

Environmental Engineering FE CBT Exam

Knowledge: Probability and Statistics with 10 Example Problems

Contents

1 Measures of Central Tendencies & Dispersions	2
1.1 Measures of Central Tendency	2
1.2 Measures of Dispersion	2
1.3 Calculator Efficiency Shortcuts for FE Exam	3
2 Probability Distributions	4
2.1 Discrete Probability Frameworks	4
2.1.1 Binomial Distribution	4
2.2 Continuous Probability Frameworks	4
2.3 Normal (Gaussian) Distribution	4
2.3.1 Standard Normal Transformation (Z-Score)	4
2.3.2 Reading NCEES Reference Handbook Unit Normal Tables	5
3 Estimation for a Single Mean	6
3.1 Point Estimation	6
3.2 Confidence Intervals for Population Mean (μ)	6
3.2.1 Case 1: Variance Known (σ Known) or Large Sample ($n \geq 30$)	6
3.2.2 Case 2: Variance Unknown (σ Unknown) and Small Sample ($n < 30$)	6
3.3 Sample Size Determination	6
4 Regression, Curve Fitting & Goodness of Fit	7
4.1 Least Squares Linear Regression	7
4.1.1 Sum of Squares Formulas	7
4.1.2 Slope and Intercept Equations	7
4.2 Goodness-of-Fit Measures	7
4.2.1 Sample Correlation Coefficient (R)	7
4.2.2 Coefficient of Determination (R^2)	7
4.2.3 Standard Error of Estimate (S_e / MSE)	8
5 Hypothesis Testing, Outlier Testing & ANOVA	9
5.1 Hypothesis Testing Framework	9
5.2 Outlier Testing: Dixon's Q-Test	9
5.2.1 Dixon's Q-Test Procedure	9
5.3 Analysis of Variance (ANOVA)	9
5.3.1 ANOVA Sum of Squares Partitioning	9
5.3.2 One-Way ANOVA Table Format	10
6 10 Example Problems	11

1 Measures of Central Tendencies & Dispersions

In environmental engineering applications, monitoring air, water, and soil parameters generates sample datasets that exhibit statistical variability. Summarizing these datasets using measures of central tendency (location) and dispersion (spread) is fundamental to regulatory compliance, monitoring network design, and risk assessment.

1.1 Measures of Central Tendency

Measures of central tendency identify the single value that best represents the center of a data distribution.

- **Sample Mean ($\bar{x}$):** The arithmetic average of n observations. Unbiased estimator of population mean μ .

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \cdots + x_n}{n} \quad (1)$$

- **Population Mean (μ):** The true arithmetic average of all N elements in the entire population.

$$\mu = \frac{1}{N} \sum_{i=1}^N X_i \quad (2)$$

- **Weighted Mean ($\bar{x}_w$):** Used when individual data points carry different relative importance, weights (w_i), or represent composite flow-weighted sampling in stream monitoring.

$$\bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} \quad (3)$$

- **Median:** The middle value when data points are arranged in ascending order. If n is odd, the median is the value at position $(n+1)/2$. If n is even, the median is the average of the two central values at positions $n/2$ and $(n/2)+1$. The median is robust against extreme environmental outliers (e.g., occasional toxic spill spikes).
- **Mode:** The value that occurs with the highest frequency in the dataset. A dataset may be unimodal, bimodal, or multimodal.

1.2 Measures of Dispersion

Measures of dispersion quantify the spread, variability, or uncertainty associated with environmental measurements.

- **Sample Variance (s^2):** The average squared deviation from the sample mean, utilizing Bessel's correction ($n-1$) to ensure s^2 is an unbiased estimator of population variance σ^2 .

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}}{n-1} \quad (4)$$

- **Population Variance (σ^2):** The true variance of the entire population.

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N} \quad (5)$$

- **Sample Standard Deviation (s):** The positive square root of sample variance, expressed in the original physical units of measurement (e.g., mg/L, NTU, ppm).

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}} \quad (6)$$

- **Coefficient of Variation (CV):** A dimensionless ratio of the standard deviation to the mean, expressed as a percentage. Used to compare relative precision across different measurement techniques or environmental parameters with distinct scales.

$$CV = \frac{s}{\bar{x}} \times 100\% \quad (7)$$

1.3 Calculator Efficiency Shortcuts for FE Exam

Examinees using the **TI-36X Pro** or **Casio fx-115ES Plus** can compute descriptive statistics in seconds:

1. Enter data list menu (**data** on TI-36X Pro; **MODE** → **STAT** → **1-VAR** on Casio).
2. Input data values into Column L_1 .
3. Select 1-Variable Statistics (**2nd** → **stat-reg/distr** → **1-Var Stats**).
4. Read output values directly: $\bar{x}$ (sample mean), S_x (sample standard deviation s), and σ_x (population standard deviation σ).

2 Probability Distributions

Probability distributions model the likelihood of observing specific environmental events or continuous concentration ranges.

2.1 Discrete Probability Frameworks

A discrete random variable X takes on distinct, countable values. Its Probability Mass Function (PMF) $p(x) = P(X = x)$ satisfies:

$$0 \leq p(x) \leq 1 \quad \text{and} \quad \sum_{\text{all } x} p(x) = 1 \quad (8)$$

2.1.1 Binomial Distribution

The Binomial distribution models the number of successes x in n independent, identical trials (Bernoulli trials), where each trial has a constant probability of success p and probability of failure $q = 1 - p$.

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1 - p)^{n-x} \quad (9)$$

$$\text{Mean: } \mu = np, \quad \text{Variance: } \sigma^2 = np(1 - p) \quad (10)$$

Environmental Application: Daily compliance monitoring where a treatment plant either satisfies $(1 - p)$ or exceeds (p) daily discharge limits over an n -day monitoring window.

2.2 Continuous Probability Frameworks

A continuous random variable X takes on any value within an interval. Its Probability Density Function (PDF) $f(x)$ satisfies:

$$f(x) \geq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} f(x) dx = 1 \quad (11)$$

The probability that X falls within the interval $[a, b]$ corresponds to the area under the PDF curve:

$$P(a \leq X \leq b) = \int_a^b f(x) dx \quad (12)$$

2.3 Normal (Gaussian) Distribution

The Normal distribution is symmetric and bell-shaped, fully defined by its mean μ and variance σ^2 , denoted $X \sim N(\mu, \sigma^2)$.

2.3.1 Standard Normal Transformation (Z-Score)

To evaluate normal probabilities using NCEES Reference Handbook tables, any normal variable X is standardized to the Unit Normal distribution $Z \sim N(0, 1)$:

$$Z = \frac{X - \mu}{\sigma} \quad (13)$$

2.3.2 Reading NCEES Reference Handbook Unit Normal Tables

The NCEES Handbook provides standard unit normal tables defining three cumulative area functions:

- $F(z) = P(Z \leq z)$: Cumulative distribution function from $-\infty$ to z .
- $R(z) = P(Z > z) = 1 - F(z)$: Upper-tail probability (right-tail exceedance).
- $W(z) = P(-z \leq Z \leq z) = 2F(z) - 1$: Central area bounded between $-z$ and $+z$.

Z-score	$F(z) = P(Z \leq z)$	$R(z) = P(Z > z)$	$W(z) = P(-z \leq Z \leq z)$
1.00	0.8413	0.1587	0.6826
1.28	0.8997	0.1003	0.7994
1.645	0.9500	0.0500	0.9000
1.960	0.9750	0.0250	0.9500
2.326	0.9900	0.0100	0.9800
2.576	0.9950	0.0050	0.9900

3 Estimation for a Single Mean

Point estimation provides a single numerical value as an estimate of an unknown population parameter, while confidence intervals construct an interval around the point estimate that contains the true population parameter with a specified probability $(1 - \alpha)$.

3.1 Point Estimation

- The sample mean $\bar{x}$ is the minimum-variance unbiased point estimator of the population mean μ .
- The sample variance s^2 is the unbiased point estimator of the population variance σ^2 .

3.2 Confidence Intervals for Population Mean (μ)

The confidence interval for a population mean depends on sample size n and whether the population standard deviation σ is known or unknown.

3.2.1 Case 1: Variance Known (σ Known) or Large Sample ($n \geq 30$)

When σ is known or sample size is large ($n \geq 30$), the Central Limit Theorem guarantees that the sampling distribution of $\bar{x}$ is approximately normal.

$$CI_{1-\alpha} : \quad \bar{x} \pm Z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right) \quad (14)$$

Where $Z_{\alpha/2}$ is the critical standard normal value corresponding to upper-tail probability $\alpha/2$.

- 90% Confidence Level ($\alpha = 0.10$): $Z_{0.05} = 1.645$
- 95% Confidence Level ($\alpha = 0.05$): $Z_{0.025} = 1.960$
- 99% Confidence Level ($\alpha = 0.01$): $Z_{0.005} = 2.576$

3.2.2 Case 2: Variance Unknown (σ Unknown) and Small Sample ($n < 30$)

When σ is unknown and sample size is small ($n < 30$), the sample standard deviation s replaces σ , and Student's t -distribution with $\nu = n - 1$ degrees of freedom is used:

$$CI_{1-\alpha} : \quad \bar{x} \pm t_{\alpha/2, n-1} \left(\frac{s}{\sqrt{n}} \right) \quad (15)$$

The standard error of the mean is defined as:

$$SE = \frac{s}{\sqrt{n}} \quad (16)$$

3.3 Sample Size Determination

To estimate a population mean μ within a specified maximum allowable error margin $E = |\bar{x} - \mu|$ at a $(1 - \alpha)$ confidence level, the minimum required sample size is:

$$n = \left(\frac{Z_{\alpha/2} \cdot \sigma}{E} \right)^2 \quad (17)$$

If n is fractional, always round up to the nearest integer.

4 Regression, Curve Fitting & Goodness of Fit

Linear regression establishes mathematical relationships between an independent environmental variable x (e.g., reagent dosage, time, temperature) and a dependent variable y (e.g., pollutant removal, concentration, absorbance).

4.1 Least Squares Linear Regression

The linear regression model assumes a linear relationship:

$$\hat{y} = \hat{a} + \hat{b}x \quad (18)$$

Where $\hat{a}$ is the y -intercept and $\hat{b}$ is the slope of the regression line.

4.1.1 Sum of Squares Formulas

To compute least-squares parameters, calculate the following fundamental sums of squares from n paired data points (x_i, y_i) :

$$S_{xx} = \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} = \sum_{i=1}^n (x_i - \bar{x})^2 \quad (19)$$

$$S_{yy} = \sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n} = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (20)$$

$$S_{xy} = \sum_{i=1}^n x_i y_i - \frac{(\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (21)$$

4.1.2 Slope and Intercept Equations

$$\text{Slope: } \hat{b} = \frac{S_{xy}}{S_{xx}} \quad (22)$$

$$\text{Intercept: } \hat{a} = \bar{y} - \hat{b}\bar{x} \quad (23)$$

4.2 Goodness-of-Fit Measures

4.2.1 Sample Correlation Coefficient (R)

The Pearson correlation coefficient R measures the strength and direction of linear association between x and y ($-1 \leq R \leq 1$):

$$R = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \quad (24)$$

4.2.2 Coefficient of Determination (R²)

R^2 represents the proportion of total variance in the dependent variable y explained by the linear regression on x ($0 \leq R^2 \leq 1$):

$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = (R)^2 \quad (25)$$

4.2.3 Standard Error of Estimate (Se / MSE)

The residual variance or Mean Square Error ($MSE = S_e^2$) measures the scatter of data points around the regression line with $n - 2$ degrees of freedom:

$$S_e^2 = \text{MSE} = \frac{S_{yy} - \hat{b}S_{xy}}{n - 2} = \frac{S_{xx}S_{yy} - S_{xy}^2}{S_{xx}(n - 2)} \quad (26)$$

5 Hypothesis Testing, Outlier Testing & ANOVA

Hypothesis testing provides a decision-making framework to determine whether empirical environmental data supports a specific scientific claim or regulatory compliance standard.

5.1 Hypothesis Testing Framework

1. **Null Hypothesis (H_0):** The baseline statement of no effect, no difference, or compliance (e.g., $H_0 : \mu = \mu_0$).
2. **Alternative Hypothesis (H_1):** The statement indicating a true change or violation (two-tailed $H_1 : \mu \neq \mu_0$; one-tailed $H_1 : \mu > \mu_0$ or $H_1 : \mu < \mu_0$).
3. **Decision Errors:**
 - **Type I Error (α):** Rejecting H_0 when H_0 is true (significance level).
 - **Type II Error (β):** Failing to reject H_0 when H_0 is false. Power of test = $1 - \beta$.
4. **Test Statistic for Single Mean (Variance Unknown):**

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \quad (27)$$

Reject H_0 if $|t_0| > t_{\alpha/2, n-1}$ (for two-tailed) or $t_0 > t_{\alpha, n-1}$ (for upper one-tailed).

5.2 Outlier Testing: Dixon's Q-Test

Dixon's Q -test screens suspected single outliers in small laboratory datasets ($3 \leq n \leq 30$).

5.2.1 Dixon's Q-Test Procedure

1. Arrange sample values in ascending order: $x_1 \leq x_2 \leq \dots \leq x_n$.
2. Compute the experimental test statistic Q_{exp} :

$$Q_{\text{exp}} = \frac{|x_{\text{suspect}} - x_{\text{nearest}}|}{x_n - x_1} = \frac{\text{Gap}}{\text{Range}} \quad (28)$$

3. Compare Q_{exp} against the critical value Q_{crit} from the appropriate table.
If $Q_{\text{exp}} > Q_{\text{crit}}$, flag the suspected value as a statistical outlier; review QA/QC context before exclusion.

5.3 Analysis of Variance (ANOVA)

One-Way ANOVA tests whether population means across k independent treatment groups are equal ($H_0 : \mu_1 = \mu_2 = \dots = \mu_k$) vs. at least one mean differs.

5.3.1 ANOVA Sum of Squares Partitioning

$$SST = SS_{\text{Tr}} + SSE \quad (29)$$

Where SST is Total Sum of Squares, SS_{Tr} is Treatment Sum of Squares (between-group), and SSE is Error Sum of Squares (within-group).

5.3.2 One-Way ANOVA Table Format

Source	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F-Test Statistic
Treatments	SS_{Tr}	$k - 1$	$MST = \frac{SS_{Tr}}{k-1}$	$F_0 = \frac{MST}{MSE}$
Error	SSE	$N - k$	$MSE = \frac{SSE}{N-k}$	
Total	SST	$N - 1$	–	–

Where $N = \sum_{i=1}^k n_i$ is the total number of observations across all k groups. Reject H_0 if $F_0 > F_{\alpha, k-1, N-k}$.

6 10 Example Problems

Example Problem 1: Descriptive Statistics for Effluent BOD₅

Problem Statement: An environmental technician monitors daily 5-day Biochemical Oxygen Demand (BOD₅) concentrations in the effluent discharge of a municipal wastewater plant. The measured values for eight consecutive sampling days (mg/L) are:

$$22.4, \quad 18.6, \quad 25.1, \quad 20.0, \quad 18.6, \quad 24.2, \quad 21.5, \quad 19.2$$

Determine: (a) the sample mean ($\bar{x}$), (b) the median, (c) the mode, and (d) the sample standard deviation (s).

Solution

1. Step 1: Calculate Sample Mean ($\bar{x}$)

$$\sum_{i=1}^8 x_i = 22.4 + 18.6 + 25.1 + 20.0 + 18.6 + 24.2 + 21.5 + 19.2 = 169.6 \text{ mg/L}$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{169.6}{8} = \mathbf{21.20 \text{ mg/L}}$$

2. Step 2: Determine Median Arrange the dataset in ascending order ($n = 8$):

$$18.6, \quad 18.6, \quad 19.2, \quad \mathbf{20.0}, \quad \mathbf{21.5}, \quad 22.4, \quad 24.2, \quad 25.1$$

Since $n = 8$ is even, the median is the average of the 4th and 5th values:

$$\text{Median} = \frac{20.0 + 21.5}{2} = \mathbf{20.75 \text{ mg/L}}$$

3. Step 3: Determine Mode The value 18.6 mg/L occurs twice, while all other observations occur once.

$$\text{Mode} = \mathbf{18.60 \text{ mg/L}}$$

4. Step 4: Calculate Sample Standard Deviation (s) Calculate squared deviations from the mean ($\bar{x} = 21.20$):

$$\begin{aligned} \sum x_i^2 &= 22.4^2 + 18.6^2 + 25.1^2 + 20.0^2 + 18.6^2 + 24.2^2 + 21.5^2 + 19.2^2 \\ &= 501.76 + 345.96 + 630.01 + 400.00 + 345.96 + 585.64 + 462.25 + 368.64 = 3640.22 \\ s^2 &= \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1} = \frac{3640.22 - \frac{(169.6)^2}{8}}{8-1} = \frac{3640.22 - 3595.52}{7} = \frac{44.70}{7} = 6.3857 \text{ (mg/L)}^2 \\ s &= \sqrt{6.3857} = \mathbf{2.53 \text{ mg/L}} \end{aligned}$$

Example Problem 2: Relative Precision via Coefficient of Variation

Problem Statement: An environmental laboratory compares the measurement precision of two turbidity sensors. Filter A sensor yields a sample mean of $\bar{x}_A = 4.20$ NTU with standard deviation $s_A = 0.55$ NTU. Filter B sensor monitoring raw inlet water yields a sample mean of $\bar{x}_B = 85.0$ NTU with standard deviation $s_B = 6.80$ NTU. Calculate the Coefficient of Variation (CV) for both sensors and identify which sensor demonstrates higher relative precision.

Solution**1. Step 1: Calculate CV for Sensor A**

$$CV_A = \frac{s_A}{\bar{x}_A} \times 100\% = \frac{0.55 \text{ NTU}}{4.20 \text{ NTU}} \times 100\% = \mathbf{13.10\%}$$

2. Step 2: Calculate CV for Sensor B

$$CV_B = \frac{s_B}{\bar{x}_B} \times 100\% = \frac{6.80 \text{ NTU}}{85.0 \text{ NTU}} \times 100\% = \mathbf{8.00\%}$$

3. Step 3: Compare Relative Precision Sensor B has a lower Coefficient of Variation ($8.00\% < 13.10\%$), demonstrating **higher relative measurement precision** despite recording a larger absolute standard deviation.

Example Problem 3: Binomial Emission Exceedance Probability

Problem Statement: An industrial manufacturing facility operates a scrubbed combustion stack. Historical monitoring shows that the probability of exceeding the daily stack gas SO₂ emission limit on any given operating day is $p = 0.12$. Assuming daily exceedance events are independent Bernoulli trials:

- (a) What is the probability that exactly 2 days exceed the emission limit in a 10-day operating period?
- (b) What is the probability that at least 1 day exceeds the emission limit during the 10-day period?

Solution

1. **Step 1: Apply Binomial PMF for $x = 2$ ($n = 10, p = 0.12$)**

$$P(X = 2) = \binom{10}{2} (0.12)^2 (1 - 0.12)^{10-2} = \frac{10!}{2!8!} (0.12)^2 (0.88)^8$$

$$\binom{10}{2} = \frac{10 \times 9}{2} = 45$$

$$P(X = 2) = 45 \times (0.0144) \times (0.35963) = \mathbf{0.2330} \quad (\mathbf{23.30\%})$$

2. **Step 2: Calculate Probability of At Least 1 Exceedance ($P(X \geq 1)$)** Utilize the complement rule $P(X \geq 1) = 1 - P(X = 0)$:

$$P(X = 0) = \binom{10}{0} (0.12)^0 (0.88)^{10} = 1 \times 1 \times (0.88)^{10} = 0.2785$$

$$P(X \geq 1) = 1 - 0.2785 = \mathbf{0.7215} \quad (\mathbf{72.15\%})$$

Example Problem 4: Standard Normal Distribution for Groundwater Contamination (Topic B)

Problem Statement: Monitoring well data indicates that trichloroethylene (TCE) concentrations in a regional aquifer are normally distributed with a mean concentration of $\mu = 3.80 \mu\text{g/L}$ and standard deviation $\sigma = 0.75 \mu\text{g/L}$. The EPA Maximum Contaminant Level (MCL) for TCE is $5.00 \mu\text{g/L}$. Determine the probability that a randomly sampled monitoring well yields a TCE concentration exceeding the EPA limit ($P(X > 5.00)$).

Solution

1. **Step 1: Compute Z-Score for $X = 5.00 \mu\text{g/L}$**

$$Z = \frac{X - \mu}{\sigma} = \frac{5.00 - 3.80}{0.75} = \frac{1.20}{0.75} = \mathbf{1.60}$$

2. **Step 2: Read Upper-Tail Exceedance Probability $R(Z = 1.60)$** From the NCEES Unit Normal table for $Z = 1.60$:

$$\begin{aligned} F(1.60) &= P(Z \leq 1.60) = 0.9452 \\ P(X > 5.00) &= R(1.60) = 1 - F(1.60) = 1 - 0.9452 = \mathbf{0.0548} \quad (\mathbf{5.48\%}) \end{aligned}$$

Example Problem 5: Confidence Interval for Landfill Leachate Lead

Problem Statement: An environmental investigation collects $n = 9$ leachate samples from a hazardous landfill. Laboratory analysis for lead concentration yields a sample mean of $\bar{x} = 1.45$ mg/L with sample standard deviation $s = 0.36$ mg/L. Assuming lead concentration is normally distributed with unknown population variance, construct a 95% confidence interval for the true population mean lead concentration (μ).

Solution

1. **Step 1: Identify Statistical Distribution and Critical Value** Since σ is unknown and $n = 9 < 30$, use Student's t -distribution with degrees of freedom $\nu = n - 1 = 9 - 1 = 8$. For a 95% confidence level ($\alpha = 0.05$, two-tailed $\alpha/2 = 0.025$):

$$t_{0.025,8} = \mathbf{2.306}$$

2. **Step 2: Calculate Standard Error of the Mean (SE)**

$$SE = \frac{s}{\sqrt{n}} = \frac{0.36}{\sqrt{9}} = \frac{0.36}{3} = \mathbf{0.12} \text{ mg/L}$$

3. **Step 3: Compute Margin of Error (E)**

$$E = t_{\alpha/2, n-1} \cdot SE = 2.306 \times 0.12 = \mathbf{0.2767} \text{ mg/L}$$

4. **Step 4: Calculate 95% Confidence Bounds**

$$\text{Lower Bound} = \bar{x} - E = 1.45 - 0.2767 = \mathbf{1.173} \text{ mg/L}$$

$$\text{Upper Bound} = \bar{x} + E = 1.45 + 0.2767 = \mathbf{1.727} \text{ mg/L}$$

The 95% confidence interval for the mean lead concentration is $1.173 \leq \mu \leq 1.727$ **mg/L**.

Example Problem 6: Spectrophotometer Least Squares Calibration Curve**(Topic D)**

Problem Statement: A laboratory spectrophotometer is calibrated for nitrate determination using five standard solutions (x , mg/L) and recording peak absorbance (y , arbitrary units):

Nitrate x_i (mg/L)	2.0	4.0	6.0	8.0	10.0
Absorbance y_i	0.12	0.22	0.35	0.44	0.57

Determine the least-squares linear regression line equation ($\hat{y} = \hat{a} + \hat{b}x$).

Solution**1. Step 1: Compute Summary Terms ($n = 5$)**

$$\begin{aligned}\sum x_i &= 2.0 + 4.0 + 6.0 + 8.0 + 10.0 = 30.0, & \bar{x} &= 30.0/5 = 6.0 \\ \sum y_i &= 0.12 + 0.22 + 0.35 + 0.44 + 0.57 = 1.70, & \bar{y} &= 1.70/5 = 0.34 \\ \sum x_i^2 &= 2.0^2 + 4.0^2 + 6.0^2 + 8.0^2 + 10.0^2 = 4 + 16 + 36 + 64 + 100 = 220.0 \\ \sum x_i y_i &= (2 \times 0.12) + (4 \times 0.22) + (6 \times 0.35) + (8 \times 0.44) + (10 \times 0.57) \\ &= 0.24 + 0.88 + 2.10 + 3.52 + 5.70 = 12.44\end{aligned}$$

2. Step 2: Calculate Sums of Squares S_{xx} and S_{xy}

$$\begin{aligned}S_{xx} &= \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 220.0 - \frac{(30.0)^2}{5} = 220.0 - 180.0 = \mathbf{40.0} \\ S_{xy} &= \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 12.44 - \frac{30.0 \times 1.70}{5} = 12.44 - 10.20 = \mathbf{2.24}\end{aligned}$$

3. Step 3: Compute Slope ($\hat{b}$) and Intercept ($\hat{a}$)

$$\begin{aligned}\hat{b} &= \frac{S_{xy}}{S_{xx}} = \frac{2.24}{40.0} = \mathbf{0.0560} \\ \hat{a} &= \bar{y} - \hat{b}\bar{x} = 0.34 - (0.0560 \times 6.0) = 0.34 - 0.336 = \mathbf{0.0040}\end{aligned}$$

The least-squares calibration regression line is $\hat{y} = 0.0040 + 0.0560x$.

Example Problem 7: Goodness of Fit Coefficient of Determination R² (Topic**D)**

Problem Statement: Using the spectrophotometer calibration data from Problem 6 ($n = 5$, $S_{xx} = 40.0$, $S_{xy} = 2.24$), where $\sum y_i^2 = 0.12^2 + 0.22^2 + 0.35^2 + 0.44^2 + 0.57^2 = 0.7038$: Calculate (a) the total sum of squares S_{yy} , (b) the Pearson correlation coefficient (R), and (c) the coefficient of determination (R^2).

Solution

1. **Step 1: Calculate S_{yy}**

$$S_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 0.7038 - \frac{(1.70)^2}{5} = 0.7038 - 0.5780 = \mathbf{0.1258}$$

2. **Step 2: Calculate Correlation Coefficient (R)**

$$R = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{2.24}{\sqrt{40.0 \times 0.1258}} = \frac{2.24}{\sqrt{5.032}} = \frac{2.24}{2.2432} = \mathbf{0.9986}$$

3. **Step 3: Calculate Coefficient of Determination (R^2)**

$$R^2 = (R)^2 = (0.9986)^2 = \mathbf{0.9972} \quad (\mathbf{99.72\%})$$

The linear model accounts for **99.72%** of the total variance in absorbance data.

Example Problem 8: One-Sample t-Test for Soil Remediation Compliance**(Topic E)**

Problem Statement: A soil remediation site must satisfy a regulatory standard requiring mean heavy metal concentration to be below $\mu_0 = 50.0$ mg/kg. Following bio-remediation, six independent core samples ($n = 6$) yield a sample mean of $\bar{x} = 44.5$ mg/kg with sample standard deviation $s = 5.2$ mg/kg. Perform a one-tailed Student's t -test at significance level $\alpha = 0.05$ to evaluate whether the site complies with regulatory standards ($H_0 : \mu = 50.0$ vs. $H_1 : \mu < 50.0$).

Solution**1. Step 1: State Hypotheses and Significance Level**

$$H_0 : \mu = 50.0 \text{ mg/kg}, \quad H_1 : \mu < 50.0 \text{ mg/kg}, \quad \alpha = 0.05$$

2. Step 2: Calculate t_0 Test Statistic

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{44.5 - 50.0}{5.2/\sqrt{6}} = \frac{-5.5}{5.2/2.4495} = \frac{-5.5}{2.1229} = -\mathbf{2.591}$$

3. Step 3: Determine Critical Value $t_{\alpha, n-1}$ For a lower one-tailed test with degrees of freedom $\nu = 6 - 1 = 5$ and $\alpha = 0.05$:

$$t_{\text{crit}} = -t_{0.05, 5} = -\mathbf{2.015}$$

4. Step 4: Decision Rule and Conclusion Since $t_0 = -2.591 < -2.015$, the test statistic falls into the rejection region. **Reject H_0 .** There is sufficient statistical evidence at the $\alpha = 0.05$ level to conclude that the mean soil contamination is significantly below 50.0 mg/kg

Example Problem 9: Outlier Rejection via Dixon Q-Test

Problem Statement: An analytical laboratory measures water hardness (CaCO_3 mg/L) across six repeated titrations:

$$142.5, \quad 143.1, \quad 142.8, \quad 143.0, \quad 142.6, \quad 151.2$$

The value 151.2 mg/L is suspected to be an analytical outlier. Perform Dixon's Q-test at significance level 0.05 to determine whether the value is statistically flagged as an outlier. (Qcrit for $n = 6$ at significance level 0.05 is 0.621).

Solution**1. Step 1: Sort Sample Data in Ascending Order**

$$x_1 = 142.5, \quad x_2 = 142.6, \quad x_3 = 142.8, \quad x_4 = 143.0, \quad x_5 = 143.1, \quad x_6 = 151.2$$

2. Step 2: Calculate Experimental Q-Statistic (Q_{exp}) The suspect value is the upper extreme $x_6 = 151.2$, and its nearest neighbor is $x_5 = 143.1$:

$$\text{Gap} = |x_6 - x_5| = 151.2 - 143.1 = 8.1 \text{ mg/L}$$

$$\text{Range} = x_6 - x_1 = 151.2 - 142.5 = 8.7 \text{ mg/L}$$

$$Q_{\text{exp}} = \frac{\text{Gap}}{\text{Range}} = \frac{8.1}{8.7} = \mathbf{0.9310}$$

3. Step 3: Compare Qexp with Critical Value Qcrit

Since $Q_{\text{exp}} = 0.9310 > Q_{\text{crit}} = 0.621$, 151.2 mg/L is statistically flagged as an outlier at the 0.05 significance level. Review analytical QA/QC and documented justification before deciding whether to exclude it from the final reported statistics.

Example Problem 10: One-Way ANOVA for Watershed pH Comparison (Topic**E)**

Problem Statement: An environmental study compares stream pH levels across three watershed monitoring stations ($k = 3$). Each station is sampled four times ($n_i = 4, N = 12$). The ANOVA sum of squares calculations yield:

$$\text{Total Sum of Squares } (SST) = 2.84, \quad \text{Treatment Sum of Squares } (SS_{\text{Tr}}) = 1.92$$

Compile the complete One-Way ANOVA table and determine whether stream pH differs significantly across stations at significance level $\alpha = 0.05$. (F_{crit} for $df = 2, 9$ at $\alpha = 0.05$ is 4.26).

Solution**1. Step 1: Calculate Error Sum of Squares (SSE)**

$$SSE = SST - SS_{\text{Tr}} = 2.84 - 1.92 = \mathbf{0.92}$$

2. Step 2: Determine Degrees of Freedom

$$df_{\text{Tr}} = k - 1 = 3 - 1 = \mathbf{2}$$

$$df_{\text{Error}} = N - k = 12 - 3 = \mathbf{9}$$

$$df_{\text{Total}} = N - 1 = 12 - 1 = \mathbf{11}$$

3. Step 3: Calculate Mean Squares (MST and MSE)

$$MST = \frac{SS_{\text{Tr}}}{df_{\text{Tr}}} = \frac{1.92}{2} = \mathbf{0.960}$$

$$MSE = \frac{SSE}{df_{\text{Error}}} = \frac{0.92}{9} = \mathbf{0.1022}$$

4. Step 4: Compute F_0 Test Statistic

$$F_0 = \frac{MST}{MSE} = \frac{0.960}{0.1022} = \mathbf{9.393}$$

5. Step 5: Complete ANOVA Table and Make Decision

Source	SS	df	MS	F_0	F_{crit}
Treatments	1.92	2	0.960	9.393	4.26
Error	0.92	9	0.1022	—	—
Total	2.84	11	—	—	—

Since $F_0 = 9.393 > F_{\text{crit}} = 4.26$, **reject** H_0 . Stream pH levels differ significantly across the three monitoring stations at $\alpha = 0.05$.