M2: HW-1
Due: 03.01.2021 23:59 PM

Note: Data sets for the questions can be found in HW1.xlsx.

Question-1:

The accuracy of a scale is to be controlled by weighing the same product (of size 10-kg) 25 times. Assume that different weighings are independent of one another and the reading on the scale is normally distributed with $\sigma = 0.2$ kg. Let μ denote the actual average weight reading on the scale. Answer the following:

a) What hypotheses should be tested?
b) Suppose the scale needs recalibrating if either $\bar{X} \geq 10.132$ or $\bar{X} \leq 9.8968$. What is the probability that recalibration is carried out when it is in fact unnecessary?
c) What is the probability that recalibration is judged unnecessary when in fact $\mu = 10.1$? Answer the same question when $\mu = 9.8$.
d) Assume $z = \frac{(\bar{X}-10)}{(\sigma/\sqrt{n})}$. For which value of c is the rejection region of part (b) equivalent to the "two-tailed" region *either* $z \geq c$ *or* $z \leq -c$?
e) If the sample size were only 10 rather than 25, how should the procedure of part (d) be altered so that $\alpha = 0.05$.
f) Using the test in part (e), what would you conclude from the following sample data:

9.981	10.006	9.857	10.107	9.888
9.728	10.439	10.214	10.190	9.793

g) Rewrite the test procedure of part b) in terms of the standardized test statistic $z = \frac{(\bar{X}-10)}{(\sigma/\sqrt{n})}$.

Question-2:

You are given an octehadral dice and after several trials, you believe that the dice is biased. In order to prove your belief, you design an experiment by throwing the dice 80 times and record the outcomes. The result of the experiment is summarized in the table below:

Score	1	2	3	4	5	6	7	8
Frequency	7	10	11	9	12	10	14	7

Construct NHST framework to show that your belief is correct; and using the above data test your claim.

Question-3:

A farm has a herd consisting of 145 cows, which are eligible for milking. Following table presents data from the farm, where cows are grouped according to their weights (three weight categories are used and indicated in column headers). Rows of the table indicate whether a certain bacteria exist in the milk collected from the cows of the given categories.

	Weight Categories		
	Weight Group-1 (<600 kg)	Weight Group-2 (600-700 kg)	Weight Group-3 (>700 kg)
Bacteria (+)	12	13	31
Bacteria (-)	35	40	14

Do you think that the more of the positive bacteria cases occur in heavier cows than do negative bacteria cases?

Construct the NHST framework, describe the framework explicitly; and solve.

Question-4:

Suppose new incoming cohort of Data Science Certificate program are given a multiple choice test before they take the lecture on specific material and then again after they complete the lecture on the subject. We want to find out whether teaching the material leads to improvement in students' knowledge (i.e. test score). The before and after tests results are provided in the table below. Using the data, draw conclusion on whether the lectures have an impact on the performance.

score-before	score-after
18	22
21	25
16	17
22	24
19	16
24	29
17	20
21	23
23	19
18	20
14	15
16	15
16	18
19	26
18	18
20	24
12	18
22	25
15	19
17	16

Question-5:

200 blood samples are collected and level of certain protein in each sample is detected using a special blood test. The results provided in Sheet Question5 of HW1.xlsx.

Can we assume that the protein level is normally distributed? Name your method, describe it in details and answer the question using the selected method of yours.

Hint: Considering clustering the data into equal-sized (expectedly) bins. You may use 10 bins.

Question-6:

Assume that Erhan Yazıcıoğlu hosts the game show "Seç Bakalım". There are three doors: one hides a car and two hide goats. The contestant picks a door, which is not opened. Yazıcıoğlu then opens another door which has nothing behind it. Finally, contestant must decide whether to stay with her original choice or switch to the other unopened door. The problem asks which is the better strategy: staying or switching?

To be precise, let's label the door that contestant picks first by A, and the other two doors by B and C. Hypothesis A is that the car is behind door A, and similarly for hypotheses B and C.

(a) In the usual formulation, Yazıcıoğlu is sober and knows the locations of the car and goats. So if the contestant picks a door with a goat, Yazıcıoğlu always opens the other door with a goat. And if the contestant picks the door with a car, Yazıcıoğlu opens one of the other two doors at random. Suppose that sober Yazıcıoğlu opens door B, revealing a goat (this is the data). Make a Bayes' table with prior, likelihood and posterior. Use the posterior probabilities to determine the best strategy.

(b) Now suppose that Yazıcıoğlu is drunk, i.e. he has completely forgotten where the car is and is only aware enough to open one of the doors not chosen by the contestant at random. It's entirely possible he might accidentally reveal the car, ruining the show.

Suppose that drunk Yazıcıoğlu opens door B, revealing a goat. Make a Bayes' table with prior, likelihood and posterior. Use the posterior probabilities to determine the best strategy. (Hint: the data is the same but the likelihood function is not.)

(c) Based on Yazıcıoğlu's pre-show behavior, contestant thinks that Yazıcıoğlu is sober with probability 0.7 and drunk with probability 0.3. Repeat the analysis from parts (a) and (b) in this situation.

Question-7:

How well do exams given during the semester predict performance on the final (represented with variable *s*)? One class had three tests during the semester. Computer output of the regression gives the following:

Predictor	Coeff	SE(Coeff)	t	P-value
Intercept	−6.72	14.00	−0.48	0.636
Test1	0.2560	0.2274	1.13	0.274
Test2	0.3912	0.2198	1.78	0.091
Test3	0.9015	0.2086	4.32	<0.0001

Analysis of Variance

Source	DF	SS	MS	F	P-value
Regression	3	11961.8	3987.3	22.02	<0.0001
Error	19	3440.8	181.1		
Total	22	15402.6			

a) Write the equation of the regression model.
b) What can you conclude about the significance of the regression model? Explain.
c) Is following correct, explain: "Unit increase in final score means an increase of 0.2560 in Test1 score."
d) Calculate R^2 value.
e) How much of the variation in final exam scores is accounted for by the regression model?
f) Explain in context what the coefficient of *Test3* scores means.
g) A student argues the first exam doesn't help to predict final performance, and suggests that this exam not be given at all. Does *Test1* have no effect on the final exam score? Can you tell from this model? (Hint: Do you think test scores are related to each other?)
h) Examine following plots and discuss whether the assumptions for the regression seem reasonable.