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.

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`) 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).

Compiled 14/12/2020 at 15:51

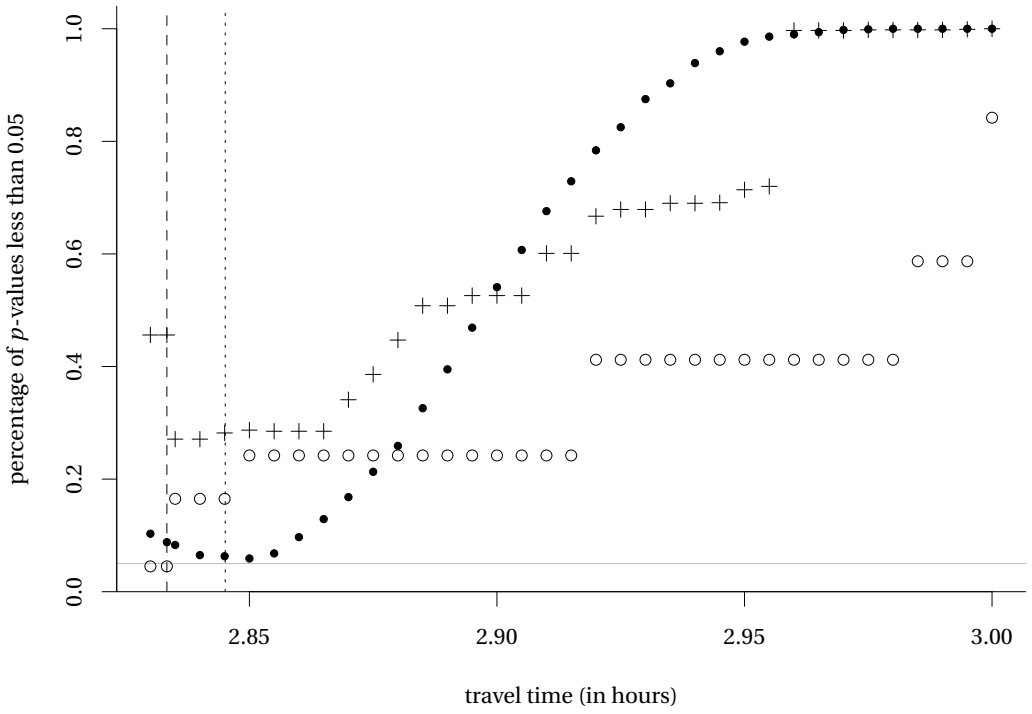

Figure 1: Power curve for three tests of location, either one-sample t-test (full dots), Wilcoxon signed rank test (crosses) or signed test (empty circles), as a function of travel time. The grey horizontal line is at 0.05, the dashed vertical line indicates the true median ν and the vertical dotted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

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.

1.4 The `insurance` data contains (simulated) records for health insurance premium in the United States for 1338 individuals, including information about
- `age`: age (in years)
- `sex`: sex (male or female)
- `bmi`: body mass index (in kg/m^2)
- `children`: number of dependent children
- `smoker`: logical indicator, `yes` if the person is a smoker and no otherwise
- `region`: geographical location with the US, one of `southwest`, `southeast`, `northwest` and `northeast`
- `charges`: individual medical costs billed by health insurance (in USD)

The data come from

> Lantz, Brett (2013), *Machine Learning with R*, Packt Publishing, 396p.

The World Health Organization (WHO) uses the classification of body mass index (BMI) presented in Table 1 to assess health status of population.

Use the `insurance` data to answer the following questions.

(a) Perform an explanatory data analysis of the `insurance` data: what are important features explaining medical costs?

Classification	BMI (kg/m^2)
< 18.5	Underweight
18.5–24.9	Normal range
25.0–29.9	Overweight
30.0–34.9	Obesity class I
35.0–39.9	Obesity class II and III

Table 1: International classification of adult underweight, overweight and obesity according to BMI (WHO)

(b) Do smokers pay on average the same charges as non-smokers? Justify your answer.

(c) Do obese smoker (i.e., `bmi` ≥ 30) face higher charges relative to other smokers? Give 90%, 95% and 99% confidence intervals for the mean difference between obese and non-obese smokers. Compare the intervals and explain how they change with varying confidence levels.

1.5 **Calculating the power of a statistical test**

The SAS program `power.sas` contains code by Rick Wicklin from SAS Institute Inc. to perform a simulation study for computing the power of the two-sample t-test, which is the same as that of the binary variable indicating group in a simple linear regression model.

(a) Briefly explain in your own words the steps of the simulation study.

(b) Plot and comment on the graph produced with parameters $n_1 = n_2 = 10$, $\sigma = 1$ and $B = 10\,000$.

(c) Vary the number of simulations from $B = 100$ to $B = 10\,000$. What do you notice when the number of simulations is small? Explain why this effect vanishes as the number of simulations increase.

(d) Modify the code so that the size of the groups is (a) $n_1 = 10, n_2 = 30$ and (b) $n_1 = 20, n_2 = 20$. In which of these two scenarios is the power highest and why?

(e) Modify the code to simulate $n = 20$ observations in each group from a normal distribution with different standard deviations, $\sigma_1 = 1$, $\sigma_2 = 5$ with $B = 100\,000$. Report the estimated level of the test along with the number of simulations and a 90% confidence interval. Explain how you derived the latter.

Compiled 14/12/2020 at 15:51