UNIVERSITY OF TWENTE.

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

Date : Monday 30 March, 2026
Time : 08:45 – 11:00 (2 hours and 15 minutes)
Reference: Design of Interactive Technologies TCS/BIT (2025-1B)

B

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 first central moment of a random variable X equals standard deviation, i.e., $\mathbb{E}(X - \mu) = \sigma$. The second central moment equals the variance, i.e., $\mathbb{E}[(X-\mu)^2] = \mathrm{Var}(X) = \sigma^2$. The third central moment $\mathbb{E}[(X-\mu)^3]$ provides information about the skewness of a distribution, and the fourth central moment is used to define kurtosis: $\gamma_2 = \frac{\mathbb{E}[(X-\mu)^4]}{\sigma^4}$.

(b) Consider the following two estimators for the variance σ^2 based on a random sample $X_1, \ldots, X_n$: $T_1 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$, $T_2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$. Since T_2 is an unbiased estimator of σ^2, while T_1 is biased, T_2 is always the better estimator than T_1.

(c) For a fixed confidence level, increasing the sample size n results in a greater confidence interval for the population mean μ.

(d) A box plot and a histogram contain the same information about the distribution of a dataset, but they display it in different graphical formats.

(e) Decreasing the significance level α of a hypothesis test makes it easier to reject the null hypothesis H_0.

(f) When comparing the means of two paired populations with unknown but equal variances, the appropriate test uses a pooled estimate of the common variance.

(g) When testing whether two categorical variables are independent in a contingency table, the appropriate procedure is a chi-square test of independency.

(h) In the simple linear regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are independent random variables.

(i) A team of CS and BIT students compare two caching strategies for a web server. They collect many response times for each strategy and assume the data are normally distributed. They want to determine whether the response times vary more for one strategy than for the other. To test this, they decide to perform a two-sample $t - test$.

2. Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a population with mean μ and variance σ^2. Consider the following two estimators for μ:$T_1 = \overline{X}$ and $T_2 = \frac{2}{3}X_1 + \frac{1}{3}X_2$

I. Which of the following statements about bias is correct? [1 pt]

(a) T_1 is an unbiased estimator of μ and T_2 is a biased estimator of μ.
(b) T_1 is a biased estimator of μ and T_2 is an unbiased estimator of μ.
(c) T_1 is an unbiased estimator of μ only if $n > 2$.
(d) Both T_1 and T_2 are biased estimators of μ.
(e) Both T_1 and T_2 are unbiased estimators of μ.
(f) T_2 is an unbiased estimator of μ only if $n = 2$.
(g) It cannot be determined whether T_1 and T_2 are biased or unbiased estimators of μ.

II. Which estimator is preferable for estimating μ? *(More than one answer may be correct.)* [1 pt]

(a) It cannot be determined which estimator is better.
(b) T_2, because it uses fewer observations.
(c) T_1, because it is always unbiased.
(d) T_1, because it uses all observations in the sample.
(e) T_1, because it has smaller variance.
(f) T_2, because its coefficients sum to 1.
(g) T_2, because it is a weighted estimator.
(h) Both estimators are equally good.

3. A research group studies the average execution time of a sorting algorithm used in a large-scale data-processing system. Previous experiments indicate that execution times are approximately normally distributed with known standard deviation $\sigma = 4$ milliseconds. To evaluate the algorithm, the researchers record the execution time for $n = 36$ independent runs. The observed sample mean execution time is $\bar{x} = 52$ milliseconds. The researchers want to determine a range of plausible values for the true mean execution time μ.

I. Which statistical procedure should the researchers compute? [1 pt]

(a) A test of independence
(b) A two-sample t-test
(c) A Shapiro–Wilk test for normality
(d) A regression model
(e) A chi-square test for the variance
(f) A confidence interval for the mean μ
(g) A goodness-of-fit test
(h) A hypothesis test for the mean μ

II. Which distribution should be used to obtain the critical value for constructing the interval? [1 pt]

(a) Standard normal Z
(b) Binomial distribution
(c) F-distribution
(d) Student's t
(e) Chi-square χ^2
(f) Uniform distribution
(g) Shapiro–Wilk table
(h) Poisson distribution

III. Which of the following expressions gives the correct 95% confidence interval for μ? [1 pt]

(a) $52 \pm Z_{\alpha/2} \frac{4}{\sqrt{35}}$
(b) $52 \pm Z_{\alpha} \frac{4}{\sqrt{36}}$
(c) $52 \pm t_{\alpha,36} \frac{4}{\sqrt{36}}$
(d) $52 \pm Z_{\alpha/2} \frac{4}{36}$
(e) $52 \pm \chi^2_{\alpha/2,35} \frac{4}{\sqrt{36}}$
(f) $52 \pm Z_{\alpha/2} \frac{4}{\sqrt{36}}$
(g) $52 \pm Z_{\alpha/2} \frac{4}{35}$
(h) $52 \pm t_{\alpha/2,35} \frac{4}{\sqrt{36}}$

4. A system processes independent tasks, each of which is either successful or unsuccessful. Let X denote the number of successful tasks out of $n = 200$ independent trials. Historically, the probability of success is $p = 0.80$. A researcher wants to test whether the success probability differs from the historical value. Which of the following statements about the appropriate model are correct? *(More than one answer may be correct.)* [1 pt]

(a) The normal approximation may be used because np and $n(1-p)$ are sufficiently large.
(b) The Central Limit Theorem justifies approximating the binomial distribution by a normal distribution.
(c) $X \sim N(160, 32)$ can be used with a continuity correction when approximating binomial probabilities.
(d) $X \sim N(160, 32)$ can be used by Chebyshev's rule because n is large enough.
(e) A chi-square distribution should be used because counts are observed, and chi-square distributions are commonly used for categorical data.
(f) A Student's t-distribution should be used because the sample mean of a normally distributed population with unknown variance follows a t-distribution.
(g) $X \sim \text{Bin}(200, 0.80)$, because X counts the number of successes in independent Bernoulli trials.

5. A mechanical engineer wants to compare the average repair time of electric motors serviced at Garage 1 and Garage 2. During one month, a number of motors are sent to Garage 1 and another group of motors is sent to Garage 2. Each motor is repaired in only one garage. Let $X_1, \ldots, X_n$ be the repair times (in hours) for motors serviced at Garage 1, and $Y_1, \ldots, Y_m$ be the repair times (in hours) for motors serviced at Garage 2.
Assume repair times follow normal distributions with equal variances, and the samples are independent. The engineer wants to test whether the average repair time differs between the two garages.

I. Which of the following is the correct null and alternative hypothesis? *(More than one answer may be correct.)* [1 pt]

(a) $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$
(b) $H_0 : \sigma_1^2 = 0$ vs. $H_1 : \sigma_1^2 \neq 0$
(c) $H_0 : \sigma_1^2 = \sigma_2^2$ vs. $H_1 : \sigma_1^2 \neq \sigma_2^2$
(d) $H_0 : \mu_1 \leq \mu_2$ vs. $H_1 : \mu_1 > \mu_2$
(e) $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_1 : \mu_1 - \mu_2 \neq 0$
(f) $H_0 : \mu_1 = 0$ vs. $H_1 : \mu_1 \neq 0$

II. Which of the following is the proper test statistic? [1 pt]

(a) $T = \frac{\overline{X}-\overline{Y}}{S_p}, \quad S_p^2 = \frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}$

(b) $T = \frac{\overline{X}-\overline{Y}}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}}, \quad S_p^2 = \frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}$

(c) $T = \frac{\overline{D}}{S_D/\sqrt{n}}, D_i = X_i - Y_i, S_D = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(D_i-\overline{D})^2}$

(d) $F = \frac{S_X^2}{S_Y^2}$

(e) $T = \frac{\overline{X}-\overline{Y}}{\sqrt{S_X^2+S_Y^2}}$

(f) $Z = \frac{\overline{X}-\overline{Y}}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}$

(g) $\chi^2 = \sum \frac{(O-E)^2}{E}$

III. What is the correct distribution of the test statistic under H_0? [1 pt]

(a) χ^2_{n+m-2}

(b) t_{n-1}

(c) t_{n+m-1}

(d) t_{m-1}

(e) $F_{n-1,m-1}$

(f) $N(\mu, \sigma^2)$

(g) t_{n+m-2}

(h) $N(0,1)$

6. Researchers in Human–Computer Interaction (HCI) study whether the type of interface used is associated with the type of usability problem reported. A sample of 150 users is observed. Each user interacts with either a touch interface or a voice interface. After the test, two categorical variables are recorded: (1) the interface used (touch or voice), and (2) the main usability problem reported (navigation, input, or feedback). The observed frequencies n_{ij} are summarized in the contingency table below.

Problem type (i)	**Touch interface**	**Voice interface**	**Row total**
Navigation	30	20	50
Input	25	35	60
Feedback	15	25	40
Column total	70	80	150

I. Which statistical test should be applied? [1 pt]

(a) Binomial test

(b) Chi-square test of independence

(c) Chi-square test of homogeneity

(d) Test for equality of variances

(e) Linear regression

(f) Paired samples t-test

(g) Independent samples t-test

(h) Shapiro–Wilk test

II. Which of the following are the correct hypotheses? [1 pt]

(a) H_0: All usability problem types occur with the same probability.
H_1: Not all problem types occur with the same probability.

(b) H_0: The probabilities of the Input and Feedback problem types are the same for the touch interface and the voice interface.
H_1: At least one of these probabilities differs between the two interfaces.

(c) H_0: The probability of each usability problem type is the same for the touch interface and the voice interface.
H_1: At least one of these probabilities differs between the two interfaces.

III. Under H_0, the expected frequencies are $E_{ij} = \frac{n_{i\bullet}\, n_{\bullet j}}{n}$. Fill in the expected frequencies E_{ij} for the table given in your answer sheet. *Please write your answers clearly in the provided space.* [2 pt]

IV. Which of the following statements are correct? *(More than one answer might be correct)* [1 pt]

(a) $df = 3 + 2 - 2$

(b) $\chi^2 \approx 5.20$

(c) $\chi^2 \approx 5.52$

(d) $df = (3-1)(2-1)$

(e) $\chi^2 \approx 4.80$

(f) $\chi^2 \approx 8.27$

7. A researcher investigates whether the number of hours students study for a statistics exam influences their exam score. A sample of 40 students is analyzed using simple linear regression. The estimated regression equation is: $\hat{y} = 48 + 5x$. The regression output also reports: $R^2 = 0.64$.

I. Which of the following correctly identifies the variables in the regression model? [1 pt]

(a) Independent variable: exam score
Dependent variable: number of students

(b) Independent variable: exam score
Dependent variable: hours studied

(c) Independent variable: number of students
Dependent variable: hours studied

(d) Independent variable: hours studied
Dependent variable: exam score

II. Which interpretation of the regression equation is correct? [1 pt]

(a) The predicted exam score decreases by 5 points for each hour studied.

(b) A student who studies 5 hours is predicted to score 48 points.

(c) Each additional hour of studying increases the predicted exam score by 48 points.

(d) Each additional hour of studying increases the predicted exam score by 5 points.

III. Which hypotheses are used to test whether studying time has a linear effect on exam score? [1 pt]

(a) $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$

(b) $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_0 \neq 0$

(c) $H_0 : R^2 = 0$ vs. $H_1 : R^2 > 0$

(d) $H_0 : \beta_1 = 1$ vs. $H_a : \beta_1 \neq 1$

IV. The coefficient of determination is $R^2 = 0.64$. What does this mean? *Please write your short interpretation clearly in the provided space in the answer sheet.* [2 pt]

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. From a random sample of 24 employees in a large organization, it was found that the average gross (full-time) annual income is 178,000 and the standard deviation is 54,000. The five-number summary is: 112,000, $Q_1 = 145{,}000$, $m = 160{,}000$, $Q_3 = 190{,}000$, 394,000.
(All data are from several years ago.)

I. Based on this information, what can you say about the shape of the histogram of annual incomes? [1 pt]

II. What percentage of the employees in this study has an annual salary higher than 190,000? [1 pt]

III. If we wanted to determine a 95% confidence interval for that percentage with a width of at most 8%, how many employees should be included in a new sample? [2 pt]

IV. When performing the Shapiro–Wilk test, a value of 0.908 was obtained for the test statistic. What does this value imply for the normality assumption? Give only: [2 pt]

(i) the hypotheses,

(ii) the critical region,

(iii) the test decision, and

(iv) the conclusion of the normality assumption in context, for the test with $\alpha = 10\%$.

9. A company produces two types of instant adhesive glue: Type A glue and Type B glue. The drying (setting) time of the glue is important for the efficiency of the assembly process. It is suspected that Type A glue sets faster than Type B glue.

For both types of glue, several measurements of the drying time were obtained under as similar conditions as possible. The measured drying times, in minutes, are as follows:

Type A glue	119	132	123	122	116	110	121	107			$\bar{x}_A = 118.8$	$s_A = 7.85$
Type B glue	127	124	140	125	109	130	126	117	129	120	$\bar{x}_B = 124.7$	$s_B = 8.27$

The question is whether Type A glue indeed sets faster than Type B glue.

I. Determine a 95% confidence interval for the variance of the drying time of Type A glue. [2 pt]

II. Explain why, without performing a formal test, it is reasonable to assume that the variances of the drying times of both types of glue are equal. [1 pt]

III. Perform a hypothesis test to determine whether Type A glue sets faster than Type B glue. Assume normal distributions and equal variances. Use a significance level of 5%. Follow the eight-step testing procedure (see formula sheet). [4 pt]

10. A company investigated whether employees use AI tools (such as chatbots) in their daily work. Employees were asked whether they regularly use AI tools or do not use AI tools. The responses were grouped by age.

Age group	Uses AI tools	Does not use AI	Total
Younger than 40	187	195	382
40 or older	116	76	192
Total	303	271	574

I. Using a 1% significance level, test the hypothesis that employees aged 40 and older are more likely to use AI tools than employees under 40. Give only: [3 pt]

(i) the model,

(ii) the hypotheses,

(iii) the observed value, and

(iv) the test decision, for the test with $\alpha = 1\%$.

II. Determine a 95% confidence interval for the difference between the proportions of employees under 40 and 40 or older who use AI tools. [2 pt]

11. An engineer wants to study the relationship between machine operating time x (in hours) and energy consumption y (in kWh) for a certain industrial machine. A random sample of 12 observations is collected and a simple linear regression model is assumed: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where the errors ε_i are independent with $E(\varepsilon_i) = 0$ and $\text{Var}(\varepsilon_i) = \sigma^2$. Using statistical software, the regression output is obtained as below.

width=0.75

COEFFICIENTS

Model	B	Std. Error	t	Sig.
(Constant)	4.25	1.10	3.86	0.003
OperatingTime	1.72	0.28	6.14	0.0001

MODEL SUMMARY

Model	R	R Square	N
1	0.889	0.791	12

ANOVA

Source	df	Sum of Squares	Mean Square	F	Sig.
Regression	1	210.4	210.40	37.84	0.0001
Residual	10	55.6	5.56		
Total	11	266.0			

I. Write down the estimated regression line (equation) and interpret the slope coefficient $\hat{\beta}_1$ in the context of this problem. [2 pt]

II. Test whether there is a linear relationship between operating time and energy consumption. Use significance level $\alpha = 0.05$ and provide only the steps given below. [2 pt]

(i) State the hypotheses.

(ii) Give the test statistic.

(iii) Determine the critical region or p-value.

Total: 45 pt