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.

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 empirical one. This illustrates the robustness of the test to

Compiled 14/12/2020 at 15:51

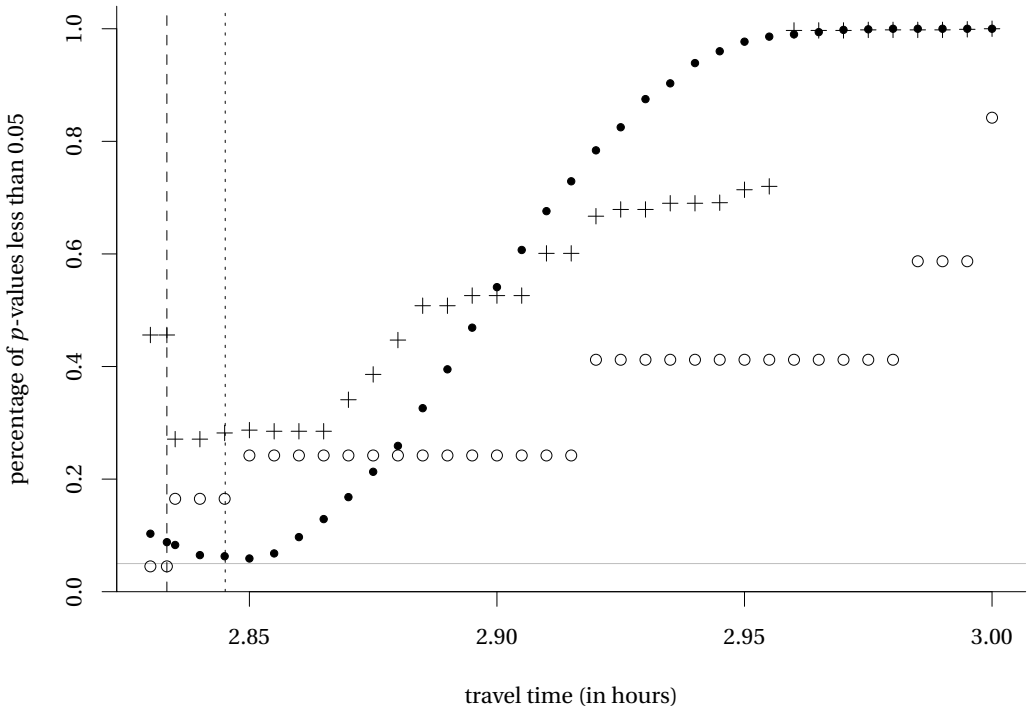

Figure 1: Power curve for three tests of location, either one-sample t-test (full dots), Wilcoxon signed rank test (crosses) or signed test (empty circles), as a function of travel time. The grey horizontal line is at 0.05, the dashed vertical line indicates the true median ν and the vertical dotted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

departures from normality.

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

(e) The empirical level 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`) 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

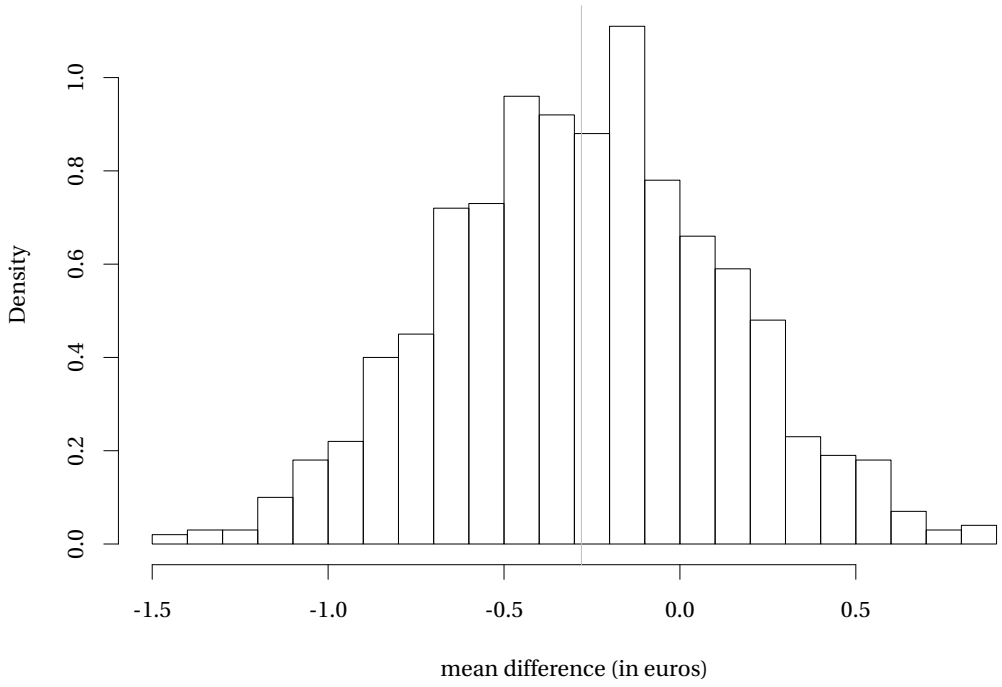

Figure 2: Histogram of the mean difference price for high-speed train tickets from Madrid to Barcelona versus Barcelona to Madrid, along with average (gray vertical line).

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.

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.

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)

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?

Solution

i. The distribution of `charges` is strictly positive, right skewed and asymmetric, meaning that the mean is larger than the median. We also note the presence of potential outliers. In contrast, the distribution of `bmi` is symmetric and centered around 30.

ii. Graphics presented in Figure 3 show a clear linear increase of charges with age, but evidence of three clusters, with higher heterogeneity for people facing higher charges. All individuals in the top payment category are smokers. The body mass index covariate is only important when combined with smoking status. There is a clear dichotomy smoker/non-smoker and a jump in charges for smokers (red) with a body mass index larger than 30kg/m^2, corresponding to obesity according to the WHO definition (bottom panel).

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

Solution

No, they pay significantly more as evidenced by Figure 3, which shows high heterogeneity and unequal variance; Levene's test confirms this finding, with a 95% confidence interval for the ratio of variance of $[0.22, 0.32]$ derived using Levene's test for equality of variance. We reject the null hypothesis that $\mu_S = \mu_N$ in favor of the alternative $\mu_S \neq \mu_N$, where μ_S (μ_N) is the average health medical costs (in USD) for smokers (non-smokers). Welch's test statistic is 32.75 and the p-value is less than 10^{-15}; even if the data are not normal, the conclusion is unequivocal given the sample size and the difference in mean of 23616$.

(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.

Solution

Yes, obese smokers face significantly higher medical costs than non-obese, as can be seen from the bottom right panel of Figure 3. We can assess this formally by testing $\mathscr{H}_0 : \mu_0 \leq \mu_1$ against the alternative $\mathscr{H}_1 : \mu_0 > \mu_1$,

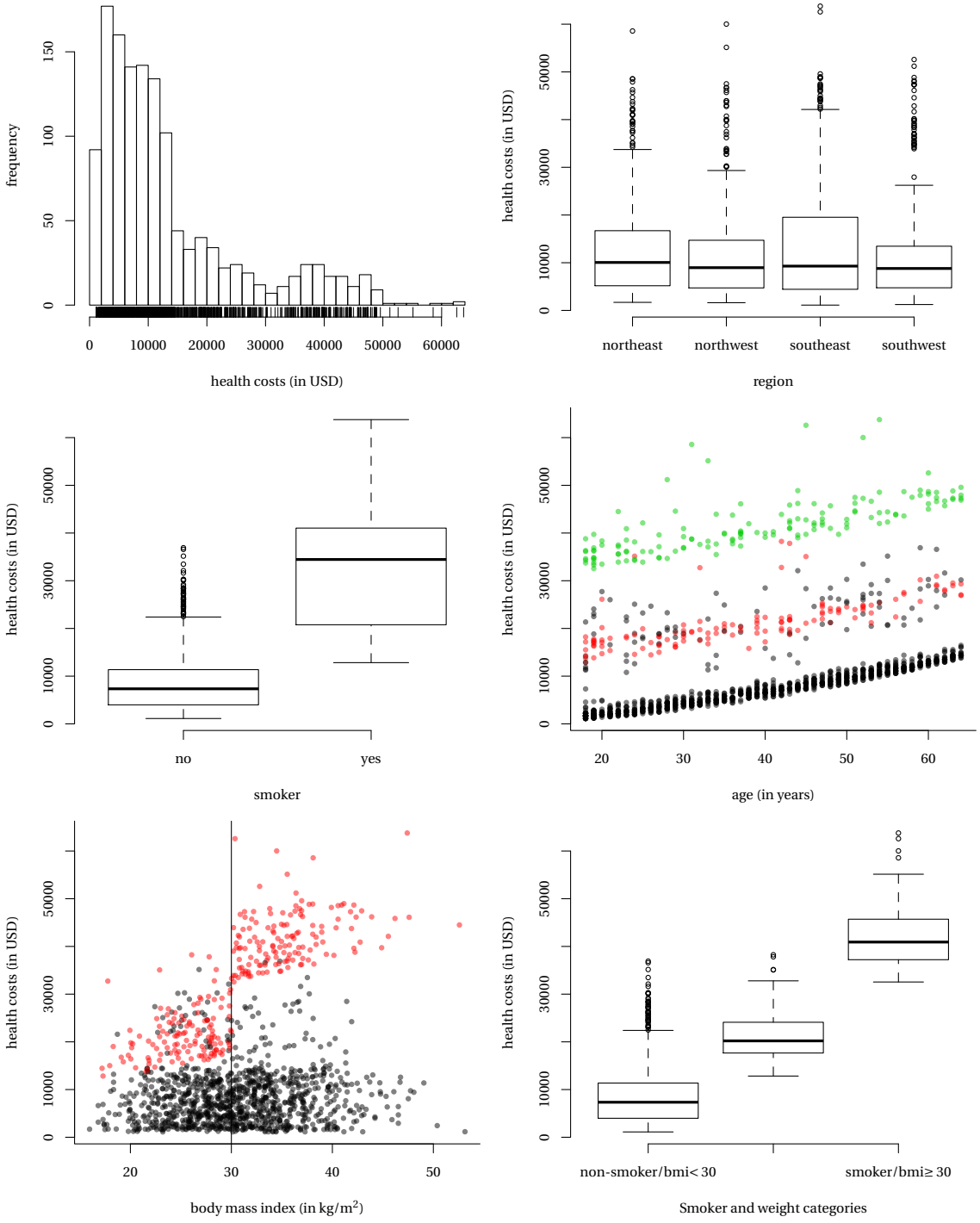

Figure 3: Histogram of annual medical costs in USD (top left), box-and-whiskers plots of charges per sex (top right) and per smoker status (middle left). Scatterplot of charges versus age (middle right), showing a clear linear trend for three separate clusters corresponding to non-smokers non-obese individuals (black), non-obese smokers (red) and obese smokers (green), also shown with boxplots (bottom right). Scatterplot of charges as a function of body mass index with smokers (red) and non-smokers (black); the vertical line indicates the obesity threshold (bottom left).

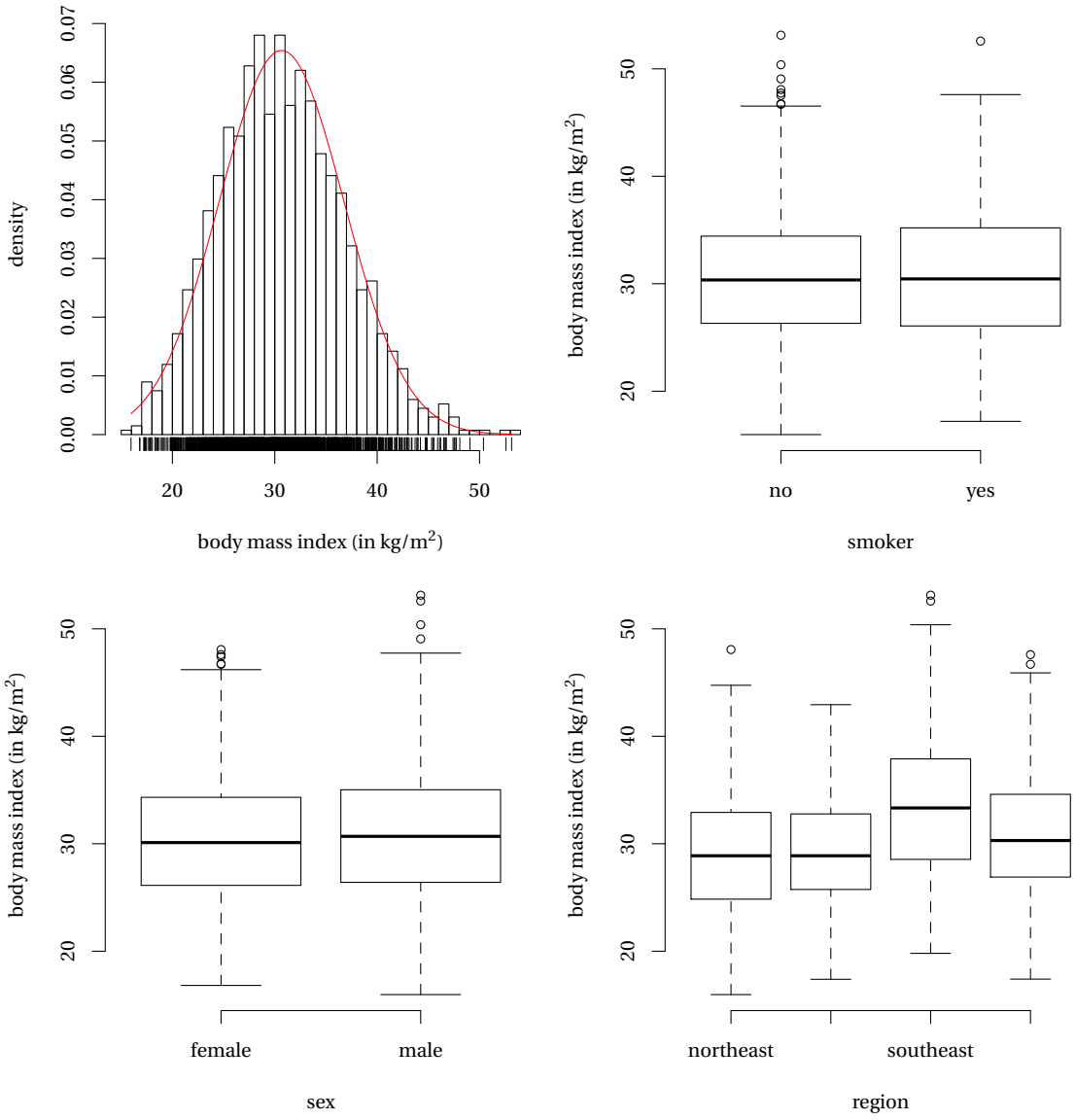

Figure 4: Histogram of body mass index (in kg/m^2) with rugs (top left), and box-and-whiskers plots showing interactions of body mass index with smoking status (top right), sex (bottom left) and region (bottom right).

where μ_0 (μ_1) is the average health charges for a obese (non-obese) smoker. The confidence intervals for the mean difference derived using Welch's test statistic for confidence level 90%, 95% and 99% are respectively $(-\infty, -19333.08)$, $(-\infty, -19087.71)$ and $(-\infty, -18625.11)$. The larger the significance level α, the larger the upper bound and the shorter the confidence interval, which are nested.

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.

Solution

The Monte Carlo study is used to compute the power of a two-sample t-test for samples $\boldsymbol{Y}$ and $\boldsymbol{Z}$ as a function of their mean difference. More specifically, we are interested in testing $\mathscr{H}_0 : \Delta = 0$ against two-sided alternatives, for different values of Δ. For each of these values of Δ, two independent samples of size $n = 10$ are drawn from normal distributions, with $Y_i \sim \mathsf{No}(\mu, 1)$ and $Z_i \sim \mathsf{No}(\mu + \Delta, 1)$. For each of these B replications, we simulate the sample, compute the t-test statistic, and then a binary indicator equal to 1 if we reject the null hypothesis at level 5% based on the null distribution, here $\mathsf{St}(n-2)$. We repeat this calculation B times for each Δ and return an estimate of the power, namely the proportion of rejection of $\mathscr{H}_0$: this is simply the sample mean of the binary variables.

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

Solution

The V-shape curve is symmetric around $\Delta = 0$. If $\Delta = 0$, the null hypothesis is true: we should obtain the level of the test, 5%, as percentage of rejection since all of the postulates are true, with normal homoscedastic data. The power increases with $|\Delta|$ up until about 1 when $|\Delta| > 2$.

(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.

Solution

We are trying to approximate the true proportion of rejection π_i for each value of Δ_i based on a sample from a binomial distribution with B trials; the sample mean is an unbiased estimator of π_i. With $B = 100$, there is more variability (and more so if we are close to 0.5 as the variance of the sample mean is $\pi_i(1-\pi_i)/B$). The curve obtained from $B = 100$ changes from one replication to the next, is discontinuous and fluctuates with Δ. With $B = 10\,000$, the curve is smooth and monotonically increases as we move away from zero.

(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?

Solution

The power is higher in the cases of balanced (i.e., equal) samples, even if the overall sample size, $n_1 + n_2 = 40$, is the same in both cases. The mean and variance estimators are unbiased, regardless of the sample size, because the data are sampled from homoscedastic normal distributions. The different powers are due to the pooled variance estimator,

$$S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

Since $1/10+1/30 = 4/30 > 3/30 = 1/20+1/20$, the power is higher in the second case (smaller variance, leading to higher statistics values for the same set of samples). This statement holds more generally provided all samples have the same variance.

(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.

Solution

Since each replication is independent and the probability of success is the same, the number of success follows a binomial distribution with B trials and success probability p. The number of success is random: with 100 000 simulations, I got an estimate of $\widehat{p} = 0.05512$, far from the postulated level 5%. A Wald-based 90% confidence interval is

$$\widehat{p} \pm \Phi^{-1}(0.95)\sqrt{\widehat{p}\cdot(1-\widehat{p})/B} = 0.05512 \pm 1.64485\sqrt{0.05512 \cdot 0,94488/100\,000} = [0.05280.0574]$$

We thus conclude, at level that the type I error is significatively different from the postulated level of the test (the null distribution is derived under the hypothesis of equal variance, so it is possible the approximation is off when the latter is false). We could fit a generalized linear model for the binomial distribution. The likelihood ratio test-based confidence interval for my sample is $[0.0537; 0.0565]$.

Compiled 14/12/2020 at 15:51