Homework 4

Statistical Inference, Fall 2021

1- Determine if the following statements are true or false, and explain your reasoning for the statements you identify as false.

a. ANOVA is robust with violations to assumptions of homogeneity of variance, so if you fail Levene's test (Normality condition) by a small amount, you're still ok.

b. F ratio has no Negative values and is positively skewed.

c. Large mean differences and small sample variances in ANOVA, are most likely to produce a large value for the F-ratio.

d. In ANOVA, the More variance within the groups, the greater the noise (variance).

e. A researcher reports an F-ratio with df = 3, 36 for an independent-measures experiment. 39 individual subjects participated in the experiment?

f. Let's assume that the main sample is (39, 61, 43, 55, 48); a bootstrap sample would be like: (43, 39, 56, 43, 61)

g. Consider this hypothesis test, $\bar{x}_1 = 3.75, s_1 = 5.25251, n_1 = 20, \bar{x}_2 = 5.25, s_2 = 2.33735, n_2 = 25$; the exact 95% CI would be $1.5 \pm 2.09 \times 1.26408$.

h. Someone wants to investigate the visual memory in boys and girls under 17; the valid result is the outcome of a paired t-student test.

i. If we use not paired t student test for two correlated groups, we are decreasing the power of the test.

j. When comparing means of two samples where $n_1 = 20$ and $n_2 = 40$, we can use the normal distribution for the difference in means since $n_2 \geq 30$.

k. Suppose the null hypothesis that the means of four groups are all the same is rejected using ANOVA at a 5% significance level. In that case, The pairwise analysis will identify at least one pair of means that are significantly different.

2- Consider the following ANOVA table for between-subjects with 25 total subjects. Fill in missing parts. Are the results significant? What is the conclusion?

	SS	Df	MS	F
Between	?	4	?	?
Within	?	20	5	
Total	120	?		

Homework 4

Statistical Inference, Fall 2021

3- **(R)** Student t-test is used to compare two groups, ANOVA generalizes the t-test beyond two groups, so it is used to compare three or more groups. You have to install the built-in **"penguins"** dataset in R.

a. We only need two-column. Select flipper length and the species.

b. Plot(using ggplot2) flipper length of each species.

c. Check assumptions of ANOVA. Plot normality assumption.

d. Plot Equality of variances (instead of boxplot, use dotplot).

e. Summarize descriptive statistics(mean and sd of each species based on their flipper_length) as a table.

f. Do ANOVA test. Interpret and show the result using the **"report"** library.

g. How many comparisons do we have if we want to use multiple comparisons? What is Bonferroni correction for alpha?

4- A committee wants to investigate the perfect average age in Iran's national basketball team; the perfect interval of average age is an interval that guarantees that the players are not too old or too rookie for the job. Let us assume they gathered 50 samples from the national team players' information. The committee wants to compare the national team data with a sample of 40 observations of a team that believes it is a perfect aged team. Here you come as a data specialist:

	Mean	Standard Deviation
Iran's national basketball team	26.9	5.6
The model of a perfect aged basketball team	29	2.4

a. What would you conclude from this data? (α=0.05)?

b. If the average age of a perfect team is 28, what is the probability of type-II error?

c. Identify if there is a possibility you made a type-II error in answering part (a)?

5- In the previous question, calculate the power of estimate to reject the committee null hypothesis. Consider the estimated perfect aged team be: 28.5, 29, 29.5, 30, 30.5, 31.
(Note: You have to do it by hand to compute the power.)

a. Plot the power curve.

b. Interpolate the powers that you calculated. (Explain the increase or decrease of the curve)

c. Assume the sample size is constant, now change the α to 0.01, without any computation estimate the curve. (Your answer may not be accurate.)

d. Change both sample sizes to 20, then repeat part (c).

Homework 4

Statistical Inference, Fall 2021

6- We recorded a sample of people in Ali Shariati's speech in two different scenarios: the events without any catering and the events with catering; assume that the sample is fair and a good representative. Use a statistical test to investigate that if the catering did attract more audience to his speeches overall. Consider $\alpha = 0.05$. If you couldn't reject the null hypothesis, determine how much can you increase α approximately? (You are not allowed to use R.)

Catering	1284	1272	1256	1254	1242	1274
No Catering	1294	1279	1274	1264	1263	1254
	1240	1232	1220	1218	1210	

7- A leading tech company has used an AI method in some of its products. We want to measure if there are any differences between the two methods so, we gathered scores from 15 users as below. Test the hypothesis.

1	User	Without	With AI
2	1	210	197
3	2	205	195
4	3	193	191
5	4	182	174
6	5	259	236
7	6	239	226
8	7	164	157
9	8	197	196
10	9	222	201
11	10	211	196
12	11	187	181
13	12	175	164
14	13	186	181
15	14	243	229
16	15	246	231

8- Find confidence interval of parameter θ for each probability density function (α *is known*):

a. $f(x|\theta) = 1;\ \frac{2\theta-1}{2} < x < \frac{2\theta+1}{2}$

b. $f(x|\theta) = \frac{2x}{\theta};\ 0 < x < \theta;\quad 0 < \theta$

9- A university in Guinea Bissau wants to estimate overall performance; Hence, some research groups' published the below table as a result. They have assumed that if the overall performance is below 35 papers per year, they must shut down the university.

a. Assuming the one-year periods are fairly typical of the volumes throughout the years, is there significant evidence that the average published paper is greater than 35? Compute the p-value for the test statistic. What is the 95% CI?

Homework 4

Statistical Inference, Fall 2021

b. **(R)** Use R to obtain 1,000 bootstrap samples from the data in the table below. Use these 1,000 samples to get the bootstrap p-value for the t-test of the hypothesis that ends up surviving university. Explain your answer.

c. **(R)** Use the bootstrap data to compute the p-value and the CI, compare it with part (a).

Published Papers	25	27	35	42	28	37	40	31	29
	33	30	26	31	28	30	15		

10- **(R)** Answer the following questions considering this Dataset.

a. Choose a numerical variable.
   i. For the mean value of this numerical variable, design a hypothesis test and by finding the p-value, confirm or reject your assumption. Interpret this p-value.
   ii. Calculate a 95% CI for this numerical variable. Based on this confidence interval, do these data support the hypothesis that you have designed? Explain it.
   iii. Calculate type II error.
   iv. Calculate the power and Explain the relationship between the power and effect size. (plot a curve)
   v. Choose a tiny sample, design a bootstrap scenario, and repeat (i) and (ii). Explain if there is a difference or not.
   vi. If you have performed part (v) via the standard error method, repeat your test via percentile method and vice versa.
   vii. Is there any difference between these two calculated CIs in part (vi)?

b. Conduct a hypothesis test for two numerical variables. Choose a random sample of 25 data points from the Dataset and choose two numerical variables, not of a corresponding quantity. We want to use this data to compare the average quantity between the two variables.
   viii. Should we use a t-test or z-test? Explain why.
   ix. Design a hypothesis test to see if these data provide convincing evidence of a difference between mean values. Does the result agree with the 95% confidence interval?

c. Are used cars with an automatic gearbox higher in the price? Is there enough evidence to prove this? (Your answer must be in detail.)

(Note: Use graphs and charts to clarify your answer and remember that the data might need a cleaning to pre-process.)