

Chapter 9: Inference for Variances

DSCC 462

Computational Introduction to Statistics

Anson Kahng

Fall 2022

Inference for Variances

- Primarily, inference tends to be focused on means (and proportions, which we will cover next)
- However, variances are essential to inference of means
- We may be interested in determining:
 - Whether a population variance is equal to a predetermined value
 - Whether two population variances are equal

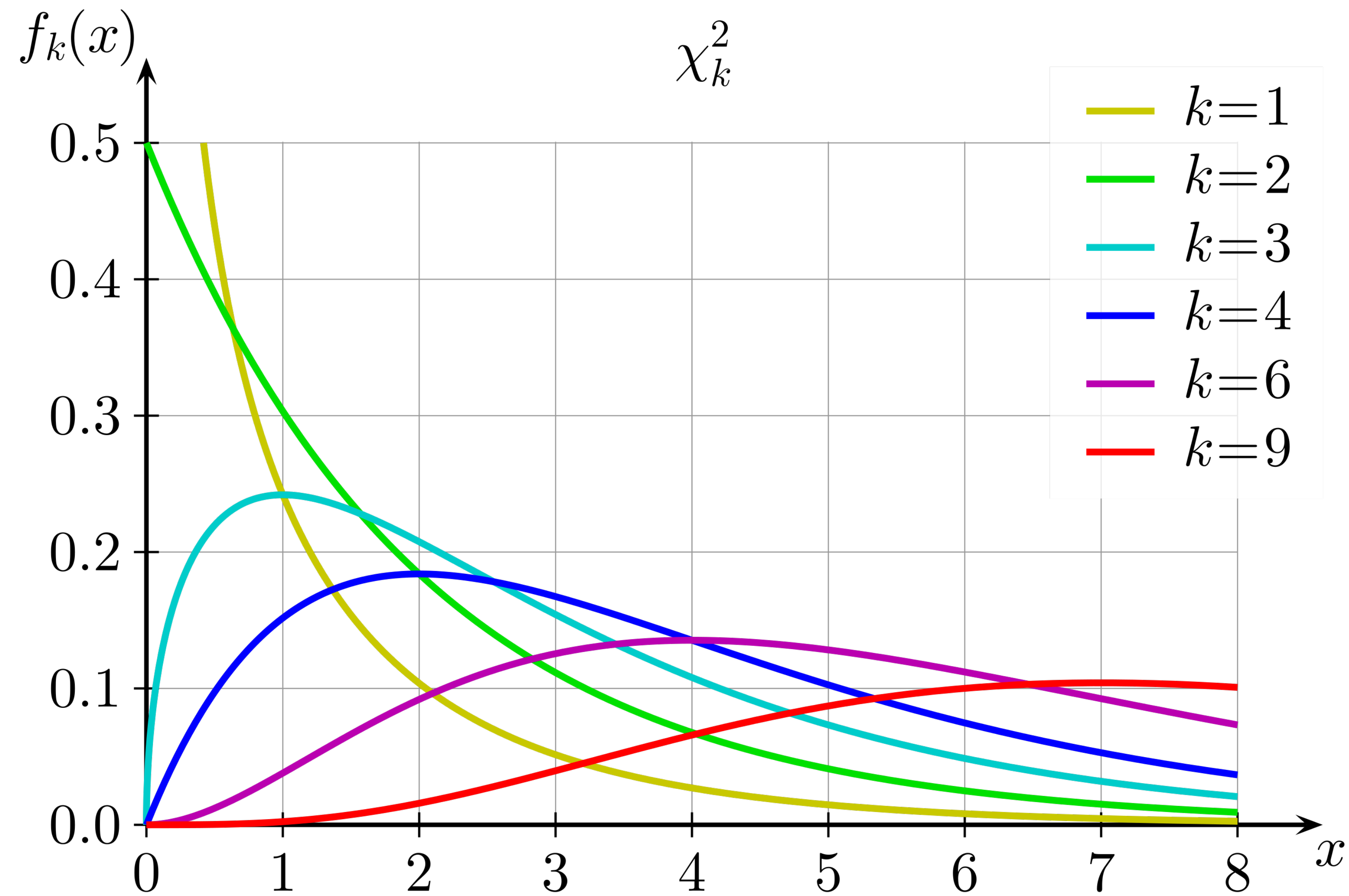
Sampling Distribution of Variance

- Recall that the sample variance of a sample x_1, \dots, x_n is defined as $s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2$
- Thought experiment: Suppose we know the population mean μ
- Rewrite the sample variance as (approximately) $s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \mu)^2$
- Now, examine s^2/σ^2
- We have $\frac{s^2}{\sigma^2} = \frac{1}{n-1} \cdot \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma} \right)^2$
- Therefore, $\frac{s^2}{\sigma^2} = \frac{1}{n-1} \cdot \sum_{i=1}^n Z_i^2$, where Z_i is a standard normal random variable

Chi-squared (χ^2) distribution

- However, $Q = \sum_{i=1}^n Z_i^2$ follows something called the **chi-squared distribution**
- $Q = \sum_{i=1}^k Z_i^2 \sim \chi_k^2$ (chi-squared distribution with k degrees of freedom)
- Mean of χ_k^2 is k (degrees of freedom)
- Variance of χ_k^2 is $2k$ (twice the degrees of freedom)

Chi-squared (χ^2) distribution



Sampling Distribution of Variance

- Returning to $\frac{s^2}{\sigma^2} = \frac{1}{n-1} \cdot \sum_{i=1}^n Z_i^2$, where Z_i is a standard normal random variable
- Recall that we assumed we knew μ ; we don't actually know this and have to use \bar{x} as an estimate
- This means that $(n-1) \cdot \frac{s^2}{\sigma^2} \sim \chi_{n-1}^2$ (chi-squared distribution with $n-1$ dof)

Hypothesis Tests for One Population Variance

- We have considered the case of comparing the *mean* of a population to a predetermined value
- We can also test whether the *variance* of a population is a specified value

Hypothesis Tests for One Population Variance

- Can run both two-sided and one-sided tests given a null σ_0^2
- One-tailed, lower hypothesis:
 - $H_0 : \sigma^2 \geq \sigma_0^2$ vs. $H_1 : \sigma^2 < \sigma_0^2$
- One-tailed, upper hypothesis:
 - $H_0 : \sigma^2 \leq \sigma_0^2$ vs. $H_1 : \sigma^2 > \sigma_0^2$
- Two-tailed hypothesis:
 - $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_1 : \sigma^2 \neq \sigma_0^2$

Hypothesis Tests for One Population Variance

- What is our test statistic?

- Recall that we know $(n - 1) \cdot \frac{s^2}{\sigma^2} \sim \chi_{n-1}^2$, so it makes sense to define

$$T = (n - 1) \cdot \frac{s^2}{\sigma^2} \text{ based on our null hypothesis}$$

- Under the null hypothesis, $\sigma^2 = \sigma_0^2$, we have the following:

$$T_{obs} = (n - 1) \cdot \frac{s^2}{\sigma_0^2} \sim \chi_{n-1}^2$$

- Once we have our test statistic, we can compare to the chi-square distribution to calculate a p-value

Hypothesis Tests for One Population Variance

- Finding p-values:
 - One-tailed, lower hypothesis: $p = \Pr(T < T_{obs})$
 - R: `pchisq(Tobs, n-1)`
 - One-tailed, upper hypothesis: $p = \Pr(T > T_{obs})$
 - R: `1-pchisq(Tobs, n-1)`
 - Two-tailed hypothesis: $p = \Pr(|T| \geq |T_{obs}|)$
 - If $T_{obs} \leq n - 1$, we have `2*pchisq(Tobs, n-1)`
 - If $T_{obs} > n - 1$, we have `2*(1-pchisq(Tobs, n-1))`

Hypothesis Tests for One Population Variance

- If the p-value is less than α , reject the null hypothesis and conclude that the population variances are unequal to each other
- If the p-value is greater than α , fail to reject the null hypothesis and conclude that the population variances are equal to each other

Hypothesis Tests for One Population Variance: Example

- Setup: In 2010, the distribution of the amount of time spent per table at The Cheesecake Factory is normally distributed with $\mu = 65$ minutes and $\sigma^2 = 121$ minutes². This year, we took a sample of $n = 45$ guests and found that the sample variance $s = 196$ minutes². **Has the variance changed?** Evaluate at $\alpha = 0.01$
- Hypotheses:
- Calculate the T statistic:
- Calculate the p-value:
- Conclusion:

Hypothesis Tests for One Population Variance: Example

- Setup: In 2010, the distribution of the amount of time spent per table at The Cheesecake Factory is normally distributed with $\mu = 65$ minutes and $\sigma^2 = 121$ minutes². This year, we took a sample of $n = 45$ guests and found that the sample variance $s = 196$ minutes². Has the variance changed? Evaluate at $\alpha = 0.01$
- Hypotheses: $H_0 : \sigma^2 = 121$ vs. $H_1 : \sigma^2 \neq 121$
- Calculate the T statistic: $T = (n - 1) \cdot \frac{s^2}{\sigma_0^2} = 44 \cdot \frac{196}{121} = 71.273$
- Calculate the p-value: $2 * (1 - \text{pchisq}(T_{\text{obs}}, n-1)) = 2 * (1 - \text{pchisq}(71.273, 44)) = 0.0115$
- Conclusion: Because $p = 0.0115 > \alpha = 0.01$, we fail to reject the null hypothesis and cannot conclude that the variance has changed since 2010 at $\alpha = 0.01$

Hypothesis Tests for One Population Variance: Example

- Setup: In 2010, the distribution of the amount of time spent per table at The Cheesecake Factory is normally distributed with $\mu = 65$ minutes and $\sigma^2 = 121$ minutes². This year, we took a sample of $n = 45$ guests and found that the sample variance $s^2 = 196$ minutes². Has the variance *increased*? Evaluate at $\alpha = 0.01$
- Hypotheses:
- Calculate the T statistic:
- Calculate the p-value:
- Conclusion:

Hypothesis Tests for One Population Variance: Example

- Setup: In 2010, the distribution of the amount of time spent per table at The Cheesecake Factory is normally distributed with $\mu = 65$ minutes and $\sigma^2 = 121$ minutes². This year, we took a sample of $n = 45$ guests and found that the sample variance $s^2 = 196$ minutes². Has the variance *increased*? Evaluate at $\alpha = 0.01$
- Hypotheses: $H_0 : \sigma^2 \leq 121$ vs. $H_1 : \sigma^2 > 121$
- Calculate the T statistic: $T = (n - 1) \cdot \frac{s^2}{\sigma_0^2} = 44 \cdot \frac{196}{121} = 71.273$
- Calculate the p-value: $1 - \text{pchisq}(T_{\text{obs}}, n - 1) = 1 - \text{pchisq}(71.273, 44) = 0.0057$
- Conclusion: Because $p = 0.0057 < \alpha = 0.01$, we can reject the null hypothesis and conclude that the variance has increased since 2010 at $\alpha = 0.01$

Confidence Intervals for One Population Variance

- We can also calculate $(1 - \alpha) \cdot 100\%$ (two-sided) confidence intervals for the population variance
- Because we have that $(n - 1) \cdot \frac{s^2}{\sigma^2} \sim \chi_{n-1}^2$, we have

$$\Pr\left(\chi_{\alpha/2, n-1}^2 \leq (n - 1) \cdot \frac{s^2}{\sigma^2} \leq \chi_{1-\alpha/2, n-1}^2\right) = 1 - \alpha$$

$$\Rightarrow \Pr\left(\frac{1}{\chi_{\alpha/2, n-1}^2} \geq \frac{\sigma^2}{(n - 1) \cdot s^2} \geq \frac{1}{\chi_{1-\alpha/2, n-1}^2}\right) = 1 - \alpha$$

$$\Rightarrow \Pr\left(\frac{(n - 1) \cdot s^2}{\chi_{1-\alpha/2, n-1}^2} \leq \sigma^2 \leq \frac{(n - 1) \cdot s^2}{\chi_{\alpha/2, n-1}^2}\right) = 1 - \alpha$$

- Therefore, the interval $\left(\frac{(n - 1) \cdot s^2}{\chi_{1-\alpha/2, n-1}^2}, \frac{(n - 1) \cdot s^2}{\chi_{\alpha/2, n-1}^2}\right)$ contains σ^2 with probability $1 - \alpha$

Confidence Intervals for One Population Variance

- Similarly, one-sided $(1 - \alpha) \cdot 100\%$ confidence intervals for the population variance are as follows

- One-sided, upper:

- $\left(0, \frac{(n - 1) \cdot s^2}{\chi^2_{\alpha, n-1}} \right)$

- One-sided, lower:

- $\left(\frac{(n - 1) \cdot s^2}{\chi^2_{1-\alpha, n-1}}, \infty \right)$

Confidence Intervals for One Population Variance: Example

- Let's say we are interested in the amount of water the average person drinks per day. We sample a group of $n = 40$ people and find that the sample variance is $s^2 = 60 \text{ oz}^2$. What is a two-tailed 95% confidence interval for the variance of this distribution?
- What is α ?
- CI:

Confidence Intervals for One Population Variance: Example

- Let's say we are interested in the amount of water the average person drinks per day. We sample a group of $n = 40$ people and find that the sample variance is $s^2 = 60$ oz². What is a two-tailed 95% confidence interval for the variance of this distribution?
- What is α ? $\alpha = 0.05$
- CI:
$$\left(\frac{(n-1) \cdot s^2}{\chi^2_{1-\alpha/2, n-1}}, \frac{(n-1) \cdot s^2}{\chi^2_{\alpha/2, n-1}} \right) = \left(\frac{39 \cdot 60}{\chi^2_{0.975, 39}}, \frac{39 \cdot 60}{\chi^2_{0.025, 39}} \right) = (40.26, 98.92)$$

Hypothesis Tests for Two Population Variances

- Consider the case of wanting to compare two populations
- We have talked about testing whether the means of the two independent populations are equal to each other
- We can also test whether the variances of the two populations are equal to each other

Hypothesis Tests for Two Population Variances

- Null hypothesis: Population variances are equal
- We attempt to see if there is a difference in variances
- One-tailed, lower hypothesis:
 - $H_0 : \sigma_1^2 \geq \sigma_2^2$ vs. $H_1 : \sigma_1^2 < \sigma_2^2$
- One-tailed, upper hypothesis:
 - $H_0 : \sigma_1^2 \leq \sigma_2^2$ vs. $H_1 : \sigma_1^2 > \sigma_2^2$
- Two-tailed hypothesis:
 - $H_0 : \sigma_1^2 = \sigma_2^2$ vs. $H_1 : \sigma_1^2 \neq \sigma_2^2$

Hypothesis Tests for Two Population Variances

- Recall that for a single population, we have $\frac{s^2}{\sigma^2} \sim \frac{1}{n-1} \cdot \chi_{n-1}^2$
- When comparing two populations, it makes sense to look at the quantity $\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$
- However, from the above, we have that $\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim \frac{\chi_{n_1-1}^2/(n_1-1)}{\chi_{n_2-1}^2/(n_2-1)}$
- It turns out that this type of distribution has a name: F distribution

Hypothesis Tests for Two Population Variances

- To test such hypotheses, we use an F test statistic:

$$F_{obs} = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F_{n_1-1, n_2-1}$$

- This F statistic follows an F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom
- Under the null hypothesis, $\sigma_1^2 = \sigma_2^2$, we have the following:

$$F_{obs} = \frac{s_1^2}{s_2^2} \sim F_{n_1-1, n_2-1}$$

- Once we have our test statistic, we can compare to the F distribution to calculate a p-value

Hypothesis Tests for Two Population Variances

- Finding p-values:
 - One-tailed, lower hypothesis: $p = \Pr(F < F_{obs})$
 - R: `pf (F_obs, n1-1, n2-1)`
 - One-tailed, upper hypothesis: $p = \Pr(F > F_{obs})$
 - R: `1-pf (F_obs, n1-1, n2-1)`
 - Two-tailed hypothesis: $p = \Pr(|T| \geq |T_{obs}|)$
 - If $F_{obs} \leq 1$, we have `2*pf (F_obs, n1-1, n2-1)`
 - If $F_{obs} > 1$, we have `2* (1-pf (F_obs, n1-1, n2-1))`

Hypothesis Tests for Two Population Variances

- If the p-value is less than α , reject the null hypothesis and conclude that the population variances are unequal to each other
- If the p-value is greater than α , fail to reject the null hypothesis and conclude that the population variances are equal to each other

Hypothesis Tests for Two Population Variances

- Instead of p-values, we can use critical values for rejection
- Let $F_{n_1-1, n_2-1, \alpha}$ be the critical value for an F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom and α in the lower tail
- One-tailed, lower hypothesis:
 - Reject if $F_{obs} \leq F_{n_1-1, n_2-1, \alpha}$
- One-tailed, upper hypothesis:
 - Reject if $F_{obs} \geq F_{n_1-1, n_2-1, \alpha}$
- Two-tailed hypothesis:
 - Reject if $F_{obs} \leq F_{n_1-1, n_2-1, \alpha/2}$ or $F_{obs} \geq F_{n_1-1, n_2-1, 1-\alpha/2}$

Hypothesis Tests for Two Population Variances

- We are interested in exploring the fuel efficiency, measured in miles per gallon (mpg), for two models of mass-produced cars: Honda Civics and Honda Accords
- In particular, we want to explore the hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 \neq \sigma_2^2$
- Take a sample of $n_1 = 10$ Civics and $n_2 = 20$ Accords
- Sample variances: $s_1^2 = 2.33$ and $s_2^2 = 9.97$
- Evaluate at $\alpha = 0.05$

Hypothesis Tests for Two Population Variances

- Calculate the test statistic:
- Conclusion:
- Similarly, we could have calculated the critical values for the F-test statistic and used them to complete the test:

$$F_{n_1-1, n_2-1, \alpha/2} = \text{qf}(0.025, 9, 19) = 0.2715$$

$$F_{n_1-1, n_2-1, 1-\alpha/2} = \text{qf}(0.975, 9, 19) = 2.880$$

- Since $F = 0.2337 < 0.2715$, we reject the null hypothesis and conclude $\sigma_1^2 \neq \sigma_2^2$

Hypothesis Tests for Two Population Variances

- Calculate the test statistic:

$$F = \frac{2.33}{9.97} = 0.2337$$

$$2 \cdot \Pr(F < 0.2337) = 2 * \text{pf}(0.2337, 9, 19) = 0.03$$

- Since $p = 0.03 < \alpha = 0.05$, we reject the null hypothesis and conclude $\sigma_1^2 \neq \sigma_2^2$
- Similarly, we could have calculated the critical values for the F-test statistic and used them to complete the test:

$$F_{n_1-1, n_2-1, \alpha/2} = \text{qf}(0.025, 9, 19) = 0.2715$$

$$F_{n_1-1, n_2-1, 1-\alpha/2} = \text{qf}(0.975, 9, 19) = 2.880$$

- Since $F = 0.2337 < 0.2715$, we reject the null hypothesis and conclude $\sigma_1^2 \neq \sigma_2^2$

Confidence Intervals for Two Population Variances

- We can also calculate confidence intervals for the ratio of two population variances
- A two-sided $(1 - \alpha) \cdot 100\%$ confidence interval for σ_1^2/σ_2^2 is given by
$$\left(\frac{1}{F_{1-\alpha/2}} \cdot \frac{s_1^2}{s_2^2}, \frac{1}{F_{\alpha/2}} \cdot \frac{s_1^2}{s_2^2} \right)$$
- A one-sided upper $(1 - \alpha) \cdot 100\%$ confidence interval for σ_1^2/σ_2^2 is given by
$$\left(0, \frac{1}{F_{\alpha}} \cdot \frac{s_1^2}{s_2^2} \right)$$
- A one-sided lower $(1 - \alpha) \cdot 100\%$ confidence interval for σ_1^2/σ_2^2 is given by
$$\left(\frac{1}{F_{1-\alpha}} \cdot \frac{s_1^2}{s_2^2}, \infty \right)$$

Confidence Intervals for Two Population Variances: Example

- Return to the fuel efficiency example (sample of $n_1 = 10$ Civics and $n_2 = 20$ Accords, sample variances: $s_1^2 = 2.33$ and $s_2^2 = 9.97$). What is a one-tailed upper 95% confidence interval for the ratio of variances?
- What is α ?
- CI:

Confidence Intervals for Two Population Variances: Example

- Return to the fuel efficiency example (sample of $n_1 = 10$ Civics and $n_2 = 20$ Accords, sample variances: $s_1^2 = 2.33$ and $s_2^2 = 9.97$). What is a one-tailed upper 95% confidence interval for the ratio of variances?
- What is α ? $\alpha = 0.05$
- CI: $\left(0, \frac{1}{F_\alpha} \cdot \frac{s_1^2}{s_2^2}\right) = \left(0, \frac{1}{0.339} \cdot \frac{2.33}{9.97}\right) = (0, 0.689)$