Biostats VHM 801 Course
Fall 2020, Atlantic Veterinary College, PEI
Henrik Stryhn

Final exam, 14 December 2020

The final exam is a 24-hour take home exam, and worth 20% of the course mark. The exam is open book, but must be completed by each student individually without assistance from other people. By handing in the exam you implicitly acknowledge to have read, accepted, and agreed to comply with the instructions for the exam, as presented in the page entitled "Instructions for home assignments, quizzes and exam" at the VHM 801 course homepage.

The exam consists of two equally-weighted questions that should both be answered. Further weights are indicated for subquestions within the two questions.

Question 1. (*10 points*)

In a study on the antibody response to vaccination with a particular vaccine, 20 adults had their antibody levels measured before the immunization (`pre`) and 4 weeks after the immunization (`post`). The table below shows the values obtained for most of the participants. Some values have been masked (by $*$) in the table so as to discourage attempts to enter the data into statistical software for analysis; instead you should use the Minitab listings provided below.

	Antibody concentration		
Subject	`pre`	`post`	`post-pre`
1	0.4	0.4	0.0
2	0.4	0.5	0.1
3	0.4	0.5	0.1
4	0.4	0.9	0.5
5	$*$	$*$	$*$
6	0.5	0.5	0.0
7	0.5	0.5	0.0
8	0.5	0.5	0.0
9	0.5	0.5	0.0
10	$*$	$*$	$*$
11	0.6	12.2	11.6
12	0.7	1.1	0.4
13	0.7	1.2	0.5
14	0.8	0.8	0.0
15	$*$	$*$	$*$
16	0.9	1.9	1.0
17	1.0	0.9	-0.1
18	1.0	2.0	1.0
19	1.6	8.1	6.5
20	2.0	3.7	1.7

In a journal article describing the study, the analysis of these data was summarized as: "$t = 1.8; P > 0.05$", and referred to as "no significant increase in antibody [...]".

a) (*3 points*)
Based on this statement, identify the type of analysis carried out and the assumptions/model on which it is based. Here, and in the following, you may use the information contained in any of the Minitab listings below.

b) (*2 points*)
Carry out a brief descriptive analysis for the variable(s) of principal interest. You may assume that the primary objective of the study was to compare antibody concentrations prior to and after immunization.

c) (*3 points*)
Carry out your own preferred analysis of the data, with the objective of obtaining a statistical comparison of antibody concentrations prior to and after immunization, and draw conclusions. Make sure to include the statistical model/assumptions behind your analysis, and the conclusions from the analysis.

d) (*2 points*)
Give a critical evaluation of the summary of the analysis in the above quote from the journal article. For example, you may address whether it is correct and whether it is an adequate summary of the analysis performed.

Minitab listings for Question 1:

ANTIBODY.MTW

Descriptive Statistics: pre, post, post-pre

Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
pre	20	0	0.7450	0.0933	0.4174	0.4000	0.5000	0.6000	0.9000	2.0000
post	20	0	1.925	0.669	2.993	0.400	0.500	0.850	1.725	12.200
post-pre	20	0	1.180	0.638	2.853	-0.100	0.000	0.100	0.875	11.600

ANTIBODY.MTW

One-Sample T: pre, post, post-pre

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean	95% CI for μ
pre	20	0.7450	0.4174	0.0933	(0.5497, 0.9403)
post	20	1.925	2.993	0.669	(0.524, 3.326)
post-pre	20	1.180	2.853	0.638	(-0.155, 2.515)

μ: population mean of pre, post, post-pre

Test

Null hypothesis H_0: $\mu = 0$
Alternative hypothesis H_1: $\mu \neq 0$

Sample	T-Value	P-Value
pre	7.98	0.000
post	2.88	0.010
post-pre	1.85	0.080

ANTIBODY.MTW

Two-Sample T-Test and CI: pre, post

Method

μ_1: population mean of pre
μ_2: population mean of post
Difference: $\mu_1 - \mu_2$

Equal variances are not assumed for this analysis.

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
pre	20	0.745	0.417	0.093
post	20	1.93	2.99	0.67

Estimation for Difference

Difference	95% CI for Difference
-1.180	(-2.595, 0.235)

Test

Null hypothesis H_0: $\mu_1 - \mu_2 = 0$
Alternative hypothesis H_1: $\mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
-1.75	19	0.097

ANTIBODY.MTW

Mann-Whitney: pre, post

Method

η_1: median of pre
η_2: median of post
Difference: $\eta_1 - \eta_2$

Descriptive Statistics

Sample	N	Median
pre	20	0.60
post	20	0.85

Estimation for Difference

Difference	CI for Difference	Achieved Confidence
-0.100000	(-0.5, 0.0000000)	95.01%

Test

Null hypothesis H_0: $\eta_1 - \eta_2 = 0$
Alternative hypothesis H_1: $\eta_1 - \eta_2 \neq 0$

Method	W-Value	P-Value
Not adjusted for ties	356.50	0.152
Adjusted for ties	356.50	0.145

ANTIBODY.MTW

Sign Test for Median: pre, post, post-pre

Method

η: median of pre, post, post-pre

Descriptive Statistics

Sample	N	Median
pre	20	0.60
post	20	0.85
post-pre	20	0.10

Test

Null hypothesis H_0: $\eta = 0$
Alternative hypothesis H_1: $\eta \neq 0$

Sample	Number < 0	Number = 0	Number > 0	P-Value
pre	0	0	20	0.000
post	0	0	20	0.000
post-pre	1	8	11	0.006

ANTIBODY.MTW

Wilcoxon Signed Rank Test: pre, post, post-pre

Method

η: median of pre, post, post-pre

Descriptive Statistics

Sample	N	Median
pre	20	0.675
post	20	0.925
post-pre	20	0.250

Test

Null hypothesis H_0: $\eta = 0$
Alternative hypothesis H_1: $\eta \neq 0$

Sample	N for Test	Wilcoxon Statistic	P-Value
pre	20	210.00	0.000
post	20	210.00	0.000
post-pre	12	76.00	0.004

Question 2. (*10 points*)

In this question, we consider the body mass index (BMI) for humans. It is calculated as weight divided by squared height, where weight is measured in kg and height is measured in m. For example, a person of weight 70 kg and height 180 cm has $\texttt{bmi} = 70/(1.8 \cdot 1.8) = 21.6$.

As part of a study on dietary habits in a population of a Western country some 30 years back, data were collected on food intake and demographic variables. This question will explore a subset of the data, both in terms of the sample size (1311 subjects, all adults) and of the number of variables included, as shown in the table below. The variable `bmigrp` is a grouping of the BMI values constructed according to the current guidelines for interpretation of BMI, see e.g. relevant Health Canada or CDC (US) websites, with the coding

$$\begin{aligned} \texttt{bmigrp} = \; & 1 \quad \text{for} \quad \texttt{bmi} < 18.5, \\ & 2 \quad \text{for} \quad 18.5 \leq \texttt{bmi} < 25, \\ & 3 \quad \text{for} \quad 25 \leq \texttt{bmi} < 30, \\ & 4 \quad \text{for} \quad \texttt{bmi} \geq 30, \end{aligned}$$

A few values of `bmi` were missing; these are coded by a negative value. Our interest here is in describing relationships between `bmi` and the demographic variables etc. (including also the energy intake).

Variable	Description	Values
`bmi`	body mass index (BMI)	(kg/m^2), negative values correspond to missing values
`bmigrp`	BMI group	$1-4$, negative values correspond to missing values
`age`	age of subject	(years)
`gender`	gender of subject	$1 =$ male, $2 =$ female
`energy`	daily energy intake	(kJ/day)
`urban`	urbanity of subject residence	$1 =$ metropole, $2 =$ city, $3 =$ rural
`socgrp`	social group of subject	$1-5$, where 1 is highest and 5 is lowest

Datafiles in both Minitab (`.mtw`) and comma-separated (`.csv`) file formats for import into Minitab or other statistical software are available from the Moodle account of the course.

Because a dataset of this size may offer challenges beyond what has been discussed in the course, you are required to select a random sub-dataset of size 500 (valid observations for `bmi`). Use your student ID as the seed/base in a reproducible random selection procedure in statistical software to construct such a sub-dataset, and include documentation for how you arrived at the dataset. Include also the dataset itself (in a suitable format, e.g. Minitab or `.csv`) with your submission for the exam.

Although the interest would generally be in modelling the relationships between BMI and all the demographic etc. variables simultaneously, for the purpose of the exam all of the following explorations are considered as equally valuable (in no prioritized order):

(*continues on next page*)

a. `bmi` versus age,

b. `bmi` versus gender,

c. `bmi` versus `energy`,

d. `bmi` versus `urban`,

e. `bmi` versus social group,

f. `bmi` versus gender and social group,

g. `bmigrp` versus gender,

h. `bmigrp` versus social group,

i. `bmi` versus all variables (simultaneously).

Your task for the exam is to conduct full statistical analyses according to **two of the above** specifications. You are free to choose any two analyses from the above list as you want, but they must be for different specifications; this means that two alternative analyses for the same specification only counts as one analysis. The two analyses will be worth **5 points each**, for a total of 10 points for this question. It is allowed (but *not recommended*) to include one extra analysis, in which case the mark will be for the two best analyses of the three. If more than three analyses are included, only the first three as they appear in the text will be considered.

Recall that a full statistical analysis should include: a stated specific objective for the analysis, relevant descriptive analyses of the outcome variable, a statistical model describing the model assumptions, assessment of the model assumptions, estimates for the (relevant) model parameters with confidence intervals (whenever suitable), relevant statistical test(s), and conclusions worded both in statistical terms and non-technical terms.

Note also, for your information, and without any pressure for you to utilize this extra information, that a potential transformation of the outcome (BMI) worth exploring is the inverse BMI, that is, the variable: 1/`bmi` (not included with the dataset).