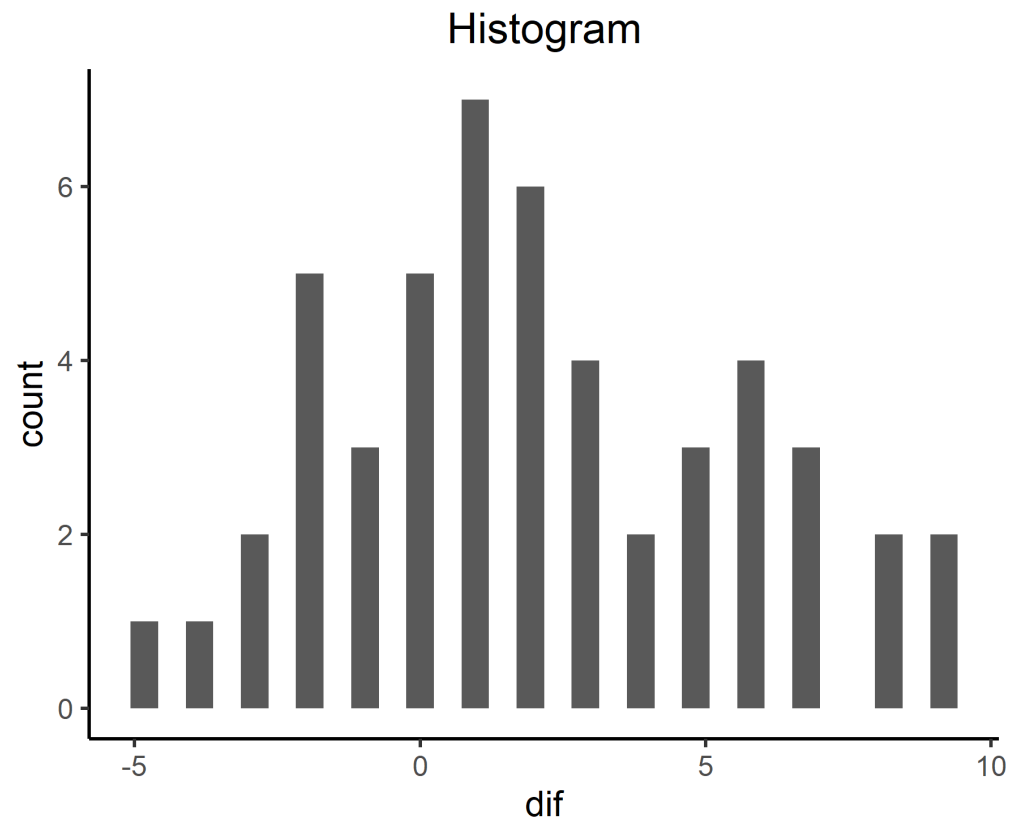
```
library(psych)
exam_wide %>%
  summarise(
    skew = round(skew(dif), 2)
  )
```

```
## # A tibble: 1 × 1
##   skew
##   <dbl>
## 1  0.18
```

- Skew is low (< 1), so we would conclude that it is not problematic.

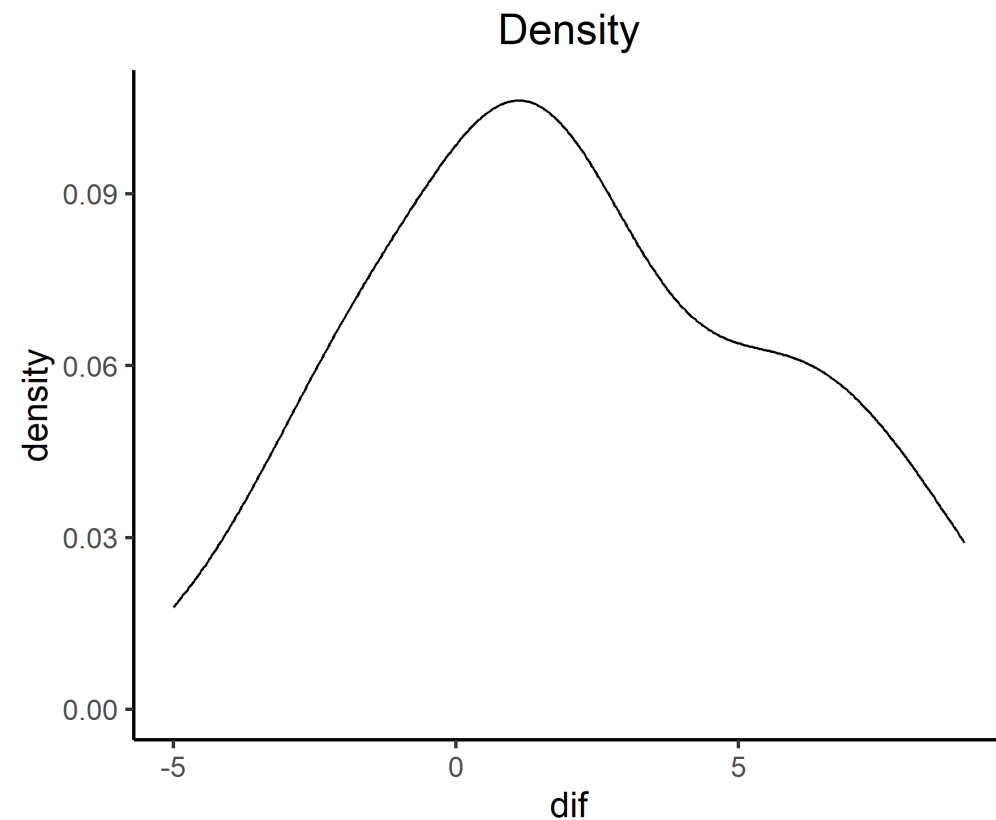
Normality: Histograms

```
ggplot(exam_wide, aes(x=dif)) +  
  geom_histogram() +  
  labs(title = "Histogram")
```



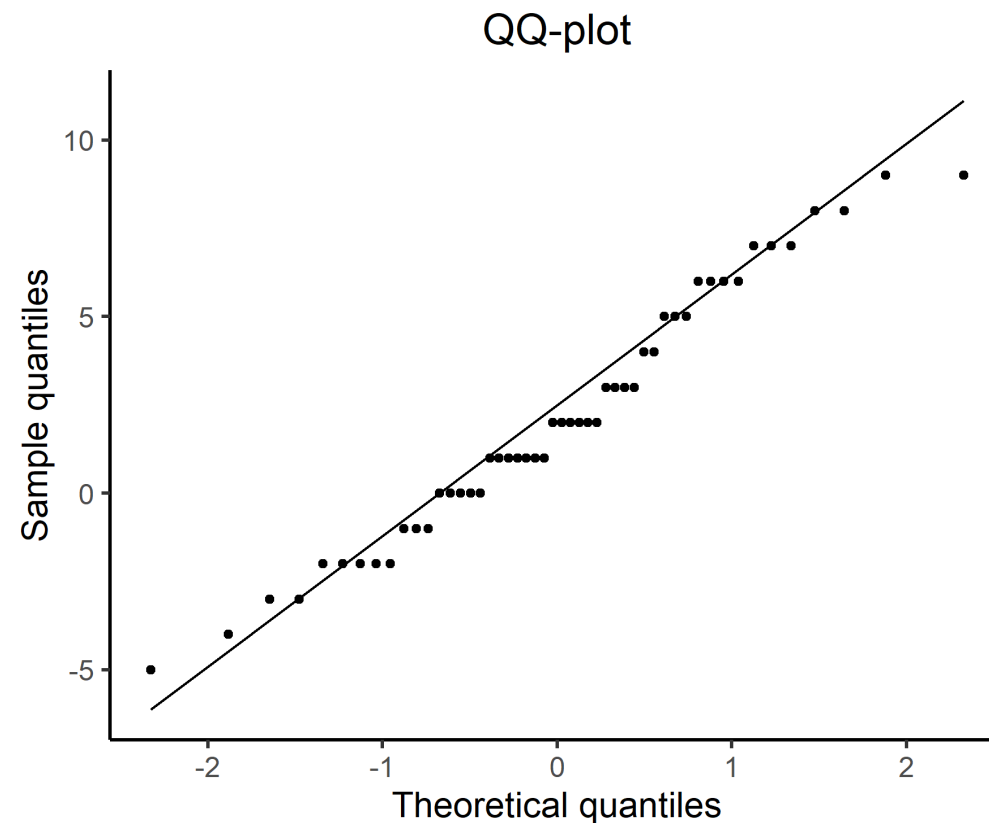
Normality: Density

```
ggplot(exam_wide, aes(x=dif)) +  
  geom_density() +  
  labs(title = "Density")
```



Normality: QQ-plots

```
ggplot(exam_wide, aes(sample = dif)) +  
  stat_qq() +  
  stat_qq_line() +  
  labs(title="QQ-plot",  
        x = "Theoretical quantiles",  
        y = "Sample quantiles")
```



Normality: Shapiro-Wilks in R

```
shapiro.test(exam_wide$dif)
```

```
##  
##      Shapiro-Wilk normality test  
##  
## data:  exam_wide$dif  
## W = 0.97142, p-value = 0.264
```

- Fail to reject the null, $p = .264$, which is $> .05$
- Normality of the differences is met.

Cohen's D: Paired t-test

- Paired-sample t -test:

$$D = \frac{\bar{d} - \mu_{d_0}}{s_d}$$

- \bar{d} = mean of the difference scores (d_i)
- μ_{d_0} is the hypothesised population difference in means in the null hypothesis
- s_d = standard deviation of the difference scores (d_i)
- So in our example:
 - $\bar{d} = 2.1$
 - $\mu_{d_0} = 0$
 - $s_d = 3.55$

$$D = \frac{2.1 - 0}{3.55} = 0.59$$

Cohen's D in R

- Wide Format Data

```
library(effectsize)
cohens_d(exam_wide$t1, exam_wide$t2,
         paired = TRUE,
         mu = 0,
         alternative = "two.sided",
         ci = 0.99)
```

```
## Cohen's d |          99% CI
## -----
## 0.59      | [0.19, 0.99]
```

- Long Format Data

```
library(effectsize)
cohens_d(exam$stress ~ exam$time,
         paired = TRUE,
         mu = 0,
         alternative = "two.sided",
         ci = 0.99)
```

```
## Cohen's d |          99% CI
## -----
## 0.59      | [0.19, 0.99]
```


Write up: Assumptions

The DV of our study, Stress, was measured on a continuous scale. Independence of observations can be assumed based on the study design. Data comprised matched pairs of observations as participants were assessed twice, pre- and post- time management course. The assumption of normality was visually assessed (via histograms, density plots, and a QQplot) as well as statistically via a Shapiro-Wilks test. The QQplot did not show much deviation from the diagonal line, and the Shapiro-Wilks test suggested that the difference scores were normally distributed ($W = 0.97, p = .264$). This was inline with the histogram and density plots, which suggested that the difference in scores between the two assessment times was normally distributed (and where $skew < 1$). The size of the effect was found to be medium-large ($D = 0.59$).

Summary

- Today we have covered:
 - Basic structure of the paired-sample t -test
 - Calculations
 - Interpretation
 - Assumption checks
 - Effect size measures

Tasks

- Read the lab worked example and complete your lab
- Complete any assigned readings
- Go to office hours if you have questions
 - Emma's Office Hours = Tuesdays 10:30-11:30 in G15, 7 George Square
- Complete the weekly quiz
 - Opens Monday at 9am
 - Closes Sunday at 5pm