# PES University, Bengaluru-85

(Established under Karnataka Act No. 16 of 2013)

**UE19CS203**

## NOVEMBER 2020: IN SEMESTER ASSESSMENT, B.TECH, III-SEMESTER

## TEST – 2

## UE19CS203 – STATISTICS FOR DATA SCIENCE
## SCHEME & SOLUTION

Time: 1.30 Hrs | Answer All Questions | Max Marks: 40

| Q. No. | | | Marks |
|---|---|---|---|
| 1. | a) | A statistics instructor believes that fewer than sixty percentages of her university students attended the opening ceremony of a university fest. She surveys 64 of her students and finds that 36 attended the fest. Write appropriate null and alternative hypothesis.<br><br>Solution:<br><br>Null Hypothesis : $p \geq 0.6$ (Each hypothesis 1 mark)<br>Alternate Hypothesis : $p < 0.6$ | 2 |
| | b) | Globally the long-term proportion of newborns who are female is 51.46%. A researcher believes that the proportion of girls at birth changes under severe economic conditions. To test this belief randomly selected birth records of 5,000 babies born during a period of economic recession were examined. It was found in the sample that 52.55% of the newborns were girls. Determine whether there is sufficient evidence to support the researcher's belief.<br><br>Solution:<br><br>$H_0 : p = 0.5146$<br>$H_1 : p \neq 0.5146$ **( 1 mark)**<br>Let $p_0 = 0.5146$<br>$q_0 = 1- p_0 = 1- 0.5146 = 0.4854$<br><br>P hat = 0.5255<br>Z statistic =( p hat – $p_0$ ) / sqrt( $p_0$ *$q_0$ /n)<br>= (0.5255 – 0.5146) / sqrt( 0.5146 * 0.4854 / 5000) **( 1 mark)**<br>**= 1.5352 or 1.54**<br><br>Since it is a two tailed test, p-value = 2*P(Z<-1.5352) = 2*0.0618 **= 0.1236 ( 1 mark)**<br>Since p-value=0.1236 $\leq$ 0.05 is false**, we Don't reject $H_0$. ( 1 mark)** | 4 |
| | c) | To compare customer satisfaction levels of two competing service providers of a particular service, 174 customers of Service Provider1 and 355 customers of Service Provider2 were randomly selected and were asked to rate their service providers on a five-point scale, with 1 being least satisfied and 5 most satisfied. Average rating obtained by service provider 1 was 3.51 with a standard deviation of 0.51 from their customers and the average rating obtained by service provider 2 was 3.24 with a standard deviation of 0.52. Test at the 5% level of significance whether the data provide sufficient evidence to conclude that Service provider1 has a higher mean satisfaction rating than does Service provider 2.<br><br><br>Solution: | 4 |

| | | | |
|---|---|---|---|
| | | Given x bar = 3.51 ; $s_X = 0.51$ and y bar = 3.24 ; $s_Y = 0.52$ and $n_x = 174$ ; $n_y = 355$<br>Let $\mu_X$ = Mean service rating of Service Provider 1<br>Let $\mu_Y$ = Mean service rating of Service Provider 2<br><br>$H_0 : \mu_X - \mu_Y \le 0$<br>$H_1 : \mu_X - \mu_Y > 0$ **( 1 mark)**<br>Let $\mu_X - \mu_Y = 0$<br><br>Z statistic =( (x bar – y bar ) – ($\mu_X - \mu_Y$) ) / sqrt( $s_x^2/n_x + s_y^2/n_y$)<br>= (3.51-3.24) – 0 / sqrt($0.51^2$/174 + $0.52^2$ / 355 ) = 5.684 **( 1 mark)**<br><br>Since it is a right tailed test, p –value = P( Z > 5.684) = 1 – P(Z ≤ 5.684)<br>= 1 – 1 =0 **( 1 mark)**<br>Since p-value = 0 ≤ 0.05, we reject $H_0$. **( 1 mark)** | |
| | | | |
| 2. | a) | A man who got transferred to a new place is trying to determine which of the two routes to work has the shorter average driving time. Times in minutes for six trips on route A and five trips on route B are follows:<br>A: 16.0 15.7 16.4 15.9 16.2 16.3<br>B: 17.2 16.9 16.1 19.8 16.7<br>Can you conclude that the mean time is less for route A?<br><br>μ X – represents mean time for route B.<br>μ Y – represents mean time for route A.<br>H0 : μ X ≤ μ Y<br>H 1 : μ X > μ Y or **1 mark**<br><br>H 0 : μ X − μ Y ≤ 0<br>H 1 : μ X − μ Y > 0 | 5 |

| Value | Rank | Sample |
|---|---|---|
| 15.7 | 1 | *Y* |
| 15.9 | 2 | *Y* |
| 16 | 3 | *Y* |
| 16.1 | 4 | *X* |
| 16.2 | 5 | *Y* |
| 16.3 | 6 | *Y* |
| 16.4 | 7 | *Y* |
| 16.7 | 8 | *X* |
| 16.9 | 9 | *X* |
| 17.2 | 10 | *X* |
| 19.8 | 11 | *X* |

**1 mark**

Here, W = 42 (Counting all X(route B) ranks) **1 mark**
Since m = 5 and n = 6, we find that the area to the left of W = 42 is 0.0152 **1 mark**
Since P < 0.05, we reject H0 and conclude that mean lifetime for route A is less.
**1 mark**

b) Given the following contingency table for hair colour and eye colour. Find the value of Chi-Square and is there any good association between the two. **5**

| Hair colour / Eye colour | Fair | Brown | Black |
|---|---|---|---|
| Grey | 20 | 10 | 20 |
| Brown | 25 | 15 | 20 |
| Black | 15 | 5 | 20 |

Solution:

| Hair colour / Eye colour | Fair | Brown | Black | Row Sum |
|---|---|---|---|---|
| Grey | 20(Observed)<br>20(Expected) | 10(Observed)<br>10(Expected) | 20(Observed)<br>20(Expected) | 50 |
| Brown | 25<br>24(Expected) | 15<br>12(Expected) | 20<br>24(Expected) | 60 |
| Black | 15<br>16(Expected) | 5<br>8(Expected) | 20<br>16(Expected) | 40 |
| Column Sum | 60 | 30 | 60 | **150** |

**(Expected values – 2 marks)**

Dof = (r-1)*(c-1) = (3-1)*(3-1)=2*2=4 **0.5 mark**

Calculated $\psi$ 2 value = $\sum\sum$(Observed – Expected)$^2$ / Expected

$=(20-20)^2/20 + (10-10)^2/10 + (20-20)^2/20+(25-24)^2/24+(15-12)^2/12 + (20-24)^2/24 + (15-16)^2/16 + (5-8)^2/8 + (20-16)^2/16$

$= 1/24 + 9/12 + 16/24 + 1/16 + 9/8 + 1 = 3.6458$ **1.5 marks**

Calculated $\psi$ 2 value = 3.6458

Tabulated Value = 9.488 **0.5 mark**

( at 5% level of significance with 4 degrees of freedom)

Since, Calculated value < Tabulated value,

**Don't Reject Ho** (Null hypothesis) **0.5 mark**

3. a) Given the null hypothesis: that a process is producing no more than the maximum allowable rate of defective items. Write the Type II Error(as a statement). **2**

Solution:

Ho = process producing no more than k defectives
Ho false means process producing more than k defectives.

Type II error = P(do not reject Ho when Ho is false)

**Type II error: The process is not producing too many defectives when it actually is.**

**b)** A copper smelting process is supposed to reduce the arsenic content of the copper to less than 1000 ppm. Let $\mu$ denote the mean arsenic content for copper treated by this process, and assume that the standard deviation of arsenic content is σ = 100 ppm. The sample mean arsenic content $X$ of 75 copper specimens will be computed, and the null hypothesis $H0 : \geq 1000$ will be tested against the alternate $H1 : \mu < 1000$. **4**

i) A decision is made to reject $H0$ if $X \leq 980$. Find the level of this test.
ii) Find the power of the test in part (i) if the true mean content is 965 ppm.

Solution:

Given n= sample size = 75
True mean = 965
$\sigma = 100$ ; $\alpha = 5\% = 0.05$
$H0: : \mu \geq 1000$ $H1: \mu < 1000$

i) $z = (\bar{x} - \mu)/ \sigma/\sqrt{n} = (980 - 1000) / (100/\sqrt{75}) \approx -1.73$
The level of the test is the probability of rejecting the null hypothesis.
Level = P(Z < - 1.73) = 0.0418 **2 marks**

ii) The power is the probability of rejecting the null hypothesis when the alternative hypothesis is true
$z = (\bar{x} - \mu)/( \sigma/\sqrt{n} ) = (980 - 965) /(100/\sqrt{75}) \approx 1.30$
Power= P(Z < 1.30)=0.9032 **2 marks**

**c)** A certain type of seed has always grown to a mean height of 9.5 inches, with a standard deviation of 1 inch. Based on past experiment, the mean height of a seed is known to be distributed approximately normal. A researcher wishes to find out whether some new enriched conditions would improve the mean height. He wants to use α = .01 test and would like to have a 96% chance of rejecting the null hypothesis if the mean height is 10.5 inches. Determine the sample size for the test. **4**

Solution:
Given α = .01 ; 1 – β = 0.96 i.e. β = 0.04 **1 mark**

$\mu_{H0}$ = 9.5 and
$\mu_{HA}$ = 10.5 **1 mark**

We know that $n = (\sigma^2 * (Z_\alpha + Z_\beta)^2) / (\mu_{HA} - \mu_{H0})^2$ **1 mark**
n = 12 * (2.33 + 1.75)2 / (10.5 – 9.5 ) 2 = 16.65 ≈ 17
n = 17 **1 mark**

**4. a)** Find the least-squares regression line for the given data: **5**

| | | | | | |
|---|---|---|---|---|---|
| x | 2 | 2 | 6 | 8 | 10 |
| y | 0 | 1 | 2 | 3 | 3 |

Solution:

| x | y | x – x bar | y – y bar | (x – x bar)$^2$ | (x – x bar)*( y – y bar) |
|---|---|---|---|---|---|
| 2 | 0 | -3.6 | -1.8 | 12.96 | 6.48 |
| 2 | 1 | -3.6 | -0.8 | 12.96 | 2.88 |
| 6 | 2 | 0.4 | 0.2 | 0.16 | 0.08 |
| 8 | 3 | 2.4 | 1.2 | 5.76 | 2.88 |

| | | | | | |
|---|---|---|---|---|---|
| 10 | 3 | 4.4 | 1.2 | 19.36 | 5.28 |

**(Table 1 mark)**

x bar = 5.6 and y bar = 1.8 **( 0.5 mark)**

$\sum(x\text{-xbar})^2 = 51.2$ ; **( 0.5 mark)**

$\sum(x\text{-xbar})*(y\text{-ybar}) = 17.6$ **( 0.5 mark)**

$\beta_1$ hat = $\sum(x\text{-xbar})*(y\text{-ybar}) / \sum(x\text{-xbar})^2$ = 17.6 / 51.2 = 0.34375 **( 1 mark)**

$\beta_0$ hat = y bar - $\beta_1$ hat * x bar = 1.8 – 0.34375 * 5.6 = -0.125 **( 1 mark)**

Least Squares Regression Line is y hat = 0.34375 x – 0.125 **( 0.5 mark)**

b) Find the uncertainties in the estimates of $\beta_0$ hat and $\beta_1$ hat of part a). — 5

$\sum(y\text{-ybar})^2 = (-1.8)^2 + (-0.8)^2 + (0.2)^2 + (1.2)^2 + (1.2)^2 = 6.8$ **(1 mark)**

Correlation Coefficient r = $\sum(x\text{-xbar})*(y\text{-ybar})$ / (sqrt($\sum(x\text{-xbar})^2$ )* sqrt($\sum(y\text{-ybar})^2$ ))

r = 17.6 / sqrt(51.2) * sqrt(6.8 ) = 0.9432 **(1 mark)**

Error Uncertainty s = sqrt( ( (1-$r^2$)* ($\sum(y\text{-ybar})^2$)/(n-2) ) = 0.5002 **(1 mark)**

n=5

$s_{\beta 0\ hat}$ = s* sqrt ((1/n) +( xbar $^2$ / $\sum(x\text{-xbar})^2$ ) ) = 0.4509 and **(1 mark)**

$s_{\beta 1\ hat}$ = s / sqrt($\sum(x\text{-xbar})^2$) = 0.5002 / sqrt(51.2) = 0.0699 **(1 mark)**