Biostats VHM 801 Course
Fall 2018, Atlantic Veterinary College, PEI
Henrik Stryhn

Final exam, 10 December 2018

All aids are allowed, except a computer-like device (including tablets and smartphones) and personal assistance. 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 the analysis, specify what information you need, how you would obtain it using Minitab, and how you would use it in the analysis.

Question 1. (*15 points*)

This question contains three completely separate parts that can be answered independently of each other, and should all be answered for a full mark.

a) (*4 points*)
A series of recent studies carried out by Trevor Jain, Director of UPEI's Program in Paramedicine, focused on the time needed to triage a disaster scene. The triage procedure aims to determine an order of priority for actions to be taken on the scene, such as attending to victims or neutralizing hazards. In one study, students from the paramedicine program were presented with a simulated disaster scene and asked to triage victims following one of two protocols; the purpose of the study was to compare the two protocols. The listing below gives the times (in seconds, for convenience ordered from lowest to highest) taken for 13 students in one group to triage the scene, with corresponding descriptive statistics shown in the Minitab listing below:

429, 517, 527, 530, 555, 557, 561, 606, 608, 622, 700, 819, 888.

Descriptive Statistics: time

Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
time	13	0	609.2	35.1	126.5	429.0	528.5	561.0	661.0	888.0

i) Use the information provided to sketch a suitable graphical display, of your own choice, for the distribution of triage completion times.

ii) Use your graphical display and the descriptive statistics provided to (briefly) characterize the distribution, while focusing on the features of the distribution that would be relevant in preparation for a statistical analysis. (*Note:* The primary analysis of interest is a comparison of triage completion times with the other protocol group, but its data are for simplicity not included here.)

b) (*6 points*)
According to one source, the weights of newborn children in Canada may overall be described by a normal distribution with mean 3.4 kg and standard deviation 570 g. A child is classified as being born with low birthweight if it weighs less than 2.5 kg.

i) What proportion of babies are born with low birthweight, according this information?

ii) For three (unrelated) babies chosen at random, what is the probability that their average birthweight is below 2.5 kg?

iii) A recent government publication shows a low birthweight rate (proportion) on PEI of about 5% (making PEI one of the regions in Canada with lowest rate of low birthweight). Based on this information, what is the probability that among three (unrelated) babies born on PEI, none of them have low birthweight?

c) (*5 points*)
A lab technician performs measurements of specific gravity of a compound, and has obtained the following four values for a particular specimen:

3.84, 3.93, 3.67, 3.76 – with sample mean 3.80 and standard deviation 0.11.

According to the lab manual, repeated measurements will vary according to a normal distribution with standard deviation $\sigma = 0.15$. The interest is in the true (mean) specific gravity of the specimen.

i) The lab manual asks for a 95% confidence interval for the true specific gravity. Compute such an interval and interpret it (briefly) in a non-technical manner.

ii) The lab manual also asks whether the data show convincingly that the true specific gravity is less than 3.9. Carry out a statistical analysis to assess this, and draw conclusions (again, phrased in a non-technical manner).

Question 2. (*15 points*)

Meat inspection control of swine carcasses at slaughterhouses (or abbattoirs) in Denmark involve among other things a bacteriological examination of selected carcasses that ultimately leads to each carcass being either accepted or condemned (discarded). To investigate whether these bacteriological examinations were carried out similarly at different slaughterhouses the bacteriological samples for a total of 670 carcasses (from different slaughterhouses) were analyzed at one laboratory, and the results were used to categorize each carcass as accepted or condemned. The number of samples submitted to the laboratory depended on the size of the slaughterhouse, and the data given here in the table below distinguish between two large slaughterhouses, denoted as I and II, and the remaining (smaller) slaughterhouses considered as a single group.

	Slaughterhouse			
	I	II	other	Total
Accepted	134	275	146	555
Condemned	45	43	27	115
Total	179	318	173	670

a) (*5 points*)
Use the data of the table, and any of the attached Minitab listings, to investigate differences between the proportions of carcasses condemned at the slaughterhouses. State and motivate your statistical model and hypothesis, and draw conclusions; if necessary, carry out any additional analysis to describe the differences between slaughterhouses. (*Note:* The requested analysis here should *not* involve the additional variable (diagnosis) described below.)

In addition to the outcome of the bacteriological examination it was also recorded for each carcass involved in the study why it was selected for examination. In the data, the most common diagnosis given as the cause for bacteriological examination was "tail bite" (inflammation of the tail due to bites). A total of 173 carcasses received this diagnosis, and their bacteriological assessments are shown in the table below.

Diagnosis:	Slaughterhouse			
tail bite	I	II	other	Total
Accepted	30	82	25	137
Condemned	19	13	4	36
Total	49	95	29	173

b) (*5 points*)
Use the additional information (table and any of the attached computer listings) to further investigate differences between the slaughterhouses in their proportion of condemned carcasses. Specifically, are any differences between the slaughterhouses noted previously associated with specific diagnoses ("tail bite" or "other" diagnoses)? If necessary, adjust your conclusions from **a)** in light of the additional information.

c) (*5 points*)
For each of the incomplete statements below, indicate which of the suggested completions are correct (there may be more than one correct completion). Further explanation is only required when a chosen option indicates the need for further details.

(A) The present study can be characterized as...

i) an experimental study,

ii) an observational study,

iii) another study type (give details): .....

(B) Among the submitted samples for examination, the proportion submitted for tail bite was...

i) significantly different between the slaughterhouses, and clearly highest at slaughterhouse I,

ii) significantly different between the slaughterhouses, and clearly highest at slaughterhouse II,

iii) significantly different between the slaughterhouses, and clearly lowest at the other slaughterhouses,

iv) not significantly different between the slaughterhouses,

v) another summary statement (give details): .....

(C) To study the association between diagnosis (i.e., "tail bite" versus "other") and the subsequent result of the bacteriological examination,

i) we should compute the correlation between the two variables and test whether it is significantly different from zero,

ii) we should examine the two-way table formed by the two variables and test the significance by a chi-square test, as indicated in the listing,

iii) we should use another procedure (give details): .....

iv) we should carry out another study because the association can not be determined from the present data (give details): .....

(Minitab listings for Question 2 follow on the next page.)

Tabulated Statistics: status, slhouse

Using frequencies in ncarc

Rows: status Columns: slhouse

	I	II	other	All
accepted	134	275	146	555
	148.28	263.42	143.31	
condemned	45	43	27	115
	30.72	54.58	29.69	
All	179	318	173	670

Cell Contents
Count
Expected count

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	11.270	2	0.004
Likelihood Ratio	10.672	2	0.005

Tabulated Statistics: status, slhouse, diagn

Using frequencies in ncarc

Results for diagn = other

Rows: status Columns: slhouse

	I	II	other	All
accepted	104	193	121	418
	109.34	187.55	121.11	
condemned	26	30	23	79
	20.66	35.45	22.89	
All	130	223	144	497

Cell Contents
Count
Expected count

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	2.634	2	0.268
Likelihood Ratio	2.578	2	0.276

Results for diagn = tail bite

Rows: status Columns: slhouse

	I	II	other	All
accepted	30	82	25	137
	38.80	75.23	22.97	
condemned	19	13	4	36
	10.20	19.77	6.03	
All	49	95	29	173

Cell Contents
Count
Expected count

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	13.391	2	0.001
Likelihood Ratio	12.398	2	0.002

Tabulated Statistics: diagn, slhouse

Using frequencies in ncarc

Rows: diagn Columns: slhouse

	I	II	other	All
other	130	223	144	497
	132.78	235.89	128.33	
tail bite	49	95	29	173
	46.22	82.11	44.67	
All	179	318	173	670

Cell Contents
Count
Expected count

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	10.364	2	0.006
Likelihood Ratio	10.989	2	0.004

Tabulated Statistics: status, diagn

Using frequencies in ncarc

Rows: status Columns: diagn

	other	tail bite	All
accepted	418	137	555
	411.69	143.31	
condemned	79	36	115
	85.31	29.69	
All	497	173	670

Cell Contents
Count
Expected count

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	2.179	1	0.140
Likelihood Ratio	2.109	1	0.146

Question 3. (*20 points*)

A study was carried out to assess the accuracy of the reporting of caloric content in "diet" and "health" foods. A sample of 40 food products was collected in a restricted area within a short period of time (in 1992). The food items were classified into three food origin groups (for the purpose of comparing these groups): nationally advertised (N), regionally distributed (R) and locally prepared (L). For each of the food items, the number of kilocalories it contained was determined using a method called "bomb calorimetry", after blinding of the technician to the stated caloric content of the item. Subsequently, the deviation in caloric content was computed as the difference between the measured and stated caloric contents (i.e., measured minus stated contents), and expressed as a percentage of the stated caloric content. The values of the percent deviation in caloric content (denoted `cal_dev` in the Minitab outputs) computed this way are shown in the table below, together with a specification of each food item.

Nationally advertised food (N)		Regionally distributed food (R)		Locally prepared food (L)	
2.0	noodles and Alfredo sauce	41.0	meatless sandwich	15.0	chinese chicken
-28.0	cheese curls	46.0	oatmeal cookie	60.0	gyoza
-6.0	green beans	2.0	lemon pound cake	250	jelly diet candy-reds flavor
8.0	mixed fruits	25.0	banana cake	145	jelly diet candy-fruit flavor
6.0	cereal	39.0	brownie	6.0	Florentine manicotti
-1.0	fig bars	16.5	butterscotch bar	80.0	egg foo young
10.0	oatmeal raisin cookie	17.0	blondie	95.0	hummus with salad
13.0	crumb cake	28.0	oat bran snack bar	3.0	baba ghanoush with salad
15.0	crackers	-3.0	granola bar		
-4.0	blue cheese dressing	14.0	apricot bar		
-4.0	imperial chicken	34.0	chocolate chip cookie		
-18.0	vegetable soup	42.0	carrot muffin		
10.0	cheese				
5.0	chocolate pudding				
3.0	sausage biscuit				
-7.0	lasagna				
3.0	spread cheese				
-0.5	lentil soup				
-10.0	pasta w/shrimp & sauce				
6.0	chocolate mousse				

a) (*2 points*)
Based on the description above, characterize the study type (e.g., experimental or observational) and the statistical design, and discuss briefly what population any statistical inference could refer to (i.e., be representative for).

b) (*4 points*)
Describe the statistical model used for the analysis in the enclosed Minitab listing for

part **b)**, including a (brief) review of its assumptions. Use the information provided to critically discuss whether the data meets the model assumptions. In the paper published on the study, the authors reported the statistical methods as follows:

> "As expected, data violated major assumptions of parametric statistical inference (ie, normality, independence of observations, and homogeneity of variance). Thus, we used nonparametric [...] tests."

Do you think that is a fair summary of the data's compliance with model assumptions and a proper justification for the use of nonparametric procedures? For any critical points you might have, make sure to indicate how in your view the corresponding part of the summary should read instead.

c) (*4 points*)
Explain (briefly) the statistical analysis shown in the Minitab listing for part **c)**; make sure to include how the assumptions and hypotheses differ from the analysis in part **b)**. Draw conclusions from the results shown, and discuss whether further analysis seems to be needed to fully compare the different food groups with respect to reporting accuracy. If so, outline either how you could carry out such additional analysis from the information provided (without doing any actual calculations) or how you would carry out the such additional analysis in Minitab.

d) (*5 points*)
One approach proposed (elsewhere, i.e. not in the publication) for analysis of these data is a parametric analysis after suitable transformation. The proposed transformation takes the form $y^* = y/(100 + y)$, where y is the reported deviation in caloric content from the table. (*Note:* The specific form of the transformation is not too important for the questions to follow because you are not expected to backtransform estimates to original scale.) The results of an analysis of the transformed variable (denoted `transf_caldev`) are shown in the enclosed Minitab listing for part **d)**. Review the analysis and draw conclusions. Pay particular attention to any improvement in the data's compliance with the model assumptions. Include also a detailed assessment of what differences might exist between food products from the national, regional and local food origin groups (based on this analysis).

e) (*5 points*)
An additional objective for analysis of the collected data was to assess whether there was evidence of systematic under- or over-reporting of caloric content in any of the three food groups. Formulate a statistical model and carry out a statistical analysis to investigate this question for at least one of the three food groups (of your own choice), and draw conclusions. If time permits, carry out the analysis for all three food groups and draw conclusions. (*Hint:* You need to express any systematic under- or over-reporting in terms of features of the distribution(s) of the reporting accuracy.)

Minitab listings/output for Question 3, part b):

Descriptive Statistics: cal_dev

Statistics

Variable	group	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum	Skewness
cal_dev	N	20	0.13	10.52	-28.00	-5.50	2.50	7.50	15.00	-1.08
	R	12	25.13	16.07	-3.00	14.63	26.50	40.50	46.00	-0.42
	L	8	81.8	84.0	3.0	8.3	70.0	132.5	250.0	1.22

One-way ANOVA: cal_dev versus group

Method

Null hypothesis	All means are equal
Alternative hypothesis	Not all means are equal
Significance level	α = 0.05

Equal variances were assumed for the analysis.

Factor Information

Factor	Levels	Values
group	3	N, R, L

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
group	2	38096	19048	12.98	0.000
Error	37	54301	1468		
Total	39	92396			

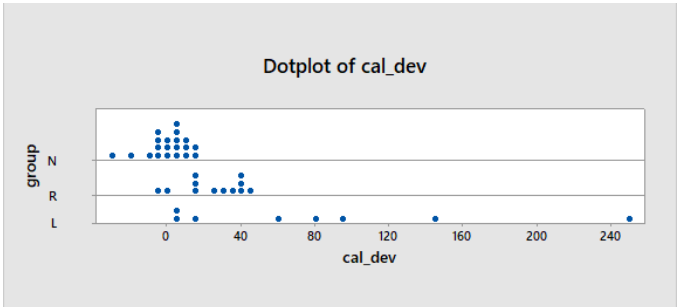
Model Summary

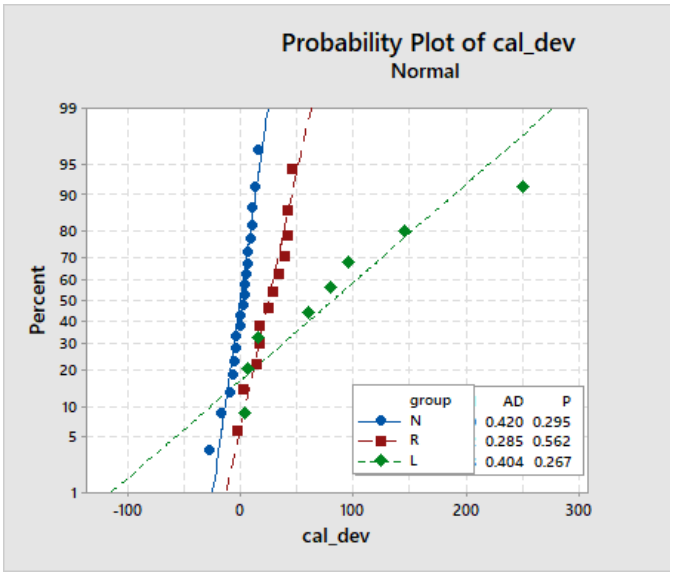
S	R-sq	R-sq(adj)	R-sq(pred)
38.3090	41.23%	38.05%	24.05%

Means

group	N	Mean	StDev	95% CI
N	20	0.13	10.52	(-17.23, 17.48)
R	12	25.13	16.07	(2.72, 47.53)
L	8	81.8	84.0	(54.3, 109.2)

Pooled StDev = 38.3090





Minitab listing for Question 3, part c):

Kruskal-Wallis Test: cal_dev versus group

Descriptive Statistics

group	N	Median	Mean Rank	Z-Value
N	20	2.5	12.5	-4.33
R	12	26.5	26.9	2.26
L	8	70.0	30.9	2.82
Overall	40		20.5	

Test

Null hypothesis	H_0: All medians are equal
Alternative hypothesis	H_1: At least one median is different

Method	DF	H-Value	P-Value
Not adjusted for ties	2	19.31	0.000
Adjusted for ties	2	19.33	0.000

Minitab listings/output for Question 3, part d):

One-way ANOVA: transf_caldev versus group

Method

Null hypothesis	All means are equal
Alternative hypothesis	Not all means are equal
Significance level	α = 0.05

Equal variances were assumed for the analysis.

Factor Information

Factor	Levels	Values
group	3	N, R, L

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
group	2	0.8340	0.41701	17.55	0.000
Error	37	0.8793	0.02376		
Total	39	1.7133			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.154157	48.68%	45.91%	37.20%

Means

group	N	Mean	StDev	95% CI
N	20	-0.0109	0.1226	(-0.0807, 0.0590)
R	12	0.1878	0.1115	(0.0976, 0.2779)
L	8	0.3536	0.2554	(0.2432, 0.4640)

Pooled StDev = 0.154157