STATISTICS MASTERS/Ph.D.-QUALIFYING EXAM: TAKE HOME

August, 2020

General directions

Complete both problems in this exam. Your report is to be typed, double spaced, no smaller than ten-point font with one-inch margins, and should be identified by your "CODE WORD" in the header on each page; *do not include your name or UNM ID number.* The report for each problem should be no longer than 4 pages. Each problem is to be no longer than four pages, and an additional four-page appendix is allowed for each problem but will be examined only at the discretion of the graders; the better constructed your appendix with cross-references from the text, the more likely it is to get examined. In your data analysis, raw uninterpreted computer output will be graded as the dross it is.

Write your answers completely, but concisely. Insert tables and figures to support your points. Tables and figures should be well-labelled and cross-referenced from text, such as, "in Table 1 . . . ", or if in the appendix, "in Table A1 . . . " and each should have a caption that describes it and briefly tells the reader why it is of value. Figures should include appropriate symbols suitable for black-and-white reproduction (that is, avoid use of color if possible; consider symbols, line types, and distinct shades of gray to distinguish categories or values).

Organize your sections to justify the validity of what you uncovered and the methods you used to uncover it. We want a summary of what you think is important, not a diary of how you spent your time. Remember that even the best data analysis is worthless if your reader does not understand it, so you are being graded on presentation as well as statistical content.

As necessary:

1. Plot and describe the data (that is, plot all the individual observations, in addition to summaries of data you might present with the results, such as the mean and confidence intervals).
2. Clearly define population parameters and sample statistics.
3. Clearly specify hypotheses tested and explicitly state the associated model at least once (i.e., write the model equation).
4. Define and assess method assumptions.
5. Write a coherent evidence-based conclusion that a layperson can understand.

You may use your course notes as well as any available books or web resources on general statistical methods for the exam. You may not consult any other person when working on this exam or discuss your exam with anyone else, regardless of whether or not the person is taking the exam nor are you allowed to use the internet to find analyses of these data.

Any points of clarification can be directed to Prof. Erik Erhardt, erike@stat.unm.edu.

Email solutions by **3 PM, Fri Aug 14, 2020** to Ana Parra Lombard, aparra@math.unm.edu, Department of Mathematics and Statistics, University of New Mexico. Please do not turn in a physical copy of your solutions.

Problem 1

An experiment was carried out to test the effect of two metals for pistons (metal 1 and metal 2), the amount of primary initiator (5mg and 10mg) and packing pressure (12K and 28K psi) on the "firing time of explosives". For each combination of metal, initiator and pressure, two pistons were randomly selected and tested. Table 1 gives the firing time of explosives in milliseconds from the pistons.

Table 1: Data for Problem 1

	Initiator			
	5mg		10mg	
Metal	Pressure		Pressure	
	12K psi	28K psi	12K psi	28K psi
1	61 56	66 57	62 56	60 59
2	57 56	59 63	67 64	71 65

Describe the statistical experimental design. Fit the simplest linear model that adequately explains the data in this experiment. List all the assumptions and explain the terms in the model. Examine the main effects, interaction terms, and do multiple comparisons of the interested treatments or combination of treatments. Make sure that you carefully assess all the assumptions and write a succinct, coherent, and complete summary of your analysis.

Problem 2

Download the data from

`https://math.unm.edu/sites/default/files/files/qual-exams/stat/unm_exam_202008_stat_qual-takehome_dat2.txt`.

Goal

Develop an explanatory model for the "Tonn/Hect" response variable relating to the other variables that influence growth (this would exclude other "response variables" mentioned below). Limit the analysis to Varieties that have at least 100 rows of data. You might consider reducing the rainfall variables to a single value, such as averaging or summing the rainfall for a relevant set of months. State and assess model assumptions. (If assumptions are not met, try to address that. If you can not satisfy model assumptions, mention this and continue anyway, but indicate which results this could impact and how they would be impacted.) Write a succinct, coherent, and complete summary of your analysis.

Description

This data gives sugar cane yields for each paddock in the Mulgrave area of North Queensland for the 1997 sugar cane season. It was obtained by David Gregory and Nick Denman for their MS305 data project at The University of Queensland in 1998.

Mulgrave is a region in North Queensland around the Mulgrave river and the city of Cairns. Sugar cane is the primary industry in Mulgrave, and all sugar cane from the area is processed through the Mulgrave Central Mill. The data was provided by the Bureau of Sugar Experimental Stations (BSES) on behalf of the Mulgrave Central Mill.

Response Variables

The response variables are the tonnes per hectare of cane produced by each paddock, the fibre per rake and the commercial sugar content per rake produced. There is a payoff between quantity (tonnage) and sugar contenct (quality). Some varieties of sugar cane have been developed to have higher sugar content under some soil conditions, and other varieties to give a better tonnage.

District

The Mulgrave area has been divided by the BSES into fifteen districts. The BSES has also divided the districts into five larger regions based on physical position and average rainfall. Another simpler grouping of the fifteen districts into compass directions was used by Denman and Gregory (1998).

Soil Type

Soil is an important factor that will determine crop's performance. There are numerous types of soils, each having separate characteristics such as nutrient content, acidity and drainage. The name of the soil type of each paddock is provided. There is also a more detailed soil ID available. Farmers also use fertilisers to complement or offset a soil's nutrient content, but no information on fertilising regime is available.

Area

Larger paddocks will have longer rows, which means that more cane can be grown per hectare.

Variety

Some varieties of cane are designed to be able to survive in drier soils without much rain and some are designed to take advantage of nutrient full volcanic soils.

Age

As sugar cane is a grass, it will grow again if cut. A farmer may choose to "plough out" a paddock of cane once it has been harvested. This requires the farmer to plant new cane. Cane planted the year before may be regarded as having age zero. Cane let to grow for one year after being cut (this is, the cane is first ratoon) can be considered to have an age of one. The greatest age in this data is eigth ratoon, and it can be expected that the sugar content and tonnage of this cane is lower than newly planted cane.

The levels of regrowth found in the original data set are 11, 1O, 1R (first ratoon or year of regrowth), 2R, 3R, 4R, 5R, 6R, 7R, 8R, F1, F2, F6, OR, PL (previously ploughed out, new growth) and RP. To create the variable Age we have coded OR, PL and RP to 0; 11, 10, 1R and F1 to 1; 2R and F2 to 2; and so on. Age therefore represents the number of years of regrowth before harvesting the cane.

Month of Harvest

The sugar cane cutting season usually begins in June and concludes in mid-November, the finishing date depending on how the season has gone with respect to rainfall and mill breakdowns. There may be some interaction between month of harvest and variety, as some varieties would be expected to give a higher sugar content at an earlier maturity level than others.

Rainfall

Monthly rainfall totals are given for each district from July 1996 through December 1997. Denman and Gregory (1998) grouped rainfall into (i) cutting season 1996: July through October 1996; (ii) wet season 1996/1997: November 1996 through February 1997; (iii) off season 1997: March through June 1997; and cutting season 1997: July through October 1997. Note that November and December 1997 rainfall was not used because most cane has already been cut by this time.

Table 2: Variable descriptions for Problem 2

Variable	Description
District	Name of district
DistrictGroup	District grouping by BSES into 5 geographical and rainfall regions
DistrictPosition	Simple grouping of districts into North, South, East, West and Central (N, S, E, W, C)
SoilID	Soil type: detailed ID number
SoilName	Soil type: general name
Area	Area of paddock (hectares)
Variety	Sugar can variety
Ratoon	Ratoon or regrowth age of cane.
Age	Number of years of regrowth before harvesting of cane. Recoded from Ratoon.
HarvestMonth	Month in which harvest was started
HarvestDuration	Duration of harvest in days
Tonn/Hect	Tonnes per hectare of cane harvested
Fibre	Fibre content per rake
Sugar	Commercial sugar content per rake
Jul-96	Rainfall for the district for July 1996
Aug-96	Rainfall for the district for August 1996
:	:
Dec-87	Rainfall for the district for December 1997

Source

These data come from http://www.statsci.org/data/oz/cane.html.

Denman, N., and Gregory, D. (1998). Analysis of Sugar Cane Yields in the Mulgrave Area, for the 1997 Sugar Cane Season. MS305 Data Analysis Project, Department of Mathematics, University of Queensland.

Data originally supplied by Doug Paton and Trevor Crooks from the Bureau of Sugar Experimental Stations, Queensland.