The Correlated Sample t Statistic

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

The Correlated Sample t Statistic

Introduction

- In this module, we study a statistic that goes by a variety of names, including "repeated measures t," "correlated sample t," "paired sample t," and "matched sample t."
- Ultimately, we discover that this statistic is nothing more than the 1-sample t applied to a column of difference scores. So, in terms of mechanics and statistical theory, there is almost nothing new to learn!
- But let's start at the beginning. . .

Analyzing Repeated Measures

- The prototypical situation in which the correlated sample t is used is the *repeated measures design* in which measurements are taken on the same subjects on two occasions.
- Suppose, for example, five people are weighed in January, undergo a weight loss diet, and are weighed again in July. Suppose the data look like these.

	Time1	Time2
1	227	209
2	156	151
3	239	213
4	220	199
5	171	154

Analyzing Repeated Measures

- The first thing we notice scanning down the two columns is that the two columns are highly *correlated*.
- This means that a score that is high in one column tends to be high in the same position of the next column.
- This makes perfect sense here — a person who is very heavy at Time 1 will likely still be relatively heavy at Time 2.

Analyzing Repeated Measures

- It turns out, this causes a problem. The problem derives from the fact that the two columns are *no longer from independent samples*. If scores are correlated across columns, the standard error of the difference between means is no longer

$$\sigma_{M_1-M_2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{n_2}}$$

- Rather, the formula under the square root sign needs an additional term

$$\sigma_{M_1-M_2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{n_2} + Cov}$$

- This additional Cov term is called a "covariance."
- One way of fixing the t test in this situation so that it works correctly is to estimate this term from sample data. If we don't do that, the statistic will be incorrect because the denominator will be either too large or too small, depending on circumstances.
- We'll bypass that idea here, because it turns out there is a much simpler strategy we can use.

Analyzing Repeated Measures
A Simpler Way

- Ultimately, we are interested in the difference in means between the two columns of data, and whether that is statistically significant.
- Our null hypothesis is that the mean weight loss is zero.
- There is a simple but important rule in statistics: *The mean of the difference scores is equal to the difference between the means.*
- You can easily demonstrate using our sample data that this is true, and we can prove it in one line of summation algebra.

Analyzing Repeated Measures

- So suppose we reduce the two columns of weights in our example to one column of difference scores, called D.
- D represents weight loss in our study.

```
> D <- Time1 - Time2
> data.frame(Time1,Time2,D)
  Time1 Time2  D
1   227   209 18
2   156   151  5
3   239   213 26
4   220   199 21
5   171   154 17
```

- If the original two columns of numbers have a *bivariate normal distribution*, then the new column of difference scores will have a normal distribution.
- Moreover, recalling that the mean of the difference is the difference between the means, we have

$$\mu_D = \mu_{Time1} - \mu_{Time2}$$

.
- So we can test the null hypothesis that mean weight loss is zero by testing whether $\mu_D = 0$.
- But this is just a one sample t test on the difference scores!

The Correlated Sample t

- We just compute the difference scores (very easy to do in R), then compute the mean and standard deviation of the difference scores, and substitute in the formula below.
- Notice that this is really just the 1-sample t test that $\mu = 0$ with a slightly modified notation.

$$t_{n-1} = \frac{M_D}{s_D/\sqrt{n}} \tag{1}$$

- Note that the degrees of freedom are $n - 1$, where n is the number of pairs of scores, *not* the total n.

The Correlated Sample t

```
> n <- length(Time1)
> D <- Time1 - Time2
> M.D <- mean(D)
> s.D <- sd(D)
> df <- n - 1
> t <- sqrt(n)*M.D /s.D
> t.crit <- qt(.975,df)
> t

[1] 5.010437

> df

[1] 4

> t.crit

[1] 2.776445
```

The Correlated Sample t

- This is a very easy calculation, but we can easily automate it.
- Next to these lecture notes, I've posted code for the 1-sample t, and you can simply apply that to the difference scores to compute the correlated sample t.

The Correlated Sample t

```
> t.1.sample <- function(xbar,mu_0,s,n,alpha=0.05,tails=2)
+ {
+   ## compute t statistic
+   df <- n-1
+   t <- sqrt(n)*(xbar - mu_0)/s
+   # compute critical value
+   if(tails == -1) p <- alpha
+   if(tails == 1) p <- 1-alpha
+   if(tails == 2) p <- c(alpha/2,1 - alpha/2)
+   crit <- qt(p,df)
+ # create a list of named quantities and return it
+   res <- list(t.statistic = t, df = df, alpha = alpha,
+               critical.t.values = crit)
+   return(res)
+ }
> t.1.sample(M.D,0,s.D,5)

$t.statistic
[1] 5.010437

$df
[1] 4

$alpha
[1] 0.05

$critical.t.values
[1] -2.776445  2.776445
```