Exam Statistical Data Analysis 30.06.2016

1. 2

a.) Dataset x: median: 3
location ($\frac{1}{2}$)
min: 1
scale ($\frac{1}{2}$)
max: 13.5
IQR: 3.5
shape ($\frac{1}{2}$)
skewed to right
extremes ($\frac{1}{2}$)
3 extremes right

2

b.) (1) x: median < mean, because large values rise the mean
(1) y: median > mean, because low values lower the mean.

3

c.) Best option: KS-test, regarding the different shapes. ($1\frac{1}{2}$)
2^{nd} option: median test, that doesn't assume any shape ($1\frac{1}{2}$)

2. 2

a) (1) choice
Exponential family. The QQ-plot against exp(1) is closest to a straight line
(1) argument

2

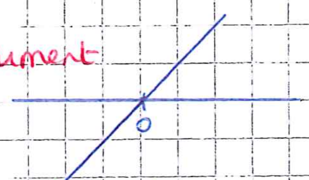
b.) points on the line $y = a + bx$: $(0, 8)$ and $(3, 14\frac{1}{2})$.
(1) a, (1) b
$a = 8$, $b = \frac{6.5}{3} = 2.67$ a is location, b is scale.

2

c.) The smaller the sample size, the less informative a QQ-plot due to higher variability around the line. This is a rather small sample, so we cannot be confident at all. Even the normal QQ-plot is within the common variability of QQ-plots for normal samples of this size. (2) correct ~~...~~ reasoning

3.

"correct"
($1\frac{1}{2}$) argument
a) ($\frac{1}{2}$) Correct, it looks like:



nonsensical
b.) ($\frac{1}{2}$) Condition indices are irrelevant for leverage points. ($1\frac{1}{2}$) argument
Cook's distances are indeed useful for potential leverage points

c.) ($\frac{1}{2}$) Incorrect. Power depends on the underlying distribution and in what way that distributions deviates from H_0. Sometimes KS is better, sometimes χ^2-test is better. ($1\frac{1}{2}$) argument

4. 2

a.) (1) $T = \#(X_i > 10) \overset{m=10}{\sim} bin(15, \frac{1}{2})$ (1)

2

b.) ($\frac{1}{2}$) for $T=4$
$T = t = 4$ is observed. $P(T \leq 4) = 0.059$ (1) correct p-value

Conclusion: H_0 is not rejected (one-sided p-value = 0.059)
($\frac{1}{2}$) correct conclusion based on found p-value

this should be in the answer

4. c.) A t-test (one-sided or two-sided) assumes normality.
(1) That is an invalid assumption here, because of the QQ-plot in figure 2. (Although the sample size is small, it is better not to use it).

"symmetry" has to be in the answer

d.) The Wilcoxon signed rank test is for symmetric distributions.
(1) That is not appropriate, regarding the histogram in fig. 2.

5. a.) One (1) sample of size $n=178$. One multinomial distribution with
(3) parameters $(178, p_{11}, p_{12}, \dots, p_{33})$ with $\sum_{i,j} p_{ij} = 1$. (1) correct parameters
(1) for H_0
$H_0: p_{ij} = p_{i\cdot}\, p_{\cdot j}$ with $p_{i\cdot} = \sum_j p_{ij}$, $p_{\cdot j} = \sum_i p_{ij}$.

b.) We need $EN_{ij} > 1$ in all entries, and $EN_{ij} > 5$ in 80% of the entries.
(2) (1) definition of rule
$EN_{ij} = \frac{N_{i\cdot} N_{\cdot j}}{n}$. The lowest one is $i=1, j=3$. $EN_{13} = \frac{45 \cdot 24}{178} = 6.1$ (1) checking rule
So the rule of thumb is fulfilled, all $EN_{ij} > 5$.

(1) c.) Compute the normalized contributions and compare them to $\Phi^{-1}(\frac{\alpha}{2})$ and $\Phi^{-1}(1-\frac{\alpha}{2})$ to find deviating cells.

(1) d.) Use a bootstrap test, with a sensible test statistic, which could be the χ^2-statistic again.

6. a.) Denote T_i^* for the i^{th} bootstrap value for T, $i=1,\dots,B$ ($\frac{1}{2}$) notation
(2) The interval is $\left(2T - T^*_{\lfloor(1-\alpha)B\rfloor+1},\ 2T - T^*_{\lfloor\alpha B\rfloor+1}\right)$ ($1\frac{1}{2}$) interval

b.) MAD is more robust, and varies less. Sample sd is influenced
(2) (2) for correct answer
a lot by the extremes. So the smaller interval [122, 333] is for MAD

c.) Due to the outliers, and regarding the length of the bootstrap
(2) intervals, the MAD is to be preferred. It is more accurate.
(2) correct answer

7. (3) a) (2) + b). see previous exam.

c) Leverage point is "single": use hatvalues, Cook's distances (1)
(2) Outlier point is "crime": use mean shift outlier model. (1)