Differences of Means

David Gerard
Many slides borrowed from Linda Collins and Yali Amit

2017-11-07

Learning Objectives

- Paired t-tests.
- Two-sample t-tests.
- Sections 5.2 and 5.3 in DBC.

Paired Data

Matched Paired t-test

In a matched pairs study, there are 2 measurements taken on the same subject (or on 2 similar subjects). For example,

- 2 rats from the same litter
- before and after observations on the same subject
- adjacent plots on a field

To conduct statistical inference on such a sample, we analyze the *difference* using the one-sample procedures described above.

Weight Data

```
library(tidyverse)
load(file="w.Rdata")
glimpse(weight)
```

```
Observations: 20
Variables: 4
$ Subject    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 1...
$ weighta    <int> 187, 175, 158, 160, 130, 170, 165, 1...
$ weightb    <int> 160, 153, 150, 148, 127, 160, 150, 1...
$ difference <dbl> 27, 22, 8, 12, 3, 10, 15, -1, 10, 6,...
```

```
t=(mean(weight$difference)-0)/(sd(weight$difference)/sqrt(20))
p=1-pt(t,19)
c(t,p)
```

```
[1] 4.8842514 0.0000515
```

Equivalent to single variable methods

```
diff_vec <- weight$weighta - weight$weightb
```

Now just perform inference on `diff_vec`.

Matched Paired t-test

To ascertain whether the diet reduces weight, we test

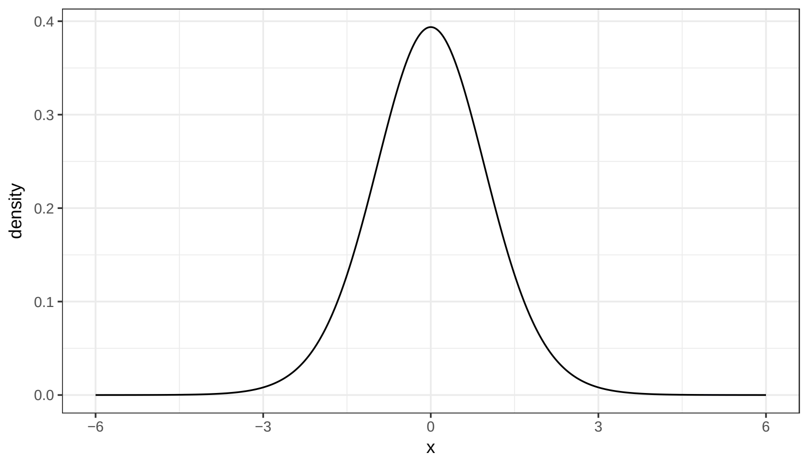
$$H_0 : \mu = 0 \qquad H_a : \mu > 0$$

where μ is the mean weight difference.

```
xbar  <- mean(diff_vec)
s     <- sd(diff_vec)
n     <- length(diff_vec)
tstat <- (xbar - 0) / (s / sqrt(n))
```

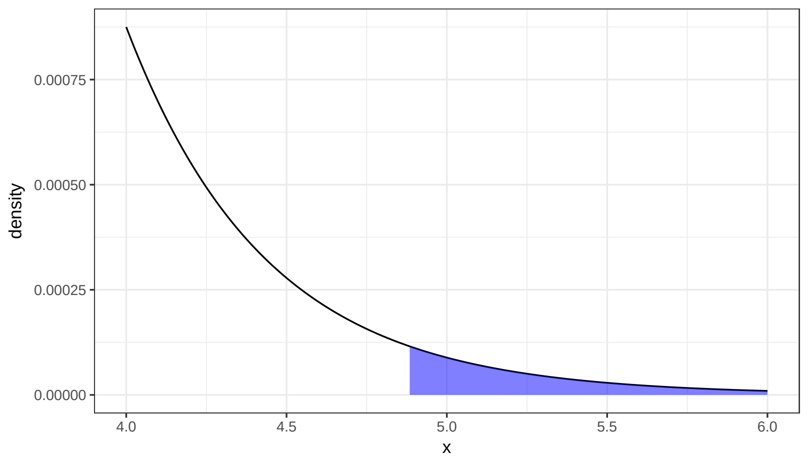
T-statistic: $t = \frac{9.35-0}{8.56/\sqrt{20}} = 4.88$

Paired t-test



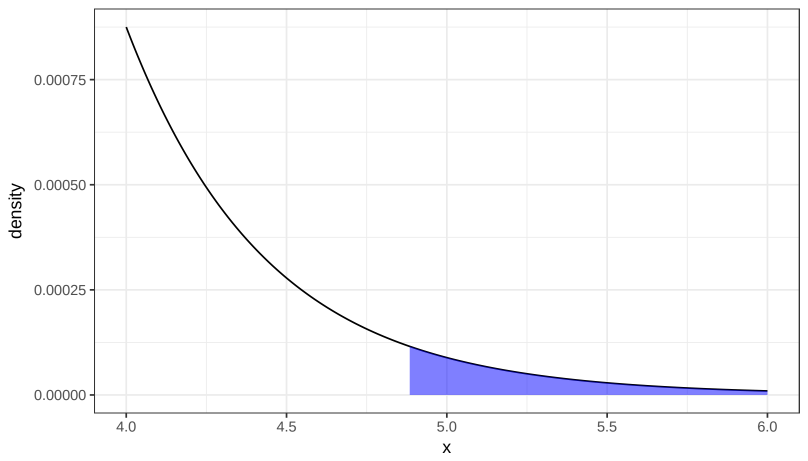
Paired t-test

Zooming in



Paired t-test

Zooming in



p-value: $p = P(t_{19} \geq 4.8843) = 0.000052$

Unpaired data (Two-sample data)

Two sample problems

- The goal of two-sample inference is to compare the responses in two groups.
- Each group is considered to be a sample from a distinct population.
- The responses in each group are independent of those in the other group (in addition to being independent of each other).

For example, Suppose we have a SRS of size n_1 drawn from a $N(\mu_1, \sigma_1)$ population and an independent SRS of size n_2 drawn from a $N(\mu_2, \sigma_2)$ population.

The first sample might be heights of male students and the second heights of female students.

We might test $H_0 : \mu_1 = \mu_2$ against $H_a : \mu_1 \neq \mu_2$.

Two sample problems

How is this different from the *matched pairs design*?

1. There is no matching of the units in two samples.
2. The two samples may be of different size.

Comparing Two Means when σ's are Known

Suppose we have a SRS of size n_1 drawn from a $N(\mu_1, \sigma_1)$ population (with sample mean $\bar{x}_1$) and an independent SRS of size n_2 drawn from a $N(\mu_2, \sigma_2)$ population (with sample mean $\bar{x}_2$). Suppose σ_1 and σ_2 are known.

The **two-sample z-statistic** is

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

Why the denominator? Since the two samples are independent, their averages are independent so:

$\mathrm{var}(\bar{X}_1 - \bar{X}_2) = \mathrm{var}(\bar{X}_1) + \mathrm{var}(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

Inference when σ's are known

- A $(1-\alpha)$ CI for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

  where $z^* : P(Z > z^*) = \alpha/2$.

- To test the hypothesis $H_0 : \mu_1 = \mu_2$, we use

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \text{ under } H_0$$

  The p-value is calculated as before

Comparing Two Means with σ's Unknown

We define $S_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(X_{1,i} - \bar{X}_1)^2, S_2^2 = \frac{1}{n_2}\sum_{i=1}^{n_2}(X_{2,i} - \bar{X}_2)^2$.

The **Two-sample t-statistic** is

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \ \dot{\sim} \ t_\nu$$

The T statistic only has an **approximate** t_ν distribution with

$\nu = \frac{(w_1+w_2)^2}{w_1^2/(n_1-1)+w_2^2/(n_2-1)}, \ w_1 = s_1^2/n_1, \ w_2 = s_2^2/n_2$.

This is called Satterthwaite's approximation.

Inference when σ's are unknown

- A $(1-\alpha)$ CI for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \text{ where } t^* : P(T_\nu > \alpha/2).$$

- To test the hypothesis $H_0 : \mu_1 = \mu_2$, we use

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \;\dot{\sim}\; t_\nu \text{ under } H_0$$

  The p-value is calculated as before.

Setting $\nu = \min(n_1 - 1, n_2 - 1)$ is simpler and yields a more conservative approximate procedure. That is, the CIs are longer than the true CI and p-values are larger than the true p-values.

Pooled two-sample t procedures

In the previous procedure, we assumed that $\sigma_1 \neq \sigma_2$. What if we have reason to believe $\sigma_1 = \sigma_2 = \sigma$ (even though we don't know either value)?

We can gain information (i.e. power) by *pooling* the two samples together for estimating the variance:

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$

If the two populations are normal this is the exact distribution of T.

Inference for pooled two sample t tests

- A $(1-\alpha)$ CI for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

  where $t^* : P(T_{n_1+n+2-2} > t^*) = \alpha/2$.

- To test the hypothesis $H_0 : \mu_1 = \mu_2$, we use

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)} \text{ under } H_0$$

  The p-value is calculated as before.

Example

Weight gains (in kg) of babies from birth to age one year are measured. All babies weighed approximately the same at birth.

Group A	5	7	8	9	6	7	10	8	6
Group B	9	10	8	6	8	7	9		

Assume that the samples are randomly selected from independent normal populations. Is there any difference between the true means of the two groups?

i) Assume $\sigma_1 = \sigma_2 = 1.5$ is known

ii) Assume σ_1 and σ_2 are unknown and unequal.

iii) Assume σ_1 and σ_2 are unknown but equal

State the hypothesis:

$$H_0 : \mu_1 = \mu_2 \qquad H_a : \mu_1 \neq \mu_2$$

Observed Statistics

$$\bar{x}_1 = 7.33 \quad \bar{x}_2 = 8.14$$
$$s_1 = 1.58 \quad s_2 = 1.35$$
$$n_1 = 9 \qquad n_2 = 7$$

Known variances

i) Assume $\sigma_1 = \sigma_2 = 1.5$ is known. Then, the two-sample z statistic is

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_1\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$
$$= \frac{7.33 - 8.14}{1.5 \times \sqrt{\frac{1}{9} + \frac{1}{7}}} = -1.07$$

The two-sided p-value is

$$2P(Z \geq |z|) = 2P(Z \geq 1.07) = 0.28$$

where $Z \sim N(0, 1)$.

So there is no difference between the true population mean of these two group at the significance level 0.1.

Known variances

A 90% confidence interval for $\mu_1 - \mu_2$ is:

$$
\begin{aligned}
(\bar{x}_1 - \bar{x}_2) \pm z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} & \\
&= (7.33 - 8.14) \pm 1.645 \times 1.5 \times \sqrt{\frac{1}{9} + \frac{1}{7}} \\
&= (-2.05\ ,\ 0.43)
\end{aligned}
$$

As expected, the 90% confidence interval covers 0. Thus, we have 90% confidence that there is no difference between the true population means.

Unknown unequal variances

ii) Assume σ_1 and σ_2 are unknown and unequal. Then, the two-sample t statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{7.33 - 8.14}{\sqrt{\frac{1.58^2}{9} + \frac{1.35^2}{7}}} = -1.10$$

The two-sided p-value is

$$2P(T \geq |z|) = 2P(T \geq 1.10) = 0.31$$

where $T \sim t_6$.

Unknown unequal variances

A 90% confidence interval for $\mu_1 - \mu_2$ is given by

$$
\begin{aligned}
(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\
&= (7.33 - 8.14) \pm 1.94 \times \sqrt{\frac{1.58^2}{9} + \frac{1.35^2}{7}} \\
&= (-2.23\ ,\ 0.61)
\end{aligned}
$$

where $P(|T| < t^*) = 0.90$. That is, $P(T > t^*) = 0.05$ or $t^* = t_{\nu,.05}$.

Unknown Equal variances

iii) Assume σ_1 and σ_2 are unknown but equal.

The pooled two-sample estimator of σ is

$$
\begin{aligned}
s_p &= \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \\
&= \sqrt{\frac{(9-1) \times 1.58^2 + (7-1) \times 1.35^2}{9 + 7 - 2}} \\
&= 1.49
\end{aligned}
$$

Thus, the pooled two-sample t statistic is

$$
t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{7.33 - 8.14}{1.49\sqrt{\frac{1}{9} + \frac{1}{7}}} = -1.08
$$

Unknown Equal variances

The two-sided p-value is given by

$$2P(T \geq |t|) = 2P(T \geq 1.08) = 0.30 \qquad \text{where } T \sim t_{14}.$$

A 90% confidence interval for $\mu_1 - \mu_2$ is

$$\begin{aligned}(\bar{x}_1 - \bar{x}_2) &\pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\ &= (7.33 - 8.14) \pm 1.76 \times 1.49 \times \sqrt{\frac{1}{9} + \frac{1}{7}} \\ &= (-2.12\ ,\ 0.51)\end{aligned}$$

Where $P(|T| < t^*) = 0.90$. That is, $P(T > t^*) = 0.05$.

How to actually do this in practice

Software

- It's important to understand the logic of a procedure.
- But you don't want to hard-code a *t*-test every time you need one — this is a recipe for human error!
- Use `t.test`.

Set up data:

```
x <- c(5, 7, 8, 9, 6, 7, 10, 8, 6)
y <- c(9, 10, 8, 6, 8, 7, 9)
```

Arguments

`x`	a (non-empty) numeric vector of data values.
`y`	an optional (non-empty) numeric of data values.
`alternative`	a character string specifying the alternative hypothesis, must be one of ‘two.sided” (default), “greater” or “less”. You can specify just the initial letter.
`mu`	a number indicating the true value of the mean (or difference in means if you are performing a two sample test).
`paired`	a logical indicating whether you want a paired t-test.
`var.equal`	a logical variable indicating whether to treat the two variances as being equal. If TRUE then the pooled variance is used to estimate the variance otherwise the Welch (or Satterthwaite) approximation to the degrees of freedom is used.

Assume $\sigma_1 = \sigma_2 = 1.5$

- People never use two-sample z-tests in practice.
- So there isn't a base R function that does this.
- Just hard-code this for HW and never do in practice.

Assume σ_1 and σ_2 are unknown and unequal.

```
t.test(x = x, y = y, alternative = "two.sided",
       var.equal = FALSE, conf.level = 0.9)



Welch Two Sample t-test


data:  x and y
t = -1.1, df = 14, p-value = 0.3
alternative hypothesis: true difference in means is not equ
90 percent confidence interval:
 -2.1004  0.4814
sample estimates:
mean of x mean of y
    7.333     8.143
```

The df

The degrees of freedom it actually used was not exactly 14, but they used Satterthwaite's approximation:

```
tout <- t.test(x = x, y = y, alternative = "two.sided",
               var.equal = FALSE, conf.level = 0.9)
tout$parameter
```

```
   df
13.84
```

Assume σ_1 and σ_2 are unknown but equal

```
t.test(x = x, y = y, alternative = "two.sided",
       var.equal = TRUE, conf.level = 0.9)



Two Sample t-test


data:  x and y
t = -1.1, df = 14, p-value = 0.3
alternative hypothesis: true difference in means is not equ
90 percent confidence interval:
 -2.1273  0.5082
sample estimates:
mean of x mean of y
    7.333     8.143
```