Biostatistics : End Sem Exam

Question 1: A small company has 5 employees with the following annual salaries: ₹40,000, ₹42,000,₹45,000, ₹48,000, and ₹250,000 (the CEO). The company reports its "average salary" as ₹85,000 to attract new hires. Identify the measure of central tendency used by the company. Explain why this measure might be considered misleading in this context, and suggest a more appropriate measure to represent the typical employee's salary, justifying your choice. Answer:

The company reports an average salary of ₹85,000 across 5 employees with the following annual salaries:
- ₹40,000
- ₹42,000
- ₹45,000
- ₹48,000
- ₹250,000 (CEO)

To determine the "average" salary, they used the mean, calculated as:
Mean = (40,000 + 42,000 + 45,000 + 48,000 + 250,000) / 5 = ₹85,000
Why this is misleading:
The mean is sensitive to extreme values (outliers). In this case, the CEO's very high salary (₹250,000) skews the average upward, making it seem like all employees earn more than they actually do.
A better alternative: Median
Sorted salaries: ₹40,000, ₹42,000, ₹45,000, ₹48,000, ₹250,000
Median (middle value) = ₹45,000
The median is not affected by outliers and better represents the typical salary.

Question 2: You are presented with summary statistics for the daily percentage change in stock prices for two different companies (A and B) over the past year: Company A: Mean Daily Change = +0.05%, Standard Deviation = 1.5%. Company B: Mean Daily Change = +0.05%, Standard Deviation = 1.5%. Based only on this information, an analyst concludes that both stocks have identical volatility and are therefore equally risky in terms of price fluctuation. Explain why the analyst's conclusion might be premature or potentially incorrect, even though the means and standard deviations are identical. Describe at least one specific scenario or characteristic the underlying distribution of daily changes for Company B could have (that Company A might lack) which would make Company B significantly more risky or unpredictable, despite having the same standard deviation. What other measure of dispersion could potentially reveal this difference if you had access to the raw data?

Given:

- Company A: μ=+0.05%\mu = +0.05% μ=+0.05%, σ=1.5% sigma = 1.5% σ=1.5% ●

  Company B: μ=+0.05%\mu = +0.05% μ=+0.05%, σ=1.5% sigma = 1.5% σ=1.5%

Two companies have the same mean (+0.05%) and standard deviation (1.5%) in daily stock price changes. An analyst claims they are equally risky.

Why this is flawed:

Mean and standard deviation alone don't fully describe the distribution. One company might have fat tails (more extreme events), bimodal behavior, or skewness. These influence investment risk**.**

Why mean & SD can mask risk:

- **Shape matters:** Two distributions with identical μ\muμ and σ\sigmaσ can differ in
    - **Skewness** (asymmetry)
    - **Kurtosis** (tail "fatness")
    - **Multimodality** (multiple peaks)
- **Illustrative scenario for Company B:**
    - **Fat-tailed distribution**: Most days ±1%, but occasionally ±10% moves rare "crash" days.
    - **Bimodal clusters**: Half the days ~+3%, half ~–3%; while SD remains ~1.5%, investors face risk.

Alternative dispersion measures:

1. **Kurtosis:**
   Kurtosis
   High kurtosis → more frequent extreme returns.
2. **Value at Risk (VaR)** / **Conditional VaR (CVaR):**
   Estimate the worst expected loss over a given horizon at a given confidence level (e.g., 95%).
3. **Interquartile Range (IQR):**
   Q3–Q1, robust to extreme values and highlights middle 50% variability.
4. **Empirical distribution plot / histogram** to visually compare tails and modality.

Question 3: In a study examining the link between weekly hours studied (X) and final exam scores (Y, 0-100), researchers found a correlation r=0.70, a coefficient of determination r2=0.49, and derived the regression equation Y=45+3X. First, interpret the slope and intercept within the specific context of hours studied and exam scores, commenting on whether the intercept

represents a practically meaningful scenario. Then, explain what the r2 value signifies about the relationship. Finally, critically evaluate the assertion that the strong correlation (r=0.70) means the regression equation will accurately predict individual student scores. Provide two distinct reasons, grounded in the principles of correlation and regression, why such a conclusion about individual predictive accuracy might be flawed.

Results:

r = 0.70

$r^2 = 0.49$

Equation: Y = 45 + 3X

Interpretation:
Slope (3): Each extra hour studied predicts a 3-point increase in score.For each additional hour of weekly study, the predicted score increases by 3 points. Intercept (45): Predicted score if no hours are studied (may not be realistic).Most students study >0 hours; thus the intercept is a mathematical artifact.

r^2 = 0.49: 49% of the variance in scores is explained by study time 49% of the total variance in scores is explained by hours studied. The remaining 51% arises from other factors (e.g., innate aptitude, teaching quality, exam difficulty, stress).

r = 0.70 doesn't ensure precise individual predictions:

Residual Standard Error (RSE): Even if r2 is moderate, the standard deviation of residuals may still be large relative to the score range, causing wide prediction intervals.

Question 4: Imagine a researcher is planning an experiment and conducts an a priori power analysis. The analysis reveals that achieving their desired statistical power of 0.80, based on their chosen alpha level (e.g., 0.05) and a plausible estimated effect size, would necessitate recruiting 500 participants. Unfortunately, due to significant resource constraints (time and funding), the researcher can realistically only recruit a maximum of 150 participants. Describe two distinct strategies or adjustments the researcher could consider during this planning phase to modify their study design or expectations, allowing them to proceed with the feasible sample size of 150. For each strategy, clearly explain its direct consequence on the key elements of power analysis (such as the actual power level, the minimum detectable effect size, or the alpha level) and briefly discuss the main conceptual limitation or practical risk associated with implementing that adjustment. (5 marks)
Original design:

- Desired power $(1-\beta) = 0.80$
- $\alpha = 0.05$

- Required n=500
- Feasible n=150

Strategy 1: Increase α (Type I error rate)

- Set α = 0.10
- **Effect on power:** Approximate power rises (e.g., from 0.80→0.90) for same effect size and nnn.
- **Trade-off:** Doubling the false positive rate (10% vs. 5%) increases the chance of declaring a nonexistent effect.
- **Risk:** Erodes confidence in findings harder to publish or gain stakeholder buy-in.

Strategy 2: Reduce desired detectable effect size (accept only larger effects)

- Suppose the original target was Cohen's d=0.3; now aim for d=0.5d
- **Effect on power:** With n=150, you might retain 80% power to detect d=0.5d, but will **lack power** to detect subtler differences.
- **Trade-off:** Smaller, yet important effects become **undetectable**, potentially missing real-world signals.

Other options (briefly):

- **Repeated measures / within-subjects:** By measuring each subject multiple times, you reduce error variance and require fewer participants.
    - **Consequence:** Each person serves as their own control, boosting power. ○

      Limitation: Potential **carry-over** or **learning effects** between measurements.
- **Covariate adjustment:** Include strong baseline predictors (ANCOVA) to explain some outcome variance, effectively increasing power.
    - **Consequence:** Smaller residual variance → higher power at same nnn.
    - **Risk:** Model misspecification or measurement error in covariates can bias results.

Question 5: A new diagnostic screening test for a relatively rare genetic disorder (prevalence ≈ 0.5% in the population) boasts high accuracy, with a sensitivity of 98% and a specificity of 96%. If an individual from the general population receives a positive result from this screening test, explain using probabilistic reasoning why their actual probability of having the disorder is considerably lower than either 98% or 96%. Identify the key concept from probability, heavily influenced by the disorder's prevalence, that determines this positive predictive value, and briefly state why understanding this is critical when interpreting results from screening tests applied to general populations versus high-risk groups.

Answer: The disorder has 0.5% prevalence. Sensitivity = 98%, Specificity = 96%.

Why a positive test doesn't mean 98% chance of disease:

In a general population, even high-accuracy tests produce many false positives when the disease is rare. Positive Predictive Value (PPV) is low.

Bayes' Theorem helps compute the true probability. With low prevalence, PPV might be only 10–20%.

Positive Predictive Value depends heavily on prevalence. In low-prevalence settings, even tests with high sensitivity/specificity yield low PPV. Conversely, in a high-risk subgroup (e.g., family history), prevalence might be 10%, boosting PPV dramatically.

- **AUC-ROC** = probability a randomly chosen positive (death) case scores higher than a randomly chosen negative (survival) case.
    - Model A: AUC = 0.82
    - Model B: AUC = 0.88

Although B has a higher overall discrimination, **decision thresholds** matter.

Scenario favoring Model A:

- Suppose clinicians must guarantee **≥95% sensitivity** (catch nearly all high-risk patients).
- At that threshold, Model A might achieve sensitivity = 95% with specificity = 50%, while Model B might only reach 92% sensitivity even at a much lower threshold.
- Despite a lower overall AUC, Model A's **partial AUC** in the high-sensitivity region is superior for this use case.

Question 6: Consider two machine learning models developed to predict patient mortality risk

based on clinical data. Model A achieves an AUC-ROC of 0.82, while Model B achieves an AUC-ROC of 0.88. Explain what the AUC-ROC value fundamentally represents regarding a model's ability to distinguish between patients who will survive versus those who will not. While Model B has a higher AUC, describe a specific scenario related to the shape of the ROC curves or the clinical application's priorities (e.g., extreme need for high sensitivity even at the cost of specificity) where Model A might still be preferred over Model B, despite its lower overall AUC value. (5 marks)

Answer:

AUC-ROC : probability a randomly chosen positive (death) case scores higher than a randomly chosen negative (survival) case.

Model A: AUC = 0.82

Model B: AUC = 0.88

Although B has a higher overall discrimination, decision thresholds matter.

Scenario favoring Model A:

- Suppose clinicians must guarantee **≥95% sensitivity** (catch nearly all high-risk patients).
- At that threshold, Model A might achieve sensitivity = 95% with specificity = 50%, while Model B might only reach 92% sensitivity even at a much lower threshold.
- Despite a lower overall AUC, Model A's **partial AUC** in the high-sensitivity region is superior for this use case.

Question 7: In analyzing time-to-relapse data from a cancer therapy trial, researchers encounter 'right-censored' observations for participants who completed the study period without relapse or were lost to follow-up. Explain the fundamental statistical challenge introduced by these censored data points. Why would simply excluding these participants from the analysis, or alternatively, treating their censoring time as if it were a relapse time, lead to a significantly biased and likely overly pessimistic estimation of the true relapse-free survival probability curve for the therapy? Briefly mention the type of statistical method designed specifically to handle such censored time-to-event data correctly. (5 marks)

Answer

Right Censoring:
When a participant's event time is only known to exceed a certain point, e.g., no relapse by study end or loss to follow-up.

Why naïve methods fail:

1. Excluding censored observations biases toward those who relapsed early: survival curve underestimates true survival probability.

2. Counting censor times as events inflates relapse rates and again underestimates survival.

Correct approach:

- Kaplan–Meier estimator computes stepwise survival probabilities, adjusting the denominator at each event time to exclude censored subjects only after their last known time.

- Cox Proportional Hazards model further allows inclusion of covariates while properly accounting for censored data.

MCQs:
Q1: (C) Positive assortative mating : individuals with similar genotypes mate more often, reducing heterozygosity.

Q2: (C) Hazard Ratio of 1.75 means at any given time, patients on Treatment X have 1.75 times the instantaneous risk of death compared to control.

Q3: (C) Geometric Mean correct for compound growth rates.

Q4: (C) 50% of data in a box plot lies between Q1 and Q3.

Q5: (D) Wilcoxon Signed-Rank test: best for comparing paired, ordinal data.