

December 17, 2024

**EXAM IN HMET4101, 28 NOVEMBER 2024:  
SUGGESTED SOLUTIONS AND COMMENTS**

Knut R. Wangen  
k.r.wangen@medisin.uio.no

INTRODUCTION

The exam was arranged digitally on the Inspera platform and consisted of 10 questions. The first two questions were multiple choice questions in which the candidates should choose one of a set of alternative answers. The remaining questions were in an "essay" format and the answers had to be written as text.

The multiple choice questions were automatically assessed, while the "essay" questions were manually assessed. The maximum scores per question were between 10 and 18 points. In total, the maximum score was between 0 and 120 points.

The following relation between grades (A-F) and points [0-120] was used: A: [107–120], B: [91–106], C: [73–90], D: [56–72], E: [45–55], F: [0–44].

**Explanation of grades.** After the grades are published, the candidates will have access to their individual scores by logging in to Inspera. The suggested solution presented here, which will be distributed to the candidates, explains the general principles for how the scores and grades were awarded. A candidate's detailed scores on the exam questions and the general suggested solution are together considered as the formal explanation of the grade, and no further formal explanations will be given.

**Appeal against grades.** Appeals against grades must be submitted following the procedure described on the course website (that is, not submitted to the course coordinator).

**Informal feedback.** A candidate who wishes to receive additional feedback can ask the course coordinator for an informal talk (phone call or Zoom meeting). The candidate must then email the course coordinator 2-3 suggested times with at least two workdays notice. The suggested times should be within normal office hours (0900h-1700h) in the period 2–10 January 2025.

QUESTION 1

$$P(4 \leq X \leq 10) = P(X \leq 10) - P(X \leq 3) = 0.8788 - 0.0674 = \underline{\underline{0.811}}.$$

QUESTION 2

$$\text{The p-value is } 2 \cdot P(Z < -1.43) = 2 \cdot 0.0764 = \underline{\underline{0.153}}$$

## QUESTION 3

Let  $X$  denote the number of individuals in the group who have a HBV infection. a) The probability is  $P(X \geq 1) = 1 - P(X = 0) = 1 - 0.1216 = 0.878$  or 87.8%.

b) The following three assumptions should be satisfied: 1) Each individual (trial) has only two outcomes, HBV positive or not. 2) The probability of being HBV positive should be the same for each individual. In this case, we only know the prevalence in the demographic group (age and sex). Given the available data, it is reasonable to assume that the probability is the same for all recruits. 3) The probability of being HBV positive for one individual does not affect the probability for others for two reasons: First, we disregard the possibility of HBV transmission among recruits. Second, the recruits are selected from a large population, and the infection status of one recruit does not noticeably alter the prevalence within the remaining population.

## QUESTION 4

We substitute the given values into the confidence interval formula:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

which becomes:

$$0.05 \pm 2.576 \sqrt{\frac{0.05(1-0.05)}{6000}}.$$

This yields the estimated 99% confidence interval (0.043, 0.057).

## QUESTION 5

Let  $p$  represent the population proportion of radio listeners, with subscripts  $x$  and  $y$  indicating the years 2022 and 2023 respectively. Let  $n$  denote the sample size.

A main assumption is that the samples in the two years are independent. The formulas are also based on the assumption that the sample sizes are small compared to the population sizes. It follows from the central limit theorem that we can use the normal distribution as an approximation because the sample sizes are large.

Hypothesis test: The null hypothesis is that there is no change in the population proportion from 2022 to 2023, that is:  $H_0 : p_x - p_y = 0$ . The alternative hypothesis is that the population proportions differs in the two years:  $H_1 : p_x - p_y \neq 0$ . The estimate for the common proportion (under  $H_0$ ) is:

$$\hat{p}_0 = \frac{n_x p_x + n_y p_y}{n_x + n_y} = \frac{6000 \cdot 0.49 + 6000 \cdot 0.48}{6000 + 6000} = 0.485.$$

The observed test statistic is:

$$Z_{obs} = \frac{p_x - p_y}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_x} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_y}}} = \frac{0.49 - 0.48}{\sqrt{\frac{0.485(1-0.485)}{6000} + \frac{0.485(1-0.485)}{6000}}} = 1.10.$$

The p-value is  $P(Z < -Z_{obs}) + P(Z > Z_{obs}) = 2 \cdot 0.1357 = 0.2714$ . We would keep the null hypothesis at any conventional significance level, and we can conclude that the change from 2022 to 2023 is not statistically significant.

Confidence interval: The confidence interval for the difference in population proportions can be estimated based on:

$$(\hat{p}_x - \hat{p}_y) \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}_x(1 - \hat{p}_x)}{n_x} + \frac{\hat{p}_y(1 - \hat{p}_y)}{n_y}}$$

Choosing a 90% confidence level and inserting the given sample data:

$$(0.49 - 0.48) \pm 1.645 \sqrt{\frac{0.49(1 - 0.49)}{6000} + \frac{0.48(1 - 0.48)}{6000}},$$

which, when simplified is:

$$0.01 \pm 0.015.$$

This yields the estimated 90% confidence interval  $(-0.005, 0.025)$ . This interval contains the value 0, which corresponds to the case where the population proportions are identical. Thus, the estimated confidence interval does not suggest that the population proportions are different. (This implies that the observed difference in sample proportions is not statistically significant at the 10% significance level.)

#### QUESTION 6

a) The standard error is given by  $SE = s/\sqrt{n}$ . Solving for  $s$  and inserting the data given in the text we get

$$s = SE \cdot \sqrt{n} = 1399\sqrt{300} = 23192.$$

b) We apply the formula

$$n = \frac{Z_{\alpha/2}^2 \cdot \sigma^2}{ME^2}.$$

The formula contains the population standard deviation,  $\sigma$ , which is not available. Instead we use the sample standard deviation from part a), and insert values:

$$n = \frac{1.96^2 \cdot 23192^2}{500^2} = 8265.2.$$

We need a sample size of at least 8266 in order to obtain a 95% confidence interval with a margin of error equal to NOK 500. The formula is based on the assumptions that the population has a normal distribution, and that the population standard deviation (or variance) is known. In the calculations above, we have assumed that the population variance is equal to the sample variance. Given that we do not know whether the population is normal and that we do not know the population variance, care should be taken before applying the results.

## QUESTION 7

The output states that this is a Pearson correlation analysis. There is a very weak, but significant negative correlation between daily cost and dietary quality, as the estimate is -0.04 with a p-value of 0.00678 (5p). Hence, higher expenditure on food is associated with slightly lower dietary quality. For a Pearson correlation analysis, both variables should be normally distributed and there should be a linear relationship between them (5p). A 1 USD higher daily expenditure on food is associated with 0.19 points lower dietary quality (5p).

## QUESTION 8

Females have 2.98 points higher dietary quality than males (3p). 95% confidence interval for the effect of female:  $2.98 \pm 1.96 \cdot 0.44 \Rightarrow (2.12, 3.84)$  (6p). The point here is to understand from the output that there are more than 4000 observations in the data (from the d.f. in the output, 4130 to be exact). Hence assuming a normal distribution and using  $\pm 1.96$  when calculating the confidence interval should be sufficient (4p). The predicted dietary score for females aged 50, having a daily food expenditure of 7 USD would be:  $41.24 + 0.16 \cdot 7 + 0.17 \cdot 50 + 2.99 = 53.85$  points (5p).

## QUESTION 9

We cannot know if there is any confounding effects on age and gender, but comparing the coefficients for daily costs in the output to this exercise to the output in Linear regression 1, we see that the effect has changed quite a bit and also switched signs (6p). Hence, age or gender of both could be confounding factors for the association between daily costs and dietary quality. As both higher age and females vs. males are associated with higher dietary quality, one explanation for the confounding would be that there is a higher frequency of older responders and female responders among those with lower daily food costs. Without adjusting for age and gender, it would then appear as if higher daily costs is negatively associated with dietary quality. Adjusting for these factors removes this false association (6p).

## QUESTION 10

From the output, responders in the category active have 0.64 points (non-significant) higher dietary score than the reference category inactive (adjusted for age, gender and daily costs), while responders in the category very active have 1.71 points higher dietary score than the inactive group (significant) (5p). An interaction between activity level and gender would imply that the difference in dietary quality between the genders depends on the activity level. Perhaps the difference in dietary quality between males and females is smaller among the active/very active responders than in the inactive group (5p). This could be relevant; we do not know from the analyses so far whether the gender difference applies to responders in general or mainly to the inactive responders only (5p). (It would also imply that the difference in dietary quality between activity levels is greater for males than for females).