F-Tests, ANOVA and Regressions

Tsering Sherpa

2025

F-Tests

The **F-test** is used to compare variances or to evaluate overall model significance, especially in **ANOVA** and **regression**.

What Does the F-Test Measure?

The F-test evaluates whether the ratio of two variances is significantly different from 1.

$$F = \frac{\text{Variance}_1}{\text{Variance}_2}$$

A high F-value suggests the variances are not equal.

Key Assumptions

- Observations are independent
- Residuals are normally distributed
- Homogeneity of variance (equal variances across groups)

How to Interpret It

- If F is large and the p-value is small, reject H_0.
- The F-distribution is right-skewed and requires two degrees of freedom:
  - Numerator degrees of freedom: related to the number of groups or predictors
  - Denominator degrees of freedom: related to residual variation

Table 1: F-Test for Variances

Aspect	F-Test
Purpose	Compare variances of two groups
Stand. Test Statistic	$F = \frac{s_1^2}{s_2^2}$
Typical Distribution	F-distribution

Common Uses of the F-Test

- **Equality of Variances (Two Samples):**

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{vs.} \quad H_a : \sigma_1^2 \neq \sigma_2^2$$

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

  where s_1^2 and s_2^2 are the sample variances.

- **ANOVA (Analysis of Variance):**

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

$$F = \frac{\text{Between-group variance}}{\text{Within-group variance}}$$

  Tests whether the group means are all equal (no treatment effect).

- **Regression Significance:**

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

  Checks if any independent variable contributes significantly to the model.

ANOVA: Analysis of Variance

ANOVA is used to test whether the means of *three or more groups* are equal. Instead of doing multiple t-tests (which increases the risk of error), ANOVA compares all groups at once using one overall test.

When to Use ANOVA

Use ANOVA when:

- You have one categorical variable with 3 or more groups (ex., robins in 3 or more locations).
- You're comparing the means of a numerical outcome across those groups (ex., mean egg size in each location)

Hypotheses

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \quad \text{(all group means are equal)}$$
$$H_a : \text{At least one group mean is different}$$

How It Works

ANOVA splits the total variation in the data into two parts:

- **Between-group variation:** how much the group means differ from the overall mean
- **Within-group variation:** how much values vary inside each group

The formula is:

$$F = \frac{MSG}{MSE}, \quad MSG = \frac{SSG}{k-1}, \quad MSE = \frac{SSE}{N-k}$$

This test only tells us if one of the means in one or more groups is not equal to the others, not how many groups are not equal or which groups are not equal to the others.

F-statistic:

$$F = \frac{\text{Between-group variation}}{\text{Within-group variation}}$$

If the between-group variation is much larger than the within-group variation, the F-statistic will be large, which suggests a difference in means.

Decision Rule

- A large F-value and a small p-value (typically < 0.05) $\rightarrow$ reject H_0
- This means **at least one group mean is different**, but ANOVA does not tell you which one. To see which group mean differs, you'd use follow-up tests (post hoc tests)

Assumptions of ANOVA

- The data in each group are normally distributed
- The variances are roughly equal across groups
- The observations are independent

Regressions

A regression is essentially the core idea behind determining causal effects, which can be a challenging task. There are a lot of factors that may affect an outcome that you do not directly observe that influence the dependent (outcome) variable you're interested in.

At a basic level, a regression computes the expected value of a dependent variable (y) given an independent variable (x). You can have multiple independent variables. The key idea here is that we already have x and y values. What you're interested in estimating is the slope of a regression.

Regression:

$$y = \alpha + \beta_i x_i + \epsilon_i$$

y is the dependent variable, β_i is the slope associated with x_i. ϵ measures the unobserved randomness, or noise. α is the intercept term.

We interpret β in linear regressions as: "A one unit change in x is associated with a β (unit) change in y".