Vrije Universiteit Amsterdam	Statistical Data Analysis, Exam II
Faculty of Sciences	June 1, 2022

Use of a basic calculator is allowed. Graphical calculators, laptops, e-readers, mobile phones, smartphones, smartwatches etc. are not allowed. This exam consists of 3 questions on 4 pages (27 points). You have 2 hours to write the exam.

Please write all answers in English. Grade = $\frac{total+3}{3}$**.** **GOOD LUCK!**

Question 1 [2+2+2+2+1=9 points]

a. Explain in your own words (and without the use of formulas) what it means when for a certain distribution of the data the asymptotic relative efficiency of the one-sample Wilcoxon test with respect to the sign test equals 2.

b. Explain why the two-sample Wilcoxon test is nonparametric (or distribution-free) under the null hypothesis.

c. Write down the formula for the two-sample Kolmogorov-Smirnov test statistic, and briefly explain all of your notation.

d. Explain and motivate at least one advantage of Spearman's rank correlation test over Pearson's correlation test, i.e. the one based on the sample correlation.

e. State one advantage of permutation tests over bootstrap-based tests.

Question 2 [2+2+2+2=8 points]
In a study about the relationship between the popularity of board games and snack preferences, 92 people were asked which of the following three games and snacks they prefer: Poker, Trivial Pursuit, Monopoly, and pizza rolls, chips and dip, cookies, respectively. Table 1 below presents the counts of their answers.

game \ snack	pizza rolls	chips and dip	cookies	total
Poker	10	3	12	25
Trivial Pursuit	8	14	7	29
Monopoly	14	17	7	38
total	32	34	26	92

Table 1: Numbers of observations in different categories.

a. Formulate a suitable model of multinomial distribution(s) and state the corresponding null and alternative hypotheses about the categories "preferred game" and "preferred snack" that can be tested with such kind of data.
You may formulate the hypotheses in words or in formulas.

The test for hypotheses from part a. can be based on the test statistic

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

which has approximately a chi-square-distribution under H_0.

b. In order to use the above-described asymptotic chi-square-distribution as an approximation of the distribution of the test statistic under the null hypothesis, a certain rule of thumb should be satisfied. Explain what this rule of thumb is, and check whether it is satisfied for the present data.

c. Irrespective of your conclusion in part b., test the hypotheses from part a. at the significance level $\alpha = 5\%$ with the help of the chi-square test. Use that the value of the test statistic is $X^2 = 11.438$ for the present data, and choose a suitable critical value from Table 2 below. What is your conclusion?

Whenever the chi-square approximation is not reliable because the rule of thumb is violated, a certain bootstrap method can be used to find reliable p-values.

d. Explain the mechanism behind the bootstrap method for contingency tables, i.e. how the bootstrap randomly generates new contingency tables.

	γ						
k	0.025	0.050	0.330	0.500	0.670	0.950	0.975
1	0.00	0.00	0.18	0.45	0.95	3.84	5.02
2	0.05	0.10	0.80	1.39	2.22	5.99	7.38
3	0.22	0.35	1.55	2.37	3.43	7.81	9.35
4	0.48	0.71	2.36	3.36	4.61	9.49	11.14
5	0.83	1.15	3.19	4.35	5.76	11.07	12.83
6	1.24	1.64	4.05	5.35	6.90	12.59	14.45
7	1.69	2.17	4.92	6.35	8.03	14.07	16.01
8	2.18	2.73	5.80	7.34	9.15	15.51	17.53
9	2.70	3.33	6.68	8.34	10.26	16.92	19.02

Table 2: γ-quantiles of χ^2_k-distribution for indicated values of γ and k.

Question 3 [3+1+2+2+2=10 points]

We consider the dataset `fruitfly` which contains data about the lifespan of 25 male fruitflies (in days), the length of their thoraces (in mm), and the percentage of daytime they sleep. Through a linear regression analysis we would like to find a model that describes the lifespan, i.e. the `longevity` variable, with the help of a selection of the other variables `thorax` and/or `sleep`.

a. Do the first step of the *step up* method for variable selection.

The R output for the parameter estimates and their estimated standard errors in both models are to be found on the right.

	$\hat{\beta}_j$	se($\hat{\beta}_j$)
(Intercept)	-61.28	15.22
thorax	125.00	18.94

Multiple R-squared: 0.6543

	$\hat{\beta}_j$	se($\hat{\beta}_j$)
(Intercept)	40.9131	5.4428
sleep	-0.1056	0.2338

Multiple R-squared: 0.008796

Use suitable hypothesis tests with significance level $\alpha = 5\%$ for the *step up* method and motivate your decisions; you may also use Table 3 in the Appendix.

Note: do not forget to state the null and alternative hypotheses.

b. State the name of the overall test that can be used to check whether at least one variable should be included in the model.

Suppose now that the variable selection method *step down* (instead of *step up*) has been used to build the following linear model: `longevity` $= \beta_0 + \beta_1 \cdot$ `thorax` $+ e$ which is estimated by

$$\texttt{longevity} = -61.28 + 125 \cdot \texttt{thorax} + e. \qquad (1)$$

For the remainder of this question, we will focus on this linear model.
Two diagnostic plots are shown in Figure 1.

c. Does including a higher-order term (e.g. quadratic) in thorax seem appropriate based on the left plot? Motivate your answer.

d. Which conclusions can be drawn from the plot on the right? Motivate your answer.

As a further analysis, based on the mean shift outlier model, no outliers could be found.

e. Explain the idea behind testing for outlying y-values in the mean shift outlier model, i.e. explain the mean shift outlier model and the corresponding hypothesis test.

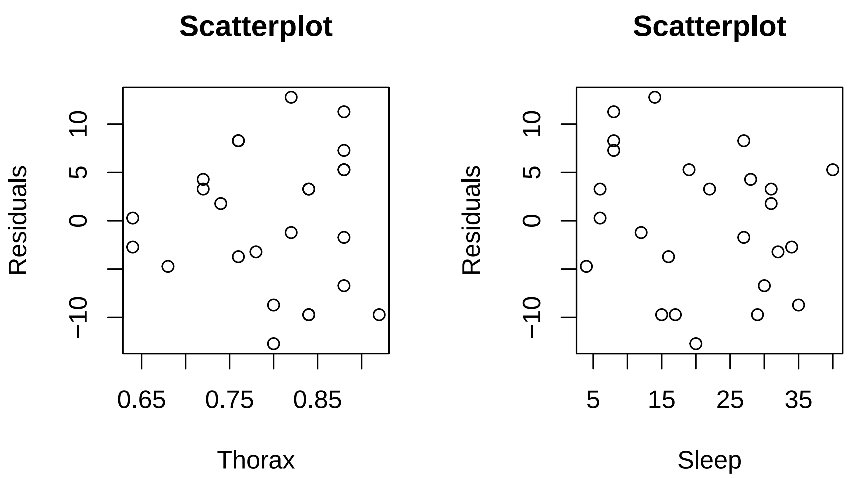

Figure 1: Two scatterplots of the residuals in the final model (1) against the `thorax` and `sleep` values, respectively.

Appendix

Table of t-quantiles

α / df	0.900	0.925	0.950	0.975	0.990	α / df	0.900	0.925	0.950	0.975	0.990
1	3.078	4.165	6.314	12.706	31.821	46	1.300	1.464	1.679	2.013	2.410
2	1.886	2.282	2.920	4.303	6.965	47	1.300	1.463	1.678	2.012	2.408
3	1.638	1.924	2.353	3.182	4.541	48	1.299	1.463	1.677	2.011	2.407
4	1.533	1.778	2.132	2.776	3.747	49	1.299	1.462	1.677	2.010	2.405
5	1.476	1.699	2.015	2.571	3.365	50	1.299	1.462	1.676	2.009	2.403
6	1.440	1.650	1.943	2.447	3.143	51	1.298	1.462	1.675	2.008	2.402
7	1.415	1.617	1.895	2.365	2.998	52	1.298	1.461	1.675	2.007	2.400
8	1.397	1.592	1.860	2.306	2.896	53	1.298	1.461	1.674	2.006	2.399
9	1.383	1.574	1.833	2.262	2.821	54	1.297	1.460	1.674	2.005	2.397
10	1.372	1.559	1.812	2.228	2.764	55	1.297	1.460	1.673	2.004	2.396
11	1.363	1.548	1.796	2.201	2.718	56	1.297	1.460	1.673	2.003	2.395
12	1.356	1.538	1.782	2.179	2.681	57	1.297	1.459	1.672	2.002	2.394
13	1.350	1.530	1.771	2.160	2.650	58	1.296	1.459	1.672	2.002	2.392
14	1.345	1.523	1.761	2.145	2.624	59	1.296	1.459	1.671	2.001	2.391
15	1.341	1.517	1.753	2.131	2.602	60	1.296	1.458	1.671	2.000	2.390
16	1.337	1.512	1.746	2.120	2.583	61	1.296	1.458	1.670	2.000	2.389
17	1.333	1.508	1.740	2.110	2.567	62	1.295	1.458	1.670	1.999	2.388
18	1.330	1.504	1.734	2.101	2.552	63	1.295	1.457	1.669	1.998	2.387
19	1.328	1.500	1.729	2.093	2.539	64	1.295	1.457	1.669	1.998	2.386
20	1.325	1.497	1.725	2.086	2.528	65	1.295	1.457	1.669	1.997	2.385
21	1.323	1.494	1.721	2.080	2.518	66	1.295	1.456	1.668	1.997	2.384
22	1.321	1.492	1.717	2.074	2.508	67	1.294	1.456	1.668	1.996	2.383
23	1.319	1.489	1.714	2.069	2.500	68	1.294	1.456	1.668	1.995	2.382
24	1.318	1.487	1.711	2.064	2.492	69	1.294	1.456	1.667	1.995	2.382
25	1.316	1.485	1.708	2.060	2.485	70	1.294	1.456	1.667	1.994	2.381
26	1.315	1.483	1.706	2.056	2.479	71	1.294	1.455	1.667	1.994	2.380
27	1.314	1.482	1.703	2.052	2.473	72	1.293	1.455	1.666	1.993	2.379
28	1.313	1.480	1.701	2.048	2.467	73	1.293	1.455	1.666	1.993	2.379
29	1.311	1.479	1.699	2.045	2.462	74	1.293	1.455	1.666	1.993	2.378
30	1.310	1.477	1.697	2.042	2.457	75	1.293	1.454	1.665	1.992	2.377
31	1.309	1.476	1.696	2.040	2.453	76	1.293	1.454	1.665	1.992	2.376
32	1.309	1.475	1.694	2.037	2.449	77	1.293	1.454	1.665	1.991	2.376
33	1.308	1.474	1.692	2.035	2.445	78	1.292	1.454	1.665	1.991	2.375
34	1.307	1.473	1.691	2.032	2.441	79	1.292	1.454	1.664	1.990	2.374
35	1.306	1.472	1.690	2.030	2.438	80	1.292	1.453	1.664	1.990	2.374
36	1.306	1.471	1.688	2.028	2.434	81	1.292	1.453	1.664	1.990	2.373
37	1.305	1.470	1.687	2.026	2.431	82	1.292	1.453	1.664	1.989	2.373
38	1.304	1.469	1.686	2.024	2.429	83	1.292	1.453	1.663	1.989	2.372
39	1.304	1.468	1.685	2.023	2.426	84	1.292	1.453	1.663	1.989	2.372
40	1.303	1.468	1.684	2.021	2.423	85	1.292	1.453	1.663	1.988	2.371
41	1.303	1.467	1.683	2.020	2.421	86	1.291	1.453	1.663	1.988	2.370
42	1.302	1.466	1.682	2.018	2.418	87	1.291	1.452	1.663	1.988	2.370
43	1.302	1.466	1.681	2.017	2.416	88	1.291	1.452	1.662	1.987	2.369
44	1.301	1.465	1.680	2.015	2.414	89	1.291	1.452	1.662	1.987	2.369
45	1.301	1.465	1.679	2.014	2.412	90	1.291	1.452	1.662	1.987	2.368

Table 3: Quantiles of t-distributions with 1 to 90 degrees of freedom (df).