STAT212: Final exam

Mustafa Sadiq (ms3035)

May 11, 2021

1 Answer to Problem 1

a : False

b : True

c : True

d : False

e : True

f : True

g : True

h : True

i : False

j : True

2 Answer to Problem 2

2.1 Part a

Let u_1, u_2, u_3, and u_4 denote the sample average for each brand Ear Light, Loud n' Clear, Sound Aid, and Crystal Clear respectively.

$$H_0 : u_1 = u_2 = u_3 = u_4 \text{ (sample average are equal)}$$

$$H_A : \text{ Not all the average are equal.}$$

We now find the test statistic:

$$k = 4$$

$$n = 12 + 17 + 16 + 15 = 22$$

The overall mean $\bar{x}$:

$$\bar{x} = \frac{\sum \bar{x}_j * n_j}{n} = \frac{330}{22} = 15$$

Earbud Type	Size, n_j	Mean, $\bar{x}_j$	$\bar{x}_j - \bar{x}$	$(\bar{x}_j - \bar{x})^2$	$n_j(\bar{x}_j - \bar{x})^2$
Ear Light	5	12	-3	9	45.00
Loud n' Clear	4	17	2	4	16.00
Sound Aid	7	16	1	1	7.00
Crystal Clear	6	15	0	0	0.00
					68.00

From the final column of the table, we see that

$$SSTR = 68.00$$

Therefore,

$$MSTR = \frac{SSTR}{k-1} = \frac{68.00}{4-1} = 22.667$$

To compute MSE, we first compute the sample variance of each sample and construct the following table:

Earbud Type	Size, n_j	Variance, $\bar{s}_j^2$	$n_j - 1$	$(n_j - 1)\bar{s}_j^2$
Ear Light	5	9.500	4	38.000
Loud n' Clear	4	10.000	3	30.000
Sound Aid	7	2.000	6	12.000
Crystal Clear	6	2.800	5	14.000
				94.000

From the final column, we see that

$$SSE = 94.000$$

Therefore,

$$MSE = \frac{SSE}{n-k} = \frac{94.000}{22-4} = 5.222$$

Finally, we determine F,

$$F = \frac{MSTR}{MSE} = \frac{22.667}{5.222} = 4.340$$

The rejection region is $F \geq 3.16$ and since the test statistic is greater than 3.16, we reject the null hypothesis. There are statistically significant differences in the sound quality of earbuds brands.

2.2 Part b

We have $\binom{4}{2}$ pairs, so individual pair-wise level of significance will be $\frac{0.05}{\binom{4}{2}} = \frac{1}{120} = 0.0083$

3 Answer to Problem 3

I : Dotplots depict the distribution of numerical data

II : The empirical quantiles of the data can be on either axis and the other axis can display the theoretical quantiles of any normal random variable.

III : None of the above

IV : The probability the confidence interval covers the true value of the parameter is $(1 - \alpha_0)$.

V : The number of observations in each group is the same.

4 Answer to Problem 4

The Central Limit Theorem (CLT) says that:

> For a relatively large sample size, the variable $\bar{x}$ is approximately normally distributed, regardless of the variable under consideration. The approximation become better with increasing sample size.

The importance of CLT is that it allows us to use statistical methods that assume normal distribtution, which otherwise we could have not used.

The assumptions that are needed for the result of CLT is that the sampling should be iid. Sampling should be random and independent of one another. Usually, a sample size of 30 or more ($n \geq 30$) is large enough.

5 Answer to Problem 5

5.1 Part a

$$\text{Byzantium} = 21 + 14 = 35 \text{ individuals}$$

$$\text{Constantinople} = 26 + 19 = 45 \text{ individuals}$$

5.2 Part b

$$\text{Percentage vaccinated} = \frac{\text{vaccinated}}{\text{total}} * 100$$

$$\text{Byzantium} = \frac{21}{35} * 100 = 60\%$$

$$\text{Constantinople} = \frac{26}{45} * 100 = 57.78\%$$

5.3 Part c

The parameters being tested are:

$$P_B = \text{proportion of Byzantium vaccinated}$$

$$P_C = \text{proportion of Constantinople vaccinated}$$

If the proportion of Byzantium vaccinated and proportion of Constantinople vaccinated is the same, then obviously $P_B = P_C$, in which case $P_B - P_C = \Delta_0 = 0$, the null value.

Consequently, writing Δ as $P_B - P_C$:

$H_0 : P_B - P_C = 0$ *or* $P_B = P_C$ (both towns have equal proportion of people vaccinated)

$H_A : P_B - P_C \neq 0$ *or* $P_B \neq P_C$ (there is difference in proportions in the two cities)

5.4 Part d

We first compute the value of the test statistic:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_p(1-\hat{p}_p)}\sqrt{(1/n_1) + (1/n_2)}}$$

where $\hat{p}_p = (x_1 + x_2)/(n_1 + n_2)$.

We know:

$$x_1 = 21$$
$$n_1 = 35$$
$$\hat{p}_1 = 0.600$$

$$x_2 = 26$$
$$n_2 = 45$$
$$\hat{p}_2 = 0.578$$

Plugging these in:

$$z_0 = 0.200$$

The critical value are:

$$\pm z_{\alpha/2}(\text{ Two tailed})$$
$$= \pm 1.28$$

The rejection region is:

$$z < -1.28 \text{ and } z > 1.28$$

The value of the test statistic does not fall in the rejection region, we do not reject the H_0. There is not a statistically significant difference in the percentage of individuals vaccinated in the two cities.

6 Answer to Problem 6

Many statistical methods are framed in terms of the population means, μ, for the following reason:

- It is practical to use one value instead of all values in the dataset and be able to make inference. This is because the mean uses every value of the data.
- The mean accounts for the distribution of values. For example a high value outliers will increase the mean and vice versa.
- An accurate sampling from the population means that on average, the sample mean equals the population mean.
- However, due to sampling error, if sample mean is not equal to the population mean, we can use statistical inference i.e. to find the accuracy of this mean using confidence intervals. This is where we can use these methods.

7 Answer to Problem 7

a : Compute a 90% Upper Bound for μ

b : Perform a Left-Tailed z-Test for μ

c : Perform a two-sided, paired hypothesis test where the null is that the common mean difference of death ages is 0

d : Compute a two-sided 95% confidence interval for the mean gas price

e : Perform a two sample, two-sided hypothesis test of equality of population standard deviations where the null is that the ratio of population standard deviations is 1.

8 Answer to Problem 8

Dataset I, since all the observations lie 'roughly' on a straight line and Q-Q plot in nearly linear. Thus for this instance, it is plausible to conclude that the 100 values recorded follow a normal distribution approximately.

Dataset II, on the other hand, does not have linear Q-Q plot.

9 Answer to Problem 9

9.1 Part 1

Yes, Dr Lumvoir can change the summary statistics to Celcius without having the raw data. Converting these summary statistic will not change the statistical distribtution and it will represent the data as if we converted every value to Celcius.

$$F^\circ = C^\circ \times 1.8 + 32$$
$$C^\circ = \frac{F^\circ - 32}{1.8}$$

Average = -0.78 °C

Median = -2.22 °C

Mode = -1.67 °C

Range = -9.44 °C

IQR = -12.22 °C

Standard Deviation = -10.28 °C

9.2 Part 2

The results will not change.

The hypotheses would change to:

$$H_0 : \mu >= 0^\circ C$$
$$H_A : \mu < 0^\circ C$$

Since the standardized test statistic formulates on the summary statistics (mean and standard deviation), it will represent the data as if we converted every value to Celcius. With this test statistic and hypotheses, we will get the same results, since the distribution will remain the same of the standardized test statistic.

10 Answer to Problem 10

Let's look at the four Assumptions (conditions) for Regression Inferences:

- Population regression line

  This assumption requires that for each year, the mean salary of all points for that year lies on the line $y = \beta_0 + \beta_1 x$. It does not appear to be case if we look at the Years of Education 10 and 24, but sustains for rest of the Years of Education. Since, inferential procedure are robust to moderate violations for this assumption, we satisfy this assumption.

- Equal standard deviations and Normal populations

  These assumptions require that the salary distribtion for the various Years of Education are all normally distribtuted with the same standard deviation σ. Similar to previous assumption, the salary distribution appears to be normally distributed and with the same standard deviation, with a few exceptions. Since, inferential procedure are robust to moderate violations for these assumption, we satisfy these assumption.

- Independent observations

  This assumption requires that the observations of the response variable are independent of one another. We do not know how the data was collected and it may be the case that the salaries were sampled over the years and thus they will have relationship to the salaries in previous year (depending on the economy). We violate this assumption.

  However, if the data was sampled on a given year and not over time, this assumption holds and we can perform inference on it.

In summary, we can perform inference if we know how the data was collected.