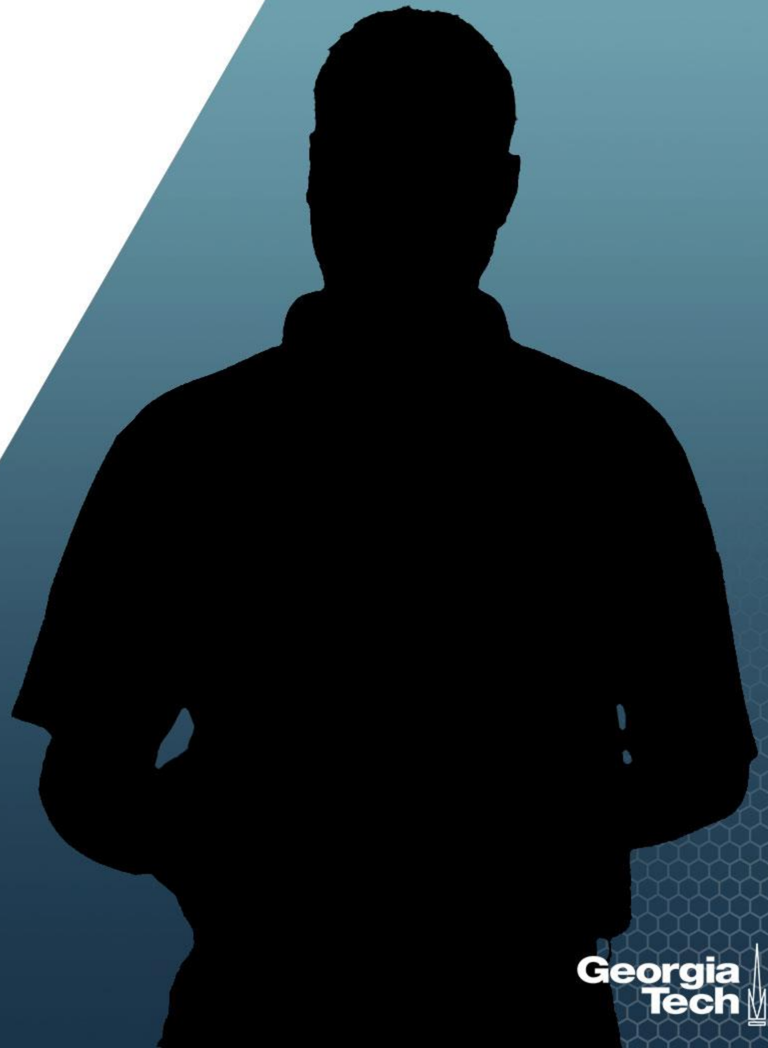
Regression Analysis

Analysis of Variance

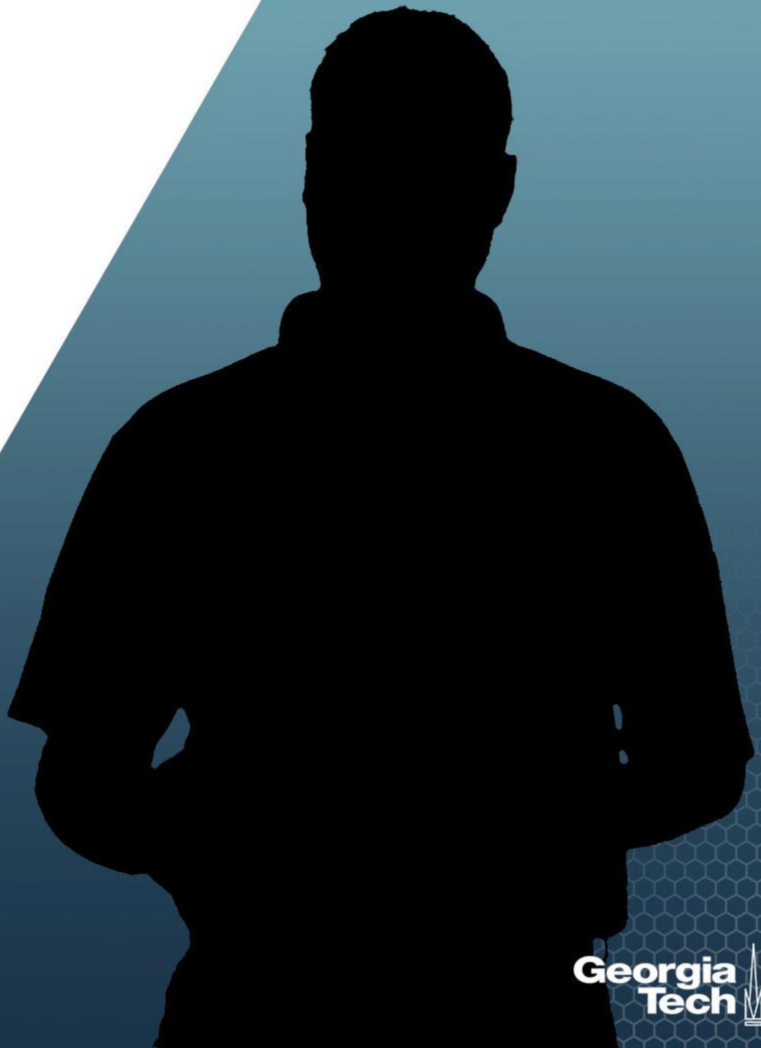
Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering



Hypothesis Test for Equal Means

Georgia Tech

About This Lesson



Hypothesis Test for Equal Means

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

H_A : some means are different

Null Hypothesis

- Under the null hypothesis, combine k samples to estimate the overall mean with the overall sample mean (grand mean) $\bar{Y}$:

$$\bar{Y}=\frac{1}{\mathrm{N}}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}$$

- Base the null hypothesis variance estimate S_0^2 on this overall sample mean:

$$\mathrm{S}_0^2=\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}\right)^2}{\mathrm{N}-1}=\frac{\mathrm{SST}}{\mathrm{N}-1}$$

- **SST** = **<u>S</u>**um of **<u>S</u>**quares **<u>T</u>**otal = $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}\right)^2$

- Because we only estimate one mean, we lose only 1 df (unlike pooled variance)

$$\frac{(\mathrm{N}-1)\mathrm{S}_0^2}{\sigma^2}=\frac{\mathrm{SST}}{\sigma^2}\sim\chi^2_{\mathrm{N}-1}$$

SST Decomposition

We can *partition* SST into two separate parts:

SST = **SSE** + $\mathbf{SST_R}$

where $\mathbf{SST_R}$ = **S**um of **S**quares of **Tr**eatments $= \sum_{i=1}^{k} n_i (\bar{Y}_i - \bar{Y})^2$, and $\bar{Y}_i$ is the i^th sample mean.

Recall:

SST $= \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}\right)^2$

SSE $= \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_i\right)^2$

1. MSE = SSE/(N-k) = *within-group variability*
2. $MSST_R$ = $\mathbf{SST_R}$ /(k-1) = *between-group variability*
3. ANOVA: comparing *between* to *within* variability
4. F = between-group variability / within-group variability

Testing Equal Variances with F-Test

$$\boxed{\frac{\mathrm{SST_R}/(k-1)}{\mathrm{SSE}/(\mathrm{N}-k)} \equiv \frac{\mathrm{MST_R}}{\mathrm{MSE}} = \mathrm{F}_0 \sim \mathrm{F}_{k-1,\mathrm{N}-k}}$$

if H_o is true

Reject H_o if $\mathbf{F_0 > F_a(k-1,N-k)}$, which is the upper α^{th} quantile of the F distribution.

$$\boxed{\text{P-value for the F-test} = P(F > F_0), \text{ where } F \sim F_{(k-1,N-k)}}$$

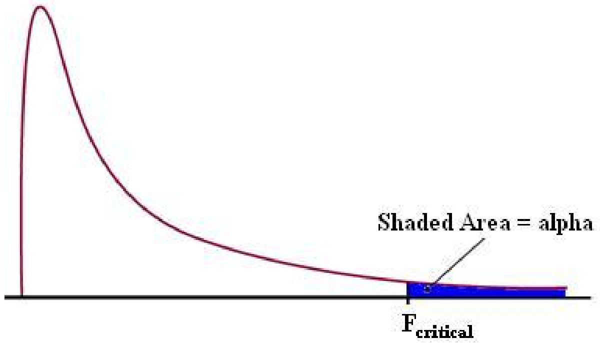
Testing Equal Variances with F-Test

$$\frac{SST_R/(k-1)}{SSE/(N-k)} \equiv \frac{MST_R}{MSE} = F_0 \sim F_{k-1,N-k}$$

if H_o is true

Reject H_o if $\mathbf{F_0 > F_a(k\text{-}1,N\text{-}k)}$, which is the upper α^{th} quantile of the F distribution.

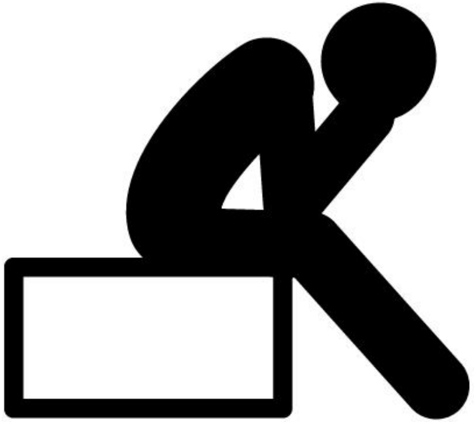
P-value for the F-test = $P(F > F_0)$, where $F \sim F_{(k-1,N-k)}$



Georgia Tech

Example 1: Global Suicide by Region

Are the mean suicide rates equal across the different country regions?



Testing for Equal Means

summary(aov(suicidesper100k ~ region, data=suicide_data))

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
region	9	1548	172.06	4.767	4.71e-05 ***
Residuals	77	2779	36.09		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

$SST_R = 1548$
$k-1 = 9$
$SSE = 2779$
$N-k = 77$
F-value = 4.767
P-value = 4.71e-05

P-value ≈ 0:
Reject the null hypothesis of equal mean heights

Example 2: Keyboard Layout

Three different keyboard layouts are being compared in terms of typing speed.

Are the mean typing times for the three keyboard layouts statistically different?

Layout 1	Layout 2	Layout 3
23.8	30.2	27.0
25.6	29.9	25.4
24.0	29.1	25.6
25.1	28.8	24.2
25.5	29.1	24.8
26.1	28.6	24.0
23.8	28.3	25.5
25.7	28.7	23.9
24.3	27.9	22.6
26.0	30.5	26.0
24.6	*	23.4
27.0	*	*

Testing for Equal Means

summary(aov(speed ~ layout))

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
layout	2	121.24	60.62	52.84	1.48e-10 ***
Residuals	30	34.42	1.15		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

SSTR = 121.24
k-1 = 2
SSE = 34.42
N-k = 30
F-value = 52.84
P-value = 1.48e-10

P-value ≈ 0:
Reject the null hypothesis of equal mean typing times

Summary