Experimental Design
Homework 1
Due May 08, 2023

- Discussion with fellow students during the preparation of the assignments does stimulate statistical thinking as well as its implementation. It is allowed to work in pairs, however, clearly indicate to whom the work belongs.

- Write your answer to each exercise in clear language. Include ONLY key R-output, such as tables, figures and analysis, necessary to answer the question. Also include key R-programming that you use to find your answer. Any output that is included need to be accompanied by a written interpretation.

- Deliver the pdf of your work into the dropbox of Assignment in the BrigthSpace Experimental Design course.

- Remark: The number of points are indicated in a box; ten are free.

1. Use the burning time data of Excel problem 2.46, see `Data.xlsx`, for two types of formulations.

   (a) $\boxed{2}$ The researcher involved is interested in the difference in means. Explain why there is no need to test for the equality of variances, if the t-test is used.

   (b) $\boxed{5}$ Test the null hypothesis that the means are equal.

   (c) $\boxed{5}$ Test the null hypothesis that the data is normally distributed for each type.
   Hint: Use the `shapiro.test`.

   (d) $\boxed{5}$ Test the null-hypothesis that the variances are equal.
   Hint: Use the `var.test`.

   (e) $\boxed{3}$ Use a visualisation to illustrate normality of the two types of measurement involved.
   Hint: Find a function in library `ggplot2` to plot the measurements of both types in a single plot.

2. Simulation study on the t-test. Suppose we want to design an experiment with measurements of Type x and of Type y. The measurements from Type x come from the $N(35, 3)$ distribution and those from type y from the $N(33, 3)$ normal distribution, where 33 is the mean and 3 the standard deviation. Obtaining the measurements is expensive, so the R& D manager wishes to know about the sample size in order to be 90% certain that the t-test correctly rejects the null hypothesis of equal means.

   (a) $\boxed{15}$ Conduct a simulation study assuming that you reject the null hypothesis of equal means if the p-value from the two-sided t-test is smaller than 0.05. Use sample sizes equal to $2, 6, \cdots, 100$.
   Hint: Use the function `rnorm,t.test` and expressions like `for (n in 2:100).`

   (b) $\boxed{5}$ Compare your result with a power computation. Use R its seek facility on its URL to find power computation functions.
   Note that the simulation is much more flexible to put ideas in!

SEE NEXT PAGE

3. From an experiment on the performance of Jet Turbine Engine so-called Trust data became available, see `table.b13` from the library `MPV`. The meaning of the variables stored as columns in the table are
y : Trust
x1 : Primary speed of rotation
x2 : Secondary speed of rotation
x3 : Fuel flow rate
x4 : Pressure
x5 : Exhaust temperature
x6 : Ambient temperature at time of test
Unfortunately, more detailed information on the experiment data, but there may be reasons for this exist due to interests of a company. The data are ver interesting from a mechanical engineering point of view. In the below we will analyze a larger an a smaller model.

```
Model 1 : y ~ x1 + x2 + x3 + x4 + x5  + x6
Model 2 : y ~ x1 + x5 + x6
```

   (a) $\boxed{5}$ Estimate the larger model and a 2 by 2 panel of diagnostic plots and evaluate these.

   (b) $\boxed{10}$ Compute the Cook's distances and evaluate whether any of these is larger than 0.5? Also compte the hat values h_i, $i = 1, \cdots, n$ and evaluate whether any of these is larger than $2p/n$, where $p = k + 1$, and k the number of predictors. Compute the variance inflation factors and evaluate whether any of these is larger than 10.
   Remark: An indicator for multicollinearity is the variance inflation factor defined as

$$\texttt{VIF} = C_{jj} = \frac{1}{1 - R_j^2},$$

   where R_j^2 is the coefficient of determination when x_j is regressed on the remaining $p - 1$ regressors, and C_{jj} is element jj from the $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. If the columns of $\boldsymbol{X}$ are orthogonal then, VIF=1. If the VIF larger than 10, then this is a clear indication that the coefficients are poorly estimated due to multicollinearity.

   (c) $\boxed{5}$ The library `car` has a very convenient function to perform a bootstrap on a linear model. Such gives an approach for constructing confidence intervals for the parameters without explicitly assuming normality for the error terms. Use 2000 replicates. Compare the 95% confidence interval from linear regression with those from the bootstrap.

   (d) $\boxed{5}$ Test the smaller model against the larger model by the F-test.
   Hint: Use the F-test from R its `anova` function.

   (e) $\boxed{10}$ Repeat the steps in (a) and (c) for the smaller model. Is the fit comparable to the larger model? What do you conclude?

SEE NEXT PAGE

4. The Feed Rate data in the Exec Problem 56 of Chapter 3 is also described in the book (e.g. Version 8 3.48).

   (a) $\boxed{5}$ Test the hypothesis of equal Feed Rate Standard Deviations by the F-test.
   Hint: Change Feed Rate from numeric into a factor.

   (b) $\boxed{5}$ Investigate the presence of outlying observations.

   (c) $\boxed{5}$ Compute the 95 percent confidence intervals for all comparisons of means. Compare the method Tukey HSD with Fisher LSD. Are the conclusions and the p-values identical?