Unit #3: The effectiveness of a language course

Stefan Evert

9 July 2016

In this exercise, we evaluate the effectiveness of a corpus-driven language course for second language teaching. The `SIGIL` package includes a data set with the results of a simulated evaluation study.

```
library(SIGIL)
LC <- simulated.language.course()
knitr::kable(LC[seq(5, 95, 10), ])
```

	id	class	pre	post
5	CGQV	A	66	69
15	VX35	A	58	60
25	EQT5	B	78	83
35	DNY6	B	58	65
45	IWZ9	C	9	19
55	MPR6	D	71	67
65	PVX0	E	44	41
75	MT34	F	42	41
85	MW08	F	32	42
95	FRT1	G	43	40

Students from seven different classes (in different schools) took a standardized language test (`pre`), then worked with the language course for one month, then took another standardized test (`post`) at the same difficulty level as the first test. The data frame `LC` lists the scores obtained by each student in the two tests (with a maximum of 100 points) together with an anonymized personal code (`id`) and an anonymized label for the student's school (`class`).

To get an overview of the study, let us check how many students participated in the study and how many students there are from each school:

```
nrow(LC)        # number of students
```

```
## [1] 102
```

```
table(LC$class) # also shows there are seven schools
```

```
##
##  A  B  C  D  E  F  G
## 15 20 10 10 14 18 15
```

Comparing the means of independent samples

Using the results of the `pre` test, we can test whether the language skills of students differ between schools. For this purpose, each class is considered to be a random sample representative of the type of students attending this school. Our goal is to make inferences about the average test score μ achieved by such students.

Since the samples are drawn from different populations, it is appropriate to apply tests for two or more independent samples.

The t-test for two independent samples

Let us begin by comparing schools A and B:

```
A <- subset(LC, class == "A")
B <- subset(LC, class == "B")
```

We can use **Student's t-test** for two independent samples to compare the means of these two schools. The null hypothesis underlying this test is

$$H_0 : \mu_1 = \mu_2$$

where μ_1 is the average test score of (the kind of) students attending school A and μ_2 the average test score of those attending school B.

```
t.test(A$pre, B$pre)
```

```
##
##  Welch Two Sample t-test
##
## data:  A$pre and B$pre
## t = 3.387, df = 32.697, p-value = 0.001855
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   7.809056 31.324278
## sample estimates:
## mean of x mean of y
##  67.26667  47.70000
```

Student's version of the t-test for independent samples makes the (often unrealistic) assumption that the variances of test scores in both populations are equal, i.e. $\sigma_1^2 = \sigma_2^2$. The R implementation automatically applies a suitable correction (Welch 1947), which adjusts the number of degrees of freedom (df) in case they are not.

Q: Is there a reason to assume that σ^2 may differ between schools A and B? (hint: `?var.test`)

The t-test above yields a highly significant result ($p \approx .0019^{**}$), so we can reject H_0 with confidence. In order to interpret this result in a meaningful way, it is essential also to look at the effect size

$$\delta = \mu_1 - \mu_2$$

(recall that $H_0 : \delta = 0$). The `t.test` function also computes a 95% confidence interval for δ, which we can access directly in the data structured returned:

```
t.test(A$pre, B$pre)$conf.int
```

```
## [1]  7.809056 31.324278
## attr(,"conf.level")
## [1] 0.95
```

Students from school A score at least 7.8 points better on average than students from school B. The confidence interval also shows how much uncertainty there is in these two small samples: the true difference δ may be as large as 31.3 points.

Multiple comparisons

In principle, we could now apply multiple t-tests in order to make pairwise comparisons between all seven schools, which results in a total of $\binom{7}{2} = 21$ tests:

```
choose(7, 2)
```

```
## [1] 21
```

There is a fundamental problem in such multiple comparisons, though. If we're willing to reject H_0 for $p < .05$, we run a 5% risk of a type I error in each individual test. At this risk, one would expect one false positive among 21 tests (under the usual assumption that H_0 is true). The risk of committing one or more type I errors in the entire family of tests is thus much higher than the nominal significance level of 5%.

Statisticians speak of the **family-wise error rate** (FWER) for such multiple comparisons. If we assume the results of tests are independent from each other, we can work out the precise distribution of the number of type I errors. Each test is like throwing a coin, with the probability of a type I error being $\pi = .05$; the total number of such false positives among n independent tests then follows a binomial distribution $B(n, \pi)$:

$$\Pr(k) = \binom{n}{k} \pi^k (1-\pi)^{n-k}$$

For example, there is a chance of 37.6% of a single false positive in the family of tests, a chance of 19.8% that there are two false positives, etc.

```
round(dbinom(0:7, size=21, p=0.05), 3)
```

```
## [1] 0.341 0.376 0.198 0.066 0.016 0.003 0.000 0.000
```

The probability of committing no type I error at all is only 34.1%. By the same token, the FWER probability of at least one type I error is almost 66%!

```
1 - dbinom(0, size=21, p=0.05)
```

```
## [1] 0.6594384
```

> **Q:** You can also compute this tail probability directly with `pbinom`. Can you work out how?

In many applications, it is more important to control the FWER rather than the risk of a type I error in each individual test. This can be achieved by using a more conservative significance threshold α' in the individual tests in order to keep the FWER below the desired significance level α.

Assuming independence of the tests, we can work out the **Šidák correction** from the binomial distribution above:

$$\alpha_S = 1 - (1-\alpha)^{\frac{1}{n}}$$

In our case, the adjusted significance level is $\alpha_S \approx 0.244\%$ for FWER $\alpha = 5\%$.

```
alphaS <- 1 - (1 - .05) ^ (1 / 21)
alphaS
```

```
## [1] 0.002439557
```

Let us confirm that the correction works as expected:

```r
1 - dbinom(0, size=21, p=alphaS)
```

```
## [1] 0.05
```

The independence assumption made by the Šidák correction is often not valid, especially for pairwise comparisons. Assume, for example, that we obtain a sample of particularly good students from one of the seven schools by coincidence. How many false positives would we observe in this situation? Would you expect such a result under Šidák's independence assumption?

Unless there are good reasons to believe that individual tests are indeed independent from each other, more conservative corrections should be applied. A simple option is the **Bonferroni correction**

$$\alpha_B = \alpha/n$$

In practice, $\alpha_B \approx 0.238\%$ is only slightly smaller than α_S. The R function `p.adjust` implements more sophisticated stepwise procedures which take the actual p-value computed by each test into account. For pairwise t-tests, there is a pre-defined convenience function:

```r
pairwise.t.test(LC$pre, LC$class) # compare pre-test scores by school)
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  LC$pre and LC$class
##
##   A       B       C       D       E       F
## B 0.00386 -       -       -       -       -
## C 3.4e-10 0.00012 -       -       -       -
## D 0.91687 0.00020 4.3e-11 -       -       -
## E 0.68078 0.40587 4.1e-07 0.10041 -       -
## F 0.01627 1.00000 4.8e-05 0.00091 0.68078 -
## G 1.00000 0.04740 8.9e-09 0.41746 1.00000 0.13106
##
## P value adjustment method: holm
```

> **Q:** Which schools can be considered to be significantly different based on this result?

Analysis of variance

A second option is to avoid multiple comparisons altogether and carry out only a single test for the null hypothesis

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_7$$

> **Q:** What exactly does this null hypothesis entail? What can you conclude from a significant rejection?

Such a null hypothesis of multiple equality can be tested with a generalization of Student's t-test known as **analysis of variance** (ANOVA). Note that the R function for ANOVA is called `aov` rather than `anova` (which has a related, but different purpose). The `aov` function supports the convenient "formula" interface:

```
res <- aov(pre ~ class, data=LC)
summary(res) # need summary() to compute p-value
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## class        6  21281    3547   15.26 3.64e-12 ***
## Residuals   95  22088     233
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

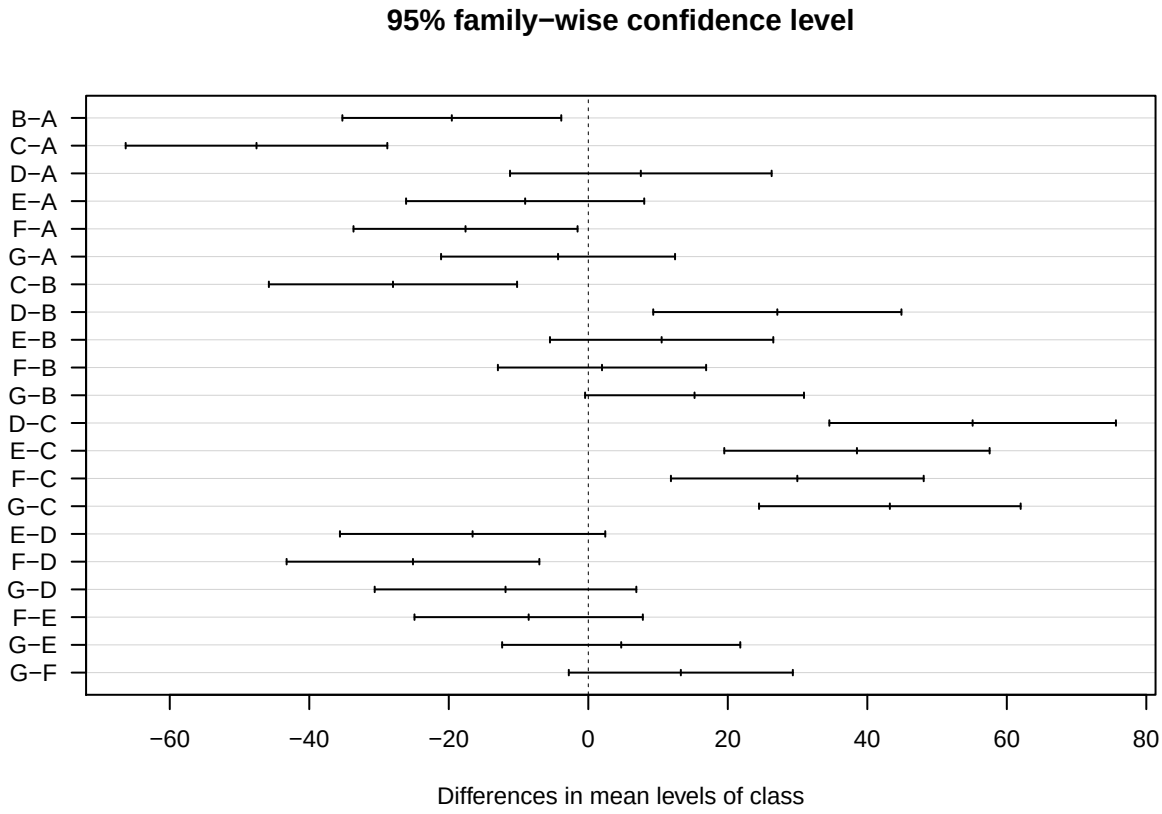
You should be able to spot the highly significant p-value in this output. The main problem of ANOVA is that it doesn't show us *where* the differences lie if H_0 has been rejected. For this purpose, a series of **post-hoc tests** have to be applied, e.g. Tukey's procedure of *honest significant differences* (HSD):

```
print(TukeyHSD(res), digits=3)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = pre ~ class, data = LC)
##
## $class
##       diff     lwr    upr p adj
## B-A -19.57 -35.256  -3.88 0.005
## C-A -47.57 -66.319 -28.81 0.000
## D-A   7.53 -11.219  26.29 0.889
## E-A  -9.05 -26.122   8.02 0.684
## F-A -17.60 -33.659  -1.54 0.022
## G-A  -4.33 -21.106  12.44 0.987
## C-B -28.00 -45.790 -10.21 0.000
## D-B  27.10   9.310  44.89 0.000
## E-B  10.51  -5.492  26.52 0.435
## F-B   1.97 -12.957  16.89 1.000
## G-B  15.23  -0.456  30.92 0.063
## D-C  55.10  34.558  75.64 0.000
## E-C  38.51  19.496  57.53 0.000
## F-C  29.97  11.850  48.08 0.000
## G-C  43.23  24.481  61.99 0.000
## E-D -16.59 -35.604   2.43 0.130
## F-D -25.13 -43.250  -7.02 0.001
## G-D -11.87 -30.619   6.89 0.481
## F-E  -8.55 -24.916   7.82 0.700
## G-E   4.72 -12.351  21.79 0.981
## G-F  13.27  -2.792  29.33 0.175
```

The HSD comparisons can also be visualized with a pre-defined `plot` method.

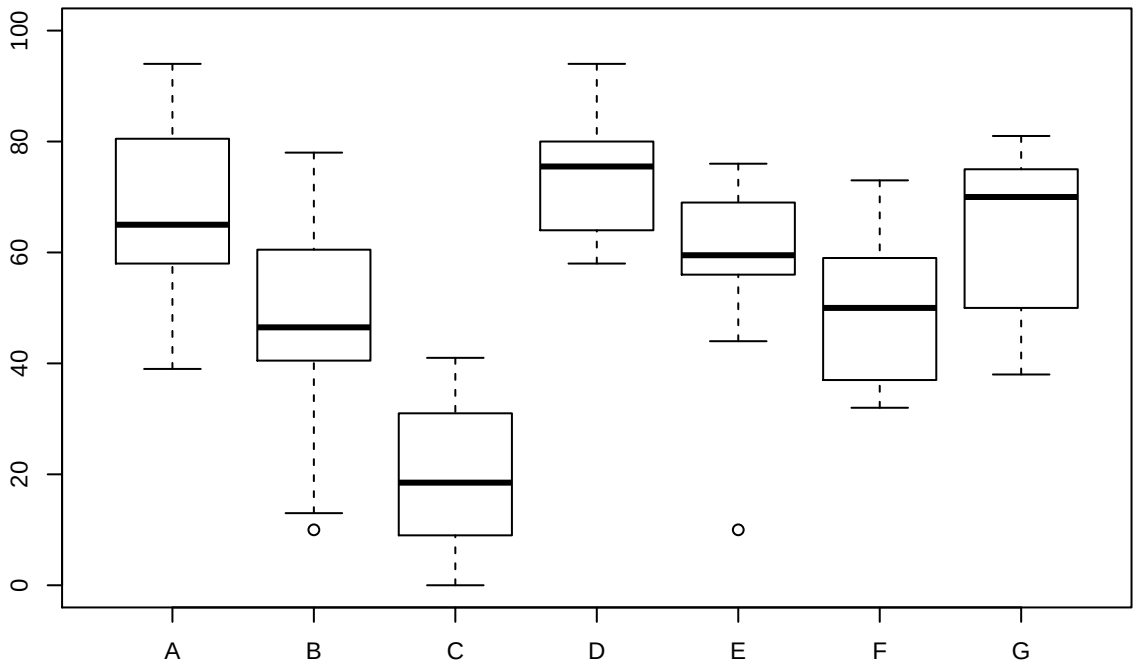
```
plot(TukeyHSD(res), las=1)
```



Q: Compare the HSD results to the output of `pairwise.t.test` above. Do the two procedures agree on which comparisons should be considered significant? Which approach seems more usefult to you?

When carrying ouy multiple comparisons, it is always a good idea to visualize the distributions in the observed samples with a side-by-side `boxplot` first, so it's easier to make sense of positive and negative effect sizes. You should also use the boxplots to check for individual outliers – e.g. a student who didn't finish his test – which might distort your results.

```
boxplot(pre ~ class, data=LC, ylim=c(0, 100))
```



Comparisons between dependent samples

Our main interest is to find out whether the language course has been effective, i.e. whether there is a significant improvement of test results from the pre-test to the post-test. One might be tempted to simply apply Student's t-test to the `pre` and `post` scores:

```
t.test(LC$post, LC$pre) # THIS IS WRONG!
```

```
##
##  Welch Two Sample t-test
##
## data:  LC$post and LC$pre
## t = 1.1099, df = 201.1, p-value = 0.2684
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.421338  8.656632
## sample estimates:
## mean of x mean of y
##  57.63725  54.51961
```

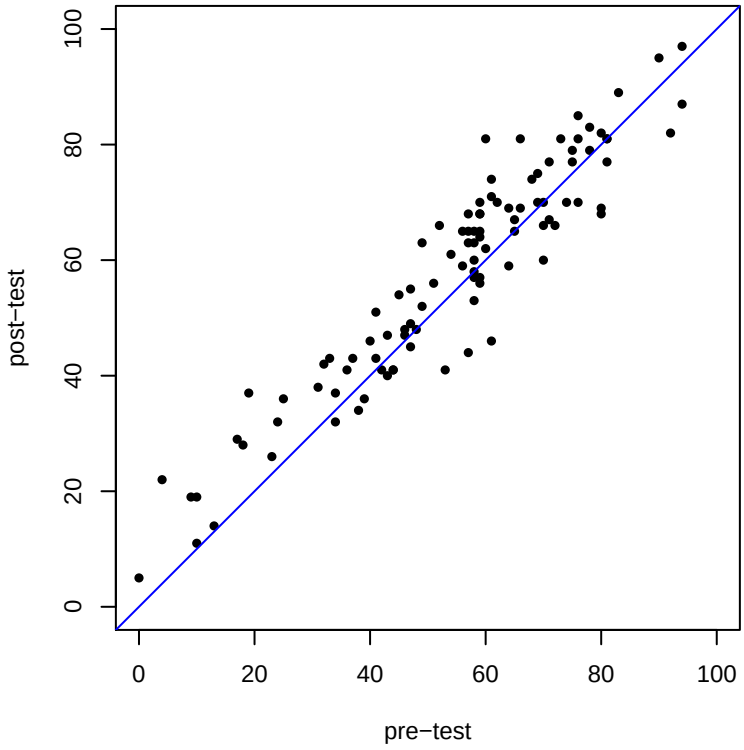
> **Q:** What is wrong with this approach? Can you explain why the test doesn't find a significant difference even though average scores in the post-test (57.6 points) are more than 3 points higher than in the pre-test (54.5 points)?

Comparing two dependent samples

The two-sample t-test assumes two independent samples from different populations, but here we have a single sample of 102 students with two "measurements" for each student. Our incorrect application of the t-test

takes the large variability of test scores betweeen schools and individual students into account, concluding that the observed difference can easily be explained by the random selection of students from the pre-test and post-test groups. In fact, however, a large part of the variability is due to the individual language skills of students. Most of the students improve between pre- and post-test, giving strong evidence for a positive effect of the course. This situation can be visualized with a scatterplot, where each point corresponds to a single student. Any student above the blue diagonal has achieved a personal improvement in the post-test.

```
plot(LC$pre, LC$post, pch=20,
     xlim=c(0, 100), ylim=c(0, 100), xlab="pre-test", ylab="post-test")
abline(0, 1, col="blue")
```



The plot above suggests that it is more meaningful to look at the differences between pre-test score x_i and post-test score y_i for each student i rather than comparing the x_i and y_i as independent samples:

$$d_i = y_i - x_i$$

We can now simply apply a one-sample t-test for $H_0 : \mu = 0$, i.e. that there is no change between pre- and post-test on average. This procedure is known as a **paired t-test**.

```
t.test(LC$post, LC$pre, paired=TRUE)
```

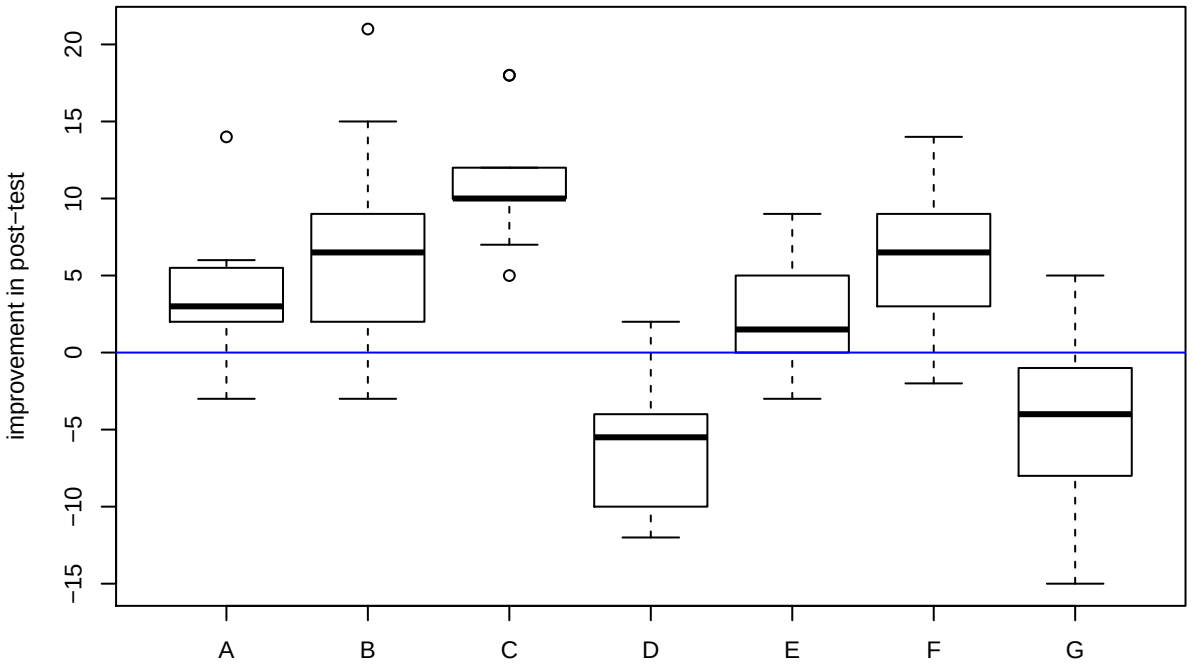
```
##
##  Paired t-test
##
## data:  LC$post and LC$pre
## t = 4.524, df = 101, p-value = 1.659e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.750575 4.484719
## sample estimates:
```

```
## mean of the differences
##                3.117647
```

The paired t-test yields a highly significant p-value $p \approx .000017^{***}$. The confidence interval $1.75 \leq \mu \leq 4.48$ shows that students improve by at least 1.75 points on average.

An interesting follow-up question would be whether the course was particularly effective in some of the schools or for certain groups of students. A boxplot of the differences d_i, grouped by school, gives a first indication:

```
boxplot((post - pre) ~ class, data=LC, ylab="improvement in post-test")
abline(h=0, col="blue")
```



Q: Can you work out whether the differences visible in the boxplot above are significant? Which test do you need for this purpose, and what are the data for the test?