Probability - HCMUS - February 17th, 2025

Semester 1 : Year 2024 - 2025 - Time Duration: 90 minutes

Question 1. (2 points) The amount of time of an algorithm on a specific computer has an average standard distribution with 250ms and 15ms for standard deviation. Let $\overline{X}_n$ be the average processing time of an algorithm in random sample running n times.

(a) Find n such that $\mathbb{P}(248 \leq \overline{X}_n \leq 252) = 0.95$

(b) Now initialize 300 times. Use calibrated standard approximation to find the probability that atleast 70 times during trials embrace the processing time longer than 260ms.

Question 2. (2 points) A software development company wants to estimate the error ratio p (bugs) in each modules in the project. Based on a large sample test, the parameter estimator $\hat{p}$ is 0.25 and margin error of 95% confident interval for p is 0.06.

(a) Find the confident interval of 99%.

(b) Find the least minimum sample such that margin error of (a) confident interval does not surpass 0.06.

Question 3. (4 points) The target of a research is to compare the efficiency of two database management systems in a cloud environment, serving for business application. Therefore, system A and B are initialized to store and query data from database management applications. Efficieny comparison is important so as to select the optimal one for business, proper access speed and sustainability. The data collected from both systems was shown below:

1. System A: sample of experiment $m = 20$, average data access speed $\overline{x} = 150$ ms, sample standard deviation $s_1 = 12$ ms.
2. System B: sample of experiment $n = 20$, average data access speed $\overline{y} = 160$ ms, sample standard deviation $s_2 = 14$ ms.

Remarkably, the time for querying data of both systems has standard normal distribution.

1. There exists a hypothesis that the average querying time for system A is less than 155 ms. Inspect this hypothesis with significant level of 5%
2. Will the average accessing time of system B be longer than of A in 5% significant level? Assuming the variances are identical.

Question 4. (2 points) The table below displays the average response time (y, unit: s) of a server system with respect to different sizes of databases (x, unit: GB).

x (GB)	0.25	0.75	1.25	1.75	2.25	2.75	3.25	3.75	4.00	4.30	4.55
y (s)	1.087	1.228	1.583	1.798	1.939	2.138	2.172	2.315	2.455	2.735	2.954

1. Setup the single linear regression line of y with respect to x. Explain the meaning of $\hat{\beta_1}$ received.
2. Predict the time response of the system when the size of database is 1.55GB.

END

SUGGESTED ANSWER

Question 1. Normal approximation on probability

$$P\left(\overline{x}_1 < \overline{X}_n < \overline{x}_2\right) \approx P\left(\frac{x_1-\mu}{\frac{\sigma}{\sqrt{n}}} < z < \frac{x_2-\mu}{\frac{\sigma}{\sqrt{n}}}\right) \iff P\left(\frac{-2}{\frac{15}{\sqrt{n}}} < z < \frac{-2}{\frac{15}{\sqrt{n}}}\right) \approx 1.65$$

The trials needed to have the rate of 95% is atleast 154 times.

$$\implies \frac{-2\sqrt{n}}{15} \approx 1.65 \iff \boxed{n \approx 153}$$

Now find the new probability of time processing given the assumption of longer than 260ms.

$$z_{\text{exp}} = \frac{x' - \mu_0}{\sigma} = \frac{260-250}{15} \approx 0.67$$

$$\implies P(X > 260) \approx P(z > 0.67) \iff 1 - P(z \leq 0.67) \approx 1 - 0.7486 \approx 0.2514$$

Let Y be the likelihood of 70 times during trials embracing less processing time.

Initialize 300 times in accordance to binomial approximation. Of all the attempts, find the probability that 70 of them running less than 260ms, which is the general approach to conventional finding.

$$Y \sim \text{Bino}\left(np, p(1-p)\right) \sim \text{Bino}(300\cdot 0.2514, 0.2514\cdot 0.7486) \implies \begin{cases} \mu = 75.42 \\ \sigma^2 = 56.459 \end{cases}$$

$$P\left(Y \geq \frac{70-75.42}{7.51}\right) \approx \boxed{P\left(z \geq -0.72\right) \approx 0.7642}$$

Question 2. 99% confident interval of parameter estimator

$$\hat{p} \in (\hat{p} - \text{ME}; \hat{p} + \text{ME}) \approx (0.17; 0.33), \quad \text{ME} = z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}$$

Least minimum sample such that margin error of 99% confident interval does not surpass 0.06

$$\text{ME} = z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \iff 0.06 = 1.96 \cdot \sqrt{\frac{0.25\cdot 0.75}{n}} \implies \boxed{n \approx 200}$$

Question 3. t-test distribution for unknown population standard deviation for system A and two-sample t-test for both systems

(a) Let μ_A be the average of time processing in a cloud environment, establish the initial assumptions:

- $H_0 : \mu_A \geq 155(\text{ms})$
- $H_1 : \mu_A < 155(\text{ms})$

$$t - \text{obs} = \frac{\overline{x_A} - \mu}{\frac{s_1}{\sqrt{n}}} = \frac{150-155}{\frac{12}{\sqrt{20}}} \approx -1.863 > -t_{\alpha,n-1} \approx -2.093$$

Hence, H_0 is accepted, which means the average querying time is larger than or equal to 155ms.

(b)
- $H_0 : \mu_B \leq \mu_A$: The average accessing time of system B is less than the one in system A
- $H_1 : \mu_B > \mu_A$: The average accessing time of system B is longer than the one in system A.

Since the variance of both systems are identical, we derive the pooled variance:

$$S_P = \sqrt{\frac{(n_A-1)S_A^2 + (n_B-1)S_B^2}{n_A+n_B-2}} = \sqrt{\frac{19\cdot 12^2 + 19\cdot 14^2}{20+20-2}} \approx 13.038(\text{ms})$$

$$\Longrightarrow t'_{\text{obs}} = \frac{\overline{x_B} - \overline{x_A}}{\sqrt{\frac{2S_P^2}{n}}} \approx 2.425 > t_{\alpha,n-1} \approx 2.093$$

Reject H_0, thereby the average accessing time of system B is longer than the one in system A.

Question 4. Linear regression line: $\boxed{y = \beta_0 + \beta_1 x}$

(a)

$$\overline{x} = 2.623,\ \overline{y} = 2.037,\ \beta_1 = \frac{n\sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \approx \frac{11 \cdot 67.32 - 28.85 \cdot 22.40}{11 \cdot 97.69 - 832.32} \approx 0.39$$

$$\beta_0 = \overline{y} - \beta_1 \overline{x} = 2.037 - 0.39 \cdot 2.623 \approx 1.01 \Longrightarrow \boxed{y = 1.01 + 0.39x}$$

$\hat{\beta}_1$ received in the equation is the expected (average) change of response time in Y associated with the increase in x (GB) size of database. Used to predict the time response given arbitrary sizes.

(b) When $x = 1.55$ (GB),

$$\boxed{y' = 1.01 + 0.39 \cdot 1.55 \approx 1.61 \text{ (s)}}$$

END