UNIVERSITY OF TWENTE.

Course : Statistical Methods for Data Analysis
Module : 06
Course code: 202001033

Date : Thursday 15 January, 2026
Time : 13:45 – 16:00 (2 hours and 15 minutes)
Reference: Design of Interactive Technologies TCS/BIT (2025-1B)

A

# Statistical Methods for Data Analysis

## Instructions

This examination comprises 11 exercises. Only use the provided *answer form* to submit your answers.

- For PART 1, exercises 1-7, you are only required to fill in the **final answer** on the *answer form*.
- For PART 2, exercises 8-11, you are required to write down a **full calculation and argumentation**.

Only hand in your *answer form*. **Any text outside the answer form will not be considered**. If you run out of space, use the extra space at the end of the *answer form* and reference it in your original answer. Do not use pencil or red pen. Do not use correction fluid or tape. Give your answers to 3 decimal places unless stated otherwise.
**Electronic devices are prohibited. Only simple calculators (non-graphing) are allowed**.

**Formulas for Linear Regression**

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ \varepsilon_i \sim N(0,\sigma^2) \qquad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}},\quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \qquad S^2 = \frac{\sum(Y_i - \hat{y}_i)^2}{n-2}$$

$$S_{xx} = \sum(x_i - \bar{x})^2,\quad S_{xy} = \sum(x_i - \bar{x})(Y_i - \bar{Y}) \qquad F = \frac{SS_R}{SS_E/(n-2)} \sim F_{1,n-2}$$

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} = r^2$$

## Part 1: Final answer questions

*Report all your answers on the answer form.*

**1.** *All correct earn 4 marks; each incorrect answer deducts 1 mark; minimum score 0.* [4 pt]
Answer the following statements (a) to (i) with either true (T) or false (F).

(a) The length of the box in a box-plot is equal to the interquartile range of the data.
(b) A negatively skewed distribution has a longer right tail than left tail.
(c) If the skewness coefficient of a sample is negative, the distribution is skewed to the left.
(d) For a symmetric distribution, the sample mean and the sample median are approximately equal.
(e) The interquartile range is a measure of dispersion that is resistant to extreme values.
(f) A distribution with heavier tails than the normal distribution has a smaller kurtosis value.
(g) A 95% confidence interval for a population mean is $(48.2, 52.7)$. In the corresponding hypothesis test of $H_0 : \mu = 55$ versus $H_1 : \mu \neq 55$ at the 5% significance level, the null hypothesis is rejected.
(h) For a fixed alternative hypothesis, reducing the probability of committing a Type II error while keeping the probability of committing a Type I error unchanged will increase the power of the statistical test.
(i) For the random variable $X$, define $Y = aX + b$, where $a, b \in \mathbb{R}$. The transformation $Y = aX + b$ changes the variance of $X$ into $\mathrm{Var}(Y) = a^2\mathrm{Var}(X) + b$.

**2.** Engineers model the daily processing times (in milliseconds) of a cloud server as normally distributed with distribution $N(\mu, 4\mu^2)$. Thus, the standard deviation is twice the mean processing time. For a given server, 10 days of processing times are observed and may be regarded as a random sample $X_1, \ldots, X_{10}$ from this distribution. Let $\overline{X} = \frac{1}{10}\sum_{i=1}^{10} X_i$ denote the sample mean. Consider the class of estimators $T = k\overline{X}$ for $\mu$, where $k \in \mathbb{R}$.

I. For which value of $k$ is $T$ an unbiased estimator of $\mu$? *(Give the value of $k$.)* [1 pt]

II. For which value of $k$ does $T$ have the smallest mean squared error? *(Give the value of $k$.)* [2 pt]

**3.** An environmental agency monitors the average concentration (in g/m$^3$) of a certain air pollutant in a city where the legal limit is 35. It is known that daily concentrations are normally distributed with known standard deviation $\sigma = 6$. A random sample of $n = 25$ days is taken and the observed mean is $\overline{x} = 37$ and the sample standard deviation is $sd = 5.7$.

I. Using the above observations, the environmental agency wants to search for a range of plausible values for the true mean concentration. They therefore construct a 95% confidence interval for $\mu$. Which of the following expressions gives the correct confidence interval? [1 pt]

(a) $37 \pm 1.645\,\frac{6}{\sqrt{25}}$

(b) $35 \pm 0.519\,\frac{6}{\sqrt{25}}$

(c) $37 \pm 1.960\,\frac{6}{\sqrt{25}}$

(d) $37 \pm 2.064\,\frac{6}{\sqrt{25}}$

(e) $35 \pm 2.060\,\frac{6}{\sqrt{25}}$

(f) $37 \pm 1.711\,\frac{6}{\sqrt{25}}$

II. The environmental agency doubts whether the average concentration is in the critical range beyond the regulatory reference level. Which of the following correctly states the hypotheses for testing this claim? [1 pt]

(a) $H_0 : \mu \leq 35$ vs. $H_1 : \mu > 35$

(b) $H_0 : \mu \geq 35$ vs. $H_1 : \mu < 35$

(c) $H_0 : \mu = 35$ vs. $H_1 : \mu \neq 35$

(d) $H_0 : \sigma^2 \leq 6^2$ vs. $H_1 : \sigma^2 > 6^2$

(e) $H_0 : \sigma^2 \geq 6^2$ vs. $H_1 : \sigma^2 < 6^2$

(f) $H_0 : \sigma^2 = 6^2$ vs. $H_1 : \sigma^2 \neq 6^2$

III. Suppose the true mean concentration is $\mu = 36$. Which of the following **expressions** equals the power of the test? [2 pt]

(a) $P(\overline{X} \geq 35 \mid \mu = 36) = P\left(Z \geq \frac{35-36}{6/\sqrt{25}}\right)$

(b) $P(\overline{X} \leq 35 \mid \mu = 36) = 1 - P\left(Z \geq \frac{35-36}{6/\sqrt{25}}\right)$

(c) $P(\overline{X} \geq 35 \mid \mu = 36) = 2 \times P\left(Z \geq \left|\frac{35-36}{6/\sqrt{25}}\right|\right)$

(d) $P(\overline{X} \geq 37 \mid \mu = 36) = P\left(Z \geq \frac{37-36}{6/\sqrt{25}}\right)$

(e) $P(\overline{X} \leq 37 \mid \mu = 36) = P\left(Z \leq \frac{37-36}{6/\sqrt{25}}\right)$

(f) $P(\overline{X} \geq 37 \mid \mu = 36) = 2 \times P\left(Z \geq \left|\frac{37-36}{6/\sqrt{25}}\right|\right)$

(g) $P(\overline{X} \geq 35 \mid \mu = 36) = P\left(Z \geq \frac{36-35}{6/\sqrt{25}}\right)$

(h) $P(\overline{X} \leq 35 \mid \mu = 36) = 1 - P\left(Z > \frac{36-35}{6/\sqrt{25}}\right)$

(i) $P(\overline{X} \geq 35 \mid \mu = 36) = 1 - P\left(Z \leq \frac{35-36}{6/\sqrt{25}}\right)$

**4.** A game studio claims that the average time needed to defeat the final boss in the game "Dragon Doom II" is 60 minutes. A random sample of $n = 20$ players gives $\bar{x} = 55$ minutes, $sd = 10$ minutes. Assuming the data are normally distributed, and at significance level $\alpha = 0.05$, we test whether players defeat the boss faster than 60 minutes with $H_0 : \mu = 60$ vs $H_1 : \mu < 60$. Which of the following statements is correct about the decision for this test? [2 pt]

(a) $z = \frac{55-60}{10/\sqrt{20}} = -2.236$; $z_{\text{critical}} = -1.645$; $-2.236 < -1.645$; hence we reject $H_0$.
(b) $z = \frac{55-60}{10/\sqrt{20}} = -2.236$; $z_{\text{critical}} = -1.960$; $-2.236 < -1.960$; hence we fail to reject $H_0$.
(c) $z = \frac{55-60}{10/20} = -10$; $z_{\text{critical}} = -1.645$; $-10 < -1.645$; hence we reject $H_0$.
(d) $t = \frac{55-60}{10/\sqrt{20}} = -2.236$; $t_{\text{critical}} = -2.093$; $-2.236 < -2.093$; hence we reject $H_0$.
(e) $t = \frac{55-60}{10/\sqrt{20}} = -2.236$; $t_{\text{critical}} = -1.729$; $-2.236 < -1.729$; hence we reject $H_0$.
(f) $t = \frac{55-60}{10/20} = -10$; $t_{\text{critical}} = -2.093$; $-10 < -2.093$; hence we fail to reject $H_0$.

**5.** A marine biologist is comparing two underwater drones that are used to film coral reefs. At the depths where these drones operate, it is impossible to replace or recharge the battery during a dive. Therefore, longer battery life allows more continuous data collection and reduces the number of costly recovery operations. The observed data are:

- **Drone A:** $n_A = 42$ dives, mean battery life $\overline{X}_A$ minutes, standard deviation $S_A$.
- **Drone B:** $n_B = 35$ dives, mean battery life $\overline{X}_B$ minutes, standard deviation $S_B$.

The biologist wants to test whether the average battery life differs between the two drone models.

I. Which of the following represents the proper test statistic for this hypothesis test? [2 pt]

(a) $\frac{\overline{X}}{S/\sqrt{77}}$ with $S^2 = \frac{42}{77}S_A^2 + \frac{35}{77}S_B^2$
(b) $\frac{\overline{X}}{S/\sqrt{76}}$ with $S^2 = \frac{42}{76}S_A^2 + \frac{35}{76}S_B^2$
(c) $\frac{\overline{X}}{S/\sqrt{75}}$ with $S^2 = \frac{41}{75}S_A^2 + \frac{34}{75}S_B^2$
(d) $\frac{\overline{X}}{S/\sqrt{76}}$ with $S^2 = \frac{41}{76}S_A^2 + \frac{34}{76}S_B^2$
(e) $\frac{\overline{X}_A - \overline{X}_B}{\sqrt{S^2\left(\frac{1}{42}+\frac{1}{35}\right)}}$ with $S^2 = \frac{41}{75}S_A^2 + \frac{34}{75}S_B^2$
(f) $\frac{\overline{X}_A - \overline{X}_B}{\sqrt{S^2\left(\frac{1}{42}+\frac{1}{35}\right)}}$ with $S^2 = \frac{42}{77}S_A^2 + \frac{35}{77}S_B^2$
(g) $\frac{\overline{X}_A - \overline{X}_B}{\sqrt{S^2\left(\frac{1}{42}+\frac{1}{35}\right)}}$ with $S^2 = \frac{41}{76}S_A^2 + \frac{34}{76}S_B^2$

II. Which of the following represents the distribution for the proper test statistic? [1 pt]

(a) $t_{34}$ given $H_0$ (b) $t_{41}$ given $H_0$ (c) $t_{75}$ given $H_0$ (d) $t_{76}$ given $H_0$ (e) $t_{77}$ given $H_0$

**6.** A tech company is developing a new wearable device. The device's older model had a well-documented population standard deviation $\sigma = 2.0$ hours for battery lifetime. However, after updates to the hardware and firmware, engineers suspect this value may no longer apply. A tester measures the battery lifetime of $n = 18$ prototypes and obtains: $\overline{X} = 204.3$ hours, $sd = 2.1$ hours. Which of the following correctly expresses the 95% confidence interval for the true mean battery lifetime? [2 pt]

(a) $\left[204.3 - t_{\alpha}\frac{2.1}{\sqrt{18}}, 204.3 + t_{\alpha}\frac{2.1}{\sqrt{18}}\right]$

(b) $\left[204.3 - z_{\alpha/2}\frac{2.0}{\sqrt{18}}, 204.3 + z_{\alpha/2}\frac{2.0}{\sqrt{18}}\right]$

(c) $\left[204.3 - z_{\alpha/2}\frac{2.1}{\sqrt{18}}, 204.3 + z_{\alpha/2}\frac{2.1}{\sqrt{18}}\right]$

(d) $\left[204.3 - t_{\alpha/2}\frac{2.1}{\sqrt{18}}, 204.3 + t_{\alpha/2}\frac{2.1}{\sqrt{18}}\right]$

(e) $\left[204.3 - z_{\alpha}\frac{2.0}{\sqrt{18}}, 204.3 + z_{\alpha}\frac{2.0}{\sqrt{18}}\right]$

(f) $\left[204.3 - t_{\alpha}\frac{2.0}{\sqrt{18}}, 204.3 + t_{\alpha}\frac{2.0}{\sqrt{18}}\right]$

**7.** A cybersecurity team is analyzing phishing-click rates in two departments of a company: [2 pt]

- **Department A:** $x_A = 3$ of $n_A = 40$ employees clicked the phishing link.
- **Department B:** $x_B = 12$ of $n_B = 200$ employees clicked the phishing link.

You want to test whether the phishing-click rate in Department A differs from that in Department B: $H_0 : p_A = p_B$ vs. $H_1 : p_A \neq p_B$, at significance level $\alpha = 0.05$. Which of the following is the correct test statistic and decision?

(a) $\hat{p} = \frac{3+12}{40+200}$ and by CLT, $\hat{p}_A - \hat{p}_B \sim N\left(0,\ \hat{p}(1-\hat{p})\left(\frac{1}{40} + \frac{1}{200}\right)\right)$
reject $H_0$ if $|\hat{p}_A - \hat{p}_B| > 1.96\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{40} + \frac{1}{200}\right)}$.

(b) $\hat{p} = \frac{3+12}{40+200}$; and by by Chebyshev's rule, $\hat{p}_A - \hat{p}_B \sim N\left(0,\ \hat{p}(1-\hat{p})\left(\frac{1}{40} + \frac{1}{200}\right)\right)$
reject $H_0$ if $|\hat{p}_A - \hat{p}_B| > 1.645\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{40} + \frac{1}{200}\right)}$.

(c) $\hat{p} = \frac{3+12}{40+200}$; and by CLT, $\hat{p}_A - \hat{p}_B \sim N\left(0,\ \hat{p}(1-\hat{p})\left(\frac{1}{40} + \frac{1}{200}\right)\right)$
reject $H_0$ if $|\hat{p}_A - \hat{p}_B| > 1.645\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{40} + \frac{1}{200}\right)}$.

(d) $\hat{p} = \frac{3}{40}$; and by CLT, $\hat{p}_A - \hat{p}_B \sim N\left(0,\ \hat{p}(1-\hat{p})\left(\frac{1}{40} + \frac{1}{200}\right)\right)$
reject $H_0$ if $|\hat{p}_A - \hat{p}_B| > 1.96\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{40} + \frac{1}{200}\right)}$.

(e) $\hat{p} = \frac{3+12}{40+200}$; and by by Chebyshev's rule, $\hat{p}_A - \hat{p}_B \sim N\left(0,\ \hat{p}(1-\hat{p})\left(\frac{1}{40} + \frac{1}{200}\right)\right)$
reject $H_0$ if $|\hat{p}_A - \hat{p}_B| > 1.96\,\hat{p}(1-\hat{p})\left(\frac{1}{40} + \frac{1}{200}\right)$.

(f) $\hat{p} = \frac{3+12}{40+200}$; and by by Chebyshev's rule, $\hat{p}_A - \hat{p}_B \sim N\left(0,\ \hat{p}(1-\hat{p})\left(\frac{1}{39} + \frac{1}{199}\right)\right)$
reject $H_0$ if $|\hat{p}_A - \hat{p}_B| > 1.645\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{39} + \frac{1}{199}\right)}$.

## Part 2: Open questions

*The full solutions to exercises 8-11 must be clearly written down on the answer form, including calculations and arguments. Points will not be awarded for achieving a correct result if this is not supported by a correct procedure and by a sound and clear argumentation.*

**8.** An artificial intelligence research group is evaluating the *inference latency* (in milliseconds) of a newly optimized deep-learning model running on specialized hardware. The latency was recorded for 25 independent test runs under identical conditions.

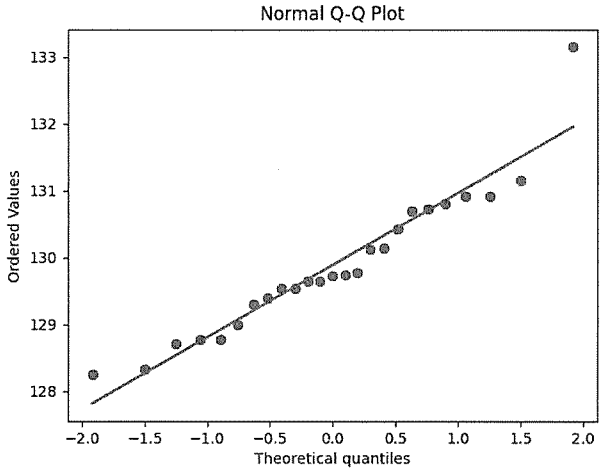
The ordered measurements are:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 128.24 | 128.33 | 128.71 | 128.77 | 128.77 | 129.00 | 129.29 | 129.39 | 129.53 | 129.54 |
| 129.64 | 129.65 | 129.73 | 129.74 | 129.78 | 130.12 | 130.14 | 130.42 | 130.69 | 130.72 |
| 130.80 | 130.91 | 130.92 | 131.15 | 133.15 | | | | | |

Using a statistical software, the following numerical summary was obtained:

| Sample size | Mean | Standard deviation | Variance | Skewness | Kurtosis |
|---|---|---|---|---|---|
| 25 | 129.885 | 1.076 | 1.157 | 0.994 | 5.151 |

The Q–Q plot of the data is given below:




I. Determine the five-number summary of the inference latencies and the $90^{th}$ percentile. [2 pt]

II. Using the Q–Q plot, comment on whether a normal distribution is a reasonable model. [2 pt]

III. Comment on the plausibility of normality on the basis of the numerical summaries. [2 pt]

IV. The Shapiro–Wilk test statistic is reported as $W = 0.932$. State the rejection region for this test at $\alpha = 5\%$ and give your conclusion regarding normality. [2 pt]

**9.** An entertainment company is comparing two different *escape-room scoring systems*. To evaluate how unpredictable each system feels to players, the company records a *score fluctuation index* after each completed room (higher values indicate greater spread). A total of 47 escape-room sessions are randomly assigned to one of the two scoring systems. The summary statistics are:

| Scoring system | $n$ | $\overline{x}$ | $s$ |
|---|---|---|---|
| System A | 21 | 37.3 | 28.3 |
| System B | 26 | 19.3 | 20.8 |

I. At $\alpha = 5\%$, assess whether the *spread* of the score fluctuation index differs between the two scoring systems. Give only the below mentioned steps: [3 pt]

1. the appropriate test statistic and its observed value,
2. the rejection region,
3. your conclusion.

II. The company now wants the 95% confidence interval for the *spread parameter* of System A with margin of error at most 150 (in index unit$^2$). Using the pilot estimate from the table, determine the smallest sample size $n$ required to achieve this margin of error. [3 pt]

**10.** A company claims that a new *blue-light screen filter* reduces eye strain. To test this, 200 volunteers are recruited. One group of 100 uses the filter-enabled screen (*treated group*) and another group of 100 uses an identical screen without the filter (*control group*). After one week, each volunteer reports whether they experienced *less eye strain*, *more eye strain*, or *no difference* compared to their usual screen use. (All participants are told they are using the new filter.)

The results are shown below:

| | **Less eye strain** | **More eye strain** | **No difference** |
|---|---|---|---|
| **Control group** | 39 | 21 | 40 |
| **Treated group** | 51 | 20 | 29 |

I. Suppose we want to test the influence of the screen filter on reported eye strain. Should we apply a test of independence to these data or a test of homogeneity? Briefly justify your choice. [1 pt]

II. Apply the test selected in part I. above, using $\alpha = 0.05$. Clearly show all steps and provide sufficient details and justification for your conclusion. [3 pt]