Statistics Exam Solutions

Question 1

(a) Stem-and-Leaf Display

Data: 359, 356, 359, 363, 375, 424, 325, 394, 402, 375, 373, 370, 364, 366, 364, 325, 339, 393, 392, 369, 374, 359, 356, 403, 334, 397

Stem-and-Leaf Display:

```
32 | 5 5
33 | 4 9
34 |
35 | 6 6 9 9 9
36 | 3 4 4 6 9
37 | 0 3 4 5 5
38 |
39 | 2 3 4 7
40 | 2 3
41 |
42 | 4
```

Interesting Features:

- The data is slightly right-skewed
- There's a gap between 383-391 (no values in this range)
- Most values cluster between 350-380 seconds
- There's an outlier at 424 seconds

(b) Nitrogen Loads Statistics

Data: 9.69, 13.16, 17.09, 18.12, 23.70, 24.07, 24.29, 26.43, 30.75, 31.54

Calculations:

- Median (Q2) = (23.70 + 24.07)/2 = 23.885 kg N/day
- Lower Fourth (Q1) = Median of lower half = 17.09 kg N/day
- Upper Fourth (Q3) = Median of upper half = 26.43 kg N/day

(c) Home Sales

Data: 590, 815, 575, 608, 350, 1285, 408, 540, 555, 679 (in $1000s)

(i) Sample Variance and Standard Deviation

Calculations:

- Mean ($\bar{x}$) = (590+815+575+608+350+1285+408+540+555+679)/10 = 640.5
- Sum of squared deviations = $(590 - 640.5)^2 + \ldots + (679 - 640.5)^2 = 486,372.5$
- Variance $= 486,372.5/(10 - 1) = 54,041.39$
- Standard Deviation $= \sqrt{54,041.39} \approx 232.47$

(ii) Correction for Data Collection Error

If all observations were multiplied by 5:

- New Variance $= 5^2\times$ Original Variance $= 25 \times 54,041.39 = 1,351,034.75$
- New Standard Deviation $= 5\times$ Original SD $= 5 \times 232.47 \approx 1,162.35$

Question 2

(a) Probability Calculations

Given:

- $P(A_1) = 0.22$, $P(A_2) = 0.25$, $P(A_3) = 0.28$
- $P(A_1 \cap A_2) = 0.11$, $P(A_1 \cap A_3) = 0.05$, $P(A_2 \cap A_3) = 0.07$
- $P(A_1 \cap A_2 \cap A_3) = 0.01$

$P(A_1' \cap A_2')$

$$\begin{aligned} &= 1 - P(A_1 \cup A_2) \\ &= 1 - [P(A_1) + P(A_2) - P(A_1 \cap A_2)] \\ &= 1 - [0.22 + 0.25 - 0.11] = 1 - 0.36 = 0.64 \end{aligned}$$

$P(A_1 \cup A_2 \cup A_3)$

$$\begin{aligned} &= P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3) \\ &= 0.22 + 0.25 + 0.28 - 0.11 - 0.05 - 0.07 + 0.01 = 0.53 \end{aligned}$$

(b) Bayes' Theorem Application

Given:

- $P(A_1) = 0.70$, $P(A_2) = 0.20$, $P(A_3) = 0.10$
- $P(\text{Spam}|A_1) = 0.01$, $P(\text{Spam}|A_2) = 0.02$, $P(\text{Spam}|A_3) = 0.05$

Bayes' Theorem Statement:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum[P(B|A_j)P(A_j)]}$$

Calculation:

$$\begin{aligned}
P(A_2|\text{Spam}) &= \frac{P(\text{Spam}|A_2)P(A_2)}{P(\text{Spam}|A_1)P(A_1) + P(\text{Spam}|A_2)P(A_2) + P(\text{Spam}|A_3)P(A_3)} \\
&= \frac{0.02 \times 0.20}{0.01 \times 0.70 + 0.02 \times 0.20 + 0.05 \times 0.10} \\
&= \frac{0.004}{0.007 + 0.004 + 0.005} = \frac{0.004}{0.016} = 0.25
\end{aligned}$$

(c) Batch Defect Probabilities

Given:

- $P(0 \text{ defect}) = 0.50$
- $P(1 \text{ defect}) = 0.30$
- $P(2 \text{ defects}) = 0.20$

Condition: One of two tested components is defective.
Calculations:
1. $P(1 \text{ defect in batch}|1 \text{ defective in sample})$:

$$\begin{aligned}
&= \frac{P(1 \text{ defect in batch AND 1 defective in sample})}{P(1 \text{ defective in sample})} \\
&= \frac{0.30 \times 1}{0.50 \times 0 + 0.30 \times 1 + 0.20 \times \frac{2\times1\times9}{10\times9}} \\
&= \frac{0.30}{0 + 0.30 + 0.20 \times \frac{18}{90}} = \frac{0.30}{0.30 + 0.04} = \frac{0.30}{0.34} \approx 0.8824
\end{aligned}$$

2. $P(2 \text{ defects in batch}|1 \text{ defective in sample})$:

$$= \frac{0.20 \times \frac{2\times1\times8}{90}}{0.34} = \frac{0.20 \times \frac{16}{90}}{0.34} \approx \frac{0.0356}{0.34} \approx 0.1047$$

3. $P(0 \text{ defects in batch}|1 \text{ defective in sample}) = 0$ (impossible since we found a defect)

Question 3

(a) Geometric Distribution

(i) PMF of X

$$P(X = k) = (1 - p)^{k-1} \times p \text{ for } k = 1, 2, 3, \ldots$$

(ii) CMF of X

$$F(k) = P(X \leq k) = 1 - (1 - p)^k$$

(b) Bernoulli Random Variable

(i) Name

Bernoulli random variable

(ii) $E(X^2)$

$$E(X^2) = 1^2 \times p + 0^2 \times (1 - p) = p$$

(iii) $V(X)$

$$V(X) = E(X^2) - [E(X)]^2 = p - p^2 = p(1 - p)$$

(c) Random Variable Calculations

Given:

- $E(X) = 5$
- $E[X(X - 1)] = 27.5$

(i) $E(X^2)$

$$E[X(X - 1)] = E(X^2) - E(X) \Rightarrow 27.5 = E(X^2) - 5 \Rightarrow E(X^2) = 32.5$$

(ii) $V(X)$

$$V(X) = E(X^2) - [E(X)]^2 = 32.5 - 25 = 7.5$$

(iii) $V(2X + 3)$

$$V(2X + 3) = 2^2 \times V(X) = 4 \times 7.5 = 30$$

Question 4

(a) Continuous Random Variable

(i) CDF

$$
\begin{aligned}
F(x) &= \int_{-2}^{x} 0.09375(4 - t^2)\, dt = 0.09375 \left[4t - \frac{t^3}{3}\right]_{-2}^{x} \\
&= 0.09375 \left(4x - \frac{x^3}{3}\right) - 0.09375 \left(-8 + \frac{8}{3}\right) \\
&= 0.09375 \left(4x - \frac{x^3}{3} + 8 - \frac{8}{3}\right) \\
&= 0.09375 \left(4x - \frac{x^3}{3} + \frac{16}{3}\right)
\end{aligned}
$$

(ii) $P(-1 < X < 1)$

$$
\begin{aligned}
F(1) - F(-1) &= \left[0.09375 \left(4 - \frac{1}{3} + \frac{16}{3}\right)\right] - \left[0.09375 \left(-4 + \frac{1}{3} + \frac{16}{3}\right)\right] \\
&= 0.09375(4 + 5) - 0.09375 \left(-4 + \frac{17}{3}\right) \approx 0.84375 - 0.15625 = 0.6875
\end{aligned}
$$

(iii) $E[X]$

$$
E[X] = \int_{-2}^{2} x \times f(x)\, dx = 0 \text{ (odd function over symmetric interval)}
$$

(iv) $V[X]$

$$
\begin{aligned}
E[X^2] &= \int_{-2}^{2} x^2 \times 0.09375(4 - x^2)\, dx = 0.09375 \int_{-2}^{2} (4x^2 - x^4)\, dx \\
&= 2 \times 0.09375 \int_{0}^{2} (4x^2 - x^4)\, dx = 0.1875 \left[\frac{4x^3}{3} - \frac{x^5}{5}\right]_{0}^{2} \\
&= 0.1875 \left(\frac{32}{3} - \frac{32}{5}\right) = 0.1875 \times \frac{160 - 96}{15} = 0.1875 \times \frac{64}{15} = 0.8 \\
V[X] &= E[X^2] - (E[X])^2 = 0.8 - 0 = 0.8
\end{aligned}
$$

(b) Normal Approximation

Given: $p = 0.25$, $n = 50$

$$
\mu = np = 12.5, \quad \sigma = \sqrt{np(1 - p)} \approx 3.0619
$$

(i) $P(X \leq 10)$

With continuity correction: $P(X \leq 10.5)$

$$Z = \frac{10.5 - 12.5}{3.0619} \approx -0.6532$$

$$P(Z \leq -0.6532) \approx 0.2569$$

(ii) $P(5 \leq X \leq 15)$

With continuity correction: $P(4.5 \leq X \leq 15.5)$

$$Z_1 = \frac{4.5 - 12.5}{3.0619} \approx -2.612$$

$$Z_2 = \frac{15.5 - 12.5}{3.0619} \approx 0.9798$$

$$P(-2.612 \leq Z \leq 0.9798) \approx 0.8365 - 0.0045 \approx 0.8320$$

(c) Exponential Distribution

Definition: A continuous probability distribution that describes the time between events in a Poisson process.

For $X \sim \text{Exp}(\lambda)$:

- Mean $= 1/\lambda$
- Standard Deviation $= 1/\lambda$

Yes, the mean and standard deviation are equal for an exponential distribution.

Question 5

(a) Poisson Distribution

Given: $p = 0.005$ per page, $n = 400$ pages

$$\lambda = np = 2$$

(i) $P(X = 1)$

$$P(X = 1) = \frac{e^{-2} \times 2^1}{1!} \approx 0.2707$$

(ii) $P(X \leq 3)$

$$P(X=0) = e^{-2} \approx 0.1353$$
$$P(X=1) \approx 0.2707$$
$$P(X=2) = \frac{e^{-2} \times 2^2}{2!} \approx 0.2707$$
$$P(X=3) = \frac{e^{-2} \times 2^3}{3!} \approx 0.1804$$
$$P(X \leq 3) \approx 0.1353 + 0.2707 + 0.2707 + 0.1804 \approx 0.8571$$

(b) Discrete Random Variable

(i) PMF from CDF

x	$P(X=x)$
1	0.30
3	0.10 (0.40-0.30)
4	0.05 (0.45-0.40)
6	0.15 (0.60-0.45)
12	0.40 (1.00-0.60)

(ii) $P(3 \leq X \leq 6)$

$$= F(6) - F(3^-) = 0.60 - 0.30 = 0.30$$

(c) Normal Distribution

Given: $\mu = 12$ lb, $\sigma = 3.5$ lb Find c such that $P(X \leq c) = 0.99$
Z-score for 0.99: ≈ 2.326

$$c = \mu + z\sigma = 12 + 2.326 \times 3.5 \approx 20.14 \text{ lb}$$

Question 6

(a) Correlation Analysis

Data:

TOST (x)	4200	3600	3750	3675	4050	2770	4870	4500	3450	2700
RBOT (y)	370	340	375	310	350	200	400	375	285	225

(i) Sample Correlation Coefficient

Calculations:

- $n = 10$
- $\sum x = 36,565$, $\sum y = 3,230$
- $\sum xy = 12,098,750$
- $\sum x^2 = 139,948,750$
- $\sum y^2 = 1,083,400$

$$
\begin{aligned}
r &= \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \\
&= \frac{10 \times 12,098,750 - 36,565 \times 3,230}{\sqrt{(10 \times 139,948,750 - 36,565^2)(10 \times 1,083,400 - 3,230^2)}} \\
&\approx \frac{120,987,500 - 118,104,950}{\sqrt{(1,399,487,500 - 1,336,999,225)(10,834,000 - 10,432,900)}} \\
&\approx \frac{2,882,550}{\sqrt{62,488,275 \times 401,100}} \approx \frac{2,882,550}{5,006,800} \approx 0.576
\end{aligned}
$$

Interpretation: Moderate positive linear relationship.

(ii) Effect of Changing RBOT Units

The correlation coefficient r would remain exactly the same because it is unitless - changing the units of measurement doesn't affect the strength or direction of the linear relationship.

(b) Regression Analysis

Data:

Rainfall (x)	5	12	14	17	23	30	40	47	55	67
Runoff (y)	4	10	13	15	15	25	27	46	38	46

(i) Regression Line

Calculations:

- $n = 10$
- $\sum x = 310$, $\sum y = 239$
- $\sum xy = 9,183$

- $\sum x^2 = 12,406$
- $\sum y^2 = 7,245$

Slope (b):

$$\begin{aligned} b &= \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} \\ &= \frac{10 \times 9,183 - 310 \times 239}{10 \times 12,406 - 310^2} \\ &= \frac{91,830 - 74,090}{124,060 - 96,100} = \frac{17,740}{27,960} \approx 0.6345 \end{aligned}$$

Intercept (a):

$$a = \bar{y} - b\bar{x} = \frac{239}{10} - 0.6345 \times \frac{310}{10} = 23.9 - 19.6695 \approx 4.2305$$

Regression equation:

$$\hat{y} = 4.2305 + 0.6345x$$

(ii) Prediction for $x = 70$

$$\hat{y} = 4.2305 + 0.6345 \times 70 \approx 4.2305 + 44.415 \approx 48.6455 \text{ m}^3$$

(c) Sampling Distribution

Given: $\mu = 4.0$ g, $\sigma = 1.5$ g, $n = 50$

(i) Center

The sampling distribution of $\bar{X}$ is centered at the population mean $\mu = 4.0$ g

(ii) Standard Deviation

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1.5}{\sqrt{50}} \approx 0.2121 \text{ g}$$

(iii) $P(3.5 \leq \bar{X} \leq 3.8)$

$$\begin{aligned} Z_1 &= \frac{3.5 - 4}{0.2121} \approx -2.357 \\ Z_2 &= \frac{3.8 - 4}{0.2121} \approx -0.9428 \\ P(-2.357 \leq Z \leq -0.9428) &\approx 0.1727 - 0.0092 \approx 0.1635 \end{aligned}$$