Biostats VHM 801 Course
Fall 2023, Atlantic Veterinary College, PEI
Henrik Stryhn

Mid-term exam, 24 October 2023

All aids are allowed, except a computer and personal assistance, as well as the use of any external information pertaining to the specific data and questions. 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 one question with six parts, or subquestions (labeled by letters **(a) – (f)**), of which you in order to achieve the maximal 15 points have to answer the first two (**(a),(b)**) and three out of the four remaining parts (**(c)-(f)**). It is allowed to answer all parts; your score will then be for the best three answers among (**(c)-(f)**). The parts can be answered independently of each other.

The mid-term exam accounts for 15% of the course mark; however, every student may choose to waive the result of the mid-term exam. The duration of the mid-term exam is 1 hour.

Generally, **statistical models and methods should be specified**, and all statistical analyses should be summarized in conclusions.

Question 1 (*15 points*)

An older study, described as "a double-blind crossover trial", investigated the effects of fiber diets on people's serum cholesterol levels. Twenty healthy subjects (4 men and 16 women, 23 to 49 years old, selected as volunteers among dietitians and other employees of a hospital) underwent three dietary regimes. After a one-week baseline period during which they consumed their usual diets, the subjects were given two types of supplements: a high-fiber oat bran and a low-fiber refined-wheat product, each for separate six-week periods. Blood samples were taken on the final days of each period and analyzed (among other things) for total cholesterol and LDL ("bad") cholesterol. Simple descriptive statistics for the 20 subjects in each period are represented as Mean ± SD (standard deviation) in the table below.

Mean ± SD (mg/dL)	Period/Diet		
	Baseline	High fiber	Low fiber
Total cholesterol	186 ± 31	172 ± 28	172 ± 25
LDL cholesterol	115 ± 23	104 ± 24	107 ± 23

In addition to the information in the table, you may also use any information you find useful from the attached Minitab listings (roughly equivalent Stata listings available upon request).

(a) (*3 points*) The standard recommendation for total cholesterol levels (at the time of the study) was that levels below 200 (mg/dL) are desirable, whereas values in the range $200-239$ are considered as borderline high and values at or above 240 are considered as high. Based on an assumed normal distribution of the cholesterol values among the subjects, compute the proportion of people (in the population the study subjects represent) who, when not given any dietary supplementation, have a cholesterol level

that is either borderline high or high. A more recently developed scale for cholesterol levels characterizes LDL cholesterol values of at least 130 (mg/dL) as borderline high. Does this new definition identify a larger part of the population as having borderline high cholesterol levels?

(b) (*3 points*) For this part we focus on how many of the subjects have high total cholesterol values. Assume that each subject has a 0.04 (or 4%) probability to have a high total cholesterol value. Compute the expected number of persons with high cholesterol levels in a sample of 20 persons. What is the probability that none of the 20 persons has a high cholesterol level? Calculate also the probability that more than one of the 20 persons have a high cholesterol level. (*Note:* if you cannot determine some of these probabilities exactly, give instead suitable bounds for them.)

Answer any **3 out of the next 4** parts (as explained in the introduction).

(c) (*3 points*) Give a 95% confidence interval for the mean total cholesterol level in the population represented by the subjects, when not given any dietary supplementation. Based on this confidence interval, would you say that the mean cholesterol level in the population is elevated (that is, at least borderline high)? Explain your reasoning.

(d) (*3 points*) One of the key research questions was whether the dietary supplementation had any effect on total cholesterol levels when compared to baseline levels. For the purpose of this part, we will focus on the high fiber supplement. Estimate the mean change in total cholesterol levels resulting from the supplement. Use the information provided to carry out as much as possible of a statistical analysis to determine whether the change was statistically significant. If you complete the analysis, draw conclusions from your results. If you think that further information is necessary in order to complete the analysis, explain what information is missing and how you would use it to complete the analysis.

(e) (*3 points*) Another key research question was whether any differences could be demonstrated between the two diets. Despite the blinding of the participants to the diets, the researchers suspected that the participants might have been able to guess the identity of the two diets. Interviews with the participants after the study revealed that 17 out of the 20 subjects had correctly identified the two diets. Give a statistical assessment (by a statistical test) of whether this finding could have happened by chance alone.

(f) (*3 points*) A journal article describing the study reported the difference between high and low fiber diets to be non-significant, and stated the conclusion as follows:

> "We conclude that oat bran has little cholesterol-lowering effect and that high-fiber and low-fiber dietary grain supplements reduce serum cholesterol levels about equally [...]."

Subsequently the journal (one of the most well-respected journals in medicine) was criticized for publishing the article with this conclusion. Suggest some reasons (at least two) that the conclusion could be considered as a misrepresentation of the findings in the study? If you identify any issues with the study, discuss briefly (based on the information provided) in what way each such issue might have impacted the results.

Minitab listings:

WORKSHEET 1

One-Sample Z

Descriptive Statistics

N	Mean	SE Mean	95% CI for μ
20	186.00	6.93	(172.41, 199.59)

μ: population mean of Sample
Known standard deviation = 31

WORKSHEET 1

One-Sample T

Descriptive Statistics

N	Mean	StDev	SE Mean	95% CI for μ
20	186.00	31.00	6.93	(171.49, 200.51)

μ: population mean of Sample

WORKSHEET 1

One-Sample T

Descriptive Statistics

N	Mean	StDev	SE Mean	95% CI for μ
20	186.0	138.6	31.0	(121.1, 250.9)

μ: population mean of Sample

WORKSHEET 1

One-Sample T

Descriptive Statistics

N	Mean	StDev	SE Mean	95% CI for μ
20	172.00	28.00	6.26	(158.90, 185.10)

μ: population mean of Sample

WORKSHEET 1

Two-Sample T-Test and CI

Method

μ_1: population mean of Sample 1
μ_2: population mean of Sample 2
Difference: $\mu_1 - \mu_2$

Equal variances are not assumed for this analysis.

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Sample 1	20	186.0	31.0	6.9
Sample 2	20	172.0	28.0	6.3

Estimation for Difference

Difference	95% CI for Difference
14.00	(-4.93, 32.93)

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
1.50	37	0.142