SDU UNIVERSITY

COMPARING TWO VARIANCES

SECTION 9.4

LEARNING OBJECTIVES

- Understand the Need for Variance Comparison
- Levene's Test for Equality of Variances
- F Test for Equality of Variances
- Practical Application and Interpretation

WHY COMPARE VARIANCES?

- Understanding Data Variability: Variance measures data spread, which impacts statistical conclusions.
- Assumptions in Statistical Tests: Many tests (e.g., t-test, ANOVA) assume equal variances.
- Fair Comparisons: Ensuring that two groups are compared under similar conditions.
- Impact on Machine Learning & Data Science: Variance differences affect model performance and inference.

APPLICATIONS IN REAL-WORLD DATA ANALYSIS

MEDICAL STUDIES

- Checking if drug responses vary between treatment & control groups.
- Variability in blood pressure levels across different age groups.

QUALITY CONTROL

- Comparing product consistency across two production lines.
- Ensuring uniform material strength in engineering applications.

FINANCE & ECONOMICS

- Analyzing volatility in stock prices across different time periods.
- Measuring risk differences between investment portfolios.

EDUCATION PSYCHOLOGY

- Comparing student performance variability in different teaching methods.
- Evaluating score dispersion in standardized testing.

TWO METHODS:

LEVENE'S TEST

F-TEST

LEVENE'S TEST

- **Assumption**: Works with any data distribution (normal or non-normal).
- **How it Works**:
    - Uses absolute deviations from the median (or mean).
    - Converts variance comparison into a **two-sample t-test**.
- **Advantages**:
    - Less sensitive to outliers and skewed data.
    - More reliable in real-world datasets.
- **Disadvantage**:
    - Slightly lower power when data is perfectly normal.

STATS.LEVENE(NONSMOKERS, SMOKERS)

F-TEST

- **Assumption**: Requires both populations to be **normally distributed**.
- **How it Works:**
  - Uses the ratio of sample variances:

$$F = \frac{S_1^2}{S_2^2}$$

  - Compares it to an **F-distribution critical value**.
- **Advantages**:
  - Simple to compute.
  - More powerful when normality holds.
- **Disadvantages**:
  - **Highly sensitive** to departures from normality.
  - Affected by outliers.

HYPOTHESIS TESTING FOR VARIANCES

- We compare two population variances, $\boldsymbol{\sigma_1^2}$ and $\boldsymbol{\sigma_2^2}$
- **Null Hypothesis (H_0):** $\sigma_1^2 = \sigma_2^2$
- **Alternative Hypothesis (H_a):**
  - $\sigma_1^2 \neq \sigma_2^2$ (two-tailed)
  - $\sigma_1^2 > \sigma_2^2$ (right-tailed)
  - $\sigma_1^2 < \sigma_2^2$ (left-tailed)

LEVENE'S TEST

Levene's test (also called the Brown-Forsythe test) is based on the idea that if $\sigma_1^2 = \sigma_2^2$, then the induced samples

$$V_{1j} = |X_{1j} - \widetilde{X}_1|, \quad j = 1, \ldots, n_1, \quad \text{and} \quad V_{2j} = |X_{2j} - \widetilde{X}_2|, \quad j = 1, \ldots, n_2,$$

where $\widetilde{X}_i$ is the sample median of $X_{i1}, \ldots, X_{in_i}$ for $i = 1, 2$, have equal population means and variances. Moreover, if $\sigma_1^2 > \sigma_2^2$, then the population mean μ_{V_1} of the V_1 sample will be larger than the population mean μ_{V_2} of the V_2 sample. Thus, testing $H_0 : \sigma_1^2 = \sigma_2^2$ versus $H_a : \sigma_1^2 > \sigma_2^2$ or $H_a : \sigma_1^2 < \sigma_2^2$ or $H_a : \sigma_1^2 \neq \sigma_2^2$ can be performed by testing the hypothesis $H_0^V : \mu_{V_1} = \mu_{V_2}$ versus

$$H_a^V : \mu_{V_1} > \mu_{V_2} \text{ or } H_a^V : \mu_{V_1} < \mu_{V_2} \text{ or } H_a^V : \mu_{V_1} \neq \mu_{V_2},$$

respectively, using the two-sample T test with pooled variance, that is, the procedure (9.2.17) based on the statistic $T_{H_0}^{EV}$ given in (9.2.14), using the two V samples.

$$T_{H_0}^{EV} = \frac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}. \tag{9.2.14}$$

The T Test Procedures for $H_0: \mu_1 - \mu_2 = \Delta_0$

(1) *Assumptions:*
(a) $X_{11}, \ldots, X_{1n_1}$, $X_{21}, \ldots, X_{2n_2}$ are independent simple random samples
(b) Either the populations are normal or n_1 and $n_2 \geq 30$

(2) *Test Statistic:*

$$T_{H_0} = \begin{cases} T_{H_0}^{EV} & \text{if } \sigma_1^2 = \sigma_2^2 \text{ is assumed} \\ T_{H_0}^{SS} & \text{regardless of whether or not } \sigma_1^2 = \sigma_2^2 \end{cases}$$

where $T_{H_0}^{EV}$ and $T_{H_0}^{SS}$ are defined in (9.2.14) and (9.2.15)

(3) *Rejection Rules for the Different H_a:*

H_a	RR at Level α
$\mu_1 - \mu_2 > \Delta_0$	$T_{H_0} > t_{\alpha,df}$
$\mu_1 - \mu_2 < \Delta_0$	$T_{H_0} < -t_{\alpha,df}$
$\mu_1 - \mu_2 \neq \Delta_0$	$\lvert T_{H_0}\rvert > t_{\alpha/2,df}$

(9.2.17)

where $df = n_1 + n_2 - 2$ if $T_{H_0} = T_{H_0}^{EV}$, or else $df = \nu$, where ν is given in (9.2.16), if $T_{H_0} = T_{H_0}^{SS}$

Example 9.4-1

Numerous studies have shown that cigarette smokers have a lower plasma concentration of ascorbic acid (vitamin C) than nonsmokers. Given the health benefits of ascorbic acid, there is also interest in comparing the variability of the concentration in the two groups. The following data represent the plasma ascorbic acid concentration measurements (μmol/l) of five randomly selected smokers and nonsmokers:

Nonsmokers	41.48	41.71	41.98	41.68	41.18
Smokers	40.42	40.68	40.51	40.73	40.91

Test the hypothesis $H_0: \sigma_1^2 = \sigma_2^2$ versus $H_a: \sigma_1^2 \neq \sigma_2^2$ at $\alpha = 0.05$.

Solution

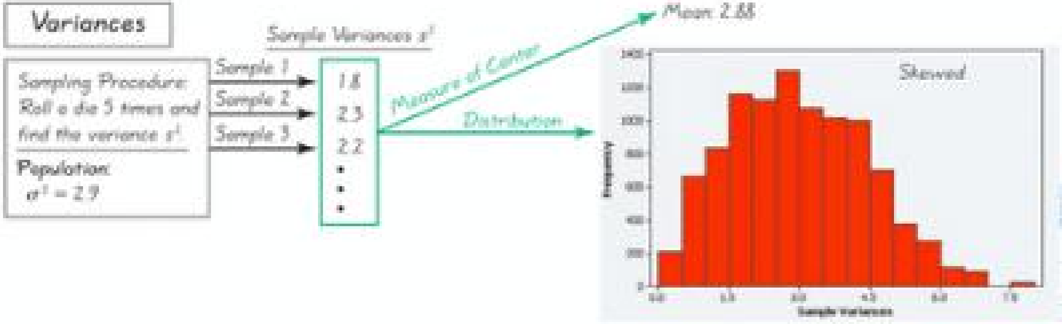
Here the two medians are $\widetilde{X}_1 = 41.68, \widetilde{X}_2 = 40.68$. Subtracting them from their corresponding sample values, and taking the absolute values, we obtain the two V samples:

V_1 Values for Nonsmokers	0.20	0.03	0.30	0.00	0.50
V_2 Values for Smokers	0.26	0.00	0.17	0.05	0.23

The two-sample test statistic $T_{H_0}^{EV}$ evaluated on the two induced V samples takes a value of 0.61. With 8 degrees of freedom, this corresponds to a p-value of 0.558. Thus, there is not enough evidence to reject the null hypothesis of equality of the two population variances. Instead of using hand calculations, the two samples can be imported into R with the commands *x1=c(41.48, 41.71, 41.98, 41.68, 41.18); x2=c(40.42, 40.68, 40.51, 40.73, 40.91)* and then we can use either of the commands in (9.4.1) or (9.4.2) to get the same p-value. ■

CHI-SQUARED DISTRIBUTION AND F DISTRIBUTIONS

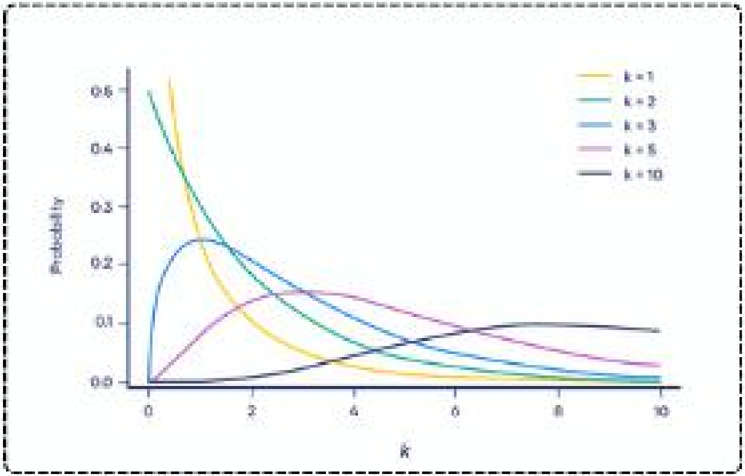
SAMPLING DISTRIBUTION OF SAMPLE VARIANCE



https://onlinestatbook.com/stat_sim/sampling_dist/

WHAT IS A CHI SQUARED DISTRIBUTION?

- The Chi-Square distribution is the distribution of the sum of squared standard normal variables.
- If $Z_1, Z_2, \ldots, Z_k$ are independent standard normal variables, then: $X = Z_1^2 + Z_2^2 + \ldots + Z_k^2$ follows a χ^2 distribution with **k degrees of freedom**.



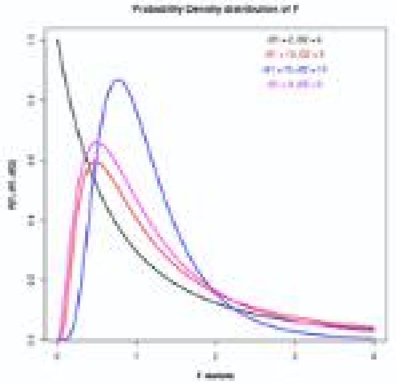
PROPERTIES OF CHI-SQUARE DISTRIBUTION

- **Shape:** Right-skewed (approaches normal as degrees of freedom increase).
- **Mean:** k (equal to the degrees of freedom).
- **Variance:** $2k$.
- **Range:** $\chi^2 \geq 0$ (only takes non-negative values).
- **Skewness:** $\frac{2}{\sqrt{k}}$, decreases as k increases.

APPLICATIONS OF THE CHI-SQUARE DISTRIBUTION

- **Goodness-of-Fit Test:** Checks if observed data fits an expected distribution.
- **Test for Independence (Chi-Square Test):** Determines if two categorical variables are independent.
- **Variance Estimation:** Used in confidence intervals for variance and standard deviation.

WHAT IS THE F-DISTRIBUTION?



- The F-distribution is the ratio of two independent Chi-Square variables divided by their degrees of freedom.
- If $X \sim \chi^2_{d_1}$ and $Y \sim \chi^2_{d_2}$, then: $F = \frac{(X/d_1)}{(Y/d_2)}$ follows an F-distribution with degrees of freedom d_1 and d_2.

PROPERTIES OF THE F-DISTRIBUTION

- **Shape:** Right-skewed.
- **Mean:** $\frac{d_2}{d_2-2}$ for $d_2 > 2$.
- **Variance:** $\frac{2d_2^2(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$ for $d_2 > 4$.
- **Range:** $F \geq 0$ (always positive).
- Used in ANOVA and regression analysis.

APPLICATIONS OF THE F-DISTRIBUTION

- **ANOVA (Analysis of Variance):** Determines if means of multiple groups are different.
- **Regression Analysis:** Tests the significance of regression models.
- **Comparing Variances:** Tests if two populations have equal variances.

F TEST UNDER NORMALITY

Theorem 9.4-1

Let $X_{11}, \ldots, X_{1n_1}$ and $X_{21}, \ldots, X_{2n_2}$ be two independent random samples from normal populations with variances σ_1^2 and σ_2^2, respectively, and let S_1^2 and S_2^2 denote the two sample variances. Then

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,\, n_2-1},$$

that is, the ratio has an F distribution with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ degrees of freedom.

This result indicates that the hypothesis $H_0: \sigma_1^2 = \sigma_2^2$ can be tested using the test statistic:

$$F_{H_0} = \frac{S_1^2}{S_2^2}$$

If H_0 is true, then F_{H_0} follows an F-distribution with degrees of freedom $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$.

- When $\sigma_1^2 > \sigma_2^2$, the value of F_{H_0} tends to be larger than expected under H_0.
- When $\sigma_1^2 < \sigma_2^2$, F_{H_0} tends to be smaller than expected under H_0. Alternatively, the reciprocal $1/F_{H_0} = S_2^2/S_1^2$ tends to be larger.

These observations lead to specific rejection rules and p-value calculations depending on whether the alternative hypothesis suggests that one variance is greater, smaller, or simply different from the other.

The F Test Procedures for $H_0 : \sigma_1^2 = \sigma_2^2$

(1) *Assumption:* $X_{11}, \ldots, X_{1n_1}$ and $X_{21}, \ldots, X_{2n_2}$ are independent samples from normal populations

(2) *Test Statistic:* $F_{H_0} = \dfrac{S_1^2}{S_2^2}$.

(3) *Rejection Rules for the Different H_a:*

$$
\begin{array}{ll}
H_a & \text{RR at Level } \alpha \\
\hline
\sigma_1^2 > \sigma_2^2 & F_{H_0} > F_{n_1-1,n_2-1;\,\alpha} \\
\sigma_1^2 < \sigma_2^2 & \dfrac{1}{F_{H_0}} > F_{n_2-1,n_1-1;\,\alpha} \\
\sigma_1^2 \neq \sigma_2^2 & \text{either } F_{H_0} > F_{n_1-1,n_2-1;\,\alpha/2} \\
 & \text{or } \quad \dfrac{1}{F_{H_0}} > F_{n_2-1,n_1-1;\,\alpha/2}
\end{array}
\tag{9.4.4}
$$

where $F_{\nu_1,\nu_2;\alpha}$ denotes the $(1-\alpha)100$th percentile of the F_{ν_1,ν_2} distribution

(4) *Formulas for the p-Value:*

$$
p\text{-value} = \begin{cases} p_1 = 1 - F_{n_1-1,\,n_2-1}(F_{H_0}) & \text{for } H_a : \sigma_1^2 > \sigma_2^2 \\ p_2 = 1 - F_{n_2-1,\,n_1-1}(1/F_{H_0}) & \text{for } H_a : \sigma_1^2 < \sigma_2^2 \\ 2[\min(p_1, p_2)] & \text{for } H_a : \sigma_1^2 \neq \sigma_2^2 \end{cases}
$$

where F_{ν_1,ν_2} denotes the CDF of the F_{ν_1,ν_2} distribution

Example 9.4-2 Consider the data and testing problem described in Example 9.4-1, and assume that the underlying populations are normal.

(a) Test $H_0 : \sigma_1^2 = \sigma_2^2$ versus $H_a : \sigma_1^2 \neq \sigma_2^2$ using the F test procedure (9.4.4).

(b) Implement the F test procedure using R commands, and report the 95% CI for σ_1^2/σ_2^2.

Solution

(a) The test statistic is

$$F_{H_0} = \frac{0.08838}{0.03685} = 2.40.$$

The value 2.4 corresponds to the 79th percentile of the F distribution with $\nu_1 = 4$ and $\nu_2 = 4$ degrees of freedom. Thus, noting also that $1 - F_{4,4}(2.4) < 1 - F_{4,4}(1/2.4)$, the p-value is $2(1 - 0.79) = 0.42$.

(b) The commands *x1=c(41.48, 41.71, 41.98,41.68, 41.18); x2=c(40.42, 40.68, 40.51, 40.73, 40.91); var.test(x1, x2)* return 2.3984 and 0.4176 for the F statistic and p-value, respectively. Rounded to two decimal places, these values match those obtained in part (a). In addition, the above commands return (0.25, 23.04) as a 95% CI for σ_1^2/σ_2^2.

CHECK FOR UNDERSTANDING

1. What is the primary purpose of Levene's Test?

A) To compare the means of two or more groups

B) To test for equality of variances between two or more groups

C) To test for normality of a dataset

D) To check for correlation between two variables

2. Which of the following assumptions does Levene's Test require?

A) The data must be normally distributed

B) The data must be categorical

C) The data should be independent within and between groups

D) The data must have equal means

3. Which of the following is a key advantage of Levene's Test over the F-test for variances?

A) It is more powerful when the data is normal

B) It is robust to non-normality

C) It tests for equality of means, not variances

D) It requires fewer samples to be valid

4. If the p-value from Levene's Test is 0.03 ($\alpha = 0.05$), what should we conclude?

A) Reject the null hypothesis and conclude that variances are equal
B) Fail to reject the null hypothesis and conclude that variances are equal
C) Reject the null hypothesis and conclude that variances are different
D) Fail to reject the null hypothesis and conclude that means are different

1. Which of the following is true about the Chi-Square distribution?

A) It is symmetric for all degrees of freedom.

B) It is always negatively skewed.

C) It is the distribution of the sum of squared standard normal variables.

D) It can take negative values.

2. If $X \sim \chi^2_5$, what is the mean of X?

A) 2

B) 5

C) 10

D) Cannot be determined

3. The F-distribution is formed by the ratio of:

A) Two normal distributions

B) Two Chi-Square distributions divided by their respective degrees of freedom

C) Two independent t-distributions

D) A normal and a Chi-Square distribution

4. The shape of the Chi-Square distribution depends on:

A) Mean and variance

B) The number of samples

C) The degrees of freedom

D) The normality of the population

5. If $F \sim F(10, 20)$, what happens to the shape of the distribution as the degrees of freedom increase?

A) It becomes more skewed

B) It becomes more symmetric

C) It does not change

D) It becomes a normal distribution

6. In an F-test for equality of variances, the null hypothesis states:

A) $\sigma_1^2 > \sigma_2^2$
B) $\sigma_1^2 < \sigma_2^2$
C) $\sigma_1^2 = \sigma_2^2$
D) $\sigma_1^2 + \sigma_2^2 = 1$

7. The Chi-Square test is often used for:

A) Comparing means of two populations

B) Testing relationships in categorical data

C) Checking normality in data

D) Testing correlation between two variables