

1. Explain the properties of the F-distribution.

ANS - Properties of the F-distribution

- **Non-negative:** The F-distribution takes only positive values.
- **Right-skewed:** The shape is right-skewed, particularly for smaller sample sizes. As the degrees of freedom increase, the distribution becomes more symmetric.
- **Asymptotic:** As the sample size increases, the distribution approaches normality.
- **Two degrees of freedom:** The F-distribution is defined by two parameters: the degrees of freedom of the numerator (df_1) and the degrees of freedom of the denominator (df_2).
- **Mean:** The mean of the F-distribution is $\frac{df_2}{df_2-2} \cdot \frac{df_1+df_2-2}{df_1}$ for $df_2 > 2$.
- **Variance:** The variance is $\frac{2 \cdot df_1^2 \cdot (df_1+df_2-2)}{df_1 \cdot (df_2-2)^2 \cdot (df_2-4)}$ for $df_2 > 4$.

2. In which types of statistical tests is the F-distribution used, and why is it appropriate for these tests?

- **ANOVA (Analysis of Variance):** The F-distribution is used to compare the variances between groups to determine if they differ significantly.
- **F-test for equality of variances:** It tests whether the variances of two populations are equal.
- **Regression analysis:** Used to compare the explained variance to the unexplained variance in models.

The F-distribution is appropriate for these tests because it models the ratio of variances, which is essential in determining if group differences are significant.

3. What are the key assumptions required for conducting an F-test to compare the variances of two populations?

Assumptions for F-test to compare variances

- **Normality:** Both populations should be normally distributed.
- **Independence:** The samples should be independent of each other.
- **Homogeneity of variances:** The variances within each group should be approximately equal (for tests like ANOVA).

4. What is the purpose of ANOVA, and how does it differ from a t-test?

Purpose of ANOVA vs. t-test

- **ANOVA:** Compares the means of three or more groups to check for significant differences between them.
- **t-test:** Compares the means of two groups to check for significant differences.

Difference: ANOVA extends the t-test by allowing comparison of multiple groups simultaneously, whereas the t-test is limited to two groups. Multiple t-tests increase the risk of Type I error (false positives), which ANOVA controls.

5. Explain when and why you would use a one-way ANOVA instead of multiple t-tests when comparing more than two groups.

One-way ANOVA is preferred over multiple t-tests when comparing more than two groups because:

- **Type I error control:** Performing multiple t-tests inflates the probability of a Type I error. ANOVA controls this error.
- **Efficiency:** ANOVA handles all group comparisons in one test, reducing complexity.

6. Explain how variance is partitioned in ANOVA into between-group variance and within-group variance. How does this partitioning contribute to the calculation of the F-statistic?

- **Between-group variance:** Measures the variation between the group means.
- **Within-group variance:** Measures the variation within each group.

The **F-statistic** is calculated as the ratio of between-group variance to within-group variance. A high F-statistic indicates that between-group variance is large relative to within-group variance, suggesting significant differences between groups.

7. Compare the classical (frequentist) approach to ANOVA with the Bayesian approach. What are the key differences in terms of how they handle uncertainty, parameter estimation, and hypothesis testing?

Classical (Frequentist) ANOVA vs. Bayesian ANOVA

- **Frequentist ANOVA:** Focuses on rejecting or failing to reject the null hypothesis. It calculates a p-value to assess whether group means differ significantly.
- **Bayesian ANOVA:** Provides a probability distribution for parameters and quantifies uncertainty. It uses prior information and updates beliefs based on the data, providing richer insights about uncertainty.

Key differences:

- **Uncertainty:** Bayesian ANOVA provides probabilistic statements, while classical ANOVA only provides a binary decision based on p-values.
- **Parameter estimation:** Bayesian approach uses prior distributions, while classical methods rely purely on data.

8. Question: You have two sets of data representing the incomes of two different professions
 Profession A: [48, 52, 55, 60, 62] Profession B: [45, 50, 55, 52, 47] Perform an F-test to determine if the variances of the two professions' incomes are equal. What are your conclusions based on the F-test? Task: Use Python to calculate the F-statistic and p-value for the given data.

```
import numpy as np
from scipy.stats import f

# Data for Profession A and B
profession_A = [48, 52, 55, 60, 62]
profession_B = [45, 50, 55, 52, 47]


# Sample variances
var_A = np.var(profession_A, ddof=1)
var_B = np.var(profession_B, ddof=1)

# F-statistic
F_statistic = var_A / var_B

# Degrees of freedom
df_A = len(profession_A) - 1
df_B = len(profession_B) - 1

# p-value (one-tailed test)
p_value = 1 - f.cdf(F_statistic, df_A, df_B)

F_statistic, p_value
```

 (2.089171974522293, 0.24652429950266952)


9. Question: Conduct a one-way ANOVA to test whether there are any statistically significant differences in average heights between three different regions with the following data
 Region A: [160, 162, 165, 158, 164] Region B: [172, 175, 170, 168, 174] Region C: [180, 182, 179, 185, 183]
 Task: Write Python code to perform the one-way ANOVA and interpret the results
 Objective: Learn how to perform one-way ANOVA using Python and interpret F-statistic and p-value

```
import scipy.stats as stats

# Data for Region A, B, and C
region_A = [160, 162, 165, 158, 164]
region_B = [172, 175, 170, 168, 174]
region_C = [180, 182, 179, 185, 183]

# Perform one-way ANOVA
F_statistic, p_value = stats.f_oneway(region_A, region_B, region_C)

F_statistic, p_value
```

 (67.87330316742101, 2.870664187937026e-07)

Interpretation:

- If the p-value is less than the significance level (e.g., 0.05), it suggests that there is a statistically significant difference in the average heights between the regions.
- If the p-value is greater than 0.05, it suggests that the differences in heights are not statistically significant.