# Guided Notes with Sample Problems for APMA 3120

By Sankalpa Banjade nxr8dq

Fall 2024

## Contents

**1 Week 1 (Aug 27 - Aug 30): Introduction and Probability Review 3**
1.1 Course Introduction . . . . . . . . . 3
1.2 Probability Basics . . . . . . . . . 3
1.3 Discrete and Continuous Random Variables . . . . . . . . . 3
1.4 The Gaussian (Normal) Distribution . . . . . . . . . 3
1.5 Central Limit Theorem (CLT) . . . . . . . . . 4
1.6 Sample Problem 1: Probability of Rolling a Die . . . . . . . . . 4
1.7 Sample Problem 2: Probability of Getting Two Heads in Three Coin Tosses 4

**2 Week 2 (Sep 3 - Sep 6): Principles of Estimation 5**
2.1 Point Estimation . . . . . . . . . 5
2.2 Properties of Estimators . . . . . . . . . 5
2.3 Sufficient Statistics . . . . . . . . . 5
2.4 Sample Problem 1: Maximum Likelihood Estimation (MLE) for the Mean of a Normal Distribution . . . . . . . . . 5

**3 Week 3 (Sep 7 - Sep 13): Confidence Intervals 7**
3.1 Confidence Interval Basics . . . . . . . . . 7
3.2 Confidence Intervals for the Mean . . . . . . . . . 7
3.3 Confidence Intervals for Proportions . . . . . . . . . 7
3.4 Confidence Intervals for Variance . . . . . . . . . 7
3.5 Sample Problem 1: Confidence Interval for the Mean (Known Variance) . 7
3.6 Sample Problem 2: Confidence Interval for Proportion . . . . . . . . . 8

**4 Week 4 (Sep 14 - Sep 20): Hypothesis Testing for One Population 9**
4.1 Hypothesis Testing Basics . . . . . . . . . 9
4.2 Test Statistics . . . . . . . . . 9
4.3 Type I and Type II Errors . . . . . . . . . 9
4.4 p-Value . . . . . . . . . 9
4.5 Hypothesis Testing Steps . . . . . . . . . 9
4.6 Sample Problem 1: Hypothesis Test for the Mean (Known Variance) . . . 10

**5 Week 5 (Sep 23 - Sep 27): Hypothesis Tests for Two Populations 11**
5.1 Comparing Two Means . . . . . . . . . . 11
5.2 Comparing Two Proportions . . . . . . . . . . 11
5.3 Comparing Two Variances . . . . . . . . . . 11
5.4 Sample Problem 1: Two-Sample t-Test (Equal Variances) . . . . . . . . . . 11

**6 Week 6 (Sep 30 - Oct 4): Two Population Tests & Other Topics 13**
6.1 Nonparametric Tests . . . . . . . . . . 13
6.2 Analysis of Variance (ANOVA) . . . . . . . . . . 13
6.3 Multiple Comparisons . . . . . . . . . . 13
6.4 Sample Problem 1: Mann-Whitney U Test . . . . . . . . . . 13

**7 Week 7 (Oct 7 - Oct 11): Categorical Data Tests 15**
7.1 Chi-Square Tests . . . . . . . . . . 15
7.2 Conditions for Chi-Square Tests . . . . . . . . . . 15
7.3 Fisher's Exact Test . . . . . . . . . . 15
7.4 Sample Problem 1: Chi-Square Test for Goodness-of-Fit . . . . . . . . . . 15

# 1 Week 1 (Aug 27 - Aug 30): Introduction and Probability Review

## 1.1 Course Introduction

- **Course Overview**: Understanding statistical methods for engineering and applied science.
- **Expectations**: Familiarize yourself with the course syllabus and homework submission policies.

## 1.2 Probability Basics

- **Sample Space (S)**: The set of all possible outcomes.
- **Events**: Subsets of the sample space.
- **Axioms of Probability**:
  1. $0 \leq P(A) \leq 1$
  2. $P(S) = 1$
  3. For mutually exclusive events $A$ and $B$, $P(A \cup B) = P(A) + P(B)$.

## 1.3 Discrete and Continuous Random Variables

- **Discrete Random Variables**
  - **Probability Mass Function (PMF)**: $P(X = x)$.
  - **Binomial Distribution**: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$.
- **Continuous Random Variables**
  - **Probability Density Function (PDF)**: $f(x)$.
  - **Cumulative Distribution Function (CDF)**: $F(x) = P(X \leq x)$.

## 1.4 The Gaussian (Normal) Distribution

- **Properties**:
  - Symmetrical about the mean $\mu$.
  - Defined by mean $\mu$ and variance $\sigma^2$.
- **Standard Normal Distribution**:
  - Mean $\mu = 0$, Variance $\sigma^2 = 1$.
  - Use Z-scores to standardize.

## 1.5 Central Limit Theorem (CLT)

- **Statement**: The sum or average of a large number of independent, identically distributed random variables tends toward a normal distribution, regardless of the original distribution.
- **Implication**: Enables approximation of sample means using the normal distribution for large sample sizes.

## 1.6 Sample Problem 1: Probability of Rolling a Die

**Problem**: What is the probability of rolling a 3 or a 6 on a fair six-sided die?
**Solution**:

- The sample space for rolling a die is: $S = \{1, 2, 3, 4, 5, 6\}$.
- The event of interest is rolling a 3 or a 6, so $A = \{3, 6\}$.
- Since the die is fair, each outcome has a probability of $\frac{1}{6}$.
- The probability of rolling a 3 or a 6 is:

$$P(A) = P(3) + P(6) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}.$$

## 1.7 Sample Problem 2: Probability of Getting Two Heads in Three Coin Tosses

**Problem**: A fair coin is tossed three times. What is the probability of getting exactly two heads?
**Solution**:

- The sample space consists of 8 outcomes: $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TT$
- We are interested in the event of getting exactly two heads: $A = \{HHT, HTH, THH\}$.
- There are 3 favorable outcomes, and each outcome has a probability of $\frac{1}{8}$.
- The probability of getting exactly two heads is:

$$P(A) = \frac{3}{8}.$$

# 2 Week 2 (Sep 3 - Sep 6): Principles of Estimation

## 2.1 Point Estimation

- **Definition**: A single value estimate of a population parameter.
- **Methods**:
  - **Method of Moments**: Equate sample moments to population moments.
  - **Maximum Likelihood Estimation (MLE)**: Choose parameters that maximize the likelihood function.

## 2.2 Properties of Estimators

- **Unbiasedness**: $E[\hat{\theta}] = \theta$.
- **Consistency**: $\hat{\theta}$ converges in probability to $\theta$ as $n \to \infty$.
- **Efficiency**: Estimator has the smallest variance among all unbiased estimators.

## 2.3 Sufficient Statistics

- **Definition**: A statistic that captures all the information about a parameter contained in the data.
- **Factorization Theorem**: Identifies sufficient statistics through factorization of the likelihood function.

## 2.4 Sample Problem 1: Maximum Likelihood Estimation (MLE) for the Mean of a Normal Distribution

**Problem**: Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. random variables from a normal distribution $N(\mu, \sigma^2)$. Find the MLE of the population mean $\mu$.

**Solution**:

- The likelihood function is:

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right).$$

- The log-likelihood function is:

$$\ln L(\mu) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2.$$

- Taking the derivative with respect to $\mu$ and setting it to zero:

$$\frac{\partial \ln L(\mu)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu) = 0.$$

- Solving for $\mu$, we get:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}.$$

- Therefore, the MLE of $\mu$ is $\hat{\mu} = \bar{X}$, the sample mean.

# 3 Week 3 (Sep 7 - Sep 13): Confidence Intervals

## 3.1 Confidence Interval Basics

- **Definition**: An interval estimate of a population parameter with a specified confidence level.
- **Interpretation**: If we repeat the sampling process, a certain percentage (confidence level) of the intervals will contain the true parameter.

## 3.2 Confidence Intervals for the Mean

- **Known Variance** ($\sigma^2$):
$$\bar{X} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$
- **Unknown Variance**:
$$\bar{X} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$$

## 3.3 Confidence Intervals for Proportions

- Formula:
$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

## 3.4 Confidence Intervals for Variance

- **Chi-Squared Distribution**:
$$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}\right)$$

## 3.5 Sample Problem 1: Confidence Interval for the Mean (Known Variance)

**Problem**: A sample of size 25 has a sample mean $\bar{X} = 50$ and the population variance $\sigma^2 = 100$ is known. Construct a 95% confidence interval for the population mean $\mu$.
**Solution**:

- The formula for the confidence interval when the variance is known is:
$$\bar{X} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right).$$
- For a 95% confidence interval, $\alpha = 0.05$ and $z_{\alpha/2} = 1.96$.
- Substituting the values:
$$50 \pm 1.96\left(\frac{10}{\sqrt{25}}\right) = 50 \pm 1.96 \times 2 = 50 \pm 3.92.$$
- Therefore, the 95% confidence interval for $\mu$ is $(46.08, 53.92)$.

### 3.6 Sample Problem 2: Confidence Interval for Proportion

**Problem**: A random sample of 200 voters is surveyed, and 120 of them say they support a particular candidate. Construct a 90% confidence interval for the true proportion of voters who support the candidate.

**Solution**:

- The sample proportion is $\hat{p} = \frac{120}{200} = 0.60$.
- The formula for the confidence interval is:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

- For a 90% confidence interval, $z_{\alpha/2} = 1.645$.
- Substituting the values:

$$0.60 \pm 1.645\sqrt{\frac{0.60(1-0.60)}{200}} = 0.60 \pm 1.645\sqrt{\frac{0.24}{200}} = 0.60 \pm 1.645 \times 0.0347 = 0.60 \pm 0.057.$$

- Therefore, the 90% confidence interval for the true proportion is $(0.543, 0.657)$.

# 4 Week 4 (Sep 14 - Sep 20): Hypothesis Testing for One Population

## 4.1 Hypothesis Testing Basics

- **Null Hypothesis** ($H_0$): The statement being tested, usually a statement of no effect or no difference.
- **Alternative Hypothesis** ($H_a$): The statement we consider if $H_0$ is rejected.

## 4.2 Test Statistics

- **For Mean (Known Variance)**:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

- **For Mean (Unknown Variance)**:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

## 4.3 Type I and Type II Errors

- **Type I Error** ($\alpha$): Rejecting $H_0$ when it is true.
- **Type II Error** ($\beta$): Failing to reject $H_0$ when $H_a$ is true.
- **Power of the Test**: $1 - \beta$, the probability of correctly rejecting $H_0$.

## 4.4 p-Value

- **Definition**: The probability of obtaining a test statistic as extreme as the one observed, assuming $H_0$ is true.
- **Decision Rule**: Reject $H_0$ if p-value $\leq \alpha$.

## 4.5 Hypothesis Testing Steps

1. State $H_0$ and $H_a$.
2. Choose Significance Level ($\alpha$).
3. Compute Test Statistic.
4. Determine p-value or Critical Value.
5. Make Decision: Reject or fail to reject $H_0$.
6. Interpret Results.

## 4.6 Sample Problem 1: Hypothesis Test for the Mean (Known Variance)

**Problem**: A machine produces bolts with an average length of 50 mm. A sample of 30 bolts is taken, and the sample mean is found to be 51 mm. The population standard deviation is known to be 2 mm. Test at the 5% significance level whether the mean length of the bolts has changed.

**Solution**:

- **Step 1: State the hypotheses**.

$$H_0 : \mu = 50 \quad \text{(null hypothesis)}$$

$$H_a : \mu \neq 50 \quad \text{(alternative hypothesis)}$$

- **Step 2: Compute the test statistic**.

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{51 - 50}{2/\sqrt{30}} = \frac{1}{0.365} = 2.74.$$

- **Step 3: Find the critical value**.

$$z_{\alpha/2} = 1.96 \quad \text{(for a two-tailed test with } \alpha = 0.05\text{)}.$$

- **Step 4: Make a decision**.

$$|Z| = 2.74 > 1.96 \quad \text{so we reject } H_0.$$

- **Conclusion**: There is sufficient evidence to suggest that the mean length of the bolts has changed.

# 5 Week 5 (Sep 23 - Sep 27): Hypothesis Tests for Two Populations

## 5.1 Comparing Two Means

- **Independent Samples**:
  - **Equal Variances (Pooled t-test)**:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

    Pooled Variance:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

  - **Unequal Variances (Welch's t-test)**: Adjust degrees of freedom using the Welch-Satterthwaite equation.

- **Paired Samples**:
  - **Paired t-test**: Compute differences $d_i = x_{1i} - x_{2i}$.

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

## 5.2 Comparing Two Proportions

- **Test Statistic**:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

  Pooled Proportion:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

## 5.3 Comparing Two Variances

- **F-Test**:

$$F = \frac{s_1^2}{s_2^2}$$

  Degrees of Freedom: $df_1 = n_1 - 1$, $df_2 = n_2 - 1$.

## 5.4 Sample Problem 1: Two-Sample t-Test (Equal Variances)

**Problem**: Two groups of students take two different teaching methods. The first group of 20 students has an average score of 75 with a standard deviation of 10. The second group of 20 students has an average score of 80 with a standard deviation of 12. Test whether the two teaching methods result in significantly different average scores at the 5% significance level.

**Solution**:

- **Step 1: State the hypotheses**.

$$H_0 : \mu_1 = \mu_2 \quad \text{(null hypothesis)}$$

$$H_a : \mu_1 \neq \mu_2 \quad \text{(alternative hypothesis)}$$

- **Step 2: Compute the pooled standard deviation**.

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(20 - 1)10^2 + (20 - 1)12^2}{20 + 20 - 2} = 121.$$

$$s_p = \sqrt{121} = 11.$$

- **Step 3: Compute the test statistic**.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{75 - 80}{11\sqrt{\frac{1}{20} + \frac{1}{20}}} = \frac{-5}{3.48} = -1.44.$$

- **Step 4: Find the critical value**.

$$t_{\alpha/2,38} = 2.024 \quad (\text{for } \alpha = 0.05).$$

- **Step 5: Make a decision**.

$$|t| = 1.44 < 2.024 \quad \text{so we fail to reject } H_0.$$

- **Conclusion**: There is no significant evidence to suggest that the two teaching methods result in different average scores.

# 6 Week 6 (Sep 30 - Oct 4): Two Population Tests & Other Topics

## 6.1 Nonparametric Tests

- **When to Use**: Data doesn't meet the assumptions required for parametric tests (e.g., normality).
- **Common Tests**:
  - **Mann-Whitney U Test**: For comparing two independent samples.
  - **Wilcoxon Signed-Rank Test**: For comparing paired samples.

## 6.2 Analysis of Variance (ANOVA)

- **Purpose**: Test for significant differences among three or more group means.
- **F-Statistic**:

$$F = \frac{\text{Between-Group Variance}}{\text{Within-Group Variance}}$$

## 6.3 Multiple Comparisons

- **Post-Hoc Tests**: Conducted after ANOVA to identify which means differ.
- **Common Methods**:
  - Tukey's HSD.
  - Bonferroni Correction.

## 6.4 Sample Problem 1: Mann-Whitney U Test

**Problem**: Two groups of students are tested for their problem-solving ability. The first group has scores: 85, 87, 89, 91, 92. The second group has scores: 78, 81, 86, 88, 90. Use the Mann-Whitney U test to determine if there is a significant difference between the two groups at the 5% significance level.

**Solution**:

- **Step 1: Rank the combined data**.

$$\text{Ranks} : 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.$$

- **Step 2: Compute the sum of ranks for each group**.

$$R_1 = 6 + 7 + 8 + 9 + 10 = 40, \quad R_2 = 1 + 2 + 3 + 4 + 5 = 15.$$

- **Step 3: Compute the U statistic for each group**.

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 = 5 \times 5 + \frac{5(6)}{2} - 40 = 25 + 15 - 40 = 0.$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2 = 5 \times 5 + \frac{5(6)}{2} - 15 = 25 + 15 - 15 = 25.$$

- **Step 4: Decision rule**.
- Use critical value tables for $U$ at $\alpha = 0.05$. If $U$ is less than the critical value, reject $H_0$.
- **Conclusion**: Check critical values and make the decision based on the ranks.

# 7 Week 7 (Oct 7 - Oct 11): Categorical Data Tests

## 7.1 Chi-Square Tests

- **Goodness-of-Fit Test**:
  - **Purpose**: Determine if observed frequencies differ from expected frequencies.
  - **Test Statistic**:
  $$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$
- **Test for Independence**:
  - **Contingency Tables**: Analyze the relationship between two categorical variables.
  - **Degrees of Freedom**:
  $$df = (r-1)(c-1)$$
  where $r$ = rows, $c$ = columns.

## 7.2 Conditions for Chi-Square Tests

- **Expected Frequencies**: Each expected cell frequency should be at least 5.
- **Independence**: Observations should be independent of each other.

## 7.3 Fisher's Exact Test

- **When to Use**: Sample sizes are small, and the chi-square test assumptions are not met.

## 7.4 Sample Problem 1: Chi-Square Test for Goodness-of-Fit

**Problem**: A six-sided die is rolled 60 times, and the observed frequencies are as follows: 1: 12, 2: 10, 3: 8, 4: 15, 5: 9, 6: 6. Test if the die is fair at the 5% significance level.
**Solution**:

- **Step 1: State the hypotheses**.
$$H_0 : \text{The die is fair (equal probabilities for all faces).}$$
$$H_a : \text{The die is not fair.}$$
- **Step 2: Compute the expected frequencies**.
$$\text{Expected frequency} = \frac{60}{6} = 10 \text{ for each face.}$$
- **Step 3: Compute the test statistic**.
$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(12-10)^2}{10} + \frac{(10-10)^2}{10} + \cdots = 1.4.$$

- **Step 4: Find the critical value**.

$$\chi^2_{\alpha,df} = 11.07 \quad \text{(for } \alpha = 0.05 \text{ and } df = 5\text{)}.$$

- **Step 5: Make a decision**.

$$\chi^2 = 1.4 < 11.07 \quad \text{so we fail to reject } H_0.$$

- **Conclusion**: There is no evidence to suggest that the die is unfair.

## Additional Resources

- **Statistical Tables**:
  - **Z-Table**: For standard normal distribution.
  - **t-Table**: For t-distribution critical values.
  - **Chi-Square Table**: For chi-square distribution critical values.
  - **F-Table**: For F-distribution critical values.
- **Software Tools**: Familiarize yourself with statistical software (e.g., R, Python, Excel) for performing computations.

## Study Tips

1. Understand Concepts: Focus on grasping the underlying concepts rather than just memorizing formulas.
2. Practice Problems: Apply what you've learned by solving practice problems.
3. Use Visual Aids: Diagrams and charts can help in understanding distributions and test statistics.
4. Form Study Groups: Discussing topics with peers can enhance understanding.
5. Seek Help When Needed: Don't hesitate to ask your instructor or TA for clarification.