Exercises 1 solutions

1. (a) Model: for the j^{th} observation in the i^{th} group (type I and type II):

$$Y_{ij} = \mu_i + e_{ij} \qquad (i = 1, 2;\ j = 1, \ldots, n_i)$$

or alternatively

$$Y_{ij} = x_{ij1}\mu_1 + x_{ij2}\mu_2 + e_{ij} \qquad (i = 1, 2;\ j = 1, \ldots, n_i)$$

with $x_{ij1} = 1(i = 1),\ x_{ij2} = 1(i = 2)$.
Assume $e_{ij} \sim N(0, \sigma^2)$ iid. $n_1 = n_2 = 10$.
$H_0 :\ \mu_1 = \mu_2,\ H_1 :\ \mu_1 \neq \mu_2$. Under H_0:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2},$$

where $\bar{Y}_1, \bar{Y}_2$ are the means of the observations in the two groups and

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

s_1^2 and s_2^2 being the sample variances of the groups.
Obtain $\bar{Y}_1 = 10.69,\ \bar{Y}_2 = 6.75,\ s_1^2 = 23.23,\ s_2^2 = 12.98,\ s_p^2 = 18.1,\ t = 2.07$. From the t_{18}-distribution get (from tables or computer) $p = 0.053$[1]. This means that there is modest but not strong evidence against H_0 (a Neyman-Pearson test with $\alpha = 0.05$ would just not reject H_0).

(b) Figure 1 shows some useful plots.
From the boxplots, no outliers can be detected. A difference between the two groups seems to be quite clear. The scatterplot of the points vs. the group number[2] shows that half of the points of type II are quite packed, while there are five points which are more scattered. (They are not indicated as outliers by the boxplot because 5 out of 10 are too many to be found by a boxplot and it's problematic to name them "outliers". The data could also be interpreted as consisting of two subpopulations with different variances. Note that the two points above 12 are about identical and it hard to see that it's actually two points.)
The equal variances assumption is unclear, particularly because of the strange pattern of the second group. The normality assumption seems to be fine for type I but the "five central points plus four or five outliers"-pattern is highlighted again in the Normal QQ-plot for type II. Therefore, the distribution of the type II data can be seen as a problem here. Note, however, that due to the small number of points quite a bit of variation can be expected in a QQ-plot even for normal data, so there is no conclusive evidence against the model assumptions here, although they look somewhat dodgy.
The overall summary "the groups seem to be different but the evidence is somewhat problematic and not very strong" is still a valid description of the situation.
Because of problems with normality, a Mann-Whitney test (Rice Sec. 11.2.3) may be seen as more appropriate[3]. It yields $p = 0.063$, not leading to any different conclusions.

[1] The R-command for this is t.test(V1,V2,var.equal=TRUE); by default, t.test carries out a more complicated test that doesn't assume equal variances but leads to about the same result for these data.

[2] It's often useful to see all points for one-dimensional data, ie, to produce a scatterplot or dotplot and not just histograms or boxplots.

[3] The R-command for this is wilcox.test; the test is also known under the name "Wilcoxon test".

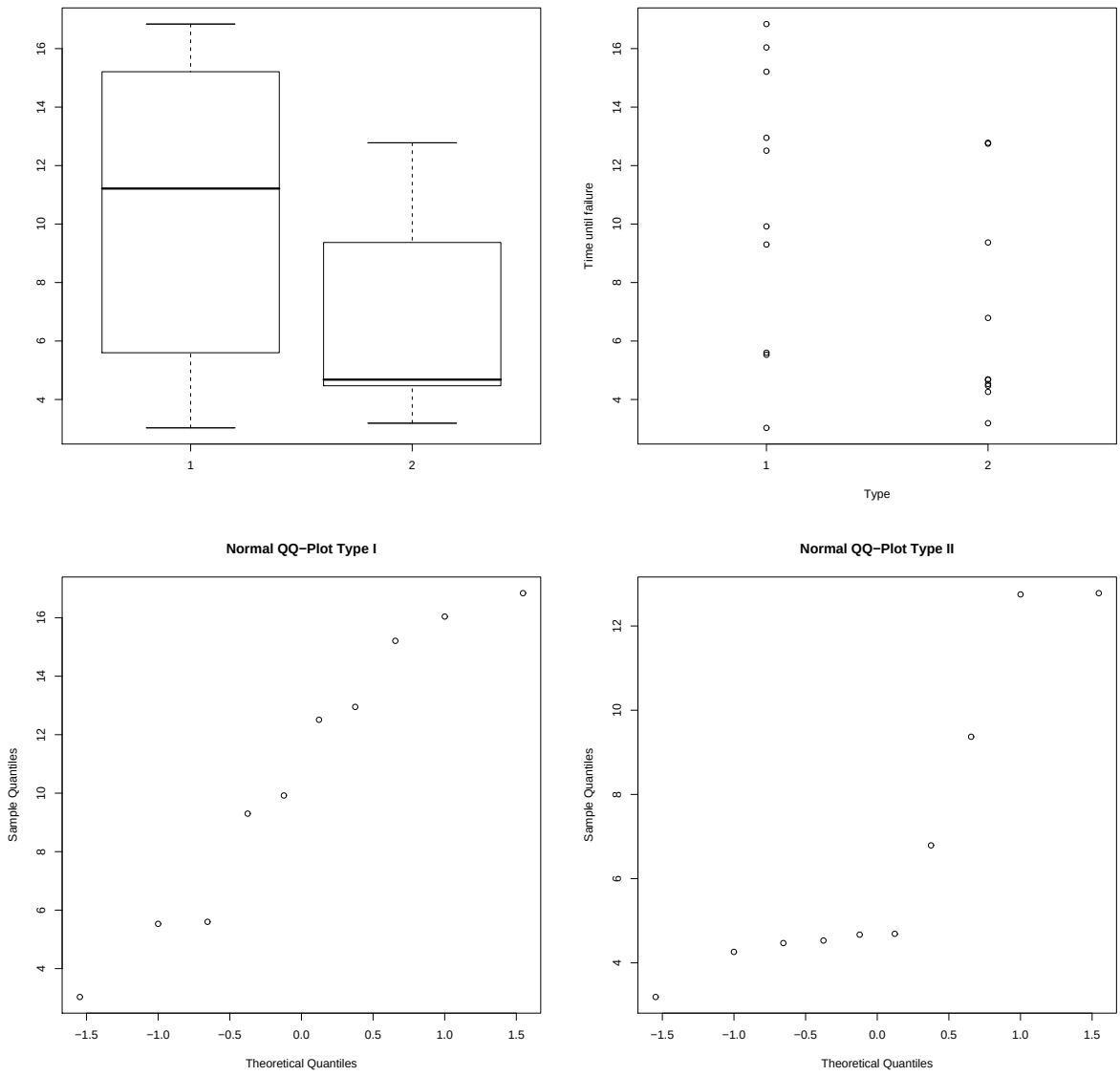

Figure 1: Diagnostic plots for engine data.

2. A few good examples were given, particularly for "explanation" (mostly from life/social sciences, such as connecting parental education to healthy eating). Note that it is always difficult to make causal statements from data that do not stem from controlled experiments (strictly establishing that smoking causes lung cancer would require forcing some random test persons to smoke; however if such experiments are not done, careful analysis of large and high quality observational datasets can provide evidence, too).

3. With two explanatory variables, the linear regression model is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i \qquad (i = 1, 2, \ldots, N).$$

(i) The sum of squares of the errors is

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2$$

Partial differentiation of $S(\boldsymbol{\beta})$ with respect to β_0 gives the first equation shown in Rice.

Partial differentiation of $S(\boldsymbol{\beta})$ with respect to β_1 and with respect to β_2 gives the equations for $k = 1$ and 2 shown in Rice.

(ii)

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & x_{N1} & x_{N2} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \qquad \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_N \end{pmatrix}$$

Then substitute these in $\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}$ and the resulting three equations are the same as those obtained from part (i).

4. Let X and Y denote the two random variables with the notation for their means, variances and correlation as in Rice. In matrix notation:

$$\mathbf{y} = \begin{pmatrix} x \\ y \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \ \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix},$$

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{(1-\rho^2)\sigma_X^2\sigma_Y^2} \begin{pmatrix} \sigma_Y^2 & -\rho\sigma_X\sigma_Y \\ -\rho\sigma_X\sigma_Y & \sigma_X^2 \end{pmatrix} = \frac{1}{1-\rho^2} \begin{pmatrix} \frac{1}{\sigma_X^2} & -\frac{\rho}{\sigma_X\sigma_Y} \\ -\frac{\rho}{\sigma_X\sigma_Y} & \frac{1}{\sigma_Y^2} \end{pmatrix}$$

Hence after some algebra you find that the alternative pdf,

$$\frac{1}{2\pi\sqrt{\det \boldsymbol{\Sigma}}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\},$$

equals the pdf in Rice section 3.3 example D *(try this if you haven't done so already!).*