Important note on t-tests

Shravan Vasishth

Universität Potsdam
vasishth@uni-potsdam.de
http://www.ling.uni-potsdam.de/∼vasishth

April 12, 2020

Some important topics regarding the t-test

In this lecture, I will discuss the following important topics:

- Two-sample t-tests
- paired t-tests
- independent vs repeated measures data
- by-subjects and by-items analyses

t-test

These are the heights of students in one of my classes at Potsdam:

```
heights <- c(173,174,160,157,158,170,172,170,
             175,168,165,170,173,180,168,162,
             180,160,155,163,173,175,176,172,
             160,161,150,170,165,184,165)
```

We can do a t-test to evaluate the null hypothesis that $H_0 : \mu = 170$ cm.

The t-distribution

The formal definition of the t-distribution is as follows:
Suppose we have a random sample of size n, say of heights, which come from a $Normal(\mu, \sigma)$ distribution. Then the quantity
$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$
has a $\mathtt{t}(\mathtt{df} = n - 1)$ sampling distribution. The distribution is defined as (r is degrees of freedom):

$$f_X(x, r) = \frac{\Gamma[(r+1)/2]}{\sqrt{r\pi}\ \Gamma(r/2)} \left(1 + \frac{x^2}{r}\right)^{-(r+1)/2}, \quad -\infty < x < \infty.$$

[Γ refers to the gamma function; in this course we can ignore what this is, but read Kerns if you are interested.]

The t-test

```
  t.test(heights,mu=170)

##
##  One Sample t-test
##
## data:  heights
## t = -1.49, df = 30, p-value = 0.15
## alternative hypothesis: true mean is not equal to 170
## 95 percent confidence interval:
##  164.95 170.80
## sample estimates:
## mean of x
##    167.87
```

Computing the p-value by hand

First, we compute the absolute observed $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$:

```
(obs_t<-abs((mean(heights)-170)/(sd(heights)/sqrt(31))))
```

```
## [1] 1.4866
```

Then we compute the probability of seeing that absolute observed t or something more extreme, assuming the null is true:

```
2*pt(-obs_t,df=30)
```

```
## [1] 0.14756
```

Two-sample t-test

This is a data-set from Keith Johnson's book (Quantitative Methods in Linguistics):

```
F1data<-read.table("data/F1_data.txt",header=TRUE)
head(F1data)

##   female male vowel  language
## 1    391  339     i  W.Apache
## 2    561  512     e  W.Apache
## 3    826  670     a  W.Apache
## 4    453  427     o  W.Apache
## 5    358  291     i CAEnglish
## 6    454  406     e CAEnglish
```

Two-sample t-test

Notice that the male and female values are paired in the sense that they are for the same vowel and language.
We can compare males and females' F1 frequencies, ignoring the fact that the data are paired.
Now, our null hypothesis is $H_0 : \mu_m = \mu_f$ or
$H_0 : \mu_m - \mu_f = \delta = 0$.

Two-sample t-test

Assuming equal variance between men and women

```
  t.test(F1data$female,F1data$male,var.equal=TRUE)

##
##  Two Sample t-test
##
## data:  F1data$female and F1data$male
## t = 1.54, df = 36, p-value = 0.13
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  -30.066 217.540
## sample estimates:
## mean of x mean of y
##    534.63    440.89
```

Two-sample t-test

Doing this "by hand": The only new thing is the SE calculation, and the the df for t-distribution $(2 \times n - 2) = 36$.

$$SE_\delta = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

```
d<-mean(F1data$female)-mean(F1data$male)
(SE<-sqrt(var(F1data$male)/19+var(F1data$female)/19))

## [1] 61.044

observed_t <- (d-0)/SE
2*(1-pt(observed_t,df=36))

## [1] 0.13339
```

The paired t-test

But this data analysis was incorrect.
This data are paired: each row has F1 measurements from a male and female for the **same vowel and language**.
For paired data, $H_0 : \delta = 0$ as before. But since each row in the data-frame is paired (from the same vowel+language), we subtract row-wise, and get a new vector d with the pairwise differences.

The paired t-test

Then, we just do a one-sample test:

```
diff<-F1data$female-F1data$male
t.test(diff)

##
##  One Sample t-test
##
## data:  diff
## t = 6.11, df = 18, p-value = 9.1e-06
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   61.485 125.989
## sample estimates:
## mean of x
##    93.737
```

Summary so far

We have worked through the

1. One sample t-test
2. Two sample t-test
3. Paired t-test

A note on paired t-tests

Note that each row of the data frame cannot have more than one row for a particular pair.
For example, doing a paired t-test on this frame would be incorrect:

female	male	vowel	language
391	339	i	W.Apache
400	320	i	W.Apache
⋮	⋮	⋮	⋮

Why? Because the assumption is that each row is independent of the others. This assumption is violated here.

A note on paired t-tests

Note that each row of the data frame cannot have more than one row for a particular pair.
Another example:

cond_a	cond_b	subject	item
391	339	1	1
400	320	1	2
⋮	⋮	⋮	⋮

Here, we have repeated measures from subject 1. The independence assumption is violated.

A note on paired t-tests

1. What to do when we have repeated measurements from each subject or each item?
2. We aggregate the data so that each subject (or item) has only one value for each condition.
3. This has a drawback: it pretends we have one measurement from each subject for each condition.
4. Later on we will learn how to analyze unaggregated data.

Example of INCORRECT pair-wise t-test

We have repeated measures data on noun pronunciation durations, in seconds.
These data are in so-called wide form.

```
dataN2<-read.table("data/dataN2.txt",header=TRUE)
head(dataN2)

##   Sentence Speaker_id N2_dur.2 N2_dur.1
## 1        1          1  0.49650  0.61444
## 2        1          2  0.47979  0.58739
## 3        1          3  0.54716  0.69451
## 4        1          4  0.37836  0.56842
## 5        1          5  0.56719  0.44040
## 6        1          6  0.51831  0.54651
```

Example of INCORRECT pair-wise t-test

```
xtabs(~Sentence+Speaker_id,dataN2)

##          Speaker_id
## Sentence 1 2 3 4 5 6 7 8 9 10 11 12 13 14
##        1  1 1 1 1 1 1 1 1 1  1  1  1  1  1
##        2  1 1 1 1 1 1 1 1 1  1  1  1  1  1
##        3  1 1 1 1 1 1 1 1 1  1  1  1  1  1
##        4  1 1 1 1 1 1 1 1 1  1  1  1  1  1
##        5  1 1 1 1 1 1 1 1 1  1  1  1  1  1
##        6  1 1 1 1 1 1 1 1 1  1  1  1  1  1
##        7  1 1 1 1 1 1 1 1 1  1  1  1  1  1
##        8  1 1 1 1 1 1 1 1 1  1  1  1  1  1
##        9  1 1 1 1 1 1 1 1 1  1  1  1  1  1
##        10 1 1 1 1 1 1 1 1 1  1  1  1  1  1
##        11 1 1 1 1 1 1 1 1 1  1  1  1  1  1
```

Example of INCORRECT pair-wise t-test

```
  ## significant effect:
  with(dataN2,
       t.test(N2_dur.2,N2_dur.1,paired=TRUE))

##
##  Paired t-test
##
## data:  N2_dur.2 and N2_dur.1
## t = 2.22, df = 335, p-value = 0.027
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  0.0023201 0.0384052
## sample estimates:
## mean of the differences
##                0.020363
```

Example of INCORRECT pair-wise t-test

- The above t-test was incorrect because we have multiple rows of (dependent) data from the same subject.
- We need to aggregate the multiple measurements from each subject until we have one data point from each subject for each combination of vowel and language.

How to figure out if we have repeated measures data? We turn to this question next.

CORRECT pair-wise t-test

Our data are in **wide form**.
First, convert data to **long form**:

```
N2dur1data<-data.frame(item=dataN2$Sentence,
                       subj=dataN2$Speaker_id,
                       cond="a",
                       dur=dataN2$N2_dur.1)
N2dur2data<-data.frame(item=dataN2$Sentence,
                      subj=dataN2$Speaker_id,
                      cond="b",
                      dur=dataN2$N2_dur.2)

N2data<-rbind(N2dur1data,N2dur2data)
```

CORRECT pair-wise t-test

```
write.table(N2data,file="N2data.txt")
  head(N2data)

##   item subj cond     dur
## 1    1    1    a 0.61444
## 2    1    2    a 0.58739
## 3    1    3    a 0.69451
## 4    1    4    a 0.56842
## 5    1    5    a 0.44040
## 6    1    6    a 0.54651
```

CORRECT pair-wise t-test

Now we can easily check that we have repeated measures by subject for each condition:

```
xtabs(~subj+cond,N2data)

##      cond
## subj  a  b
##    1  24 24
##    2  24 24
##    3  24 24
##    4  24 24
##    5  24 24
##    6  24 24
##    7  24 24
##    8  24 24
##    9  24 24
```

CORRECT pair-wise t-test

Notice that we can also easily check that we have repeated measures by **item** for each condition:

```
xtabs(~item+cond,N2data)

##      cond
## item  a  b
##    1  14 14
##    2  14 14
##    3  14 14
##    4  14 14
##    5  14 14
##    6  14 14
##    7  14 14
##    8  14 14
##    9  14 14
```

CORRECT pair-wise t-test

Then aggregate so that we have **only one data point per subject for each condition**:

```
N2data_bysubj<-aggregate(dur~subj+cond,mean,
                         data=N2data)
```

Check that we have one data point for each subject in each condition:

```
head(xtabs(~subj+cond,N2data_bysubj),n=2)

##      cond
## subj a b
##    1 1 1
##    2 1 1
```

CORRECT pair-wise t-test

Compare with the unaggregated data:

```
head(xtabs(~subj+cond,N2data),n=2)

##      cond
## subj  a  b
##    1 24 24
##    2 24 24
```

Example of CORRECT pair-wise t-test (by subject)

Create a vector for each condition:

```
conda<-subset(N2data_bysubj,cond=="a")
condb<-subset(N2data_bysubj,cond=="b")
```

Example of CORRECT pair-wise t-test (by subject)

Notice that the result is no longer significant

```
  ## not significant:
t.test(condb$dur,conda$dur,paired=TRUE)

##
##  Paired t-test
##
## data:  condb$dur and conda$dur
## t = 1.84, df = 13, p-value = 0.089
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  -0.0036046  0.0443300
## sample estimates:
## mean of the differences
##                0.020363
```

Example of CORRECT pair-wise t-test (by subj)

Alternative syntax:

```
## alternative syntax:
t.test(dur~cond,paired=TRUE,N2data_bysubj)

##
##  Paired t-test
##
## data:  dur by cond
## t = -1.84, df = 13, p-value = 0.089
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  -0.0443300  0.0036046
## sample estimates:
## mean of the differences
##               -0.020363
```

Example of CORRECT pair-wise t-test (by subj)

- There are many published papers in linguistics and psychology in which the data analysis ignores model assumptions and reports incorrect p-values.
- Some recent examples are reported in:
  Bruno Nicenboim, Timo B. Roettger, and Shravan Vasishth. Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German. Journal of Phonetics, 70:39-55, 2018.

Quick exercise: Do a by items paired t-test

Given these data (download from moodle):

```
  head(N2data,n=2)

##   item subj cond     dur
## 1    1    1    a 0.61444
## 2    1    2    a 0.58739
```

Do a by-items t-test.