# INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

## Mid-Autumn Semester Examination 2023-24

Date of Examination: 26.09.2023 Session: FN Duration: 02 hours Full Marks: 50

Subject No. : **CS 61061** Subject: **Data Analytics** Computer Science & Engineering Department

1) *Non-programmable calculator may be allowed.*
2) *Statistical tables may be allowed.*

**Special instructions:**

- *Answer to all questions.*
- *All symbols in the question, if not mentioned explicitly bear their usual meanings.*
- *You may make reasonable assumptions, if any.*

---

1. Suppose, X is a discrete uniform random variable on the consecutive integers $[a \dots b](a \leq b)$. Prove that

$$(a)\ \mu = \frac{(b+a)}{2} \qquad (b)\ \sigma^2 = \frac{(b-a+1)^2-1}{12}$$

**Solution:**

(a) There are n = (b-a+1) number of consecutive integers between b and a, each with probability $\frac{1}{b-a+1} = f(x_i) = \frac{1}{n}$

$$\mu = \sum x_i f(x_i) = \sum x_i \frac{1}{b-a+1} = \frac{1}{b-a+1} \cdot \frac{n(a+b)}{2} = \left(\frac{1}{b-a+1}\right)\left(\frac{b-a+1}{2}\right)(a+b) = \frac{\boldsymbol{(a+b)}}{\boldsymbol{2}}$$

[ *Proved*]

(b) Suppose x is an integer on the consecutive integers within [a…b] (a≤b).

If, the number of integers from a to x is = y,

$$y = x - a + 1$$

Or, $x = y + a - 1$

Now,

$E(X^2) = \frac{1}{n}\sum_{x=a}^{b} x^2$ [ here total number of elements n = (b-a+1)]

$= \frac{1}{n}\sum_{x=a}^{b} (y + a - 1)^2$

$= \frac{1}{n}\sum_{x=a}^{b} [y^2 + 2y(a-1) + (a-1)^2]$

$= \frac{1}{n}\sum_{y=1}^{b} [y^2 + 2y(a-1) + (a-1)^2]$

$= \frac{1}{n}\left[\sum \quad y^2 + 2(a-1)\sum \quad y + n(a-1)^2\right]$

The first term is $\frac{1}{n}\sum_{y=1}^{n} y^2 = 1^2 + 2^2 + 3^2 + \cdots + n^2$

$= \frac{n(n+1)(2n+1)}{6}$ // Sum of squares first n natural numbers

So,

$$E(X^2) = \frac{1}{n}\left[\frac{n(n+1)(2n+1)}{6} + 2(a-1)\frac{n(n+1)}{2} + n(a-1)^2\right]$$

$= \frac{(n+1)(2n+1)}{6} + (a-1)(n+1) + (a-1)^2$

$= (n+1)\left[\frac{2n+1}{6} + (a-1)\right] + (a-1)^2$

$=(n+1)\left[\frac{2n+1+6a-6}{6}\right] + (a-1)^2$

$= \frac{(n+1)(2n+6a-5)}{6} + (a-1)^2$

Now,

$\sigma^2 = E(X^2) - \left(\frac{a+b}{2}\right)^2$

$= \frac{(n+1)(2n+6a-5)}{6} + (a-1)^2 - \left(\frac{a+b}{2}\right)^2$

$= \frac{(n+1)(2n+6a-5)}{6} + \left[\left(a-1+\frac{a+b}{2}\right)\left(a-1-\frac{a+b}{2}\right)\right]$

$= \frac{(n+1)(2n+6a-5)}{6} + \left[\left(\frac{2a-2+a+b}{2}\right)\left(\frac{2a-2-a-b}{2}\right)\right]$

$= \frac{(n+1)(2n+6a-5)}{6} + \left[\left(\frac{3a-2+b}{2}\right)\left(\frac{a-2-b}{2}\right)\right]$

$= \frac{(n+1)(2n+6a-5)}{6} + \left[\left(\frac{-3a+2-b}{2}\right)\left(\frac{-a+2+b}{2}\right)\right]$ [ multiply both $\left(\frac{3a-2+b}{2}\right)\left(\frac{a-2-b}{2}\right)$ $with\ (-1)$]

$= (n+1)\left[\frac{2n+6a-5}{6}\right] + \left(\frac{-3a+2-b}{2}\right)\left(\frac{n+1}{2}\right)$

$= (n+1)\left[\left(\frac{2n+6a-5}{6}\right) + \left(\frac{-3a+2-b}{4}\right)\right]$

$= (n+1)\frac{4n+12a-10-9a+6-3b}{12}$

$= (n+1)\frac{4(b-a+1)+12a-10-9a+6-3b}{12}$ $[as\ n = b - a + 1]$

$= (n+1)\frac{4b-4a+4+12a-10-9a+6-3b}{12}$

$= \frac{(n+1)(b-a)}{12}$

$= \frac{(n+1)(n-1)}{12}$ $[as\ n\ =\ b\ -a\ +1]$

$= \frac{n^2-1}{12}$

$= \frac{(b-a+1)^2-1}{12}$ $[Proved]$

**[2+3]**

2. A quiz test of full marks 100 was conducted for a course where 2000 students have enrolled. From this test, it was found that $\mu = 90\ and\ \sigma = 20$.
   A random sample of size 100 was selected from this population and it was found that the mean score is 86.
   (a) What is the standard error of the sample?
   What does this value signify?
   (b) What is the probability of getting a sample whose mean score will be 86 or lower?

**[(1 +1) + 2]**

**Solution:**

**(a)** As per the **Central Limit Theorem,** the standard error is $\varepsilon = \frac{\sigma}{\sqrt{n}} = \frac{20}{10} = 2.0$

The **standard error** can be thought of as the dispersion of the sample mean estimations around the true population mean. In other words, it implies the variance of the means of different samples taken from a population. If the variance of the population is low or the sizes of samples are larger, the standard error will be smaller, indicating that the estimated sample mean value better approximates the population mean.

**(b)** The sample distribution statistics can be obtained with the **z-distribution.** For the sample,

$$z = \frac{86 - 90}{2} = -2.0$$

The probability of getting 86 or lower is $P(Z < -2.0)$. From the standard normal distribution table, it is found that $P(Z<-2.0)$ **= 0.0228.** So, we can say that 0.0228 is the probability that a sample will be whose mean score will be 86 or lower.

3. Mark the following statement as true or false. Give a brief justification to each of your answer.

   (a) If x is a value of a discrete random variable then $f(x) \leq 1$, and f(x) is called the probability mass function.
   **FALSE**.
   The probability mass function should be $0 \leq f(x) \leq 1$.

   (b) If f(x) denotes a probability density function, then $f(x) \geq 1$.
   **FALSE**.
   If f(x) denotes a probability density function, then $f(x) \geq 0$.

(c) If $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ then it represents a distribution of samples' variances with mean μ and $\sigma$ . Here, μ and $\sigma$ denote the statistics of a population from where samples are drawn.

**FALSE**.

The $\chi^2$ value represents a distribution of samples' variances with mean $v$ and variance $2v$. Here, $v$ represents the degree of freedom of the sample data.

(d) If $z_1 = \frac{X-\mu}{\sigma}$ and $z_2 = \frac{X-\mu}{\frac{\sigma}{\sqrt{n}}}$ then $z_1$ and $z_2$ are the values of z-scores according to the probability distribution of a random variable X and sampling distribution of samples drawn from a population of size *n* of mean μ and standard deviation $\sigma$, respectively.

**FALSE**.

$z_1$ and $z_2$ are the values of z-scores according to the probability distribution of a random variable X and sampling distribution of samples of size *n* drawn from a population of mean μ and standard deviation $\sigma$, respectively.

**[2 + 2 + 2 + 2]**

4. If $H_0$ and $H_1$ denote null and alternate hypothesis. Then mark the statement as true or false. Give a brief justification to each of your answer.

(a) $H_0 : \mu \leq 250$
$H_1 : \mu > 250$
Then $H_0$ and $H_1$ do not satisfy the hypothesis specification's mutually exclusive and exhaustive property.

(b) $H_0 : \mu = 250$
$H_1 : \mu < 250$

is equivalent to

$H_0 : \mu \geq 250$
$H_1 : \mu < 250$

(c) If $H_0 : \mu = 250$
$H_1 : \mu \neq 250$ then
Which of the following are not equivalent?

i. $H_0 : \mu = 250$ $H_1 : \mu < 250$

ii. $H_0 : \mu = 250$ $H_1 : \mu > 250$

iii. $H_0 : \mu \geq 250$ $H_1 : \mu < 250$

iv. $H_0 : \mu \leq 250$ $H_1 : \mu > 250$

**[2 + 1 + 4]**

## Solution:

**a) FALSE**

Here, $H_0$ and $H_1$ both cannot be true at the same time, so they are mutually exclusive. Also, $H_0$ and $H_1$ include all possible outcomes of the given random variable, ranging from $+\infty$ to $-\infty$ so, they are exhaustive.

**b) TRUE**

For the given 1-tailed test, both the given sets of hypotheses are equivalent.

**c) None of the options i), ii), iii), and iv) are equivalent to the given hypothesis.**

**Justification:**

Given,

$H_0$: $\mu = 250$

$H_1$: $\mu \neq 250$

This hypothesis set indicates a two-tailed test. This is equivalent to

$$H_0 : \mu = 250$$

$$H_1 : \mu > 250 \; or \; \mu < 250$$

Now all the given options signify one-tail tests. So these are not equivalent to the given hypothesis.

5. Suppose Y is a normally distributed random variable with mean = 10 and $sigma = 2.0$ and X is another independent random variable, also normally distributed with mean = 5 and $sigma = 5$.
Find P(Y>12 and X >4)
   (b) P(Y>12 or X>4)

## Solution:

To find $P(Y > 12 \; and \; X > 4)$ and $P(Y > 12 \; or \; X > 4)$, the individual probabilities of the events Y > 12 and X > 4 need to be calculated.

Converting Y and X, to their Z-values, we get

$$Z_Y = \frac{Y - \mu_Y}{\sigma_Y} = \frac{12 - 10}{2} = 1$$

$$Z_X = \frac{X - \mu_X}{\sigma_X} = \frac{4 - 5}{5} = -0.2$$

So the corresponding probabilities for each event are,

$$P(Y > 12) = P(Z_Y > 1) = 1 - 0.8413 = 0.1587$$

$$P(X > 4) = P(Z_X > -0.2) = 1 - 0.4207 = 0.5793$$

The probabilities corresponding to the Z-values are found using a standard normal table (Z-table).

**a) To find** $P(Y > 12 \; and \; X > 4)$:

Since Y and X are independent random variables, we can calculate this probability by multiplying the probabilities of each event separately.

So, $P(Y > 12 \; and \; X > 4) = P(Y > 12) \times P(X > 4) = 0.1587 \times 0.5793 = 0.0919$

**b) To find** $P(Y > 12 \; or \; X > 4)$**:**

Here, the probability of the union of two events needs to be calculated. So, we compute the probabilities of each event separately and then subtract the probability of their intersection since Y and X are independent. The formula for the union of two events Y and X is:

$$P(Y \cup X) = P(Y) + P(X) - P(Y \cap X)$$

We have computed $P(Y > 12) \; and \; (X > 4)$ previously. We have also computed the intersection $P(Y \cap X)$, that is, $P(Y > 12 \; and \; X > 4)$.

Using those values, we get,

$$P(Y > 12 \; or \; X > 4) = 0.1587 + 0.5793 - 0.0919 = 0.6461$$

**[4 + 2 + 2]**

6. Following Table Q.6 shows IQ measurements of 24 students based on an experiment.

**Table Q.6**

| | | | | | |
|---|---|---|---|---|---|
| 0.58 | 0.63 | 0.69 | 0.72 | 0.74 | 0.79 |
| 0.88 | 0.88 | 0.90 | 0.91 | 0.93 | 0.94 |
| 0.97 | 0.97 | 0.99 | 0.99 | 0.99 | 1.00 |
| 1.03 | 1.04 | 1.05 | 1.07 | 1.18 | 1.27 |

(a) Calculate the sample statistics.
(b) Compute the 95% confidence interval on the population mean μ .
(c) Test the hypothesis with 5% level of significance that the variance of IQs differed from 0.02. You should clearly give the necessary 5 steps in the hypothesis testing.
(d) What results you can infer if the level of significance is changed to 10% and 1%?
(e) How you can report the result in term of *p*-value?

**[4 + 4 + 6 + 2 +2]**

**Solution:**

**(a) Sample statistics**

i. Sample mean:

$$\underline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Here $n = 24$

$$\underline{x} = \frac{0.58 + 0.63 + \cdots + 1.27}{24} = \frac{22.14}{24} = 0.9225$$

ii. Sample variance:

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \underline{x})^2}{n-1}$$

$$= \frac{(0.58-0.9225)^2 + (0.63-0.9225)^2 + \cdots + (1.27-0.9225)^2}{24-1} = 0.0273$$

iii. Sample standard deviation:

$$S = \sqrt{0.0273} = 0.1652$$

**(b) Compute the 95% confidence interval on the population mean μ**

The confidence interval is 95% or 0.95
So, level of significance α=1-0.95 = 0.05
It is basically the sample statistics. So here t-test is applicable.

As, it is two tailed tests, so the population range in 95% confidence interval will be

$$\mu = \underline{x} \pm t_{\frac{\alpha}{2}} * \frac{S}{\sqrt{n}}$$

$$= (\underline{x} - t_{\frac{\alpha}{2}} * \frac{S}{\sqrt{n}}, \underline{x} + t_{\frac{\alpha}{2}} * \frac{S}{\sqrt{n}})$$

So, $\mu = \underline{x} \pm t_{\frac{\alpha}{2}} * \frac{S}{\sqrt{n}}$

$$= 0.9225 \pm t_{\frac{0.05}{2}} * \frac{0.1652}{\sqrt{24}}$$

$$= 0.9225 \pm 2.069 * \frac{0.1652}{4.89}$$

[ t value at 0.025 level of significance with 24-1 =23 degree of freedom is 2.069]

$$\mu = 0.9225 \pm 2.069 * 0.0337$$

$$= 0.9225 \pm 0.0697$$

$$= (0.8528, 0.9922)$$

**(c) Test the hypothesis with 5% level of significance that the variance of IQs differed from 0.02.**

This test is two tailed chi-square test.
The below stated five steps draw the inference from sample to population.

**Step1:**

$H_0 : \sigma^2 = 0.02$ (The population variance is equal to 0.02)

$H_1 : \sigma^2 \neq 0.02$ (The population variance is not equal to 0.02)

**Step 2:**

Here, $n = 24$

Sample variance $S^2 = 0.0273$

Population variance $\sigma^2 = 0.02$

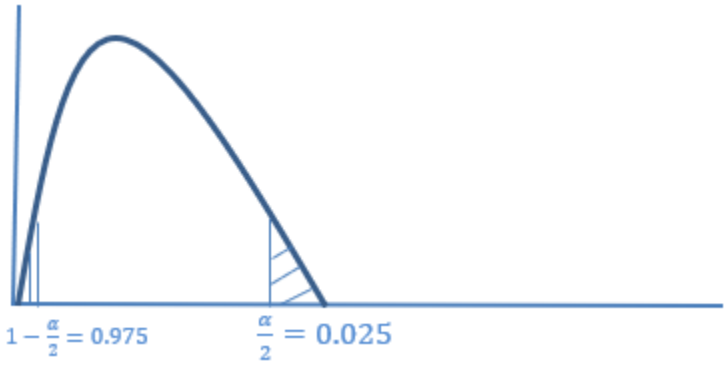
**Step 3:**

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

$$or, \chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

$$= \frac{(24-1)*0.0273}{0.02} = \frac{23*0.0273}{0.02} = 31.39$$

As, the test is two tailed, so both the tail will be condered.




The critical value for $\chi^2_{0.025,23} = 38.076$

The critical value for $\chi^2_{0.975,23} = 11.689$

[ Here for two tailed tests in right tail the level of significance will be $\frac{0.05}{2} = 0.025$ , and level of significance for left tail will be 0.975 and the degree of freedom is $(n-1) = (24-1) = 23$]

**Step 4:**

Here, $\chi^2_{0.925,23} < \chi^2 < \chi^2_{0.025,23}$

The derived value of $\chi^2$ is less than $\chi^2_{0.025,23}$ and greater than $\chi^2_{0.925,23}$. So, the derived value of $\chi^2$ lies in $H_0$ region.

So, the null hypothesis will **not be rejected**.

**Step 5:** So, variance of IQs is not differed from 0.02.

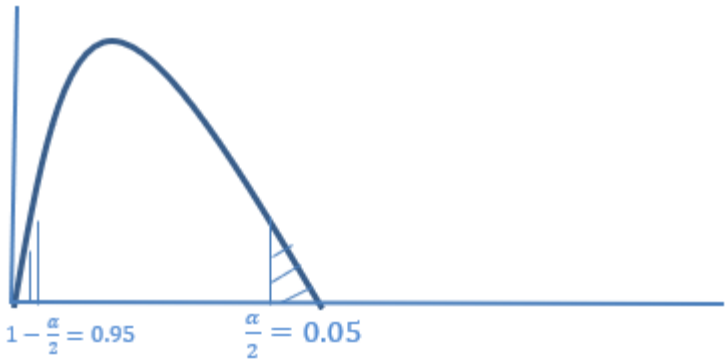
**(d) What results you can infer if the level of significance is changed to 10% and 1%**

**Case 1: level of significance is 10%**

The generated , $\chi^2 = 31.39$

Now α= 0.1

So, $\frac{\alpha}{2} = 0.05 \ \ and \ 1 - \frac{\alpha}{2} = 0.95$




Now, $\chi^2_{0.05,23} = 35.172$

$\chi^2_{0.95,23} = 13.091$

Here, $\chi^2_{0.95,23} < \chi^2 < \chi^2_{0.05,23}$

The derived value of $\chi^2$ is less than $\chi^2_{0.05,23}$ and greater than $\chi^2_{0.95,23}$. So, the derived value of $\chi^2$ lies in $H_0$ region.

The null hypothesis will **not be rejected**.

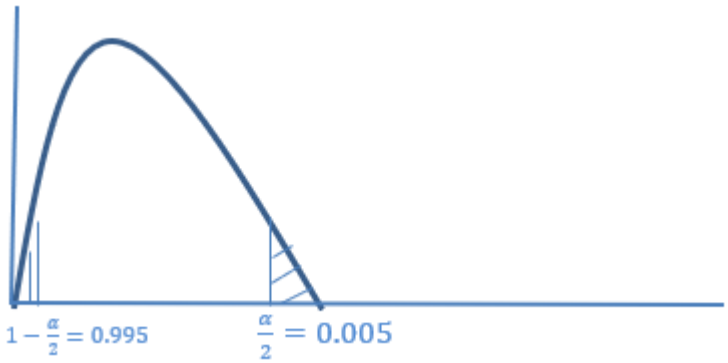
**So, variance of IQs is not differed from 0.02.**

**Case 2: level of significance is 1%**

The generated , $\chi^2 = 31.39$

Now α= 0.01

So, $\frac{\alpha}{2} = 0.005 \ \ and \ 1 - \frac{\alpha}{2} = 0.995$




Now, $\chi^2_{0.005,23} = 44.181$

$\chi^2_{0.995,23} = 9.260$

Here, $\chi^2_{0.995,23} < \chi^2 < \chi^2_{0.005,23}$

The derived value of $\chi^2$ is less than $\chi^2_{0.005,23}$ and greater than $\chi^2_{0.995,23}$. So, the derived value of $\chi^2$ lies in $H_0$ region.

The null hypothesis will **not be rejected**.

**So, variance of IQs is not differed from 0.02.**

**(e) How you can report the result in terms of p-value?**

The generated , $\chi^2 = 31.39$

when α=0.05 (5% level of significance)

p = $2 * P(\chi^2 \geq 31.39)$

= $2^* (1 - P(\chi^2 \leq 31.39))$

= 2*(1-0.9)

= 2*0.1

= 0.2

Here, $p > \alpha \ [as \ 0.2 > 0.05]$

Null hypothesis cannot be rejected.

So, **variance of IQs is not differed from 0.02.**