Final exam, 11 December 2023

All aids are allowed, except a computer and personal assistance. Restricted use of some computer-like devices (including tablets and smartphones) is permitted under the rules described at the VHM 801 course homepage. The exam consists of three questions that should all be answered. The weights for each of the three questions and also for each subquestion within a question are indicated; these weights total *50 points*. Note that questions, and often also subquestions, can be answered independently of each other. The duration of the exam is 3 hours.

Generally, **statistical models and methods should be specified**, and every statistical analysis should be summarized in a conclusion. Throughout, if you realize that you need more information than is provided to carry out an analysis, specify what information you need, how you would obtain it using statistical software, and how you would use it in the analysis.

Question 1. (*15 points*)

A study carried out at AVC investigated the properties of a particular diagnostic test, a so-called crude indirect ELISA test, for a parasite (*Ostertagia ostertagi*) affecting cattle. The test applies to milk samples, and in the study individual cow milk samples were repeatedly tested (six times). The ELISA test yields an optical density (OD) value, and higher OD values correspond to stronger presence of antibodies against the pathogen in the sample. In addition to the raw OD values, also adjusted OD values which take into the account the readings for positive and negative controls on each plate, were recorded. The present data comprise 40 cow milk samples, and hence a total of $40 \cdot 6 = 240$ raw OD values and also 240 adjusted OD values.

a) (*3 points*)
The Minitab displays on the next page show the distributions across the 40 cows for an individual adjusted OD value per cow (`adj1`) and the average adjusted (`adjmean`) OD value over the six repeated ELISA results per cow. Carry out a brief comparative, descriptive analysis of the two sets of adjusted OD values; contrast the distributions using standard descriptors such as the center, spread, distributional shape and extremes. For your analysis, you may also use the added information that the P-values for a Anderson-Darling test for the individual and average values were 0.112 and 0.012, respectively.

b) (*3 points*)
It is often desirable to interpret ELISA test results in the form of a dichotomous (or binary) outcome corresponding to high or low antibody levels in a sample. Assume for the purpose of this question that the cut-off, or threshold, for samples to be classified as having high antibody levels is 0.45. For both the individual and average OD values, estimate the proportion of cows with high antibody levels based on this cut-off. Explain your methods and justify your choice from the characteristics of the distributions. (*Note:* You are allowed to use the same method for both distributions, but you need to justify your choice of method in both cases.)

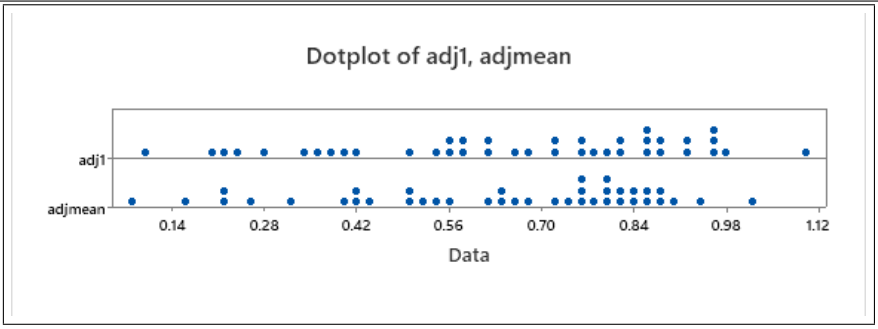
(*continues on the next page*)

WORKSHEET 1

Descriptive Statistics: adj1, adjmean

Statistics

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum	Skewness
adj1	40	0.6534	0.2541	0.1050	0.4373	0.7010	0.8603	1.0940	-0.42
adjmean	40	0.6351	0.2390	0.0770	0.4594	0.7032	0.8170	1.0232	-0.68

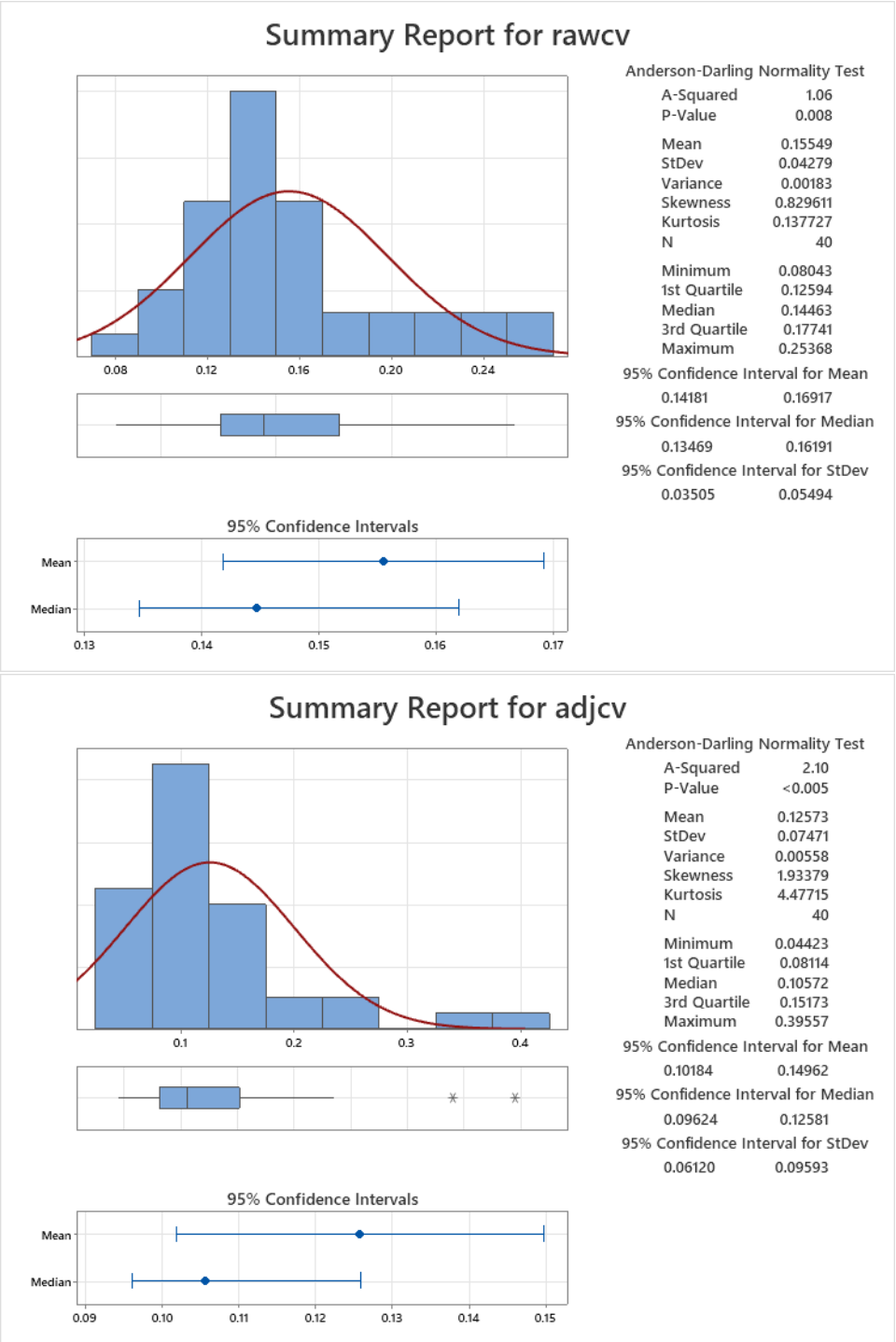


One major reason to adjust the OD values with plate reference values is that such adjusted values are expected to show less variability between repeated testing than the unadjusted (raw) values. From the six repeated test results for each sample, not only the mean but also the standard deviation in OD values between repeated testing of the same sample can be computed. A commonly used statistic for quantification of variability in a laboratory setting is the coefficient of variation (appreciated as cv), defined as the standard deviation divided by the mean. This statistic is particularly useful when the spread is expected to vary with the mean, typically with larger standard deviations being associated with larger means.

The two Minitab displays on the next page show the distributions of the coefficient of variation obtained from raw and adjusted OD values. Use the information from those displays to answer the following questions.

c) (*3 points*)
No definitive guideline exists for what constitutes an adequately low cv-value. Values below 5% are usually considered (very) good, values below 10% may be good or acceptable for some applications, whereas values above 15% may be considered as critically high (even if in some applications, the threshold for high cv-values will be higher). For **both** the cv-distributions, extract the statistical inference provided about the distribution centers from the listings (with any additional calculations needed) and use it to characterize the distributions relative to the $5\% - 10\% - 15\%$ classication points. (*Hint:* Make sure to use the most appropriate information about the distribution center in view of the data available.)

d) (*4 points*)
Contrast the two distributions descriptively, with particular focus on the questions of interest and the expectations described above. Additionally, describe how you would carry out statistical inference to compare relevant features of the distributions; give enough detail about statistical models and methods to uniquely define your proposed method. If possible, carry out some or all of the steps of the analysis.



e) (*2 points*)
In this final part, we return to the analyses from **a)** and **b)** for the individual and average adjusted OD-values. Answer **one of the two parts below** to achieve a full mark.

i) Compute a 95% confidence interval (CI) for the proportion of cows with high antibody levels, for one of the two variables (of your own choice). Note that not all methods to estimate the proportion will allow easy calculation of a CI, so you may need to alter your estimation method to compute an associated CI.

ii) Does the comparison between the two distributions conform to what one would expect if the six individual values per milk sample were independent of each other? Explain your answer.

Question 2. (*15 points*)

Cigarette labels warn pregnant women against smoking. Does nicotine actually reach the fetus, crossing the protective placental barrier? For a study published in a medical journal, researchers selected pregnant women delivering at an Egyptian hospital and categorized them as either active smokers, passive smokers or non-smokers (i.e., not exposed to smoking) during the pregnancy, with 10 women per group. They analyzed the newborn's meconium for its content of cotinine, the metabolized form of nicotine. Meconium is a newborn's first stool right after birth and is a good biological marker for fetal exposure to drugs or other chemical agents. The table below shows the meconium cotinine concentrations for the 10 mothers/babies in each group.

Cotinine (ng/ml)	Mother									
Smoking group	1	2	3	4	5	6	7	8	9	10
active smoker	490	418	405	328	700	292	292	272	240	232
passive smoker	254	219	287	257	271	282	148	273	350	293
non-smoker	158	163	153	207	211	159	199	187	200	213

The following quotes are excerpts from the publication of the study:

- "Statistical tests used were [...] F-test (ANOVA)..."
- "There was a significant difference in the cotinine levels in meconium ($F = 10.45$, $P = 0.01$)."
- "The mean concentration of cotinine in the meconium of neonates of active smokers was significantly higher than in neonates of passive and of non-smokers. But the mean concentration of cotinine in the meconium of neonates of passive smokers was not significantly higher than that of neonates of non-smokers."

Use this information as well as the Minitab listings on the next page to answer parts **a)-d)**.

a) (*2 points*)
Characterize the study type (e.g., experimental vs. observational) and design (e.g., one-sample or two-sample).

b) (*5 points*)
Determine and describe the statistical model the statistical analysis seems to have been based on. While assuming for now the validity of this model (so defer discussion of model assumptions to **c)**), assess whether the statistical significance seems to have been determined and reported properly for the analysis. (*Note:* Make sure to cover all statements about statistical significance.)

c) (*5 points*)
Discuss whether the assumptions for the model in **b)** seem to be met for these data; include all relevant model assumptions in your discussion. If you detect a problem with a model assumption, discuss how it might impact the conclusions about statistical significance presented in the paper; again, make sure to address all conclusions about statistical significance.

d) (*3 points*)
Outline what you consider to be the "best" approach for analysis of these data. If you think the "best" approach coincides with the analysis presented in the paper and/or Minitab listings, you should describe in detail how the analysis after the ANOVA table in your view should be carried out. If you think the "best" approach differs from the analysis presented in the paper and/or the Minitab listings, you should present your suggested approach as a series of steps towards reaching conclusions about statistical significance comparable to (not necessarily identical to) those in the paper. Try to avoid your answer taking the form of an unstructured collection of suggestions for analysis.

Minitab listings and graphical displays for Question 2:

COTININE.MWX

Descriptive Statistics: cotinine

Statistics

Variable	smoker	N	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
cotinine	1	10	367.2	45.4	143.7	232.0	264.0	311.5	436.0	700.0
	2	10	263.4	16.6	52.5	148.0	245.3	272.0	288.5	350.0
	3	10	185.00	7.66	24.23	153.00	158.75	193.00	208.00	213.00

Variable	smoker	Skewness	Kurtosis
cotinine	1	1.53	2.44
	2	-0.89	2.50
	3	-0.24	-2.02

COTININE.MWX

One-way ANOVA: cotinine versus smoker

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
smoker	2	167059	83530	10.45	0.000
Error	27	215900	7996		
Total	29	382959			

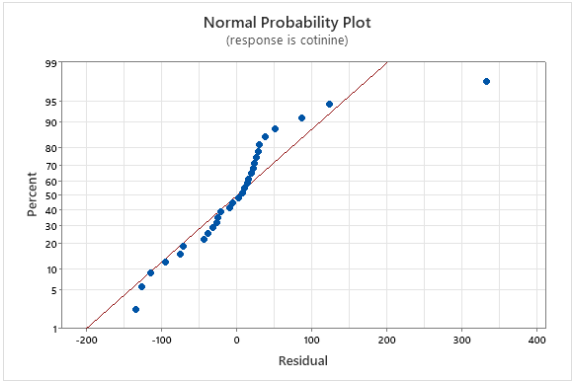
Model Summary

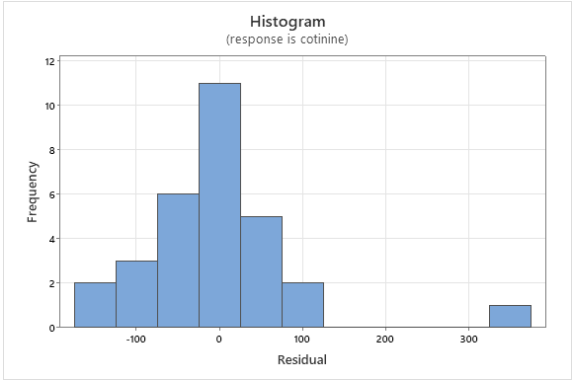
S	R-sq	R-sq(adj)	R-sq(pred)
89.4220	43.62%	39.45%	30.40%

Means

smoker	N	Mean	StDev	95% CI
1	10	367.2	143.7	(309.2, 425.2)
2	10	263.4	52.5	(205.4, 321.4)
3	10	185.00	24.23	(126.98, 243.02)

Pooled StDev = 89.4220





Question 3. (*20 points*)

From a recent "proof-of-concept" study on the effect of the drug (medication) reserpine on blood pressure in (human) patients with refractory hypertension, we consider the values obtained for six patients. Refractory hypertension is characterized by a blood pressure that remains uncontrolled on maximal or near-maximal therapy; therefore, people with this condition are very difficult to successfully treat for their high blood pressure. The participants were dispensed reserpine 0.1 *mg* to be taken orally once daily for 4 weeks in addition to their other antihypertensive medications. For each patient, measurements were taken at baseline (prior to the beginning the treatment) and after 4 weeks. The table below shows the values for systolic and diastolic blood pressures (both in *mm Hg*) for each patient, together with the achieved reductions in blood pressure. Descriptive statistics for these variables are shown in the two subsequent Minitab listings.

Patient	Baseline: before treatment systolic	diastolic	After 4 weeks of treatment systolic	diastolic	Drop (reduction) systolic	diastolic
1	173	107	163	100	10	7
2	135	81	105	62	30	19
3	142	86	116	68	26	18
4	187	113	116	66	71	47
5	173	116	145	82	28	34
6	155	95	144	88	11	7

RESERPINE

Descriptive Statistics: syst_base, diast_base, syst_4w, diast_4w, syst_drop, diast_drop

Statistics

Variable	N	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
syst_base	6	161.7	12.6	30.8	129.0	137.3	156.5	185.0	212.0
diast_base	6	100.00	7.39	18.11	78.00	83.25	100.50	112.50	129.00
syst_4w	6	132.33	5.62	13.76	114.00	117.75	135.00	145.25	146.00
diast_4w	6	78.00	4.09	10.02	62.00	68.75	80.00	87.25	88.00
syst_drop	6	29.33	9.05	22.18	10.00	10.75	27.00	40.25	71.00
diast_drop	6	22.00	6.44	15.77	7.00	7.00	18.50	37.25	47.00

RESERPINE

Correlation: syst_base, diast_base, syst_4w, diast_4w, syst_drop, diast_drop

Correlations

	syst_base	diast_base	syst_4w	diast_4w	syst_drop
diast_base	0.978				
syst_4w	0.763	0.663			
diast_4w	0.630	0.495	0.974		
syst_drop	0.916	0.947	0.440	0.270	
diast_drop	0.723	0.834	0.143	-0.067	0.916

The primary focus of the study was on whether the treatment was successful in lowering either, or both, of the systolic and distolic blood pressures, and on quantifying such effects.

a) (*5 points*)
Describe the statistical design, formulate an appropriate statistical model and analyse the data for diastolic blood pressure to quantify and obtain statistical inference for the treatment effect, and draw conclusions.

b) (*3 points*)
It is customary to measure both the systolic and diastolic blood pressure, because both are considered important for monitoring heart health. It is therefore of interest whether the treatment has similar effects on these two measurements. Which statistic can be used to quantify the association between treatment effects on these two measurements? Give its value, and carry out statistical inference to further assess the strength of association, and draw conclusions. You should ideally work from the information provided, but if you find it insufficient you can describe how you would obtain any additional information required using statistical software, such as Minitab, and how you would use it.

c) (*6 points*)
The plots and Minitab listings on the next page for parts **c)-e)** include two additional analyses of the diastolic blood pressure data, presented on the left and right hand side of the page. For each of them, explain the statistical model it is based on, estimate and interpret the model's parameters and draw conclusions from the results given. For one of the models, compute additionally 95% confidence intervals for the regression parameters.

d) (*4 points*)
For **at least two of the three models** considered in parts **a)** and **c)**, compute a 95% interval for the diastolic blood pressure after treatment of a patient whose diastolic blood pressure before reserpine treatment is 110 mm Hg. Interpret carefully the interval you give. Compare the intervals obtained from different models. If you find the given Minitab listings insufficient for one or several models, you may sketch your approach to compute the requested interval.

e) (*2 points*)
Which of the three models considered do you think is preferable for these (diastolic blood pressure) data? Include in your discussion the assumptions of the models, and whether it is possible to make preferences between the models by statistical tests. You may also include other models of potential interest in the discussion; in that case, describe also how you would carry out the analysis.

Minitab listings for Question 3, parts c)-e):

RESERPINE

Regression Analysis: diast_4w versus diast_base

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	50.6	24.4	2.08	0.106
diast_base	0.274	0.240	1.14	0.318

Model Summary

S	R-sq	R-sq(adj)
9.73489	24.49%	5.61%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	122.9	122.93	1.30	0.318
Error	4	379.1	94.77		
Total	5	502.0			

RESERPINE

Regression Analysis: diast_drop versus diast_base

Coefficients

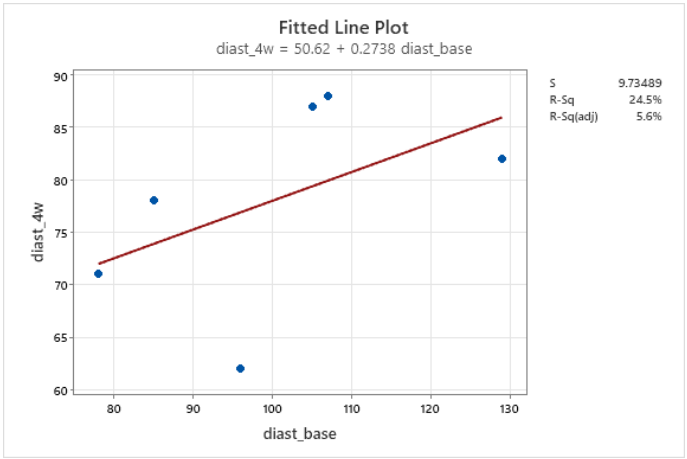
Term	Coef	SE Coef	T-Value	P-Value
Constant	-50.6	24.4	-2.08	0.106
diast_base	0.726	0.240	3.02	0.039

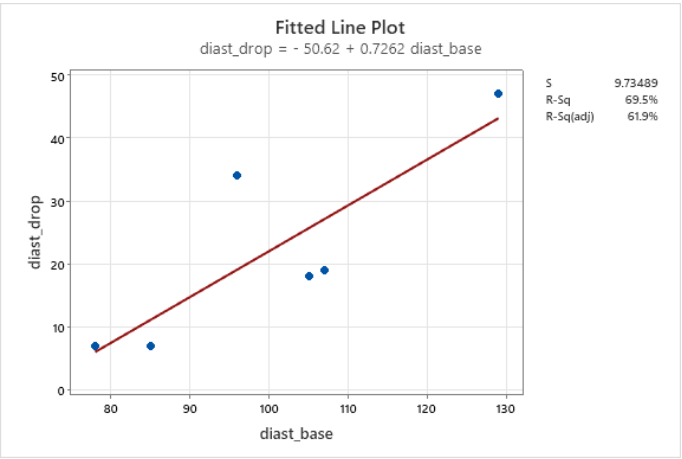
Model Summary

S	R-sq	R-sq(adj)
9.73489	69.53%	61.91%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	864.9	864.93	9.13	0.039
Error	4	379.1	94.77		
Total	5	1244.0			





RESERPINE

Prediction for diast_4w

Regression Equation

diast_4w = 50.6 + 0.274 diast_base

Settings

Variable	Setting
diast_base	110

Prediction

Fit	SE Fit	95% CI	95% PI
80.7378	4.64470	(67.8421, 93.6336)	(50.7906, 110.685)

RESERPINE

Prediction for diast_drop

Regression Equation

diast_drop = -50.6 + 0.726 diast_base

Settings

Variable	Setting
diast_base	110

Prediction

Fit	SE Fit	95% CI	95% PI
29.2622	4.64470	(16.3664, 42.1579)	(-0.685002, 59.2094)