SDS_SE_Ans

Let’s solve **Q.1 completely, step by step**, with **clear interpretation and probability calculations**, exactly at an exam-answer level.

Q.1 (A) Confidence Interval

Given

- Estimated **mean runtime** = **120 ms**
- **95% confidence interval** = **(115 ms, 125 ms)**

---

(1) Interpretation of the confidence interval

A **95% confidence interval** of **(115 ms, 125 ms)** means:

> If we were to **repeat the experiment many times** and compute a 95% confidence interval each time, **about 95% of those intervals would contain the true mean runtime** of the sorting algorithm.

In the context of this study:

- We are **reasonably confident** that the **true average runtime** of the sorting algorithm on large inputs lies **between 115 ms and 125 ms**.

⚠️ Important:
We are **not** saying the runtime will always fall in this range — this interval is about the **mean**, not individual runs.

---

(2) One common misconception about confidence intervals

❌ **Misconception:**

> “There is a 95% probability that the true mean lies between 115 ms and 125 ms.”

✅ **Why this is incorrect:**

- The **true mean is a fixed (unknown) value**, not a random variable.
- The **interval is random**, because it depends on the sample.
- Probability applies to the **method**, not the specific interval.

Correct interpretation:

> The **procedure** used to construct the interval captures the true mean **95% of the time** in repeated sampling.

Q.1 (B) Probability Using Contingency Table

Given Table

Electrical Failure	Gas Leak: Yes	Gas Leak: No
Yes	55	17
No	32	3

Step 1: Find total number of AC failures

$$\text{Total} = 55 + 17 + 32 + 3 = 107$$

(a) Probability that failure involves a gas leak

Failures involving gas leaks:

$$55 + 32 = 87$$

$$P(\text{Gas Leak}) = \frac{87}{107}$$

✅ **Answer:**

$$P(\text{Gas Leak}) = \frac{87}{107} \approx 0.813$$

(b) Probability of electrical failure given that there was a gas leak

This is a **conditional probability**.

Formula:

$$P(E \mid G) = \frac{P(E \cap G)}{P(G)}$$

Where:

- $E \cap G = 55$
- $G = 87$

$$P(E \mid G) = \frac{55}{87}$$

✅ **Answer:**

$$P(E \mid G) \approx 0.632$$

(c) Probability of gas leak given that there is electrical failure

Again, conditional probability.

Electrical failure total:

$$55 + 17 = 72$$

$$P(G \mid E) = \frac{55}{72}$$

✅ **Answer:**

$$P(G \mid E) \approx 0.764$$

Final Answers Summary

Part	Result
(a) $P(\text{Gas Leak})$	$\frac{87}{107} \approx 0.813$
(b) $P(\text{Electrical} \mid \text{Gas Leak})$	$\frac{55}{87} \approx 0.632$
(c) $P(\text{Gas Leak} \mid \text{Electrical})$	$\frac{55}{72} \approx 0.764$

If you want, I can also:

- Rewrite this **exactly in university exam format**
- Add **Venn diagram intuition**
- Or explain **why (b) ≠ (c)** conceptually

Just tell me 👍

Let's solve **Q.1 (C)** and **Q.2 (A) step by step**, slowly and clearly, exactly how it should be written in an exam.

Q.1 (C) Normal Distribution – Quality Control

Given

- Amount of ketchup per bottle is **normally distributed**
- Mean:

$$\mu = 36 \text{ oz}$$

- Standard deviation:

$$\sigma = 0.11 \text{ oz}$$

- Bottle **fails inspection** if:
  - below **35.8 oz**, or
  - above **36.2 oz**

Step 1: Convert values to Z-scores

(i) For 35.8 oz

$$z = \frac{x-\mu}{\sigma} = \frac{35.8-36}{0.11} = \frac{-0.2}{0.11} \approx -1.82$$

(ii) For 36.2 oz

$$z = \frac{36.2-36}{0.11} = \frac{0.2}{0.11} \approx +1.82$$

Step 2: Use standard normal table

From Z-table:

$$P(Z < -1.82) = 0.0344$$

$$P(Z < 1.82) = 0.9656$$

(a) Percent of bottles with less than 35.8 oz

$$P(X < 35.8) = P(Z < -1.82) = 0.0344$$

✅ **Answer:**

$$\boxed{3.44\%}$$

(b) Percent of bottles that pass inspection

Passing range:

$$35.8 \le X \le 36.2$$

$$P(\text{Pass}) = P(-1.82 \le Z \le 1.82)$$

$$= 0.9656 - 0.0344 = 0.9312$$

✅ **Answer:**

$$\boxed{93.12\%}$$

(Optional check)

Failing percentage:

$$100\% - 93.12\% = 6.88\%$$

(which matches both tails combined)

Q.2 (A) Binomial Probability

Given

- Probability a sample contains pollutant:

$$p = 0.10$$

- Number of samples:

$$n = 18$$

- Exactly **2 samples** contain pollutant

This is a **Binomial distribution**.

Step 1: Identify the formula

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Here:

- $n = 18$
- $k = 2$
- $p = 0.10$

Step 2: Substitute values

$$P(X = 2) = \binom{18}{2} (0.10)^2 (0.90)^{16}$$

Step 3: Compute each part

Combination:

$$\binom{18}{2} = \frac{18 \times 17}{2} = 153$$

Powers:

$$(0.10)^2 = 0.01$$

$$(0.90)^{16} \approx 0.185$$

Step 4: Final calculation

$$P(X = 2) = 153 \times 0.01 \times 0.185$$

$$= 0.283$$

✅ **Final Answer:**

$$\boxed{P(X = 2) \approx 0.283 \ (28.3\%)}$$

Final Answers Summary

Question	Result
% bottles < 35.8 oz	**3.44%**
% bottles passing inspection	**93.12%**
Probability exactly 2 polluted samples	**0.283**

If you want:

- Z-table reading explained line-by-line
- Graph intuition for normal distribution
- Why binomial applies here (and not Poisson)

Just tell me 👍

Let's solve **Q.2 (B)** and **Q.2 (C) step by step**, with **clear reasoning, formulas, and final numeric answers**, exactly as expected in a university exam.

Q.2 (B) Geometric Distribution (First Success)

Given

- Probability of success (successful alignment):

$$p = 0.8$$

- Trials are **independent**

Let random variable X = number of trials until the **first success**.

👉 This follows a **Geometric distribution**.

(a) Probability that the first success requires exactly 4 trials

Meaning:

- First 3 trials → **failures**
- 4th trial → **success**

Formula:

$$P(X = k) = (1 - p)^{k-1} p$$

Substitute values:

$$P(X = 4) = (0.2)^3 \times 0.8$$

$$= 0.008 \times 0.8 = 0.0064$$

✅ **Answer:**

$$\boxed{P(X = 4) = 0.0064}$$

(b) Probability that the first success requires at most 4 trials

Meaning:

$$P(X \leq 4)$$

This includes:

- Success on trial 1, 2, 3, or 4

Method 1 (Direct formula):

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

Substitute values:

$$P(X \leq 4) = 1 - (0.2)^4$$

$$= 1 - 0.0016 = 0.9984$$

✅ **Answer:**

$$\boxed{P(X \leq 4) = 0.9984}$$

(c) Probability that the first success requires at least 4 trials

Meaning:

$$P(X \geq 4)$$

This means:

- First 3 trials must be failures

Formula:

$$P(X \geq k) = (1 - p)^{k-1}$$

Substitute values:

$$P(X \geq 4) = (0.2)^3 = 0.008$$

✅ **Answer:**

$$\boxed{P(X \geq 4) = 0.008}$$

Q.2 (C) Confidence Interval for Mean (σ Known)

Given

- Population is **normally distributed**
- Known standard deviation:

$$\sigma = 20$$

General CI Formula (σ known)

$$\bar{x} \pm z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$$

(i) 95% Confidence Interval

Given

- Sample size: $n = 10$
- Sample mean: $\bar{x} = 1000$
- $z_{0.025} = 1.96$

Step 1: Standard error

$$\frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{10}} \approx 6.325$$

Step 2: Margin of error

$$1.96 \times 6.325 \approx 12.40$$

Step 3: CI

$$(1000 - 12.40,\ 1000 + 12.40)$$

✅ **95% CI:**

$$\boxed{(987.6,\ 1012.4)}$$

(ii) 99% Confidence Interval

Given

- Sample size: $n = 25$
- Sample mean: $\bar{x} = 1000$
- $z_{0.005} = 2.576$

Step 1: Standard error

$$\frac{20}{\sqrt{25}} = \frac{20}{5} = 4$$

Step 2: Margin of error

$$2.576 \times 4 = 10.304$$

Step 3: CI

$$(1000 - 10.304,\ 1000 + 10.304)$$

✅ 99% CI:

$$\boxed{(989.70,\ 1010.30)}$$

Final Answers Summary

Part	Result
Q2(B)(a)	0.0064
Q2(B)(b)	0.9984
Q2(B)(c)	0.008
95% CI	$(987.6,\ 1012.4)$
99% CI	$(989.70,\ 1010.30)$

If you want, I can:

- Show **why geometric distribution applies**
- Explain **why 99% CI is narrower here**
- Convert this into **perfect exam-ready handwriting format**

Just say 👍

Let's solve **Q.3 (A) step by step**, very cleanly, with **probability model** → **expected value** → **variance**, exactly how examiners want it.

Q.3 (A) Expected Winnings in a Card Game

Given rules

From a **standard deck of 52 cards**, you win:

- **$1** if you draw a **heart (not an ace)**
- **$5** if you draw **any ace** (including ace of hearts)
- **$10** if you draw **king of spades**
- **$0** for **any other card**

Let random variable **X = winnings (in dollars)**.

Step 1: Classify all possible outcomes correctly

We must be careful about **overlaps** (e.g., ace of hearts is both an ace and a heart).

Count cards in each category

1. **Heart (not ace)**
   - Total hearts = 13
   - Exclude ace of hearts

$$13 - 1 = 12$$

2. **Ace (all suits)**

$$4 \text{ aces}$$

3. **King of spades**

$$1 \text{ card}$$

4. **All other cards**

$$52 - (12 + 4 + 1) = 35$$

✔ No overlap remains.

Step 2: Construct the probability model

Event	X (winnings)	Number of cards	Probability
Heart (not ace)	1	12	$\frac{12}{52}$
Ace	5	4	$\frac{4}{52}$
King of spades	10	1	$\frac{1}{52}$
All else	0	35	$\frac{35}{52}$

Step 3: Expected value $E[X]$

Formula:

$$E[X] = \sum x \cdot P(X = x)$$

Substitute values:

$$E[X] = 1\left(\frac{12}{52}\right) + 5\left(\frac{4}{52}\right) + 10\left(\frac{1}{52}\right) + 0\left(\frac{35}{52}\right)$$

Simplify:

$$E[X] = \frac{12 + 20 + 10}{52}$$

$$E[X] = \frac{42}{52} = \frac{21}{26}$$

✅ **Expected winning:**

$$\boxed{E[X] = 0.8077 \text{ dollars (approx)}}$$

Step 4: Measure variation — Variance

First compute $E[X^2]$

X	X^2	Probability
1	1	$\frac{12}{52}$
5	25	$\frac{4}{52}$
10	100	$\frac{1}{52}$
0	0	$\frac{35}{52}$

$$E[X^2] = 1\left(\frac{12}{52}\right) + 25\left(\frac{4}{52}\right) + 100\left(\frac{1}{52}\right)$$

$$= \frac{12 + 100 + 100}{52} = \frac{212}{52}$$

$$E[X^2] = 4.0769$$

Step 5: Variance

Formula:

$$\text{Var}(X) = E[X^2] - (E[X])^2$$

$$= 4.0769 - (0.8077)^2$$

$$= 4.0769 - 0.6524$$

✅ **Variance:**

$$\boxed{\operatorname{Var}(X) \approx 3.4245}$$

Step 6: Standard deviation (optional but good to mention)

$$\sigma = \sqrt{3.4245} \approx 1.85$$

Final Answers (Exam-Ready)

- **Expected winning per game:**

$$\boxed{E[X] \approx \$0.81}$$

- **Variance of winnings:**

$$\boxed{\operatorname{Var}(X) \approx 3.42}$$

- **Standard deviation:**

$$\boxed{\sigma \approx 1.85}$$

Interpretation (one strong exam line)

Although the expected gain is about **$0.81 per game**, the relatively large variance shows that winnings **vary significantly** from game to game.

If you want, I can:

- Show a **tree diagram**
- Explain **why variance matters**
- Or rewrite this in **perfect handwritten-exam style**

Just say 👍

Let's answer **Q.3 (B)** exactly as expected in a **statistics exam**:
✔️ how to **construct box plots (step-by-step)**
✔️ what each plot shows
✔️ a **clear comparison/interpretation** using median, spread, skewness, and outliers.

Q.3 (B) Box Plots & Interpretation

Problem summary

- Gene expression levels are measured under:

    - **High Dose treatment**
    - **Control 1**
    - **Control 2**
    - **Control 3**
- You are asked to:
    1. **Construct a box plot for each group**
    2. **Interpret and compare the plots**
- **Mean, Median, and Standard Deviation are given** (important for interpretation).

STEP 1: What a box plot needs

A box plot is constructed using the **five-number summary**:

1. Minimum
2. First Quartile (Q1)
3. Median (Q2)
4. Third Quartile (Q3)
5. Maximum

📌 (Exact quartiles are not required here because interpretation is allowed using median, mean, spread, and range.)

STEP 2: Given summary statistics (from table)

Group	Median	Mean	Std. Dev
High Dose	45	52.65	38.60
Control 1	215.35	382.67	418.60
Control 2	355.4	460.81	475.81
Control 3	424.3	897.18	974.23

STEP 3: How to draw the box plots (exam explanation)

For **each group**:

1. Draw a **number line** (gene expression values)
2. Draw a **box** from Q1 to Q3
3. Draw a **line inside the box at the median**
4. Draw **whiskers** extending to min and max

5. Plot **outliers**, if any (large extreme values)

(You would draw **four side-by-side box plots** for comparison.)

STEP 4: Interpretation of each group

1️⃣ High Dose Group

- **Lowest median (45)** → very low gene expression
- **Small standard deviation (38.6)** → low variability
- Mean ≈ median → roughly symmetric distribution
- Few or no extreme values

📌 **Interpretation**:
High-dose treatment strongly **suppresses gene expression** and does so **consistently**.

2️⃣ Control 1

- Median = **215.35**
- Mean much larger than median → **right-skewed**
- Large standard deviation → high variability

📌 **Interpretation**:
Gene expression is higher than High Dose, but **highly variable** with some large values.

3️⃣ Control 2

- Median = **355.4**
- Mean > median → right skew
- Even larger spread than Control 1

📌 **Interpretation**:
Gene expression is **moderate to high**, with **substantial variability** and possible outliers.

4️⃣ Control 3

- **Highest median (424.3)**
- Mean ≫ median → **strong right skew**
- **Very large standard deviation (974.23)**

📌 **Interpretation**:
Gene expression is **highest and most unstable**, with **extreme outliers**.

STEP 5: Comparative interpretation (MOST IMPORTANT FOR MARKS)

Median comparison

$$\text{High Dose} < \text{Control 1} < \text{Control 2} < \text{Control 3}$$

➡️ Gene expression increases steadily from treatment to controls.

Variability comparison

$$\text{High Dose} \ll \text{Control 1} < \text{Control 2} \ll \text{Control 3}$$

➡️ Controls show **much larger variability**, especially Control 3.

Skewness

- All control groups show **right skewness**
- High Dose is relatively symmetric

FINAL EXAM-READY INTERPRETATION (WRITE THIS)

> The box plots show that the **high-dose treatment group has substantially lower gene expression levels** compared to all control groups, with **much less variability**.
> The control groups exhibit **higher medians and increasing spread**, particularly Control 3, which shows extreme variability and strong right skewness due to large outliers.
> Overall, the plots suggest that **high-dose treatment effectively reduces and stabilizes gene expression**, while untreated controls display higher and more erratic expression levels.

What the examiner is testing

- Understanding of **box plots**
- Ability to **compare distributions**
- Interpretation using **median, spread, skewness, and outliers**
- Biological insight: **treatment effect vs controls**

If you want, I can:

- Sketch an **ASCII box-plot diagram**
- Help you **write this in 6-mark answer format**
- Explain **why median is preferred over mean here**

Just say 👍

Below is a **clear, step-by-step, exam-ready answer** for **Q.3 (C)**.
This is a **4-mark theory question**, so clarity + correct interpretation matters more than long formulas.

Q.3 (C)

1️⃣ How do you know if data is skewed or not?

Method 1: Compare Mean, Median, and Mode

- **Symmetric data**:

$$\text{Mean} \approx \text{Median} \approx \text{Mode}$$

- **Right-skewed (positively skewed)**:

$$\text{Mean} > \text{Median} > \text{Mode}$$

(Long tail on the right)

- **Left-skewed (negatively skewed)**:

$$\text{Mean} < \text{Median} < \text{Mode}$$

(Long tail on the left)

Method 2: Using graphical methods

- **Histogram / box plot**
  - Longer right tail → right skew
  - Longer left tail → left skew
- In a box plot:
  - Median closer to Q1 → right skew
  - Median closer to Q3 → left skew

2️⃣ Definitions and Interpretations

Mean

Definition:
The arithmetic average of all observations.

$$\text{Mean} = \frac{\sum x}{n}$$

Interpretation:
Represents the **central value**, but is **highly affected by extreme values (outliers)**.

Median

Definition:
The **middle value** when data is arranged in ascending order.

Interpretation:
A **robust measure of central tendency**, preferred when data is **skewed**.

Mode

Definition:
The value that occurs **most frequently**.

Interpretation:
Useful for **categorical or discrete data**; may not represent center in skewed data.

Variance

Definition:
The average of the **squared deviations from the mean**.

$$\text{Variance} = \frac{\sum(x-\mu)^2}{n}$$

Interpretation:
Measures **how spread out** the data is; larger variance means greater dispersion.

Standard Deviation

Definition:
The square root of variance.

$$\sigma = \sqrt{\text{Variance}}$$

Interpretation:
Shows the **typical distance of observations from the mean** in original units.

3️⃣ Summary Table (Very Exam-Friendly)

Measure	Meaning	Key Interpretation
Mean	Average	Sensitive to outliers
Median	Middle value	Best for skewed data
Mode	Most frequent	Useful for categories

Measure	Meaning	Key Interpretation
Variance	Squared spread	Measures dispersion
Std. Deviation	Spread in same units	Typical deviation from mean

Final 2-line Conclusion (Good for Full Marks)

> Data skewness can be identified by comparing mean, median, and mode or by observing asymmetry in graphs such as box plots and histograms.
> Mean and median describe central tendency, while variance and standard deviation quantify the spread of data.

If you want:

- **One-page crisp exam answer**
- **Numerical example showing skewness**
- **Box-plot-based explanation**

Just tell me 👍