Chapter 6: Comparing Two Populations

Introductory Statistics for Engineering Experimentation

Peter R. Nelson, Marie Coffin and Karen A.F. Copeland

Slides by Douglas Bates

Outline

6.1 Paired samples

Comparative experiments

- The "single sample" types of hypothesis tests described in the last chapter, where $H_0 : \mu = \mu_0$ or $H_0 : p = p_0$, are often used to compare a modified process to a standard method.
- Even if we reject H_0 in favor of H_a all we can really conclude is that the process has changed. We don't know if the change is due to our modification (the experimental factor) or due to other environmental factors.
- If we wish to focus on a particular modification it is better to use a *comparative experiment* in which we keep environmental factors (raw materials, time, temperature, etc.) as consistent as possible and change only the factor of interest.
- In this chapter we focus on comparative experiments where the experimental factor has only two levels. In statistical terms we are comparing two populations, corresponding to the two levels of the factor.

Overview of techniques

- When the data are on a continuous scale, we wish to compare the means, written μ_1 and μ_2 of the two populations, either by forming a confidence interval on $\mu_1 - \mu_2$ or testing the hypothesis $H_0 : \mu_1 = \mu_2$ versus a one- or two-sided alternative.
- If we have controlled for a known source of variability by taking, say, before-after measurements on the same subject, we consider the data as a set of n pairs, $(y_{1i}, y_{2i}), i = 1, \dots, n$ and analyze the differences $d_i = y_{1i} - y_{2i}$ as a single sample.
- For unpaired data we take the difference in the sample means, suitably standardized, and compare to a T distribution in which we approximate the degrees of freedom.
- For binary response data we compare the observed proportions $\hat{p}_1$ and $\hat{p}_2$ using a standardized statistic.

R functions used in this chapter

- Comparison of two samples on a continuous scale is done with `t.test`, as in the previous chapter. Paired samples are indicated by the optional argument `paired = TRUE`.
- We can always use the data in the “stacked” format where all the response measurements are in one column and there is a second column, a factor with two levels, that distinguishes the two samples.
- When the sample sizes are equal, $n_1 = n_2$, and especially for paired samples, the data are often available in an “unstacked” format. That is, the responses are in two columns.
- The `t.test` function is used in both cases but the form of the arguments is different. For stacked data we can use a formula/data specification.
- Comparison of two population proportions is done with `prop.test`. All the we need for this test are the sample sizes, n_1 and n_2, and the number of successes in each sample, y_1 and y_2.

Section 6.1, Paired samples

- The trick with paired samples is recognizing that the observations in the two samples are paired.
- Obviously, if they are to be paired you must have equal sample sizes, $n_1 = n_2$.
- There must also be some other factor (subject, location, raw material, etc.) that associates the first observation in sample 1 with the first observation in sample 2, and so on.
- A scatterplot of y_{2i} versus y_{1i} should show points scattered about a line with positive slope. If it doesn’t then the pairing is unsuccessful.
- To analyze the data we take the differences, $d_i = y_{1i} - y_{2i}, i = 1, \ldots, n$ and analyze them as a single sample. The hypothesis $H_0 : \mu_1 = \mu_2$ corresponds to $H_0 : \mu_d = 0$.
- In practice we can specify `paired = TRUE` in the call to `t.test` to have the data analyzed as a paired sample.

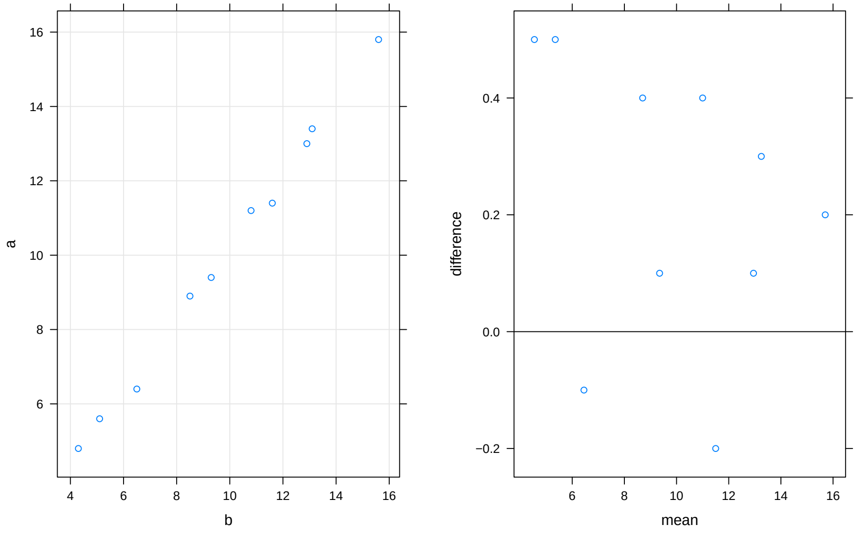
Example 5.1.1

- The `uvcoatin` data are from an experiment comparing the two UV coatings on lenses. Each pair of observations came from one pair of glasses worn by a person for 3 months, one lens with the new coating and one lens with the commercial coating.

```
> str(uvcoatin)
'data.frame': 10 obs. of  3 variables:
 $ a   : num  8.9 9.4 11.2 11.4 13 6.4 13.4 5.6 4.8 15.8
 $ b   : num  8.5 9.3 10.8 11.6 12.9 6.5 13.1 5.1 4.3 15.6
 $ diff: num  0.4 0.1 0.4 -0.2 0.1 -0.1 0.3 0.5 0.5 0.2
```

- We can plot these data as a scatterplot. Sometimes it is helpful also to rotate the scatterplot by 45^o, which corresponds to plotting the difference versus the mean. The `tmd` function automates this.

Plots of UV coating data



Paired t-test on UV coating data

```
> with(uvcoatin, t.test(a, b, alt = "g", paired = TRUE))

Paired t-test
data:  a and b
t = 2.8508, df = 9, p-value = 0.009533
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.07853456        Inf
sample estimates:
mean of the differences
                   0.22
```