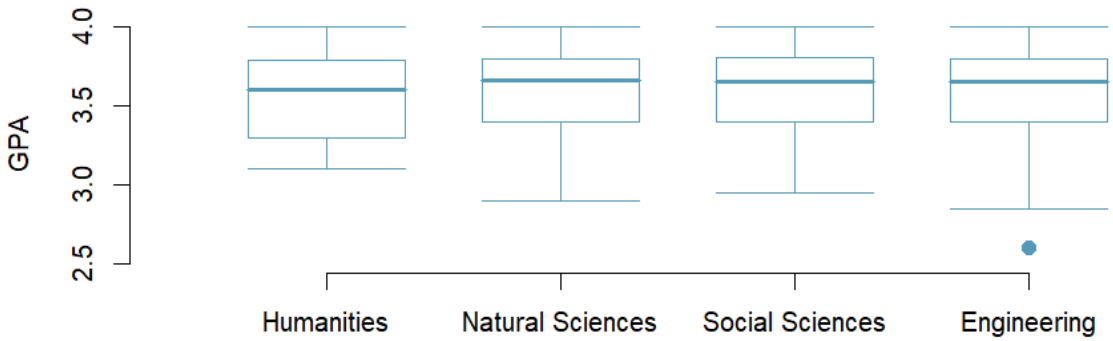
Introduction to Statistics - Quiz #4(70 minutes)

June 10, 2025 (Friday)

Section(교반): _____ Cadet Number(교번): _____ Name(성명): _____ Score: _____

- All solutions must include a detailed step-by-step explanation.
- If an answer has more than four decimal places, round to the **fourth decimal place**.

1. 268 undergraduate students taking an introductory statistics course conducted a survey on GPA and major. We conduct **a one-way ANOVA** to determine whether the average GPA differs across college majors at the significance level $\alpha = 0.05$. The side-by-side box plots show the distribution of GPA among four groups of majors. Summary statistics and ANOVA output are provided below. (Suppose that all the condnitions for conducting ANOVA are satisfied.) [30 points]



College Major

	Humanities	Natural Sciences	Social Sciences	Engineering	All
Mean	3.58	3.61	3.57	3.59	3.59
SD	0.30	0.27	0.26	0.27	0.28
n	113	52	33	70	268

(1) State the null and alternative hypothesis. (Define parameters of interest.)

Solution:

Let $\mu_1, \mu_2, \mu_3, \mu_4$ represent the true mean GPA for students majoring in one of four groups: Humanities, Natural Sciences, Social Sciences, and Engineering.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4, \qquad H_A : \text{not } H_0.$$

(2) Below is part of the output associated with this test. Fill in the empty cells.

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Major		0.0348			0.931
Residuals		20.6831			
Total	267	20.7179			

Solution: (From left to right) 3, 0.0116, 0.1481, 264, 0.0783

(3) What's the conclusion of the hypothesis test? State the conclusion in the context of data.

Solution: The p-value is 0.931, which is greater than the significance level $\alpha = 0.05$. Thus, we fail to reject the null hypothesis. There is no evidence that the average GPA differs across the four majors.

Reference Table

pt(0.8502, df = 9, lower.tail = FALSE) = 0.2086	pt(0.8502, df = 10, lower.tail = FALSE) = 0.2076
pt(4.8448, df = 9, lower.tail = FALSE) = 0.0004	pt(4.8448, df = 10, lower.tail = FALSE) = 0.0003

(Problem 2, 3) A sports league collected data on 11 randomly selected players, measuring their total game time (x, hours) and annual salary (y, USD 1M). Assume that the game time and the salary follows a bivariate normal distribution.

Player (i)	1	2	3	4	5	6	7	8	9	10	11
Game Time(x_i, hours)	10	15	20	25	30	12	18	28	22	16	24
Salary(y_i, USD 1M)	2.0	2.2	2.5	3.0	3.5	2.1	2.4	2.9	2.6	3.0	3.2

$$\bar{x} = \frac{1}{11}\sum_{i=1}^{11} x_i = 20, \quad \bar{y} = \frac{1}{11}\sum_{i=1}^{11} y_i = 2.6727, \quad s_x = \sqrt{\frac{1}{10}\sum_{i=1}^{11}(x_i - \bar{x})} = 6.4653, \quad s_y = \sqrt{\frac{1}{10}\sum_{i=1}^{11}(y_i - \bar{y})} = 0.4839,$$

$$S_{xx} = \sum_{i=1}^{11}(x_i - \bar{x})^2 = 418, \quad S_{yy} = \sum_{i=1}^{11}(y_i - \bar{y})^2 = 2.3418, \quad S_{xy} = \sum_{i=1}^{11}(x_i - \bar{x})(y_i - \bar{y}) = 26.6$$

2. We conduct **a hypothesis test for the population correlation coefficient** ρ to determine if there is a **positive** relationship between total game time and annual salary at the significance level $\alpha = 0.05$. [40 points]

(1) State the null and alternative hypotheses. (Use a **one-sided** test.)

Solution: $H_0 : \rho = 0, \quad H_A : \rho > 0$

(2) Find the null distribution of the test statistic $\dfrac{R\sqrt{n-2}}{\sqrt{1-R^2}}$, where R is the sample correlation and n is the sample size.

Solution: Under H_0, $T = \dfrac{R\sqrt{n-2}}{\sqrt{1-R^2}} = \dfrac{R\sqrt{9}}{\sqrt{1-R^2}} \sim t(9)$.

(3) Compute the observed sample correlation coefficient r.

Solution:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{26.6}{\sqrt{418 \cdot 2.3418}} = 0.8502.$$

(4) Compute the observed test statistic for problem (2).

Solution:

$$\frac{r\sqrt{11-2}}{\sqrt{1-r^2}} = \frac{0.8502 \cdot 3}{\sqrt{1-0.8502^2}} = 4.8448.$$

(5) Compute the p-value and complete the hypothesis test. State the conclusion in the context of data.

Solution:

$p\text{-value} = P(T > 4.8448) \approx \boxed{0.0004}$, where $T \sim t(9)$. Since the p-value is less than $\alpha = 0.05$, we reject the null hypothesis. Therefore, there is statistically significant evidence of a positive linear relationship between game time and salary.

3. A simple linear regression model is used to explain the relationship between total game time x and salary y of the players. Assume the linear model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$ independently for $i = 1, \cdots, 11$. [30 points]

(1) Compute the least squares estimates of the regression coefficients, $\hat{\beta}_0$ and $\hat{\beta}_1$.

Solution:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{26.6}{418} = 0.0636, \quad \text{or} \quad \hat{\beta}_1 = r\frac{s_y}{s_x} = 0.8502 \times \frac{0.4839}{6.4653} = 0.0636,$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 2.6727 - 0.0636 \cdot 20 = 1.4007.$$

(2) Compute the fitted value $\hat{y}_4$ and residual e_4 for the $i = 4$th player.

Solution:

$$\hat{y}_4 = 1.4007 + 0.0636 \cdot 25 = 2.9907, \quad e_4 = 3 - 2.9907 = 0.0093.$$

(3) Suppose Sum of Squares of Error is $SSE = \sum_{i=1}^{11}(y_i - \hat{y}_i)^2 = 0.6489$. Compute mean squared error $MSE = \hat{\sigma}^2$.

Solution: $MSE = \dfrac{SSE}{n-2} = \dfrac{0.6489}{11-2} = 0.0721$