DSA 507: Statistics for Data Science and AI

Answer all questions. Each sub-question carries equal marks. Use clear reasoning and, where relevant, refer to the lecture content: https://usjstat.github.io/courses/dsa507/

1. **Structured vs Unstructured Data**

   (a) (2 points) Define structured data and provide an example.

   > **Solution:** Structured data refers to data that is organized in a fixed format such as rows and columns, often stored in tables or data frames.
   > Example: The ‘cricket_data‘ data frame with players, runs, strike rate, and captaincy status.

   (b) (2 points) List two main types of structured data along with one example for each.

   > **Solution:** - Numeric: e.g., Number of runs scored (continuous or discrete).
   > - Categorical: e.g., Player role such as ”Bowler”, ”Batsman” (nominal), or education level (ordinal).

   (c) (3 points) Explain the significance of converting unstructured data into a structured form for statistical analysis.

   > **Solution:** Converting unstructured data into structured form enables the application of statistical methods such as summary statistics, modeling, and visualization. Structured formats are necessary for tools like R or Python to perform EDA.

2. **Estimates of Location and Variability**

   (a) (2 points) Explain the difference between mean, median, and trimmed mean using your own words.

   > **Solution:** - Mean: Arithmetic average of all values.
   > - Median: Middle value when data is sorted.
   > - Trimmed Mean: Mean calculated after removing extreme values from both ends.

   (b) (2 points) Why is the median considered a robust estimator?

   > **Solution:** The median is robust because it is not affected by outliers or extreme values. It only depends on the order of data.

(c) (2 points) Explain what the interquartile range (IQR) measures.

Solution: IQR measures the range of the middle 50% of the data. It is the difference between the 75th and 25th percentiles.

(d) (2 points) Which of the following is most affected by an outlier: mean, median, mode?

Solution: Correct answer: Mean

3. **Visualizing Distributions**

(a) (3 points) Describe what each of the following visualizations shows: histogram, density plot, boxplot.

Solution: - Histogram: Shows frequency distribution across bins.
- Density Plot: Smoothed curve showing distribution shape.
- Boxplot: Summarizes median, quartiles, and highlights outliers.

(b) (3 points) In a boxplot, what do the "whiskers" represent?

Solution: The whiskers extend to the smallest and largest values within 1.5 * IQR from the lower and upper quartiles.

(c) (2 points) Explain how outliers are identified in a boxplot.

Solution: Outliers are identified as data points that fall outside the whiskers (i.e., beyond 1.5 * IQR from Q1 or Q3).

4. **Exploring Categorical Data**

(a) (2 points) Differentiate between nominal, ordinal, and binary categorical data with one example each.

Solution: - Nominal: Categories with no order (e.g., team name).
- Ordinal: Categories with a clear order (e.g., education level).
- Binary: Two categories (e.g., captaincy status: Yes/No).

(b) (2 points) Define expected value in the context of binary outcomes.

Solution: Expected value is the long-run average outcome. For binary variables (e.g., Win=1, Loss=0), it's the proportion of Wins.

(c) (2 points) Explain why bar plots are preferred over pie charts in professional reporting.

> **Solution:** Bar plots are easier to interpret, more accurate for comparison, and suitable for publications. Pie charts are harder to read and interpret.

5. **Correlation and Multivariate Exploration**

   (a) (2 points) What does a correlation coefficient of 0.95 imply about the relationship between two variables?

   > **Solution:** A correlation coefficient of 0.95 indicates a very strong positive linear relationship — as one variable increases, the other also increases.

   (b) (2 points) Match each visualization with its primary use:

A. Violin Plot	1. Handling overplotting for dense scatterplots
B. Contour Plot	2. Distribution shape + central tendency
C. Hexbin Plot	3. Topographical density view

   Your answers: A → _____, B → _____, C → _____

   > **Solution: Correct matching:**
   > A → 2,
   > B → 3,
   > C → 1

6. **Population, Sample, and Sampling Methods**

   (a) (2 points) Define population and sample using a sports example.

   > **Solution:** Population: All matches played in a football league season.
   > Sample: 50 randomly selected matches from that season.

   (b) (2 points) Explain the importance of random sampling in data science.

   > **Solution:** Random sampling ensures each unit has an equal chance of being selected, reducing bias and improving representativeness.

   (c) (3 points) Describe the difference between sampling with and without replacement.

   > **Solution:** With replacement: a unit can be selected multiple times.
   > Without replacement: once selected, a unit cannot be selected again.

   (d) (3 points) What is sample bias? Give one historical example (mentioned in the notes).

Solution: Sample bias occurs when the sample misrepresents the population. Example: Literary Digest's 1936 election poll using car and phone owners.

7. **Selection Bias and Regression to the Mean**

   (a) (2 points) Define selection bias and give one way to avoid it.

   Solution: Selection bias arises when the sample does not represent the population due to how it's selected.
   Avoid it by using random selection and holdout validation.

   (b) (2 points) Explain the concept of regression to the mean.

   Solution: Extreme observations tend to be followed by more average ones due to random variation.

8. **Sampling Distributions and Central Limit Theorem (CLT)**

   (a) (2 points) What is a sampling distribution?

   Solution: The distribution of a statistic (e.g., mean) over many repeated samples from a population.

   (b) (2 points) What happens to the shape of the sampling distribution of the mean as sample size increases?

   Solution: It becomes more bell-shaped and less variable — closer to a normal distribution.

   (c) (3 points) State the Central Limit Theorem.

   Solution: The sampling distribution of the sample mean approaches a normal distribution as sample size increases, regardless of the population's shape (provided it's not too skewed).

   (d) (3 points) Define standard error and explain its relation to sample size.

   Solution: Standard error is the standard deviation of the sampling distribution. It decreases as sample size increases.

9. **Bootstrap and Confidence Intervals**

   (a) (2 points) What is the bootstrap method used for?

Solution: Estimating the sampling distribution of a statistic by resampling with replacement from the observed data.

(b) (2 points) Outline the steps of bootstrapping a mean.

Solution: (1) Resample data with replacement, (2) compute mean, (3) repeat many times, (4) analyze distribution of means.

(c) (3 points) How do you construct a 90% bootstrap confidence interval for the mean?

Solution: Take 1000+ bootstrap samples, compute the mean for each, then take the 5th and 95th percentiles as interval bounds.

(d) (3 points) What does a 90% confidence interval imply?

Solution: If we repeated the sampling process many times, about 90% of the constructed intervals would contain the true population mean.

10. **Normal Distribution and Z-Scores**

(a) (2 points) State two properties of the normal distribution.

Solution: It is symmetric and bell-shaped; defined by mean μ and standard deviation σ.

(b) (2 points) What is the empirical rule? List the percentages for $\pm 1\sigma$, $\pm 2\sigma$, and $\pm 3\sigma$.

Solution: 68% within $\pm 1\sigma$, 95% within $\pm 2\sigma$, 99.7% within $\pm 3\sigma$.

(c) (2 points) Convert a height of 220 cm to a z-score given $\mu = 200$, $\sigma = 10$.

Solution: $z = \frac{220-200}{10} = 2$

(d) (2 points) Does standardizing a dataset make it normally distributed? Briefly explain.

Solution: No. Standardization only rescales the data; it does not change the underlying distribution shape.

11. **Assessing Normality and Central Limit Theorem in Practice**

(a) (2 points) What does a Q-Q plot show? How do you interpret it?

> **Solution:** A Q-Q plot compares sample quantiles to theoretical normal quantiles. Points on a straight diagonal line suggest normality.

(b) (2 points) How can a histogram of sample means provide evidence for the Central Limit Theorem?

> **Solution:** Even if the original data is skewed, the histogram of sample means tends toward a normal shape as sample size increases.

(c) (3 points) Why does the sampling distribution of the mean tend to be more bell-shaped than the original data distribution?

> **Solution:** Because averaging reduces variability and extreme values, leading to a more symmetric and normal-like shape (CLT).

(d) (3 points) Suppose a right-skewed dataset has a mean of 30 and SD of 10. What is the expected shape of the sampling distribution of the mean for samples of size 50?

> **Solution:** Approximately normal-shaped due to the Central Limit Theorem.

12. **Student's t-Distribution**

(a) (2 points) Write down the formula for the t-statistic.

> **Solution:** $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$

(b) (2 points) When should we use the t-distribution instead of the normal distribution?

> **Solution:** When the sample size is small and population standard deviation is unknown.

(c) (2 points) What happens to the t-distribution as degrees of freedom increase?

> **Solution:** It approaches the standard normal distribution.

(d) (2 points) Why does the t-distribution have heavier tails than the normal distribution?

> **Solution:** To account for the additional uncertainty when estimating the standard deviation from small samples.

13. **Discrete Probability Distributions**

(a) (2 points) Define the binomial distribution and provide its mean and variance.

> **Solution:** Models number of successes in n trials with probability π. Mean = $n\pi$, Variance = $n\pi(1-\pi)$

(b) (2 points) What does the chi-square distribution test?

> **Solution:** It tests whether observed counts significantly differ from expected counts under a null hypothesis.

(c) (2 points) State the formula for the chi-square statistic.

> **Solution:** $\chi^2 = \sum \frac{(O-E)^2}{E}$

(d) (2 points) Define the F-distribution and state one common application.

> **Solution:** The ratio of two variances; used in ANOVA to compare group means.

14. **Modeling Events Over Time or Space**

(a) (2 points) When is the Poisson distribution appropriate?

> **Solution:** When modeling the number of events in a fixed time or space interval, assuming constant rate and independence.

(b) (2 points) What is the relationship between the Poisson and exponential distributions?

> **Solution:** Poisson models count of events; exponential models time between those events.

(c) (2 points) What parameter of the Weibull distribution allows for increasing or decreasing failure rates?

> **Solution:** The shape parameter β.

(d) (2 points) State two practical applications of the Weibull distribution.

> **Solution:** Reliability engineering and failure analysis.

15. **A/B Testing and Experimental Design**

(a) (2 points) Define A/B testing and explain its relevance in data science with an example.

Solution: A/B testing compares two treatments or strategies to determine which performs better. In sports analytics, it helps evaluate training programs, equipment, or tactics.

(b) (2 points) Explain the importance of randomization in experimental design.

Solution: Randomization eliminates selection bias and ensures that differences in outcomes are due to the treatment and not pre-existing differences.

(c) (2 points) Why is a control group necessary in a statistical experiment?

Solution: A control group serves as a baseline, allowing researchers to isolate the treatment effect by comparison.

(d) (2 points) What is the difference between single-blind and double-blind experiments?

Solution: In single-blind, participants do not know which treatment they receive; in double-blind, both participants and researchers are unaware.

(e) (2 points) Describe one ethical consideration when conducting experiments involving human subjects.

Solution: Informed consent is critical—participants must know and agree to the nature of the experiment, especially if it can affect their performance or well-being.

16. **Hypothesis Testing and p-Values**

(a) (2 points) Define null hypothesis and alternative hypothesis with an example.

Solution: Null: No difference exists (e.g., Strategy A = Strategy B). Alternative: A difference exists (e.g., Strategy A ¿ Strategy B).

(b) (2 points) What is a p-value?

Solution: The probability of observing a result as extreme as the actual one, assuming the null hypothesis is true.

(c) (2 points) What does it mean if a result is statistically significant at $\alpha = 0.05$?

Solution: It means the p-value is less than 0.05, suggesting the observed result is unlikely due to chance alone.

(d) (2 points) Explain the difference between a one-tailed and two-tailed test.

Solution: A one-tailed test checks for an effect in one direction only; a two-tailed test checks for any difference regardless of direction.

(e) (2 points) What is a Type I error? What is a Type II error?

Solution: Type I: Rejecting a true null hypothesis (false positive). Type II: Failing to reject a false null hypothesis (false negative).

17. **Permutation Tests and Resampling**

(a) (2 points) Describe the purpose of a permutation test.

Solution: To test the null hypothesis by evaluating how likely the observed statistic is under all possible rearrangements of group labels.

(b) (2 points) What is the difference between bootstrapping and permutation testing?

Solution: Bootstrapping estimates variability; permutation tests assess significance by reshuffling data.

(c) (2 points) In what type of scenario is a permutation test especially useful?

Solution: When the distributional assumptions (e.g., normality) of parametric tests are not met.

(d) (2 points) How is the permutation distribution generated?

Solution: By repeatedly shuffling the data labels, reassigning to groups, and computing the test statistic each time.

(e) (2 points) What does it mean if the observed difference lies in the extreme tail of the permutation distribution?

Solution: It suggests the result is unlikely due to chance, and the treatment effect may be significant.

18. **Multiple Testing and Adjustments**

(a) (2 points) What is alpha inflation, and why does it occur during multiple hypothesis testing?

> **Solution:** Alpha inflation refers to the increased probability of at least one Type I error when performing multiple tests. It occurs because each test carries its own chance of a false positive.

(b) (2 points) Calculate the probability of making at least one Type I error when performing 10 independent tests at $\alpha = 0.05$.

> **Solution:** $1 - (1 - 0.05)^{10} \approx 0.4013$

(c) (2 points) Explain how the Bonferroni adjustment modifies the significance level.

> **Solution:** It divides the overall alpha by the number of tests, reducing the significance threshold for each individual test.

(d) (2 points) What is the primary advantage of the False Discovery Rate (FDR) approach over Bonferroni in exploratory research?

> **Solution:** FDR balances false positives with discovery, allowing more findings in exploratory contexts without overly penalizing multiple tests.

(e) (2 points) Write the formula for Tukey's HSD and identify what each symbol means.

> **Solution:** $HSD = q_{\alpha,k,df_{error}} \times \sqrt{\frac{MS_{error}}{n}}$, where q is the studentized range statistic, k number of groups, df_{error} error degrees of freedom, MS_{error} mean square error, and n observations per group.

19. **One-Way ANOVA and the F-Statistic**

(a) (2 points) Define the decomposition of variance in the context of ANOVA.

> **Solution:** It breaks an observation into grand mean, treatment effect, and residual: $Y_{ij} = \bar{Y}_{\cdot\cdot} + \tau_j + \epsilon_{ij}$.

(b) (2 points) Write the formula for the F-statistic in One-Way ANOVA.

> **Solution:** $F = \frac{MSB}{MSW} = \frac{SSB/(k-1)}{SSW/(N-k)}$

(c) (2 points) What does a large F value indicate in ANOVA?

Solution: That between-group variability is large relative to within-group variability, suggesting group means differ significantly.

(d) (2 points) When do we reject the null hypothesis in ANOVA?

Solution: When the p-value associated with the F-statistic is less than the significance level α, often 0.05.

20. **Post-Hoc Testing and Tukey's HSD**

(a) (2 points) What is the purpose of a post-hoc test in ANOVA?

Solution: To determine which specific group means differ after finding a significant ANOVA result.

(b) (2 points) Write the formula for Tukey's HSD.

Solution: $HSD = q_{\alpha,k,df_{error}} \times \sqrt{\frac{MS_{error}}{n}}$

(c) (2 points) When is Tukey's HSD preferred over pairwise t-tests?

Solution: When multiple comparisons are needed after ANOVA, to control the family-wise error rate.

(d) (2 points) What does it mean if the confidence interval from Tukey's test includes zero?

Solution: The two groups being compared are not significantly different.

(e) (2 points) Why do we adjust for multiple comparisons?

Solution: To reduce the risk of Type I errors due to multiple hypothesis tests.

21. **Two-Way ANOVA and Interaction Effects**

(a) (2 points) State the full model equation for a Two-Way ANOVA with interaction.

Solution: $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$

(b) (2 points) What does a significant interaction term in Two-Way ANOVA suggest?

Solution: That the effect of one factor depends on the level of the other factor.

(c) (2 points) What are the null hypotheses tested in a Two-Way ANOVA?

Solution: $H_{0A} : \alpha_i = 0, \quad H_{0B} : \beta_j = 0, \quad H_{0AB} : (\alpha\beta)_{ij} = 0$

(d) (2 points) Why is it important to check interaction before interpreting main effects?

Solution: Because a significant interaction invalidates interpretation of main effects independently.

22. **ANOVA Assumptions and Diagnostics**

(a) (2 points) List the key assumptions of ANOVA.

Solution: Independence, normality of residuals, and homogeneity of variances.

(b) (2 points) What is the purpose of the Q-Q plot in ANOVA diagnostics?

Solution: To assess whether residuals are approximately normally distributed.

(c) (2 points) What does a Residuals vs Fitted plot help assess?

Solution: Homoscedasticity—whether residual variance is constant across fitted values.

(d) (2 points) What can you conclude if the residual plot shows a funnel shape?

Solution: Heteroscedasticity—variance increases or decreases with fitted values, violating ANOVA assumptions.

(e) (2 points) What action can be taken if ANOVA assumptions are violated?

Solution: Use transformations, robust methods, or non-parametric alternatives.

23. **Chi-Square Test of Independence**

(a) (2 points) Define the chi-square statistic and state its formula.

Solution: $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

(b) (2 points) What are the degrees of freedom in a chi-square test for an $r \times c$ table?

Solution: $(r-1)(c-1)$

(c) (2 points) In which situation should you use Fisher's Exact Test instead of a chi-square test?

Solution: When expected cell counts are very small (typically less than 5), especially in 2x2 tables.

(d) (2 points) What does a large chi-square statistic suggest?

Solution: That observed counts differ significantly from expected counts, suggesting dependence between variables.

24. **Expected Counts and Pearson Residuals**

(a) (2 points) Explain how expected counts are computed in a chi-square test.

Solution: $E_{ij} = \frac{(\text{row total})_i \times (\text{column total})_j}{\text{grand total}}$

(b) (2 points) Define a Pearson residual and provide its formula.

Solution: $R_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$

(c) (2 points) What do large absolute values of Pearson residuals indicate?

Solution: Cells that contribute significantly to the chi-square statistic.

(d) (2 points) Why is it important to check residuals even when p-value is significant?

Solution: To understand which specific combinations of categories deviate from the null model.

(e) (2 points) Calculate the expected count for Player A and "Made" given 3 players with 100 attempts each, and total made = 240.

Solution: Expected $= \frac{100 \times 240}{300} = 80$

25. **Chi-Square Test: Resampling vs. Theoretical**

(a) (2 points) What is the main idea of the permutation (resampling) test?

Solution: Randomly shuffle labels to generate a null distribution for the test statistic.

(b) (2 points) What does the p-value represent in a permutation test?

Solution: Proportion of permuted test statistics that are at least as extreme as the observed statistic.

(c) (2 points) Why might permutation p-values differ from theoretical p-values?

Solution: Permutation is data-driven and does not assume large-sample approximations.

(d) (2 points) When are theoretical chi-square tests reliable?

Solution: When all expected cell counts are at least 5.

26. **Multi-Arm Bandit(MAB) Algorithms**

(a) (2 points) What is the exploration vs. exploitation tradeoff in the context of the MAB problem?

Solution: Exploration involves trying different options to learn about their success rates; exploitation involves choosing the best-known option to maximize success.

(b) (2 points) What role does the epsilon (ϵ) parameter play in the epsilon-greedy algorithm?

Solution: It controls the probability of exploring a random option. With probability ϵ, explore; with $1 - \epsilon$, exploit the best-known option.

(c) (2 points) Define a 'win' in the context of the T20 batting pair MAB simulation.

Solution: A win is defined as the batting pair scoring at least 30 runs in the Powerplay (first 6 overs).

(d) (2 points) In Thompson Sampling, how is the posterior of Beta distribution for each arm updated after each trial?

Solution: The α parameter of Beta prior is increased by 1 for a success; the β parameter of Beta prior is increased by 1 for a failure.

27. **Power and Sample Size Analysis**

(a) (2 points) Define statistical power and its importance in hypothesis testing.

> **Solution:** Power is the probability of correctly rejecting the null hypothesis when it is false. High power reduces the risk of Type II errors.

(b) (2 points) Explain what happens to power when sample size increases, assuming fixed effect size and significance level.

> **Solution:** Power increases as sample size increases, because the test becomes more sensitive to detecting the effect.

28. **Simulation-Based Power Estimation**

(a) (2 points) Describe the basic steps of estimating power through simulation.

> **Solution:** Generate samples from two groups, run hypothesis tests repeatedly, and compute the proportion of significant p-values.

(b) (2 points) What is the advantage of using simulation-based methods over analytical methods for power analysis?

> **Solution:** They allow estimation without strict distributional assumptions and can handle complex or real-world data scenarios.

(c) (2 points) Explain why permutation tests are useful in power estimation.

> **Solution:** They provide a non-parametric way to generate null distributions and assess significance without normality assumptions.

(d) (2 points) What does a power curve illustrate?

> **Solution:** How statistical power changes as a function of sample size for a fixed effect size and significance level.