Statistics 2 - Exam 4 April 2023

You have 105 minutes. You can use all written material. You can use a pocket calculator but no device that is able to access the internet or send or receive messages of any kind.

All numbers in the questions are invented. Sometimes for tests it isn't directly indicated against what alternative the null hypothesis should be tested. In these cases you need to decide this from the "story".

Marking: I will mark on a percentage scale (1-100) and then transform results to the range 1-30 (you can only achieve a maximum of 25 marks but for achieving the maximum you don't need 100%). If I find out from your solutions that the exam was too difficult, I may take this into account for the transformation. Your marks from submitting regular exercises still count.
Here are the percentages carried by the questions: Q1 20 [(a) 4 (b) 5 (c) 8 (d) 3], Q2 32 [(a) 3 (b) 5 (c) 8 (d) 3 (e) 10 (f) 3], Q3 14 [(a) 3 (b) 4 (c) 4 (d) 3], Q4 14 [(a) 10 (b) 4], Q5 20 (4 each).

1. On bottles of a brand of alkaline drinking water, a label indicates that the pH value is 8.5. A competing company suspects that this indication is wrong. They have commissioned a laboratory to measure the exact pH values of a sample of eight bottles in order to find evidence in case the indication might be systematically wrong. The measured values are

$$8.21; 8.41; 8.48; 8.52; 8.58; 8.63; 8.28; 8.42.$$

Assume that these measurements behave like a random sample from a normal distribution.

(a) Estimate the mean and the variance of the distribution of the pH measurements.

(b) Compute a 95% confidence interval for the true mean pH value.

(c) Test at a level of 5% whether the indication of the bottles is on average correct. Give an interval as small as possible from a statistical table for the p-value. Interpret the result in a way understandable to a non-statistician from the company.

(d) What can be said about the validity of the test in case the underlying distribution is not normal?

2. Sara and Evelyn have met on a runners forum. They both run the 400m hurdles. As they are in different countries, they never run against each other, but they are comparing the times (in seconds) of their last six runs to see who is faster. These are the times:

Sara 62.4; 65.2; 62.1; 61.8; 64.3; 64.0

Evelyn 62.0; 59.8; 60.2; 61.9; 63.0; 59.7

(a) Who of the runners has the better times, on average, and by how much?

(b) Estimate from the data the sampling variance of the difference of the two sampling means as computed in (a).

(c) Use a two-sample t-test to see whether one of the runners is significantly better than the other. Interpret the result, indicating the strength of the evidence.

(d) State all the assumptions of the two-sample t-test.

(e) Using a Wilcoxon rank sum test, test the null hypothesis that the two runners are on average equal. Using a statistical table, give an interval as short as possible for the p-value, and interpret the result in a way understandable for a non-statistician, indicating the strength of the evidence.

(f) Describe a situation in which the Wilcoxon test is preferable to the two-sample t-test. Is this situation given for the data in this question?

3. Define the distribution $\mathcal{T}(\tau)$ for $0 \leq \tau \leq \frac{1}{2}$ so that if $X \sim \mathcal{T}(\tau)$, X can take the values 1,2, and 3, $P\{X = 1\} = P\{X = 2\} = \tau,\ P\{X = 3\} = 1 - 2\tau$.

Assume $X_1, \ldots, X_n$ i.i.d. according to $\mathcal{T}(\tau)$. Consider

$$T(X_1, \ldots, X_n) = \frac{1}{2n} \sum_{i=1}^{n} [1(X_i = 1) + 1(X_i = 2)]$$

as estimator for τ. $1(X_i = j)$ is 1 if $x_i = j$ and 0 otherwise, which means that $T(X_1, \ldots, X_n)$ is half the relative frequency of observing either 1 or 2. Note that the random variable $Y_i = 1(X_i = 1) + 1(X_i = 2)$ has a Bernoulli(2τ)-distribution, so that

$$\mathrm{E}(Y_i) = 2\tau,\ \mathrm{V}(Y_i) = 2\tau(1 - 2\tau).$$

(a) Is T unbiased as estimator for τ? Prove your answer.

(b) Compute the MSE of T as estimator for τ.

(c) Is T asymptotically unbiased as estimator for τ? Prove your answer.

(d) Is T consistent as estimator for τ? Prove your answer.

4. Khalid likes to have a coffee after dinner, but his wife tells him that he will sleep better without it. In order to find out whether the coffee after dinner affects his sleep, Khalid decides to observe this over 100 days, having a coffee after dinner on some days and no coffee on some others. In the next morning he always classifies his sleep as either "good" or "not so good". His results are as follows:

Sleep:	good	not so good	Total
Coffee	41	17	58
No coffee	38	4	42
Total	79	21	100

(a) At the level of 5%, test the null hypothesis that whether Khalid's sleep is good or not so good is independent on whether he has a coffee after dinner. Interpret the result understandable for a non-statistician.

(b) What are potential problems with the experiment that could invalidate the conclusion in (a)? Overall, how reliable do you think is the result from (a)?

5. Indicate whether the following statements are true or false, and give a short reason (one to two sentences).

(a) If we compare the mean squared errors of a biased and an unbiased estimator for the same parameter, it is possible that the unbiased estimator has the larger mean squared error, and it is also possible that the unbiased estimator has the smaller mean squared error.

(b) With a large sample size, a test of a certain parameter may have a large probability to reject its null hypothesis even if the difference between the true parameter value and the null hypothesis is so small that it is irrelevant in practice.

(c) A chi-square test for testing the goodness of fit of a certain distributional assumption (such as the normal distribution) for data can only be applied if the true parameters of the distribution of interest are known.

(d) Given a confidence interval C for the variance of a certain assumed distribution at level α and a null hypothesis that the variance is equal to a prespecified value, say σ_0^2, a test of level α of this null hypothesis can be defined by rejecting the null hypothesis if and only if $\sigma_0^2 \notin C$.

(e) In many situations in which statistical inference is applied, the population to which the inference refers does not really exist, but is only an idealisation.