Biostats VHM 801 Course
Fall 2019, Atlantic Veterinary College, PEI
Henrik Stryhn

Final exam, 9 December 2019

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 statistical software, and how you would use it in the analysis.

Question 1. (*15 points*)

This question contains two separate parts *i)* and *ii)*, each comprising two (sub)questions. All four (sub)questions should be answered for a full mark.

Part i)
Listeria is a bacteria that can be transmitted to humans in certain types of food such as soft cheese. In an ongoing monitoring of a cheese production, a sample of 20 cheeses were selected (randomly) from a batch of 1000 cheeses and tested for Listeria. Of the selected cheeses, 4 tested positive.

a) (*5 points*)
Estimate the proportion of Listeria positive cheeses in the batch, and supplement the estimate with a 95% confidence interval. In addition, estimate the total number of Listeria positive cheeses in the batch of 1000 cheeses, and supplement also this estimate with a confidence interval.

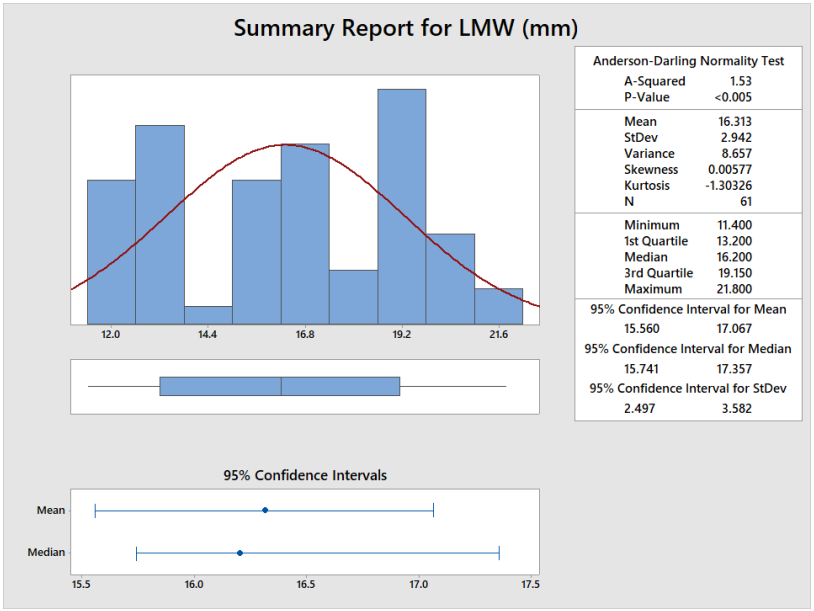
b) (*3 points*)
Monitoring of previous batches in the production had establed Listeria bacteria to be present in approximately 5% of the cheeses. The inspectors wondered whether the result of the current batch signaled an increased Listeria presence in the cheeses. Give a statistical assessment of this question; make sure to quantify the statistical strength of your assessment.

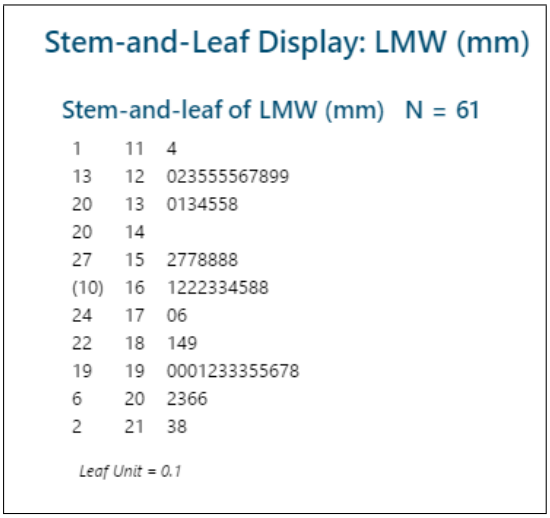
Part ii)
Biologists and ecologists record the distributions of measurements made on animal species to help study the distribution and evolution of the animals. We consider here a particular bird, the black-bellied seedcracker (*Pyrenestes ostrinus*), one of many African species of finch.

c) (*3 points*)
A study in Cameroon found that the wing length of male finches varies according to a normal distribution with mean 61.2 *mm* and standard deviation 1.8 *mm*. Based on this information, what proportion of male finches have wings longer than 65 *mm*? Determine also the wing length that only 1% of male finches exceed.

d) (*4 points*)
Another study had recorded the bill (or beak) size of 61 finches. Use the Minitab displays below to carry out a brief descriptive analysis for the distribution of the lower mandible width (related to bill size), measured in *mm*. Based on this description, discuss the choice of a parameter to represent the centre of the distribution, and the validity and meaningfulness of a 95% confidence interval for such a parameter.

Minitab listings and plots for Question 1:





Question 2. (*15 points*)

In a study of the mutual effects of the air pollutants ozone (O_3) and sulphur dioxide (SO_2), Blue Lake snap beans were grown in open-top field chambers. Some chambers were fumigated repeatedly with sulphur dioxide. The air in some chambers was carbonfiltered to remove ambient ozone. There were three chambers per combination of O_3 and SO_2 conditions, allocated at random. After 1 month of growth under these conditions, the total yield (kg) of bean pods was recorded for each chamber, with results as shown in the table.

O_3	SO_2	Yield (kg)			Mean	Stand.dev.
absent	absent	1.52	1.55	1.39	1.487	0.0850
absent	present	1.49	1.55	1.21	1.417	0.1815
present	absent	1.15	1.30	1.57	1.340	0.2128
present	present	0.65	0.76	0.69	0.700	0.0557

The question of primary interest to the investigators was whether the two air pollutants impaired the growth and yield of beans. Such effects could either be related to a single of the pollutants or their combined effect. Use the information provided above as well as in the Minitab listings (on the next page) to answer all of the following questions.

a) (*3 points*)
Characterize the statistical design (including descriptors such as: experimental unit, treatment, factor, design type, replication), and describe briefly how you would carry out a proper randomization for the experiment.

b) (*4 points*)
Suggest an appropriate statistical model for the data, including the model assumptions. Use the information provided to discuss whether the model assumptions are met to a reasonable degree. If you identify concerns with the model assumptions, describe briefly how these could be further explored or dealt with for valid analysis.

c) (*3 points*)
Discuss how the effects of the air pollutants on the yield should preferably be presented graphically. If possible from the information provided, sketch your proposed graphical display, and draw conclusions (i.e., describe your findings in the display). If you are unable to construct your proposed graphical display, explain how you would obtain it using statistical software, and explain how you would use it to draw conclusions.

d) (*5 points*)
Describe the analysis shown in the listings and use it to draw conclusions about the effects of interest. Specifically, your conclusions should include statements about whether the data show effects of O_3 and SO_2 on the yield, and whether any effect of SO_2 seems to depend on the presence of O_3 in the chamber, or vice versa. Compute 95% confidence intervals to represent the uncertainty on the estimates you consider relevant for presentation of the results. If you need further information to finalize the analysis and conclusions, explain carefully what information you need, how you would obtain it from statistical software and how you would use it.

Minitab listings and plots for Question 2:

General Linear Model: yield versus O3, SO2

Factor Information

Factor	Type	Levels	Values
O3	Fixed	2	absent, present
SO2	Fixed	2	absent, present

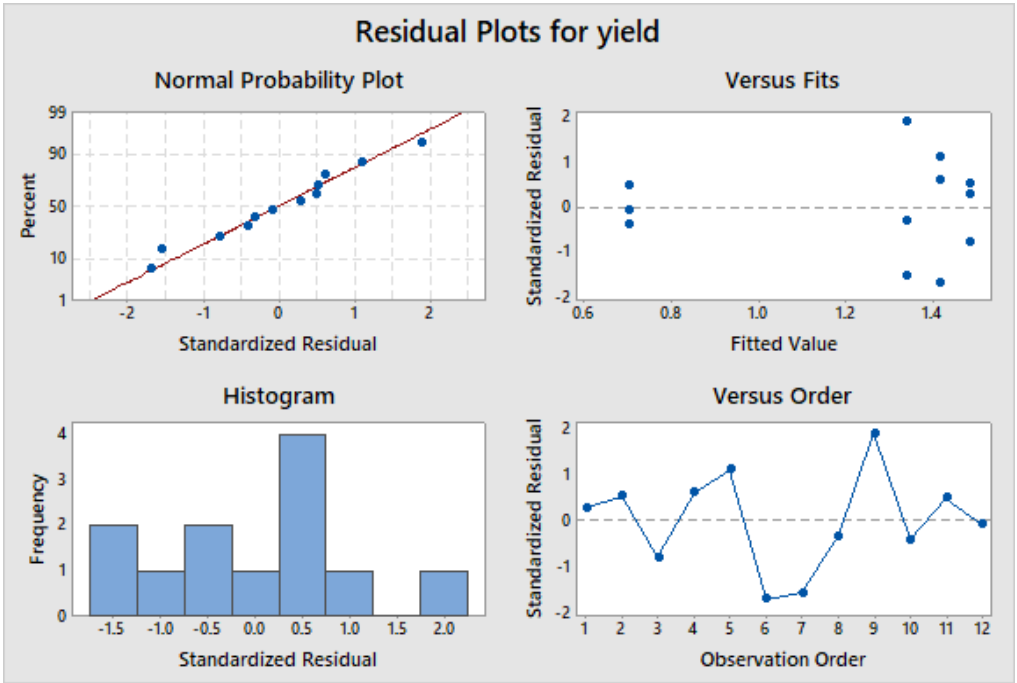
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
O3	1	0.5590	0.55901	25.25	0.001
SO2	1	0.3781	0.37807	17.08	0.003
O3*SO2	1	0.2437	0.24367	11.01	0.011
Error	8	0.1771	0.02214		
Total	11	1.3579			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.148801	86.96%	82.06%	70.65%

Residual Plots for yield



Question 3. (*20 points*)

A dataset on home insulation was collected in the 1960's by Mr. Derek Whiteside of the UK Building Research Statioin. He recorded (among other things) the weekly gas consumption (in 1000 cubic feet) and the average outside temperature (in °C) at his own house in south-east England, for 26 weeks before and 30 weeks after cavity-wall insulation had been installed. The house thermostat was set at 20 °C throughout. The questions of interest are: how is gas consumption related to outside temperature, and did this relationship change after insulation? The values recorded are listed in the table below; note that the data are not in chronological order.

Observation number	Before insulation		After insulation	
	Temperature	Gas consumption	Temperature	Gas consumption
1	-0.8	7.2	-0.7	4.8
2	-0.7	6.9	0.8	4.6
3	0.4	6.4	1.0	4.7
4	2.5	6.0	1.4	4.0
5	2.9	5.8	1.5	4.2
6	3.2	5.8	1.6	4.2
7	3.6	5.6	2.3	4.1
8	3.9	4.7	2.5	4.0
9	4.2	5.8	2.5	3.5
10	4.3	5.2	3.1	3.2
11	5.4	4.9	3.9	3.9
12	6.0	4.9	4.0	3.5
13	6.0	4.3	4.0	3.7
14	6.0	4.4	4.2	3.5
15	6.2	4.5	4.3	3.5
16	6.3	4.6	4.6	3.7
17	6.9	3.7	4.7	3.5
18	7.0	3.9	4.9	3.4
19	7.4	4.2	4.9	3.7
20	7.5	4.0	4.9	4.0
21	7.5	3.9	5.0	3.6
22	7.6	3.5	5.3	3.7
23	8.0	4.0	6.2	2.8
24	8.5	3.6	7.1	3.0
25	9.1	3.1	7.2	2.8
26	10.2	2.6	7.5	2.6
27			8.0	2.7
28			8.7	2.8
29			8.8	1.3
30			9.7	1.5
Mean	5.35	4.75	4.46	3.48

(*continues on next page*)

a) (*6 points*)
Explain and motivate the statistical models used for the analyses in the enclosed Minitab listings (on the following two pages). For each analysis, give estimates and, if possible, also confidence intervals for the model parameters. Comment briefly on the strength and statistical significance of the relationship between the temperature and gas consumption, both before and after insulation.

b) (*3 points*)
Use the information provided in the Minitab output to discuss how well the models comply with their assumptions. If you identify any issues of concern, outline how you would want to deal with them.

Regardless of your conclusions in **b)** with respect to the validity of the models and analyses, you are expected to base your answers to the remaining questions on the models considered in **a)** and **b)**.

c) (*3 points*)
As part of the discussion of how (and whether) the relation between temperature and gas consumption was affected by the insulation, we focus here on the increase in gas consumption per degree decrease in outside temperature. Identify the parameter in the statistical models that expresses the rate of change in gas consumption per degree increase in outside temperature. Use statistical inference to assess whether this rate was affected by the insulation.

d) (*4 points*)
Another way of quantifying the effect of the insulation is to compare the gas consumption at different temperatures. Estimate the gas consumption before and after insulation at temperatures 0 °C and 10 °C. If possible from the information provided, give in a similar manner as for **c)** statistical assessments of whether the gas consumptions at those two temperatures were affected by the insulation. If you need further information to complete your statistical comparison before and after insulation, explain carefully what information you need and how you would use it.

e) (*3 points*)
As a final way of quantifying the effect of the insulation, compute the range of temperatures for which the estimated gas consumption (based on these data) is higher before than after the insulation. (*Hint:* Compute first the temperature at which the estimated gas consumption is the same before and after insulation.)

f) (*1 point*)
Summarize your analysis from **a)-e)**, as well as any other quantitative comparisons of the data before and after insulation, into a short overall conclusion describing how the insulation appears to have affected the gas consumption.

Minitab listings and plots for Question 3:

Regression Analysis: Gas_Before versus Temp_Before

The regression equation is
Gas_Before = 6.854 - 0.3932 Temp_Before

Model Summary

S	R-sq	R-sq(adj)
0.281334	94.38%	94.15%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	31.9054	31.9054	403.11	0.000
Error	24	1.8996	0.0791		
Total	25	33.8050			

Fitted Line: Gas_Before versus Temp_Before

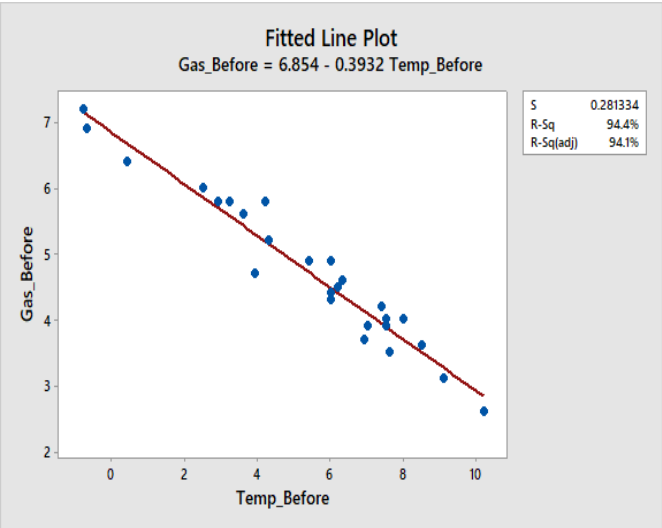
Coefficients

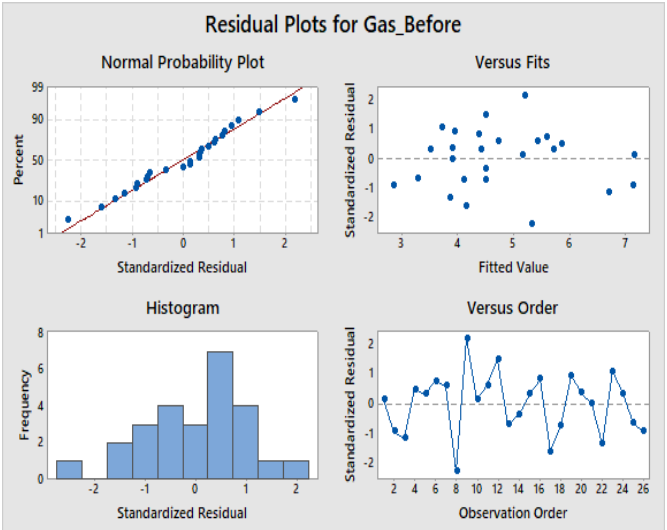
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	6.854	0.118	57.88	0.000	
Temp_Before	-0.3932	0.0196	-20.08	0.000	1.00

Fits and Diagnostics for Unusual Observations

Obs	Gas_Before	Fit	Resid	Std Resid	
8	4.7000	5.3202	-0.6202	-2.26	R
9	5.8000	5.2022	0.5978	2.17	R

R Large residual





(*continues on next page*)

Regression Analysis: Gas_After versus Temp_After

The regression equation is
Gas_After = 4.724 - 0.2779 Temp_After

Model Summary

S	R-sq	R-sq(adj)
0.354848	81.31%	80.64%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	15.3360	15.3360	121.79	0.000
Error	28	3.5257	0.1259		
Total	29	18.8617			

Fitted Line: Gas_After versus Temp_After

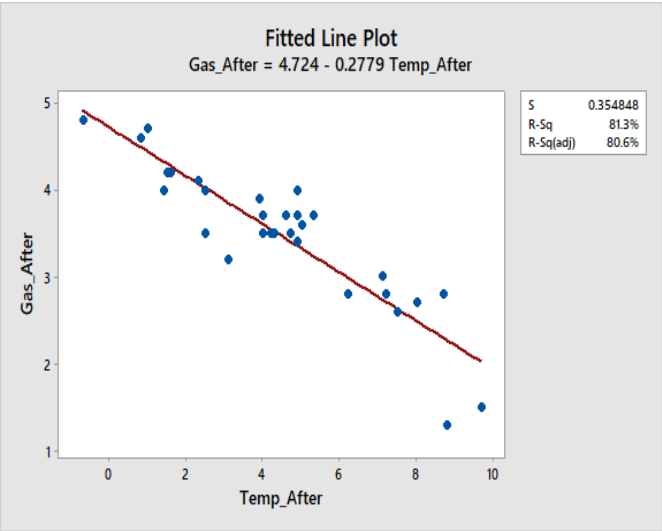
Coefficients

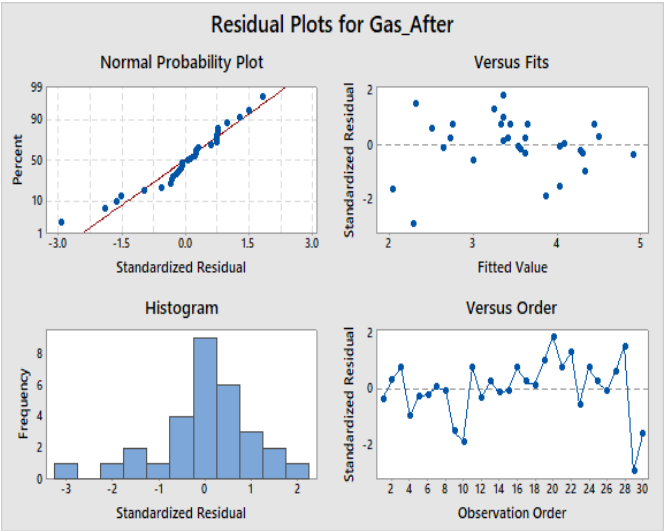
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	4.724	0.130	36.41	0.000	
Temp_After	-0.2779	0.0252	-11.04	0.000	1.00

Fits and Diagnostics for Unusual Observations

Obs	Gas_After	Fit	Resid	Std Resid	
29	1.300	2.278	-0.978	-2.95	R

R Large residual





Stata listings for final exam
(edited, roughly equivalent to Minitab listings, except as noted)

```
************************************************
* Q1
* note: some information from Minitab's graphical summary not included here

. sum lmwmm, d
                          LMW (mm)
-------------------------------------------------------------
      Percentiles      Smallest
 1%         11.4            11.4
 5%         12.3              12
10%         12.5            12.2       Obs                  61
25%         13.3            12.3       Sum of Wgt.          61

50%         16.2                       Mean           16.31311
                        Largest       Std. Dev.      2.942305
75%         19.1            20.6
90%         19.8            20.6       Variance       8.657158
95%         20.6            21.3       Skewness       .0056258
99%         21.8            21.8       Kurtosis       1.704366

. stem lmwmm
Stem-and-leaf plot for lmwmm (LMW (mm))

lmwmm rounded to nearest multiple of .1
plot in units of .1

  11* | 4
  12* | 023555567899
  13* | 0134558
  14* |
  15* | 2778888
  16* | 1222334588
  17* | 06
  18* | 149
  19* | 0001233355678
  20* | 2366
  21* | 38
```

(*continues on next page*)

Stata listings for final exam
(edited, roughly equivalent to Minitab listings, except as noted)

```
***********************************************
* Q2
* note: residual plot panel not included here

. anova yield O3##SO2

                          Number of obs =         12    R-squared     =  0.8696
                          Root MSE      =    .148801    Adj R-squared =  0.8206

                  Source | Partial SS         df         MS        F    Prob>F
              -----------+----------------------------------------------------
                   Model |  1.1807583          3   .39358611     17.78  0.0007
                         |
                      O3 |  .55900832          1   .55900832     25.25  0.0010
                     SO2 |  .37807498          1   .37807498     17.08  0.0033
                   O3#SO2 |  .24367503          1   .24367503     11.01  0.0106
                         |
                Residual |  .17713334          8   .02214167
              -----------+----------------------------------------------------
                   Total |  1.3578917         11    .1234447
```

```
***********************************************
* Q3
* note: fitted line plots and residual plot panels not included here

. regress Gas Temp if insulation=="Before"

      Source |       SS           df       MS      Number of obs   =        26
-------------+----------------------------------   F(1, 24)        =    403.11
       Model |  31.9054318         1  31.9054318   Prob > F        =    0.0000
    Residual |  1.89956816        24  .079148673   R-squared       =    0.9438
-------------+----------------------------------   Adj R-squared   =    0.9415
       Total |      33.805        25      1.3522   Root MSE        =    .28133

------------------------------------------------------------------------------
         Gas |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Temp |  -.3932388    .019586   -20.08   0.000
       _cons |   6.853828   .1184234    57.88   0.000
------------------------------------------------------------------------------
```

```
. predict fit, xb
. predict resid, res
. predict stdres, rstandard

. list Gas fit resid stdres if insulation=="Before" & abs(stdres)>=2

     +---------------------------------------+
     | Gas        fit    residual      stdres |
     |---------------------------------------|
  8. | 4.7   5.320196   -.6201963   -2.260152 |
  9. | 5.8   5.202225    .5977753    2.174129 |
     +---------------------------------------+

. regress Gas Temp if insulation=="After"

      Source |       SS           df       MS      Number of obs   =        30
-------------+----------------------------------   F(1, 28)        =    121.79
       Model |  15.3359874         1  15.3359874   Prob > F        =    0.0000
    Residual |  3.52567925        28  .125917116   R-squared       =    0.8131
-------------+----------------------------------   Adj R-squared   =    0.8064
       Total |  18.8616667        29  .650402299   Root MSE        =    .35485

------------------------------------------------------------------------------
         Gas |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Temp |   -.277935   .0251843   -11.04   0.000
       _cons |    4.72385   .1297394    36.41   0.000
------------------------------------------------------------------------------

. drop fit resid stdres
. predict fit, xb
. predict resid, res
. predict stdres, rstandard

. list Gas fit resid stdres if insulation=="After" & abs(stdres)>=2

     +---------------------------------------+
     | Gas        fit       resid      stdres |
     |---------------------------------------|
 55. | 1.3   2.278022   -.9780221   -2.951643 |
     +---------------------------------------+
```