1.1 **Price of high-speed train tickets**: the `renfe` data contains information about 10 000 train ticket sales from Renfe, the Spanish national train company. The data include:

- `price`: price of the ticket (in euros);
- `dest`: binary variable indicating the journey, either Barcelona to Madrid (`0`) or Madrid to Barcelona (`1`);
- `fare`: categorical variable indicating the ticket fare, one of `AdultoIda`, `Promo` or `Flexible`;
- `class`: ticket class, either `Preferente`, `Turista`, `TuristaPlus` or `TuristaSolo`;
- `type`: categorical variable indicating the type of train, either Alta Velocidad Española (`AVE`), Alta Velocidad Española jointly with TGV (parternship between SNCF and Renfe for trains to/from Toulouse and beyond) `AVE-TGV` or regional train `REXPRESS`; only trains labelled `AVE` or `AVE-TGV` are high-speed trains.
- `duration`: length of train journey (in minutes);
- `wday` integer denoting the week day, ranging from Sunday (`1`) to Saturday (`7`).

The goal of the analysis is to explore what factors influence the price of high speed trains. We consider travel time for high-speed (`AVE` and `AVE-TGV`) trains. The true "population" median travel time between cities is known to be $\nu = 2.833$ hours, whereas the true "population" mean is $\mu = 2.845$ hours (the instructor has access to the full dataset of more than 2.3 millions records, so these are known quantities, unlike in most practical settings).

A simulation study is performed to assess the behaviour of univariate tests under repeated sampling. The following algorithm was repeated 10 000 times

(a) Select a random subsample of size $n = 100$.

(b) Compute the one-sample t-test statistic for $\mathscr{H}_0 : \mu = \mu_0$ (versus $\mathscr{H}_0 : \mu \neq \mu_0$) for different values of μ_0.

(c) Compute the signed test for the bilateral test $\mathscr{H}_0 : \nu = \nu_0$ for different values of ν_0.

(d) Compute the Wilcoxon signed-rank test $\mathscr{H}_0 : \nu = \nu_0$ for different values of ν_0.

(e) Return the p-values of the three tests.

Note that both the signed test and Wilcoxon signed-rank test are tests for the **median**.

Figure 1 shows the percentage of the 10 000 p-values that are less than 0.05, i.e. the percentage of rejection (at the 5% level) of $\mathscr{H}_0 : \mu = \mu_0$ against the two-sided alternative at $\mu_0 \in \{2.83, \nu, 2.835, 2.84, \ldots, 2.995, 3\}$ (for the signed and signed-rank test, we are testing for the median at these values). Use the resulting power curve (Figure 1) for the three location tests to answer the following questions:

(a) Explain why the values of each test increase towards the right of the plot.

(b) Suppose we repeated the simulation study, but this time with subsamples of size $n = 1000$. How would the points compare for the one-sample t-test: should they be higher, equal, or lower than their current values?

(c) Explain why the value for the one-sample t-test around $\mu = 2.845$ **should be** approximately 0.05 (and similarly the value of the signed test and the signed-rank test **should be** approximately 0.05 around $\nu = 2.833$).

(d) According to Figure 1, how often would you reject the null hypothesis for the Wilcoxon signed-rank test at $\nu = 2.833$? Explain the consequences for your inference.

(e) Is the assumption of the one-sample t-test valid in this example? Produce a quantile-quantile plot and hence comment on the robustness of the t-test to departures from the normality assumption.

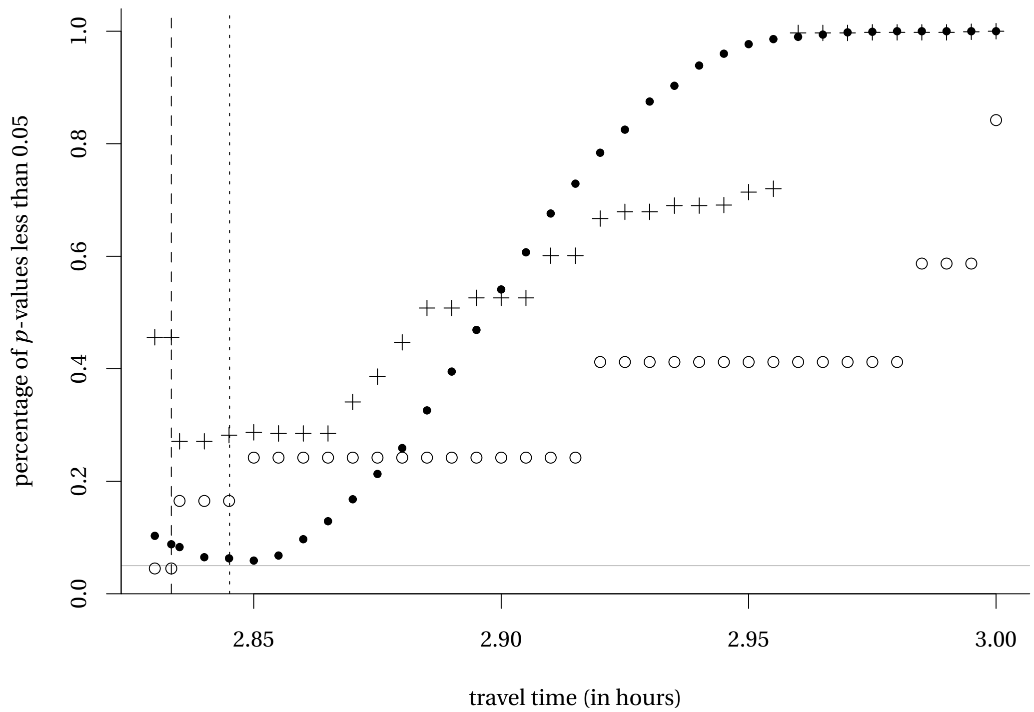

Figure 1: Power curve for three tests of location, either one-sample t-test (full dots), Wilcoxon signed rank test (crosses) or signed test (empty circles), as a function of travel time. The grey horizontal line is at 0.05, the dashed vertical line indicates the true median ν and the vertical dotted line the true mean μ.

Solution

(a) The curve is the power curve, i.e., the percentage of rejection of the null hypothesis for one-sample t-test. The further away from the true value μ, the higher the ability to detect departures from $\mathscr{H}_0$. Because we set $\alpha = 0.05$, the curve should be around 0.05 near μ and increase towards 1 as we move away from the true mean (or median).

(b) Power increases if n increases, so we expect to see the curve be higher everywhere, but at μ where it should be close to 0.05 if the test is calibrated; the value at μ for the one-sample t-test (respectively ν for the signed test and the signed-rank test) on the curve is the level α, here 5%.

(c) The data are clearly not normal and heavily discretized, yet the power curve of the one-sample t-test is steadily increasing and the nominal level matches the type I error. This illustrates the robustness of the test to departures from normality.

(d) The level of the test is 5%, so we should reject 5% of the time under the null.

(e) The type I error of the Wilcoxon signed-rank test is 0.44, very far from the (expected) nominal error rate of 0.05. Far from ν, the test behaves as expected (i.e., power increase far away from the true median ν), but the lack of symmetry and the presence of ties severely affects the conclusions of the test under $\mathscr{H}_0$, showing that nonparametric tests are not a panacea either. The consequence is a large Type I error.

1.2 Suppose we want to compare the mean fare for high-speed train tickets for the two destinations, i.e. Madrid to Barcelona versus Barcelona to Madrid. We run a simulation study where we perform a two-sided Welch test for this hypothesis repeatedly with random subsamples of size $n = 1000$. The data `renfe_simu` contains the mean difference (`meandif`), the test statistic (`Wstat`), the p-value (`pval`) and the confidence interval (`cilb` and `ciub`)

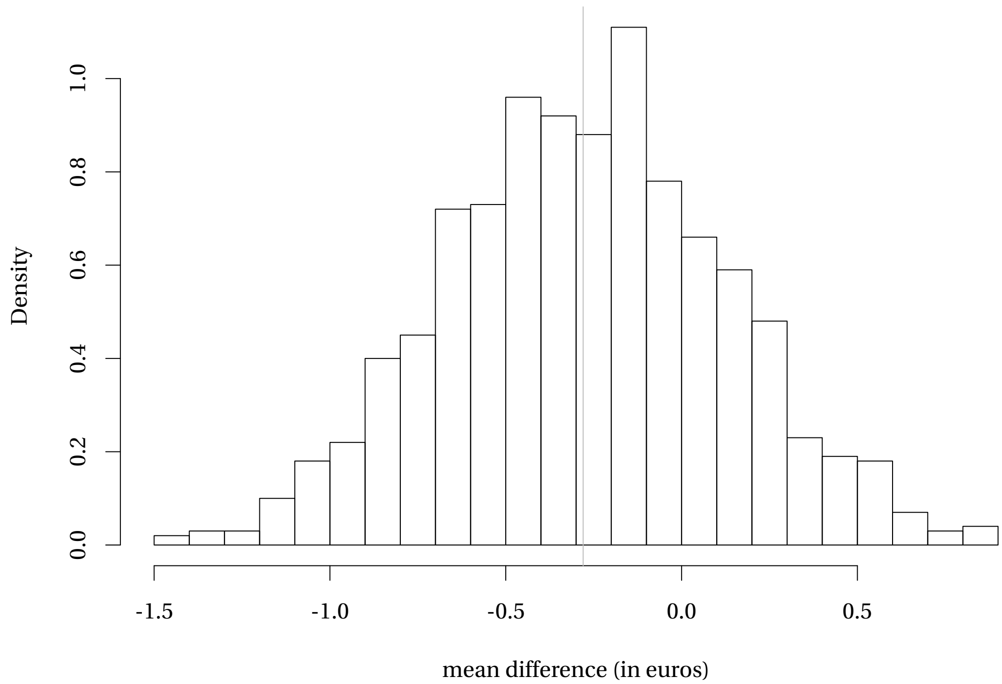

Figure 2: Histogram of the mean difference price for high-speed train tickets from Madrid to Barcelona versus Barcelona to Madrid, along with average (gray vertical line).

for these 1000 repetitions. Based on the entire database, the true mean difference is known to be -0.28€. Use the simulated data to answer the following questions and **briefly comment** on each item

(a) What is the empirical coverage of the 95% confidence intervals (i.e., the percentage of intervals covering the true mean difference value)?

(b) Plot an histogram of the mean differences and superimpose the true mean difference in the population.

(c) Compute the power of the test (percentage of rejection of the null hypothesis).

Solution

(a) The empirical coverage is 0.947. The coverage is not far from nominal coverage of 0.95, indicating the test is well calibrated.

(b) The histogram for the mean difference in Figure 2 looks normally distributed and centered around 0.28, whereas the p-values are scattered in the unit interval, with some values closer to zero.

(c) The power is 0.105. Under the alternative regime (since $\Delta = 0.28$€), we only reject 10.5% of the time. While this number is low, it is due to the small size of the true mean difference, which is hard to detect unless the sample size is enormous. The estimated mean difference for the sample is 0.274€.

1.3 Using the `renfe` data, test whether the average ticket price of `AVE-TGV` trains is different from that of Regio Express trains (`REXPRESS`). Make sure to

- State your null and alternative hypothesis.
- Carefully justify your choice of test statistic.
- Report the estimated mean difference and a 90% confidence interval for that difference.

- Conclude within the setting of the problem.

Solution

Careful here, as the price of the `REXPRESS` tickets is fixed at 43.25€. The only random sample is for the other class of train!

- The null hypothesis is $\mathscr{H}_0 : \mu_{\texttt{AVE-TGV}} = 43.25$€against the alternative $\mathscr{H}_1 : \mu_{\texttt{AVE-TGV}} \neq 43.25$€, where $\mu_{\texttt{AVE-TGV}}$ is the average `AVE-TGV` ticket price.
- Since this is a one-sample location problem, we use a one-sample t-test.
- The estimated mean difference is 45.63€= 88.88€−43.25€, with 90% confidence interval for the mean difference of $[44.14, 47.12]$.
- The t-statistic is 50.519 with 428 degrees of freedom, which has a negligible p-value.