| VU University | Statistical Data Analysis |
|---|---|
| Faculty of Sciences | 2 July 2015 |

# SOLUTION

## Question 1 [8 points]

a. [2 points] Incorrect. A two sample $QQ$-plot contains the matching order statistics of two samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, whereas a scatter plot of a bivariate sample plots the points $(X_i, Y_i)$ of a bivariate sample $(X_1, Y_1), \ldots, (X_n, Y_n)$.

b. [2 points] Incorrect, the mean is not robust, since its influence function is unbounded, whereas the a trimmed mean is robust.

c. [2 points] Incorrect. M-estimators can be not robust, like the mean, which is an M-estimator with $\psi(x) = x$.

d. [2 points] Incorrect. One cannot consider (rank) correlation of two independent samples, one can only consider correlation in bivariate data.

## Question 2 [8 points]

a. [2 points] From the histogram we see that $x$ seems to have a decreasing right tail, whereas $y$ seems to have no right tail, like a uniform distribution. Therefore, the tail of $x$ is heavier, since $y$ seems not to have any tail. This can be seen from the $QQ$-plot as well, since the dots in the top right corner of the $QQ$-plot tend towards a horizontal line, indicating more probability mass in the right tail for the distribution on the horizontal axis ($x$). (*Note: this question was harder than average, and graded very lenient*)

b. [3 points]
*i*) Kolmogorov-Smirnov test is suitable, since we want to test a simple null hypothesis.
*ii*) chi-square test for goodness of fit is suitable, since we want to test a simple null hypothesis. One has to specify the intervals for this test according to the rule of thumb.
*iii*) Shapiro-Wilk test is not suitable, since we want to test a simple null hypothesis, not the compositie hypothesis of all normal distributions.

c. [3 points]
*i*) Wilcoxon rank sum test is suitable although it is most powerful for detecting shifts between distributions, which is not the case here.
*ii*) Kendall's rank correlation test is not applicable for testing whether the underlying distributions are equal
*iii*) Two sample Kolmogorov-Smirnov test is suitable, since is can be used to test any difference between the underlying distributions.

**Question 3 [7 points]**

a. [2 points] One error is due to estimating the underlying distribution $P$ of the data by some estimate $\tilde{P}$ (either the empirical distribution of the sample or an estimated parametric distribution). The other error is due to estimating the distribution of $T_n$ by the empirical distribution of the bootstrap values $T_1^* \ldots, T_B^*$, with $B$ the number of bootstrap values.

b. [2 points] The error mentioned second can be made arbitrarily small by increasing $B$, the number of bootstrap values for $T_n$.

c. [3 points] It is incorrect. In step 4 the distribution $F$ is chosen incorrectly. It should be the normal distribution with the sample mean and sample variance of the $X^*$-sample as parameters.

**Question 4 [8 points]**

a. [1 point] Define $m$ as the median of the underlying distribution. $H_0 : m \leq 33.00$ versus $H_1 : m > 33.00$.

b. [2 points] The test statistic is $T = \#(X_i > 33.00)$ which has the binomial distribution with $n = 12$ and $p = 0.5$ if $m = 33.00$.

c. [3 points] From the data we find: $T = t = 8$, and no observations are equal to 33.00. We compute the one-sided $p$-value, because $H_1$ is one-sided.
$p_r = P_{p=0.5}(T \geq 8) = 1 - P_{p=0.5}(T \leq 7) = 1 - 0.806 = 0.194 > 0.05$.
Hence, we do not reject the null hypothesis; there is not enough evidence to state that the machine puts too much beer in the bottles.

d. [2 points] If a test statistic $T$ is distribution free over the class of distribution functions under $H_0$, then its distribution under $H_0$ does not depend on specifically which distribution under $H_0$ is the true distribution.

**Question 5 [4 points]**

a. [3 points] Define $n$ the sum of counts in the entire table. The test statistic for the chi-square test for contingency tables is

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(N_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}}$$

with $\hat{p}_{ij} = N_{\cdot j} N_{i\cdot}/n$ the estimated cell probability for cell $(i, j)$. $N_{\cdot j} = \sum_{i=1}^{r} N_{ij}$, and $N_{i\cdot} = \sum_{j=1}^{c} N_{ij}$. Under the null hypothesis of independence $X^2$ has approximately the $\chi^2_{(r-1)(c-1)}$-distribution.

b. [1 point] The estimated expectation in cell $(i, j)$, $EN_{ij} = n\hat{p}_{ij}$ should be larger than 1 in all cells, and larger than 5 in 80% of the cells.

**Question 6 [10 points]**

a. [3 points] The multiple linear regression model is:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i, \qquad i = 1, \ldots, 25,$$

with

| | | |
|---|---|---|
| $Y_i$ | – | $i^{th}$amount of nicotine in cigarette $i$ |
| $x_{i1}$ | – | amount of tar in cigarette $i$ |
| $x_{i2}$ | – | amount of carbon monoxide in cigarette $i$ |
| $\beta_0$ | – | intercept |
| $\beta_1$ | – | parameter denoting the influence of tar on nicotine |
| $\beta_2$ | – | parameter denoting the influence of carbon monoxide on nicotine |
| $e_i$ | – | error in $i^{th}$ observation |

Furthermore, we assume $e_i \sim N(0, \sigma^2)$ i.i.d. with $\sigma^2$ the unknown error variance.

b. [3 points]
A *leverage (potential) point* is a point with an outlying value in one of the explanatory variables. Hatvalues can be used to find leverage points.
An *influence point* is a point with an outlying value in one of the explanatory variables, that has a large influence on $\hat{\beta}$. Influence points can be found by computing Cook's distances.
*Collinearity* is the problem of having (nearly) linear dependency amongst columns in the design matrix. Tools to detect collinearity are variance inflation factors, condition indices and variance decomposition proportions.

c. [2 points] This model is the same as the model under part (a) expect for $i = 6$. For the $6^{th}$ observation we model

$$Y_6 = \beta_0 + \beta_1 x_{61} + \beta_2 x_{62} + \delta + e_6$$

with $\delta$ the shift of the mean.
(Alternatively, we can model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 \mathbf{1}_{i=6} + e_i, \qquad i = 1, \ldots, 25,$$

where now $\beta_3$ is the shift of the mean for the $6^{th}$ observation.)

d. [2 points] The significance of $\delta$ can be tested with a $t$-test. $H_0 : \delta \leq 0$ versus $H_1 : \delta > 0$. The test statistic is

$$T = \frac{\hat{\delta}}{\widehat{Cov}(\hat{\beta}^*)_{4,4}}$$

where $\beta^* = (\beta_0, \beta_1, \beta_2, \delta)$. $T$ has the $t_{25-4} = t_{21}$-distribution under $\delta = 0$. $H_0$ is rejected for $T > t_{21;1-\alpha}$.