Faculty of Computers &Artificial Intelligence
2nd Term (June 2022) Final Exam
Medical Informatics Program
Course Code: MBS312 **Level:** 3rd level
Subject: Biostatistics

Benha University
Date: 11 / 06 /2022
Time: 3 Hours
Total Marks: 50 Marks
Examiner : Dr. Mohamed Abdelgawad

Answer all the following questions [4 questions in 4 pages]

Question No. 1 [15 Marks]

A. To study the relationship between the student intelligence X and achievement in the biostatistics exam Y at faculty of computers and artificial intelligence, the data were as follows:

X	63.0	29.0	20.8	19.1	13.4	8.5
Y	7	3.9	2.1	2.8	1.4	1.5

Find the Correlation Coefficient between the two variables and determine its type. Also, find the equation of the regression line for the data to predict Y when X=15, 40 ?

B. Write briefly the characteristics of a good estimator and prove that the sample mean $\bar{x}$ is an unbiased estimator of μ?

Question No. 2 [15 Marks]

A. A survey conducted by Sallie Mae and Gallup of 1404 respondents found that 323 students paid for their education by student loans. Find the 90% confidence interval of the true proportion of students who paid for their education by student loans, where $Z_{0.05}$ =1.65?

B. Consider a random sample $x_1, x_2, \ldots, x_n$ from a normal distribution $N(\mu, \sigma^2)$. Find the maximum likelihood estimators for μ and σ^2 ?

Question No. 3 [10 Marks]

A. Check the following data set for outliers: 5, 6, 12, 13, 15, 18, 22, 50 ?

B. Traveling between two campuses of a university in a city via shuttle bus takes, on average, 28 minutes with a standard deviation of 5 minutes. In a given week, a bus transported passengers 40 times. What is the probability that the average transport time was more than 30 minutes? Assume the mean time is measured to the nearest minute, where P(Z<3.16)=0.9992 ?

Question No. 4 [10 Marks]

A. Find the value corresponding to the 60th percentile of the data:

18, 15, 12, 6, 8, 2, 3, 5, 20, 10 .

B. Choose the correct answer for the following:

1. The mean of the data a, a, a, a will be…

A. 0 B. a C. 2 D. none of the above.

2. When "n" is an odd number then median is defined as…

A. Middle value B. Median of two middle values C. Sum of the values

D. Most repeated value.

3. The sample mean $\bar{x}$ is known as the point estimator of the population…

A. Median B. Variance C. Mean μ D. Mode.

4. Student t-test is used to test population mean when population variance is always unknown and the sample size is….

A. Less than 30 B. More than 30 C. Any size D. None of them.

5. Z-score is calculated for…..

A. Chi-quire distribution B. Standard normal distribution C. Normal distribution

D. T-distribution.

6. A type of graphical presentation data used to explain correlation between dependent and independent variable is…..

A. Frequency polygon B. Scatter plot C. Frequency curve D. Histogram.

7. 95% confidence interval refers to….

A. considering 1 out of 20 chances are taken to be wrong.

B. considering 1 out of 100 chances are taken as wrong.

C. considering 95 out of 100 chances are taken as wrong.

D. considering 5 out of 20 chances are taken as wrong.

8. A statistic which describes the interval of scores bounded by the 25th and 75th percentile ranks is…..

A. Inter-quartile range B.Confidence Interval C.Standard deviation D. Variance.

9. All possible out comes of an experiment is known as sample space. When a coin is tossed 3 times then total sample space is……

A. 0 B. 8 C. 10 D. 6.

10. The listed observations 1,2,3,4,100, suggest the distribution…..

A. is positively skewed B. is negatively skewed C. has zero skewness D.is left-skewed.

Model answer

Solution Question No. 1 [15 Marks]

A. The Correlation Coefficient between the two variables

Company	Cars x (in 10,000s)	Revenue y (in billions of dollars)	xy	x^2	y^2
A	63.0	7.0	441.00	3969.00	49.00
B	29.0	3.9	113.10	841.00	15.21
C	20.8	2.1	43.68	432.64	4.41
D	19.1	2.8	53.48	364.81	7.84
E	13.4	1.4	18.76	179.56	1.96
F	8.5	1.5	12.75	72.25	2.25
	$\Sigma x = 153.8$	$\Sigma y = 18.7$	$\Sigma xy = 682.77$	$\Sigma x^2 = 5859.26$	$\Sigma y^2 = 80.67$

Step 3 Substitute in the formula and solve for r.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

$$= \frac{(6)(682.77) - (153.8)(18.7)}{\sqrt{[(6)(5859.26) - (153.8)^2][(6)(80.67) - (18.7)^2]}} = 0.982$$

The linear correlation coefficient suggests a strong positive linear relationship between the number of cars a rental agency has and its annual revenue. That is, the more cars a rental agency has, the more annual revenue the company will have.

To find the equation of the regression line for the data to predict Y when X=15, 40, we get

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{(18.7)(5859.26) - (153.8)(682.77)}{(6)(5859.26) - (153.8)^2} = 0.396$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{6(682.77) - (153.8)(18.7)}{(6)(5859.26) - (153.8)^2} = 0.106$$

Hence, the equation of the regression line $y' = a + bx$ is

$$y' = 0.396 + 0.106x$$

To graph the line, select any two points for x and find the corresponding values for y. Use any x values between 10 and 60. For example, let $x = 15$. Substitute in the equation and find the corresponding y' value.

$$y' = 0.396 + 0.106x$$
$$= 0.396 + 0.106(15)$$
$$= 1.986$$

Let $x = 40$; then

$$y' = 0.396 + 0.106x$$
$$= 0.396 + 0.106(40)$$
$$= 4.636$$

Then plot the two points (15, 1.986) and (40, 4.636) and draw a line connecting the two

B. 1- Characteristics of a good estimator: (unbiased)

The point estimator $\hat{\Theta}$ is an **unbiased estimator** for the parameter θ if

$$E(\hat{\Theta}) = \theta$$

If the estimator is not unbiased, then the difference

$$E(\hat{\Theta}) - \theta$$

is called the **bias** of the estimator $\hat{\Theta}$.

2- Characteristics of a good estimator: (with the least variance)

Suppose that $\hat{\theta}_1$ and $\hat{\theta}_2$ are unbiased estimator of $\hat{\theta}$. Since $\hat{\theta}_1$ has smaller variance than $\hat{\theta}_2$. Then $\hat{\theta}_1$ is minimum variance than $\hat{\theta}_2$ is the best. To prove that the sample mean $\bar{x}$ is an unbiased estimator of μ. If we have $X_1, X_2, ..., X_n$, i.i.d. random variable with sample size n taken from population mean is μ, lead to $EX_1 = EX_2 = ... = EX_n = \mu$ and the sample mean define as:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$E\overline{X} = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} EX_i$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mu = \mu.$$

The sample mean $\bar{x}$ is an unbiased estimator of μ, so $E\overline{X} = \mu$.

Question No. 2 [15 Marks]

A.

Step 1 Determine $\hat{p}$ and $\hat{q}$.

$$\hat{p} = \frac{X}{n} = \frac{323}{1404} = 0.23$$

$$\hat{q} = 1 - \hat{p} = 1.00 - 0.23 = 0.77$$

Step 2 Determine the critical value.

$$\alpha = 1 - 0.90 = 0.10$$

$$\frac{\alpha}{2} = \frac{0.10}{2} = 0.05$$

$$z_{\alpha/2} = 1.65$$

Step 3 Substitute in the formula

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$0.23 - 1.65\sqrt{\frac{(0.23)(0.77)}{1404}} < p < 0.23 + 1.65\sqrt{\frac{(0.23)(0.77)}{1404}}$$

$$0.23 - 0.019 < p < 0.23 + 0.019$$

$$0.211 < p < 0.249$$

or

$$21.1\% < p < 24.9\%$$

Hence, you can be 90% confident that the percentage of students who pay for their college education by student loans is between 21.1 and 24.9%.

B.

Let X be normally distributed with mean μ and variance σ^2, where both μ and σ^2 are unknown. The likelihood function for a random sample of size n is

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/(2\sigma^2)}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

and

$$\ln L(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Now,

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0$$

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial (\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

The solutions to the above equations yield the maximum likelihood estimators

$$\hat{\mu} = \bar{X} \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Conclusion: Once again, the maximum likelihood estimators are equal to the moment estimators.

Solution Question No. 3 [10 Marks]

A.

The data value 50 is extremely suspect. These are the steps in checking for an outlier.

Step 1 Find Q_1 and Q_3. $Q_1 = \frac{(6+12)}{2} = 9$; $Q_3 = \frac{(18+22)}{2} = 20$.

Step 2 Find the interquartile range (IQR), which is $Q_3 - Q_1$.

$$\text{IQR} = Q_3 - Q_1 = 20 - 9 = 11$$

Step 3 Multiply this value by 1.5.

$$1.5(11) = 16.5$$

Step 4 Subtract the value obtained in step 3 from Q_1, and add the value obtained in step 3 to Q_3.

$$9 - 16.5 = -7.5 \quad \text{and} \quad 20 + 16.5 = 36.5$$

Step 5 Check the data set for any data values that fall outside the interval from -7.5 to 36.5. The value 50 is outside this interval; hence, it can be considered an outlier.

B. $\bar{x} = 28, n = 40, s = 5$

$$P(\bar{x} > 30) = P\left(\frac{\bar{x} - 30}{5/\sqrt{40}}\right) = P(Z > 3.16) = 1 - P(Z < 3.16) = 1 - 0.9992 = 0.0008$$

Solution Question No. 4 [10 Marks]

C.

- **Step 1**. Arrange data.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20.

- **Step 2**.

$$c = \frac{n.p}{100} = \frac{10.6}{100} = 6$$

Step 3. Count over to the value that corresponds to 6. Find the mean of the 6th and 7th value 2, 3, 5, 6, 8, 10, 12, 15, 18, 20.

$$\frac{10 + 12}{2} = 11$$

This value corresponds to the 60th percentile!

D. <u>Choose the correct answer for the following</u>:

1. A
2. A
3. C
4. A
5. B
6. B
7. A
8. A
9. B
10. A

نموذج اجابة امتحان الاحصاء العضوية المستوى الثالث معلوماتية طبية برامج خاصة

كلية الحاسبات والذكاءالاصطناعى

د/ محمد عبدالجواد احمد عبدالجواد

مدرس – كلية العلوم – قسم الرياضيات

تاريخ الإمتحان ٢٠٢٢/٦/١١ الزمن ٣ ساعات