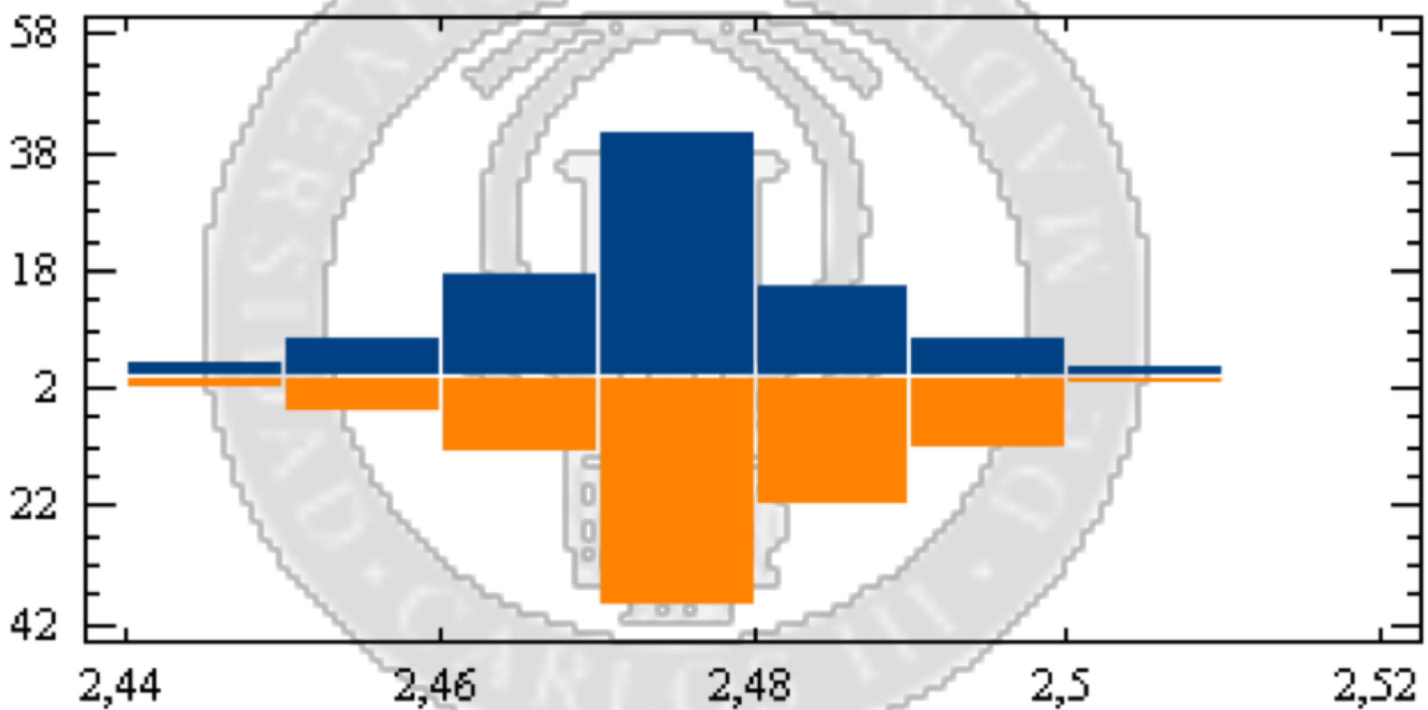
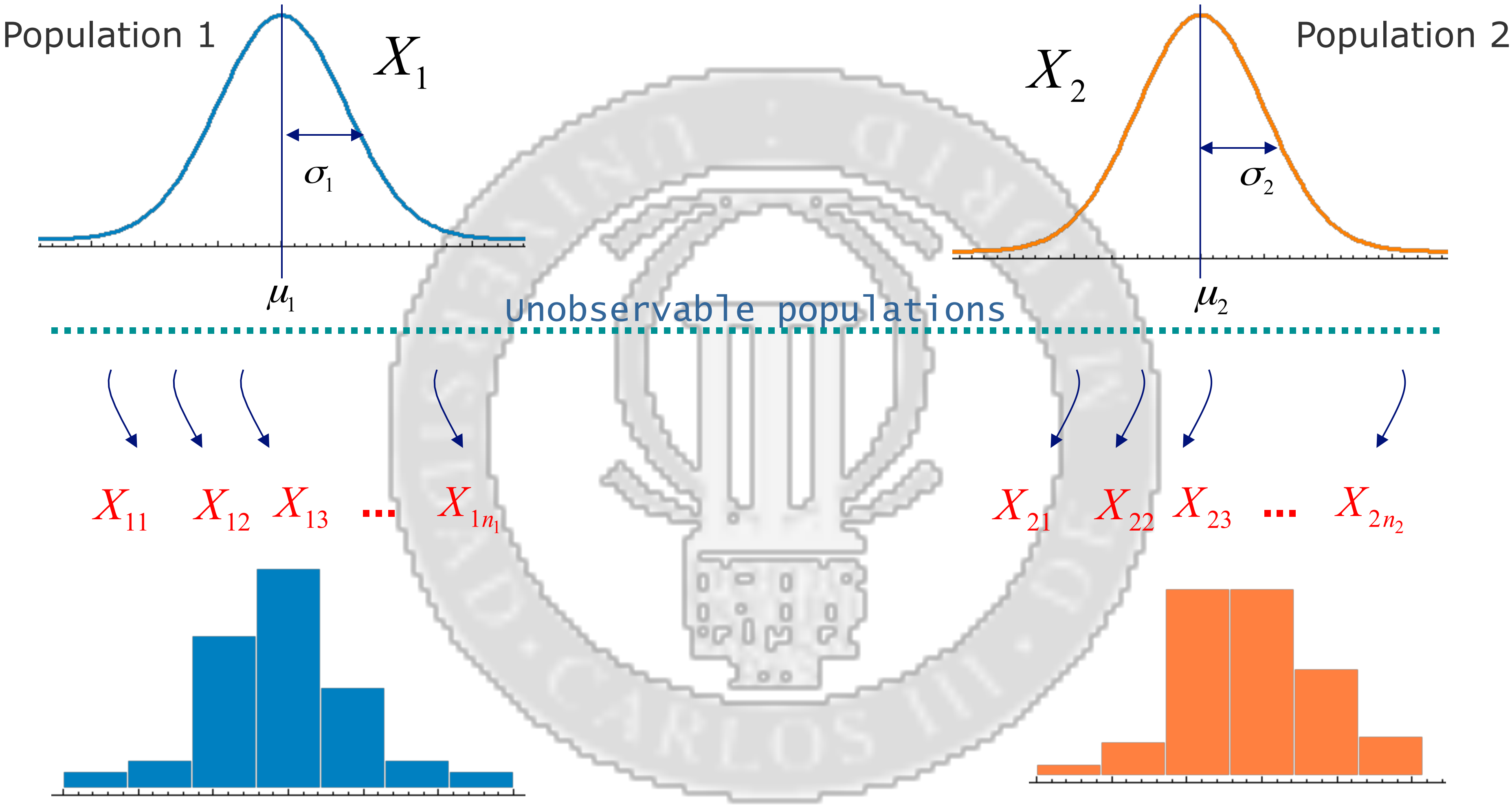
8. Comparison of Populations



Chapter 8: Comparison of Populations

1. Introduction
2. Comparing two populations means: Independent samples
3. Comparing two populations means: Paired data
4. Comparing two population proportions
5. Comparing two populations variances (normal populations)



$\bar{x}_1$

$\hat{s}_1^2$

$$¿\mu_1 = \mu_2?$$

$$¿\sigma_1^2 = \sigma_2^2?$$

$\bar{x}_2$

$\hat{s}_2^2$

Example

We consider two samples of bearings from two different manufactures and we measure their resistance.

Are these two types of bearings different?

Example

There are two different access systems for connecting to a network. We measure some connecting times from each system.

Which system is the fastest one?

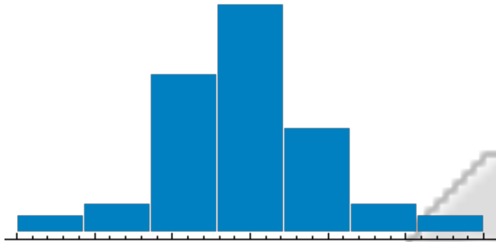
Example

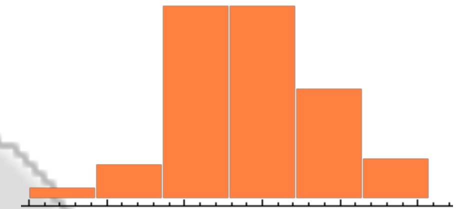
We analyze the weight of newborns in a Hospital during the eastern holidays. Looking at these data:

Are male newborns heavier than females?

Chapter 8: Comparison of Populations

2. Comparing two populations means: Independent samples





$$\bar{X}_1 = \frac{\sum_{i=1}^{n_1} X_{1i}}{n_1} \qquad \hat{S}_1^2 = \frac{\sum_{i=1}^{n_1} \left(X_{1i} - \bar{X}_1\right)^2}{n_1 - 1} \qquad \bar{X}_2 = \frac{\sum_{i=1}^{n_2} X_{2i}}{n_2} \qquad \hat{S}_2^2 = \frac{\sum_{i=1}^{n_2} \left(X_{2i} - \bar{X}_2\right)^2}{n_2 - 1}$$

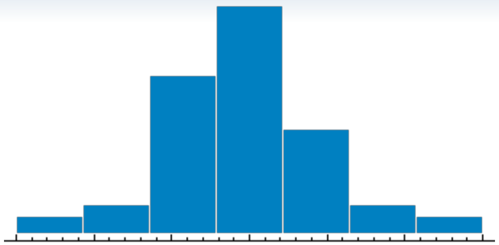
$$¿\mu_1 - \mu_2?$$

Considering normal populations or large samples...

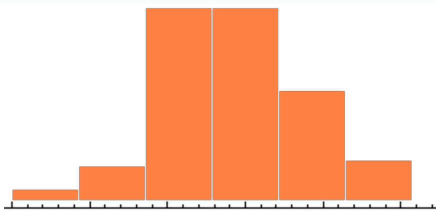
$$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \qquad \bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

$$\boxed{\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}$$

2. Comparing two populations means: Independent samples



$$¿\mu_1 - \mu_2?$$



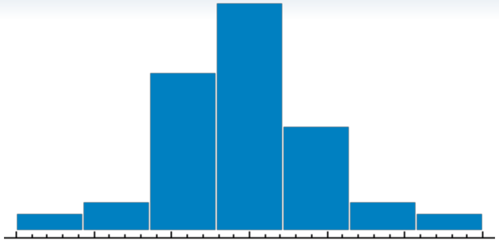
$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \longrightarrow \boxed{\frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)}$$

$$\Downarrow$$

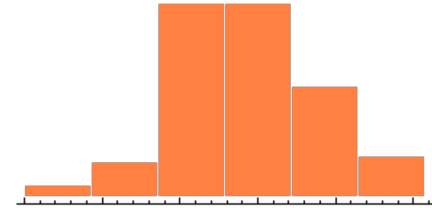
Confidence interval

$$\boxed{IC(1-\alpha): \underbrace{\mu_1 - \mu_2}_{\text{parameter}} \in \left\{ \underbrace{\bar{x}_1 - \bar{x}_2}_{\text{estimators}} \pm \underbrace{z_{\alpha/2}}_{\text{tables}} \underbrace{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}_{\text{Standard deviation of the estimator}} \right\}}$$

2. Comparing two populations means: Independent samples



$$¿\mu_1 - \mu_2?$$



$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \longrightarrow \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

$$\Downarrow$$

Confidence interval

$$\boxed{IC(1-\alpha): \mu_1 - \mu_2 \in \left\{\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right\}}$$

$$\boxed{IC(1-\alpha): \mu_1 - \mu_2 \in \left\{\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right\}} \longleftarrow \text{If } \sigma_1^2 = \sigma_2^2$$

Large samples

$$IC(1-\alpha): \mu_1 - \mu_2 \in \left\{ \bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}} \right\}$$

$$IC(1-\alpha): \mu_1 - \mu_2 \in \left\{ \bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2} \hat{s}_T \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right\}$$

$$\hat{S}_T^2 = \frac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2}$$

With large samples, the normal approximation is still valid if we replace the parameters by their estimators.

Normal populations
(small samples)

$$\sigma_1^2 \neq \sigma_2^2$$

$$IC(1-\alpha): \mu_1 - \mu_2 \in \left\{ \bar{x}_1 - \bar{x}_2 \pm t_{v;\alpha/2} \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}} \right\}$$

$$v \approx \frac{\left(\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2} \right)^2}{\frac{1}{n_1 - 1}\left(\frac{\hat{S}_1^2}{n_1} \right)^2 + \frac{1}{n_2 - 1}\left(\frac{\hat{S}_2^2}{n_2} \right)^2}$$

$$\sigma_1^2 = \sigma_2^2$$

$$IC(1-\alpha): \mu_1 - \mu_2 \in \left\{ \bar{x}_1 - \bar{x}_2 \pm t_{n_1+n_2-2;\alpha/2} \hat{s}_T \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right\}$$

Example

We want to choose between two types of textile materials to produce mooring systems. To this end, we measure the breaking stress of several ropes made of two different materials. We consider a sample of 24 ropes made of material M1, resulting in $\bar{x}_1 = 87$ and $\hat{s}_1 = 2$. Similarly 30 ropes made of material M2 are analyzed resulting $\bar{x}_2 = 75$ and $\hat{s}_2 = 2.3$. Additionally, we know that the breaking stresses follow a normal distribution and it is assume that the two variance populations are the same.

If the variances are the same and we consider a small sample, but the populations are normal then:

$$IC(1-\alpha): \mu_1 - \mu_2 \in \left\{ \bar{x}_1 - \bar{x}_2 \pm t_{n_1+n_2-2;\alpha/2} \hat{s}_T \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right\}$$

$$t_{n_1+n_2-2;\alpha/2} = t_{52;0.025} = 2.0$$

$$\hat{s}_T^2 = \frac{23 \times 2^2 + 29 \times 2.1^2}{23+29} = 4.2 \Rightarrow \hat{s}_T = 2.06$$

$$IC(0.95): \mu_1 - \mu_2 \in \left\{ 87 - 75 \pm 2 \times 2.06 \times \sqrt{\frac{1}{24} + \frac{1}{30}} \right\} = (12 \pm 1.13)$$

- There is evidence in favor of M1 (the interval does not include 0)
- On average, M1 excels M2 a value between 10.87 and 13.13 units (95% confidence level).

Hypothesis test:

Step 1:

$$H_0 : \mu_1 = \mu_2 ; H_1 : \mu_1 \neq \mu_2$$
(a)

$$H_0 : \mu_1 \leq \mu_2 ; H_1 : \mu_1 > \mu_2$$
(b)

$$H_0 : \mu_1 \geq \mu_2 ; H_1 : \mu_1 < \mu_2$$
(c)

Step 2:

$\sigma_1^2 = \sigma_2^2$

$$Z_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightarrow N(0,1)$$

$$T_0 = \frac{\bar{X}_1 - \bar{X}_2}{\hat{S}_T \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightarrow t_{n_1+n_2-2}$$

$\sigma_1^2 \neq \sigma_2^2$

$$Z_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightarrow N(0,1)$$

$$T_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}} \rightarrow t_v$$

Step 3:

Large samples $\rightarrow$ N(0,1)

Normal populations

N(0,1)

$t_{n_1+n_2-2}$

N(0,1)

t_v

$$v \approx \frac{\left(\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{\hat{S}_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{\hat{S}_2^2}{n_2}\right)^2}$$

Hypothesis test

Step 4: Rejection region

Step 1:

$H_0 : \mu_1 = \mu_2 ; H_1 : \mu_1 \neq \mu_2$

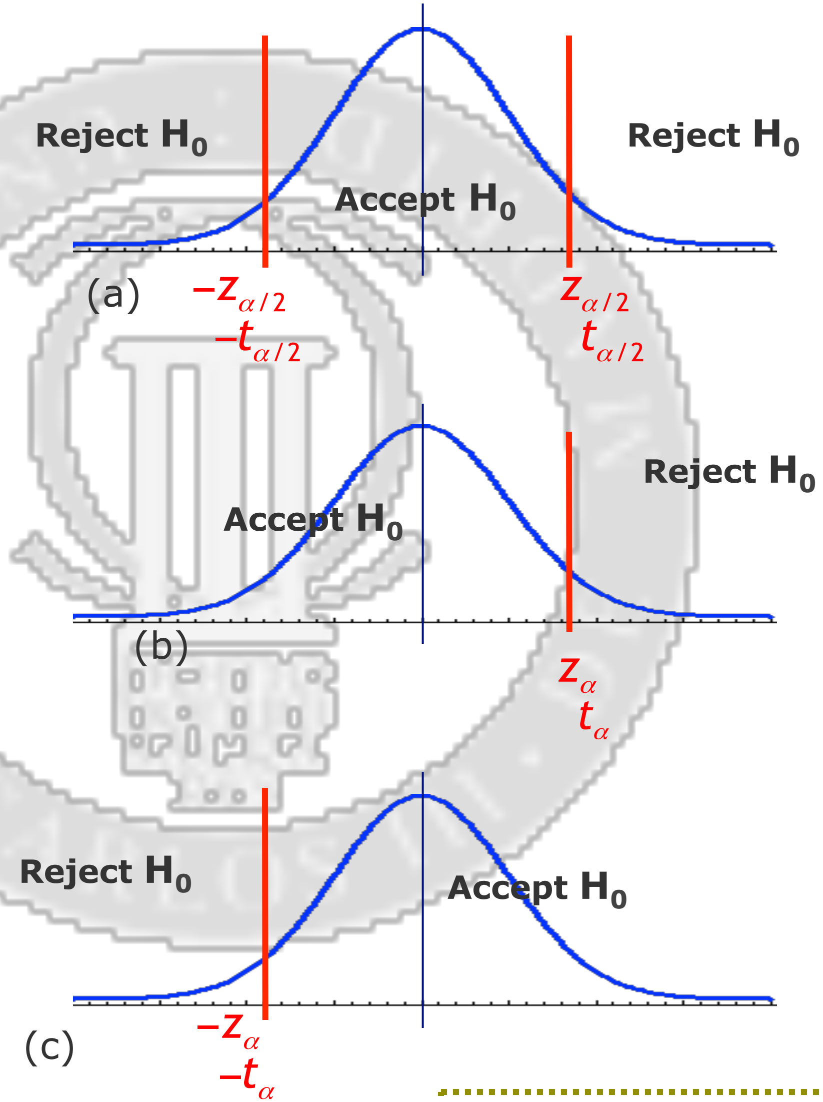
(a)

$H_0 : \mu_1 \leq \mu_2 ; H_1 : \mu_1 > \mu_2$

(b)

$H_0 : \mu_1 \geq \mu_2 ; H_1 : \mu_1 < \mu_2$

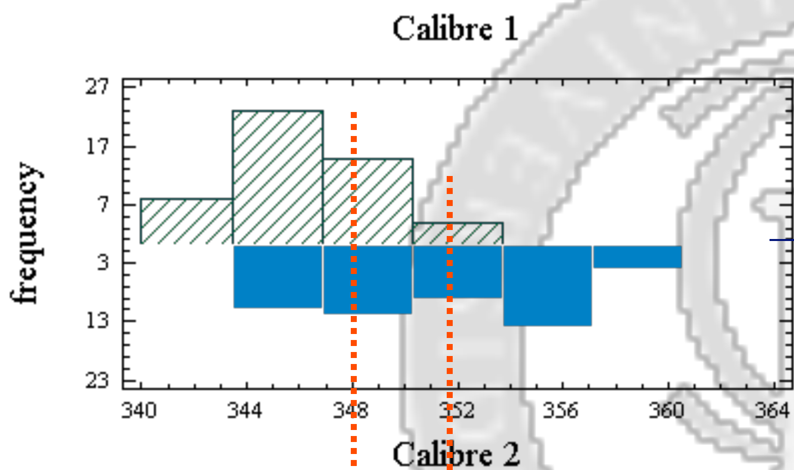
(c)



The rejection region is where H_1 is accepted

Example

We seek to compare the precision of two different calibers. To this end, we compare the measurements of 100 nails which belong to the same production batch. 50 nails are measured with one caliber and 50 nails are measured with the other one. How are the average measurements resulting from each of the calibers?



All the nails are of the same type. Thus, the observed differences are not caused by the nails' characteristics.

Is this difference important?

$\bar{x}_1 \neq \bar{x}_2 \longrightarrow$ ¿$\mu_1 \neq \mu_2$?

$$\boxed{H_0 : \mu_1 = \mu_2 ; H_1 : \mu_1 \neq \mu_2}$$

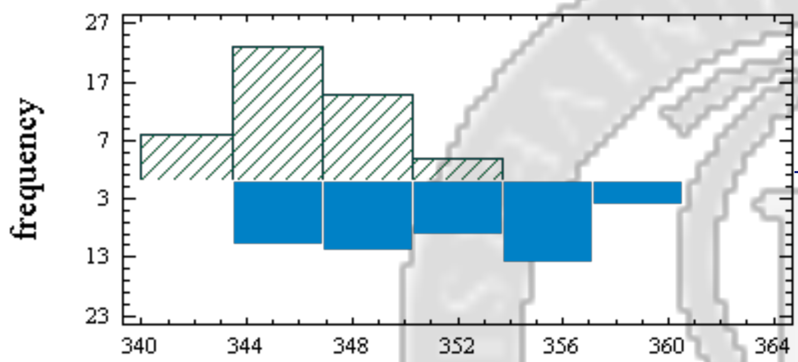
Large samples.

The variances might be different

$$\boxed{T_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}}} \qquad \boxed{T_0 \sim N(0,1)}$$

Example

We seek to compare the precision of two different calibers. To this end, we compare the measurements of 100 nails which belong to the same production batch. 50 nails are measured with one caliber and 50 nails are measured with the other one. How are the average measurements resulting from each of the calibers?



All the nails are of the same type. Thus, the observed differences are not caused by the nails' characteristics.

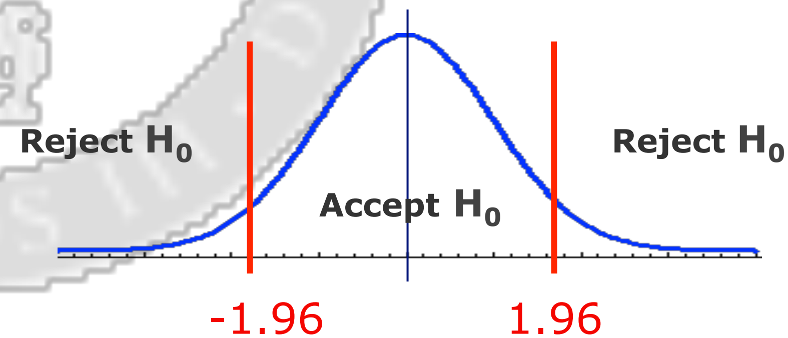
Is this difference important?

$$H_0 : \mu_1 = \mu_2; H_1 : \mu_1 \neq \mu_2$$

$$\bar{x}_1 = 346.16; \hat{s}_1^2 = 7.40$$

$$\bar{x}_2 = 351.12; \hat{s}_2^2 = 21.90$$

$$t_0 = \frac{346.16 - 351.12}{\sqrt{\frac{7.40}{50} + \frac{21.90}{50}}} = -6.48.$$



Reject H_0

The difference between the means is sufficiently relevant

Chapter 8: Comparison of Populations

1. Introduction
2. Comparing two populations means: Independent samples
3. Comparing two populations means: Paired data
4. Comparing two population proportions
5. Comparing two populations variances (normal populations)

3. Comparing two populations means: Paired data

From each element: 2 data

Examples:
- Before/after introducing a modification
- Before/after a treatment
- Different measure devices

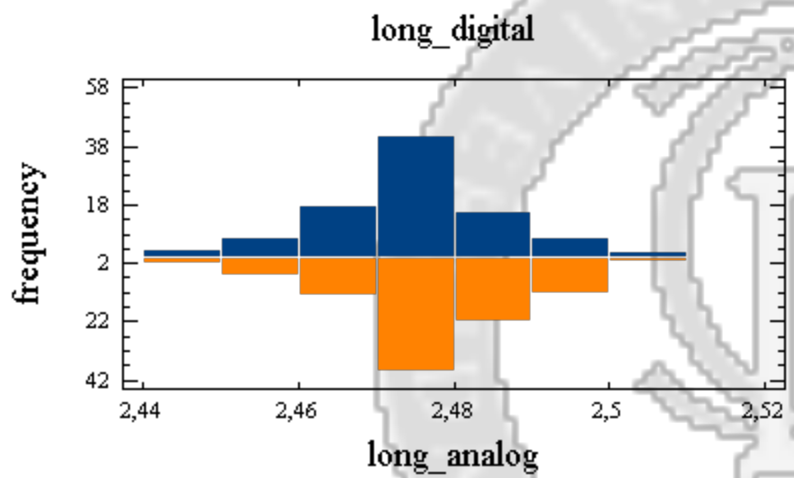
	X_1	X_2	Y
• Element 1	X_{11}	X_{21}	$Y_1 = X_{11} - X_{21}$
• Element 2	X_{12}	X_{22}	$Y_2 = X_{12} - X_{22}$
• Element 3	X_{13}	X_{23}	$Y_3 = X_{13} - X_{23}$
…	…	…	…
• Element n	X_{1n}	X_{2n}	$Y_n = X_{1n} - X_{2n}$
	$\downarrow$	$\downarrow$	$\downarrow$
	μ_1	μ_2	μ_Y

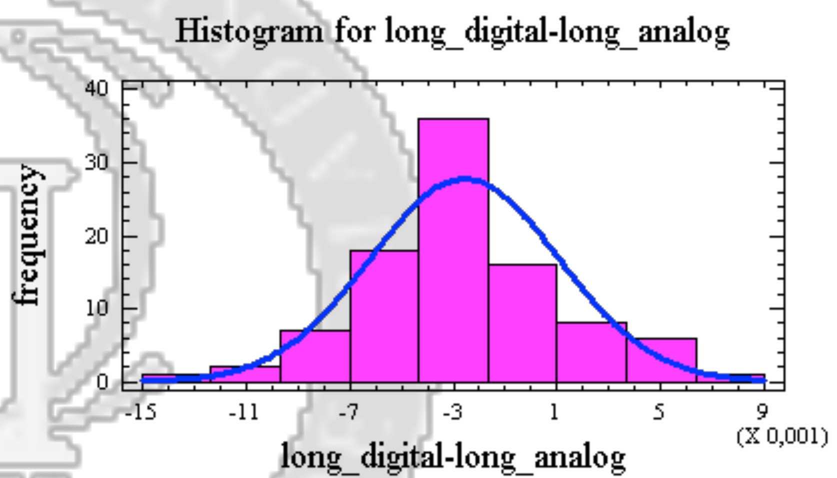
$$\boxed{¿\mu_1 = \mu_2? \longrightarrow ¿\mu_Y = 0?} \longrightarrow \text{Like the previous sections}$$

Ejemplo

We seek to compare the precision of two different calibers; an analogic one and a digital one. To this end, we measure the length of 95 nails of the same type. Note that each nail has been measure two times, one with the analogic one (less accurate) and one with the digital one (very accurate).

Are there differences in the measurements?





Y= difference between the digital and the analogical measurements

$$\boxed{H_0 : \mu_y = 0; H_1 : \mu_y \neq 0}$$

Test statistic $$\boxed{T_0 = \frac{\bar{y} - 0}{\hat{S}_y / \sqrt{n}}}$$

Since it is a large sample $\longrightarrow$ $\boxed{T_0 \sim N(0,1)}$

Ejemplo

We seek to compare the precision of two different calibers; an analogic one and a digital one. To this end, we measure the length of 95 nails of the same type. Note that each nail has been measure two times, one with the analogic one (less accurate) and one with the digital one (very accurate).

Are there differences in the measurements?

$$H_0 : \mu_y = 0; H_1 : \mu_y \neq 0$$

$$T_0 = \frac{\bar{y} - 0}{\hat{S}_y / \sqrt{n}}$$

$$\bar{y} = -0.00256; \hat{s}_y' = 0.00364$$

$$t_0 = \frac{\bar{y} - \mu_0}{\hat{s}_y / \sqrt{n}} = \frac{-0.00256 - 0}{0.00364 / \sqrt{95}} = -6.8$$

$$T_0 \sim N(0,1)$$

$$z_{0.025} = 1.96$$

Since $|t_0| > 1.96$ Reject H_0

The observe mean difference is small but relevant.

Chapter 8: Comparison of Populations

1. Introduction
2. Comparing two populations means: Independent samples
3. Comparing two populations means: Paired data
4. **Comparing two population proportions**
5. Comparing two populations variances (normal populations)

4. Comparing two population proportions

Population 1

Proportion of elements with a certain characteristic

p_1

Population 2

Proportion of elements with a certain characteristic

p_2

sample n1

$\hat{p}_1$

sample n2

$\hat{p}_2$

¿$p_1 = p_2$?

4. Comparing two population proportions

Population 1

Proportion of elements with a certain characteristic

p_1

sample n1

$\hat{p}_1$

$$\hat{p}_1 \sim N\left(p_1, \frac{p_1 q_1}{n_1}\right)$$

Population 2

Proportion of elements with a certain characteristic

p_2

sample n2

$\hat{p}_2$

$$\hat{p}_2 \sim N\left(p_2, \frac{p_2 q_2}{n_2}\right)$$

$$\boxed{\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right)}$$

Confidence interval

$$IC(1-\alpha): p_1 - p_2 \in \left\{\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}\right\}$$

Hypothesis test

Step 1:

$H_0: p_1 = p_2; H_1: p_1 \neq p_2$

$H_0: p_1 \leq p_2; H_1: p_1 > p_2$

$H_0: p_1 \geq p_2; H_1: p_1 < p_2$

Step 2:

$$Z_0 = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}_0\hat{q}_0\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

with

$$\hat{p}_0 = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Step 3:

Large samples $\rightarrow$ N(0,1)

Step 4:

The rejection region is where H_1 is accepted

Example

Is the proportion of males and females students from industrial engineering that pass the statistics course the same?

We take a sample of students: June 2003 exam

Students from 1º year of I. Eng.

270 students

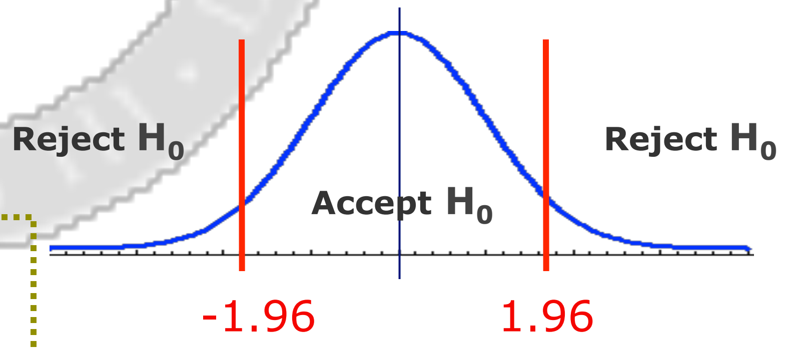
225 males. 30% pass the exam

45 females. 42% pass the exam

$$\boxed{\hat{p}_0 = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{225 \times 0.30 + 45 \times 0.42}{225 + 45} = 0.32}$$

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0 \hat{q}_0 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{0.30 - 0.42}{\sqrt{0.32 \times 0.68 \left(\frac{1}{225} + \frac{1}{45} \right)}} = -1.57$$

Since $|z_0| < z_{0.025} = 1.96$



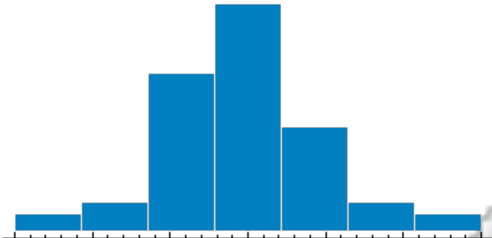
The samples difference is not relevant enough (considering 5% confidence level).

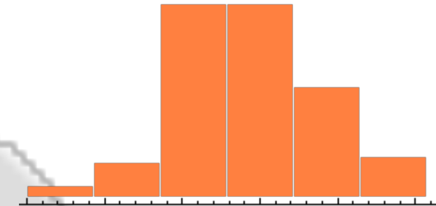
We cannot reject that the two genders have the same probability to pass the exam.

Chapter 8: Comparison of Populations

1. Introduction
2. Comparing two populations means: Independent samples
3. Comparing two populations means: Paired data
4. Comparing two population proportions
5. **Comparing two populations variances (normal populations)**

5. Comparing two populations variances (normal populations)





$$\hat{S}_1^2 = \frac{\sum_{i=1}^{n_1} \left(X_{1i} - \bar{X}_1 \right)^2}{n_1 - 1}$$

$$\hat{S}_2^2 = \frac{\sum_{i=1}^{n_2} \left(X_{2i} - \bar{X}_2 \right)^2}{n_2 - 1}$$

$$¿\sigma_1^2 / \sigma_2^2 ?$$

Considering normal populations

$$F = \frac{\hat{S}_1^2 / \sigma_1^2}{\hat{S}_2^2 / \sigma_2^2} \sim \boxed{F_{n_1 - 1, n_2 - 1}}$$

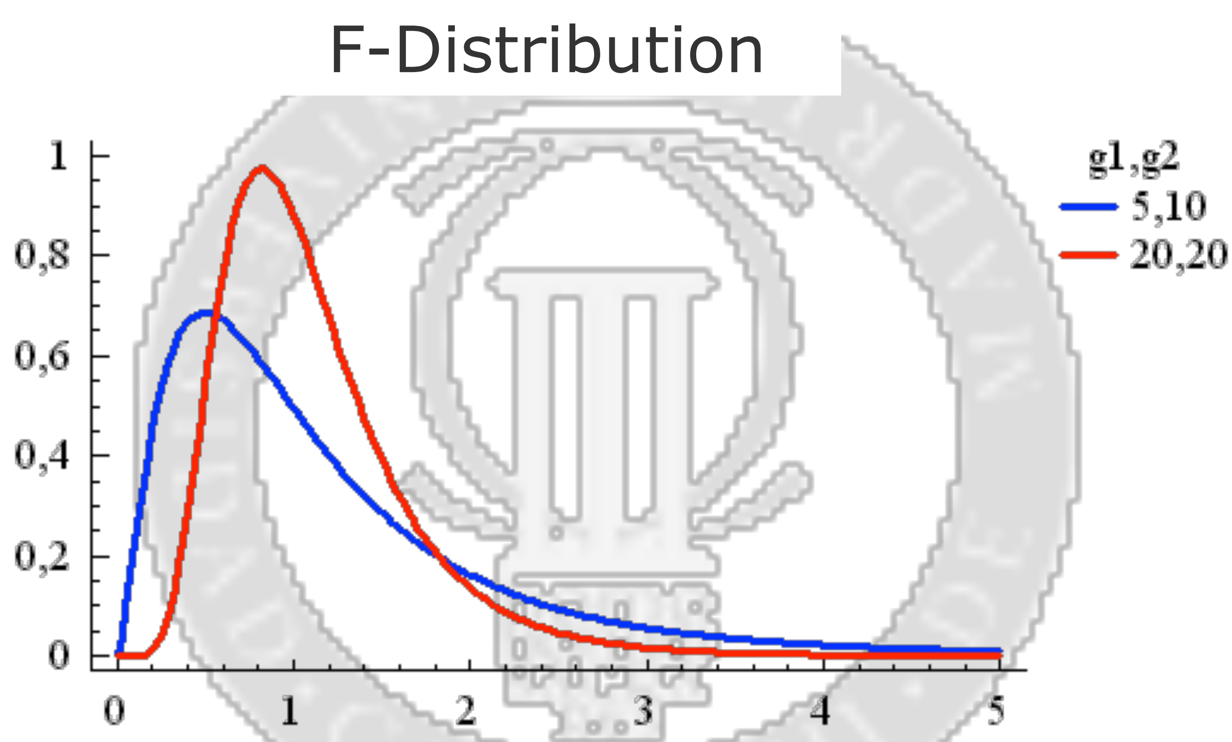
Fisher's F-Distribution

$$F_{g_1, g_2}$$

Numerator's degrees of freedom

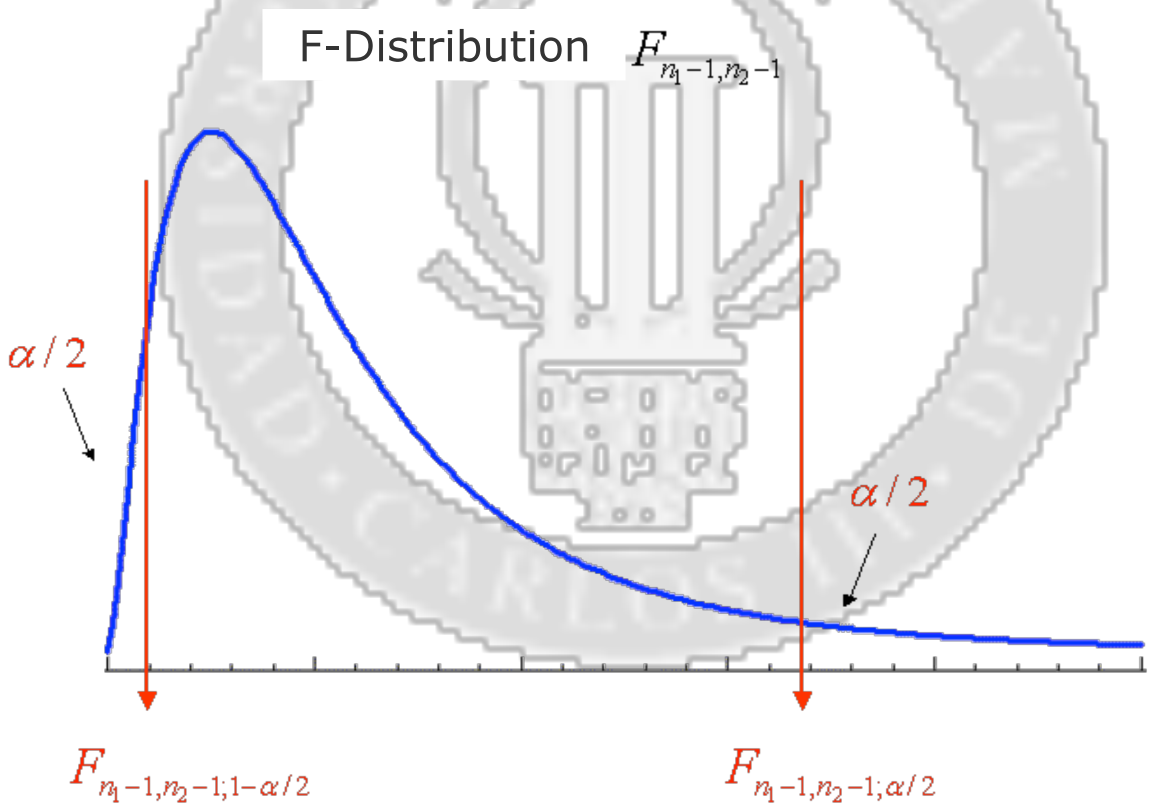
Denomiator's degrees of freedom

5. Comparing two populations variances (normal populations)



- Similar to Chi-square distribution.
- The asymmetry increases with the degrees of freedom.
- The mode is close to 1.

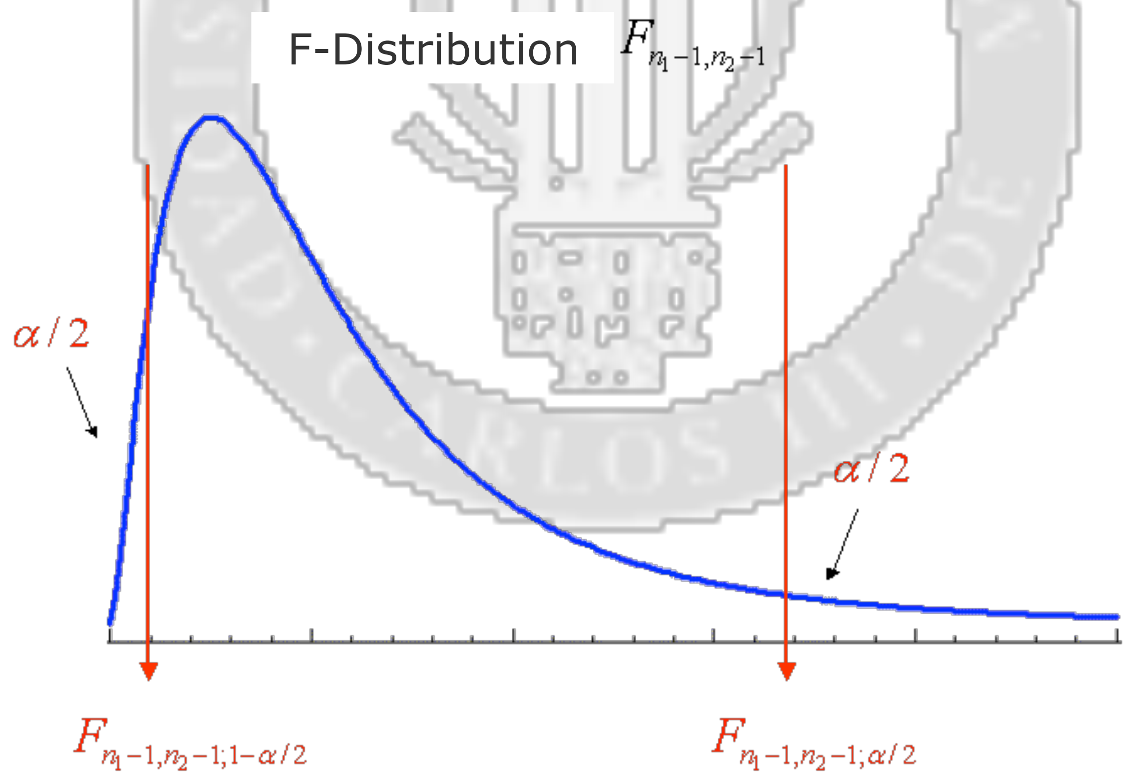
$$F = \frac{\hat{S}_1^2 / \sigma_1^2}{\hat{S}_2^2 / \sigma_2^2} \sim F_{n_1-1, n_2-1}$$



5. Comparing two populations variances (normal populations)

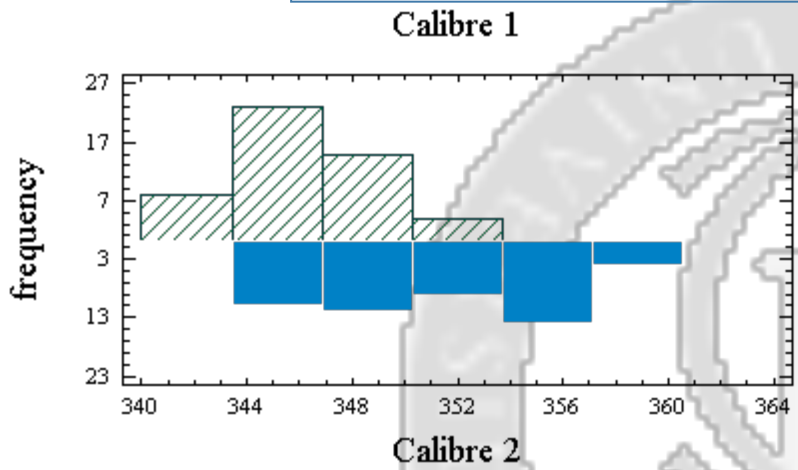
Confidence intervals

$$IC(1-\alpha):\ \frac{\sigma_1^2}{\sigma_2^2} \in \left\{ \frac{\hat{S}_1^2}{\hat{S}_2^2} F_{n_2-1,n_1-1;1-\alpha/2};\ \frac{\hat{S}_1^2}{\hat{S}_2^2} F_{n_2-1,n_1-1;\alpha/2} \right\}$$



Example

We seek to compare the precision of two different calibers. To this end, we compare the measurements of 100 nails which belong to the same production batch. 50 nails are measured with one caliber and 50 nails are measured with the other one. How are the average measurements resulting from each of the calibers?



$$\hat{s}_1^2 = 7.4 \qquad \hat{s}_2^2 = 21.9$$

$$F_{49,49;0.975} = 0.57$$

$$F_{49,49;0.025} = 1.76$$

$$IC(0.95) : \frac{\sigma_1^2}{\sigma_2^2} \in \left\{ \frac{7.4}{21.9} 0.57 ; \frac{7.4}{21.9} 1.76 \right\} = \{0.193 ; 0.595\}$$

- The interval is far from including the value 1.
- There is a lot evicence showing that the variances are different.
- The precision of the first caliber if much better that the second one.

5. Comparing two populations variances (normal populations)

Hypothesis test

$$F = \frac{\hat{S}_1^2 / \sigma_1^2}{\hat{S}_2^2 / \sigma_2^2} \sim F_{n_1-1, n_2-1}$$

Test statistics

$$F_0 = \frac{\hat{S}_1^2}{\hat{S}_2^2}$$

Reference distribution

$$F_0 \sim F_{n_1-1, n_2-1}$$

Step 1:

$$H_0:\sigma_1^2=\sigma_2^2; H_1:\sigma_1^2\neq\sigma_2^2$$

(a)

$$H_0:\sigma_1^2\leq\sigma_2^2; H_1:\sigma_1^2>\sigma_2^2$$

(b)

$$H_0:\sigma_1^2\geq\sigma_2^2; H_1:\sigma_1^2<\sigma_2^2$$

(c)

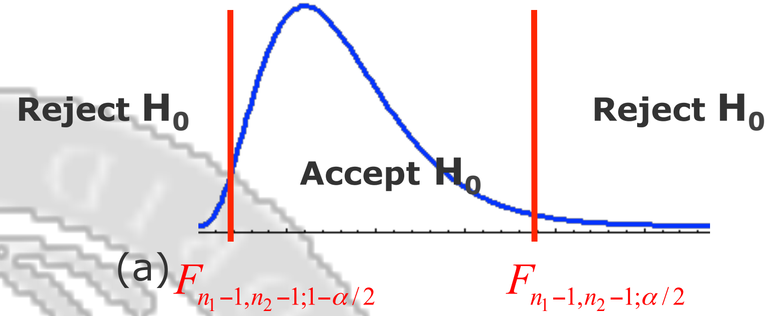
Step 2:

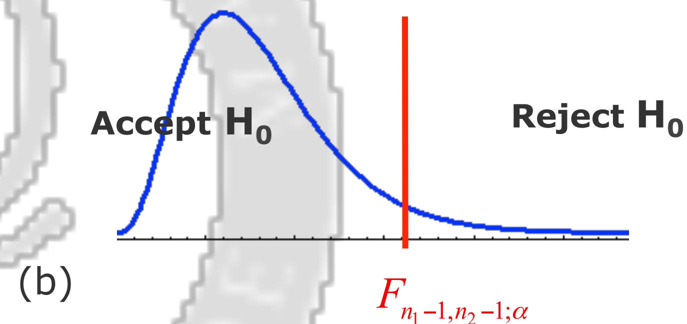
$$F_0=\frac{\hat{S}_1^2}{\hat{S}_2^2}$$

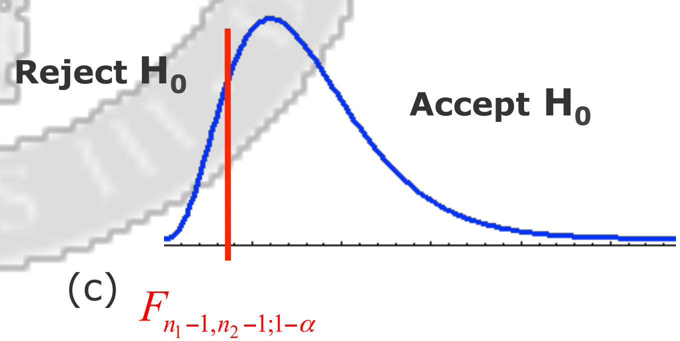
Step 3:

$$F_0\sim F_{n_1-1,n_2-1}$$

Step 4:







The rejection region is where H_1 is accepted

Example

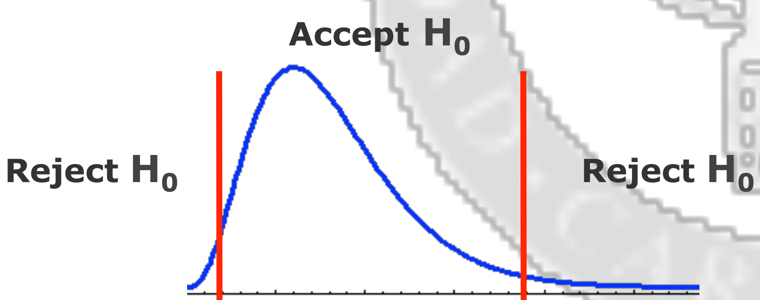
In the previous example about choosing between two types of textile materials to produce mooring systems, we assumed that the variances were the same. Knowing that the populations are normal, test this hypothesis.

$$H_0 : \sigma_1^2 = \sigma_2^2 ; H_1 : \sigma_1^2 \neq \sigma_2^2$$

Material M1: 24 data, $\hat{s}_1 = 2$

Material M2: 30 data, $\hat{s}_2 = 2.3$

$$f_0 = \frac{2^2}{2.3^2} = 0.76$$



$F_{23,29;0.975} = 0.44$ $F_{23,29;0.025} = 2.17$

It is accepted, with a significance level of 5%, that the variances are the same.

The observed difference between the variances is not relevant.