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

Course Overview

Exploratory Data Analysis	Research design and data
	Describing categorical data
	Describing continuous data
	Describing relationships
	Functions
Probability	Probability theory
	Probability rules
	Random variables (discrete)
	Random variables (continuous)
	Sampling

Foundations of inference	Confidence intervals
	Hypothesis testing (p-values)
	Hypothesis testing (critical values)
	Hypothesis testing and confidence intervals
	Errors, power, effect size, assumptions
Common hypothesis tests	One sample t-test
	Independent samples t-test
	Paired samples t-test
	Chi-square tests
	Correlation

Learning Objectives

- Understand when to use an paired samples t-test
- Understand the null hypothesis for an paired samples t-test
- Understand how to calculate the test statistic
- Know how to conduct the test in R

T-Test: Paired Samples

Purpose

- The paired samples 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)

t-statistic

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

- Where
  - $\bar{d}$ = 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 t-test on the difference scores

Hypotheses

Two-tailed

$$H_0 : \mu_d = \mu_{d_0} \qquad \text{vs} \qquad H_1 : \mu_d \neq \mu_{d_0} \qquad\qquad H_0 : \mu_d - \mu_{d_0} = 0 \qquad \text{vs} \qquad H_1 : \mu_d - \mu_{d_0} \neq 0$$

One-tailed

$$H_0 : \mu_d = \mu_{d_0} \qquad \text{vs} \qquad H_1 : \mu_d < \mu_{d_0} \qquad\qquad H_0 : \mu_d - \mu_{d_0} = 0 \qquad \text{vs} \qquad H_1 : \mu_d - \mu_{d_0} < 0$$

$$H_0 : \mu_d = \mu_{d_0} \qquad \text{vs} \qquad H_1 : \mu_d > \mu_{d_0} \qquad\qquad H_0 : \mu_d - \mu_{d_0} = 0 \qquad \text{vs} \qquad H_1 : \mu_d - \mu_{d_0} > 0$$

Questions?

Example

Time Management & Stress

- 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: 100 × 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
##  7 ID7       11 t1
##  8 ID8        9 t1
##  9 ID9       10 t1
## 10 ID10      14 t1
## # ℹ 90 more rows
```

Hypotheses

- I want to be quite sure the intervention has worked and stress levels are different pre- and post- time management course
- I elect to use a two-tailed test with alpha (α) of .01
- I specify the following hypotheses:

$$H_0 : \mu_d = \mu_{d_0}$$
$$H_1 : \mu_d \neq \mu_{d_0}$$

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 $(SE_{\bar{d}})$

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

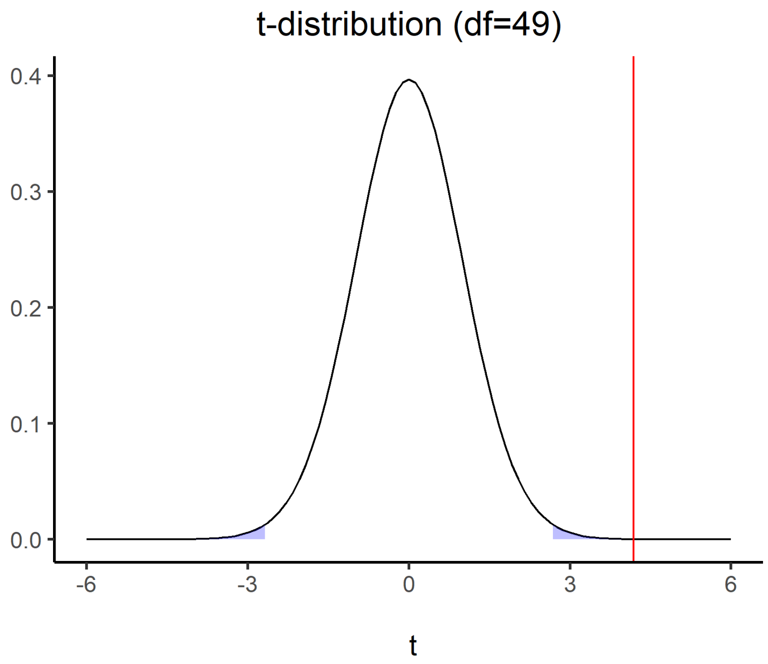
$$t = \frac{\bar{d} - \mu_{d_0}}{SE_{\bar{d}}} = \frac{2.1 - 0}{\frac{3.55}{\sqrt{50}}} = \frac{2.1}{0.5} = 4.20$$

- 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 our Test Significant?

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

Is our Test Significant?




```
tibble(
  LowerCrit = round(qt(0.005, 49),2),
  UpperCrit = round(qt(0.995, 49),2),
  Exactp = round(2*(1-pt(4.186, 49)),5)
)
```

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

Is our Test Significant?

- The critical value is 2.68, and our t-statistic (4.19) is larger than this
- We found that $p < .001$, which is $< \alpha$
- Thus, we **reject the null hypothesis**

Paired Samples T-Test in R

```r
# must have 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
```

Write Up

A paired samples t-test was conducted in order to determine a if a statistically significant $(\alpha = .01)$ mean difference in self-reported 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 stress scores were between 0.76 and 3.44 points lower than pre-intervention stress scores. Thus, we reject the null hypothesis of no difference.

Questions?

Data Requirements & Assumptions

Data Requirements

- A numeric variable

Assumption Checks Summary

	Description	One-Sample t-test	Independent Samples t-test	Paired Samples t-test
Normality	Numeric variable (or difference) is normally distributed OR sample size is sufficiently large.	Yes (variable). Sample size guideline: $n \geq 30$	Yes (variable in each group). Sample size guideline: $n1 \geq 30$ and $n2 \geq 30$	Yes (difference). Sample size guideline: number of pairs ≥ 30
Tests:	Descriptive Statistics and Plots; QQ-Plot; Shapiro-Wilks Test			
Independence	Observations are sampled independently.	Yes	Yes (within and across groups)	Yes (across pairs)
Tests:	None. Design issue.			
Homogeneity of variance	Population 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.			

* Welch t-test is available if this is not met

Data Requirements & Assumptions: How to Check/Test

- DV is numeric
    - The dependent variable should be measured on a interval/ratio/integer scale
- Normality of the difference scores (d_i)
    - Can be checked with descriptive statistics, visually with plots, and with a Shapiro-Wilks test for each group separately
- Independence of observations **across pairs**
    - More of a study design issue, and cannot directly test
- Data are matched pairs
    - More of a study design issue, and cannot directly test

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

- Skew is a descriptive statistic informing us of both the direction and magnitude of asymmetry
    - Below are some rough guidelines on how to interpret skew
    - No strict cuts for skew - these are loose guidelines

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

Skew in R

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

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

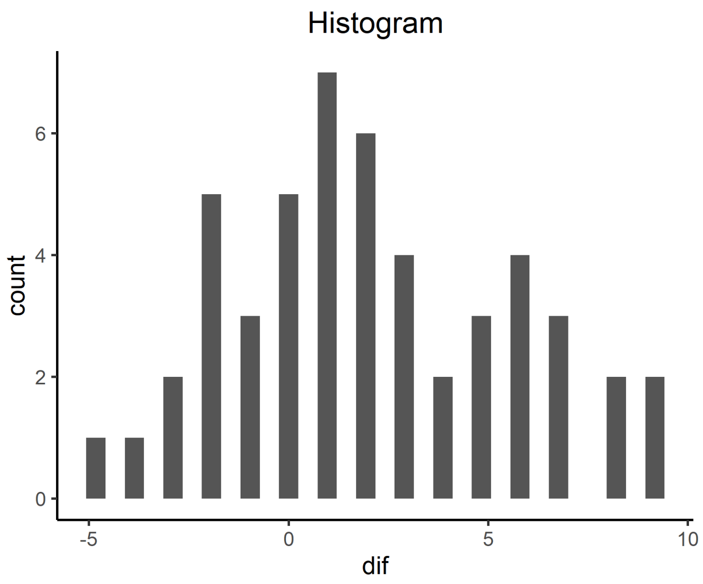
- Skew is low (< 1), so we would conclude that it is not problematic

Normality: Visual Assessment

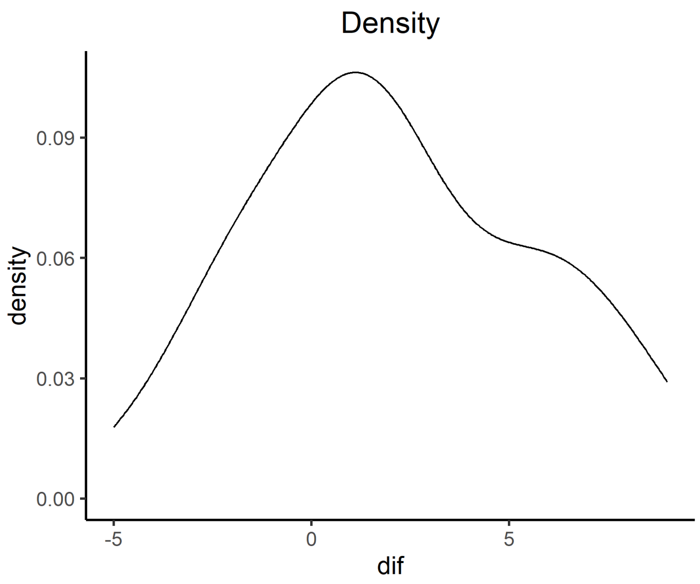
- We can visually assess normality by plotting the distribution of the difference scores (d_i)
  - Histograms
    - The count (or frequency) of data points that fall within specified intervals/bins
  - Density Plots
    - The probability density (or proportion of values) of data points at each value of the observed variable
  - QQ-Plots (Quantile-Quantile plot):
    - Plots the sorted quantiles of one data set (distribution) against sorted quantiles of data set (distribution)
    - Quantile = the percent of points falling below a given value
    - For a normality check, we can compare our own data to data drawn from a normal distribution

Histogram & Density Plots in R

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

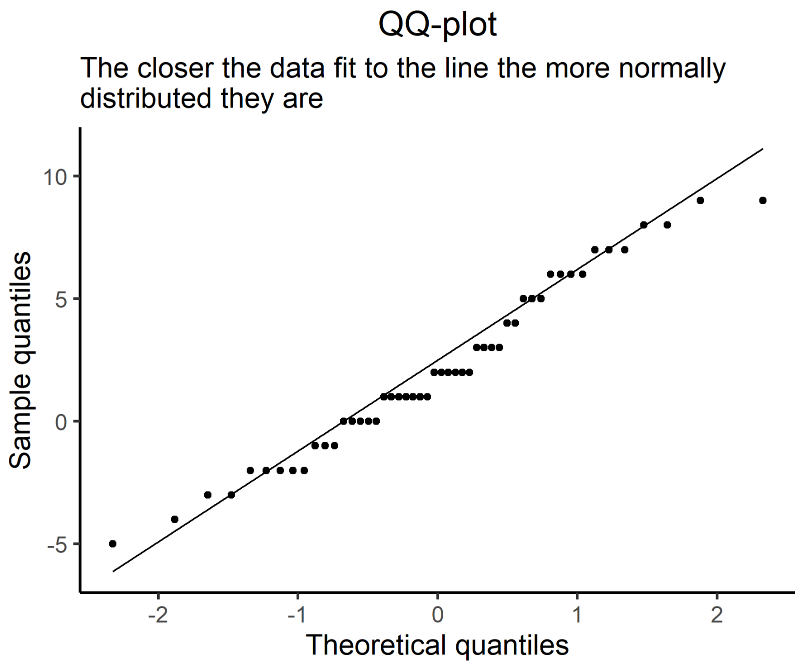



```r
ggplot(exam_wide, aes(x=dif)) +
  geom_density() +
  labs(title = "Density")
```



QQ-Plots **in R**

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



Normality: Shapiro-Wilks Test

- Shapiro-Wilks test:
  - Checks properties of the observed data against properties we would expected from normally distributed data.
  - Statistical test of normality.
  - H_0: data = a normal distribution.
  - p-value $< \alpha$ = reject the null, data are not normal.
    - Sensitive to n as all p-values will be.
    - In very large n, normality should also be checked with QQ-plots alongside statistical test.

Shapiro-Wilks Test in R

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

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

- $W = 0.97, p = .264$
- Fail to reject the null since p = .264, which is > α (.05)
- Normality of the difference scores is met

Questions?

Effect Size

Cohen's D: Paired T-Test

- Paired samples t-test:

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

- where
  - $\bar{d}$ = mean of the difference scores (d_i)
  - μ_{d_0} = the hypothesised population difference in means in the null hypothesis
  - s_d = standard deviation of the difference scores (d_i)

- In our example:
  - $\bar{d}$ = 2.1
  - μ_{d_0} = 0
  - s_d = 3.55

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

Cohen's D: Paired Samples T-Test **in R**

```r
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]
```

Write Up: Data Requirements, Assumptions, & Effect Size

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 $(D = 0.59\ [0.19, 0.99])$.

Summary

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

This Week

Tasks

- Attend both lectures
- Attend your lab and work on the assessed report with your group (due by 12 noon on Friday 27th of March 2026)
- Complete the weekly quiz
  - Opened Monday at 9am
  - Closes Sunday at 5pm

Support

- **Office Hours**: for one-to-one support on course materials or assessments
  (see LEARN > Course information > Course contacts)
- **Piazza**: help each other on this peer-to-peer discussion forum
- **Student Adviser**: for general support while you are at university
  (find your student adviser on MyEd/Euclid)