Math 362: Mathematical Statistics II

Le Chen
le.chen@emory.edu
chenle02@gmail.com

Emory University
Atlanta, GA

Last updated on Spring 2021
Last compiled on January 15, 2023

2021 Spring

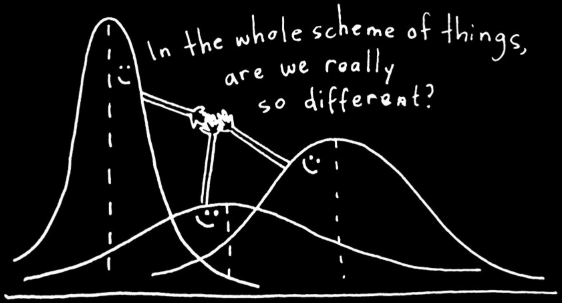
Chapter 12. The Analysis of Variance

§ 12.1 Introduction

§ 12.2 The F Test

§ 12.3 Multiple Comparisons: Turkey's Method

§ 12.4 Testing Subhypotheses with Contrasts


In the whole scheme of things,
are we really
so different?

E.g. 1 Study the relation between smoking and heart rates.

Generations of athletes have been cautioned that cigarette smoking impedes performance. One measure of the truth of that warning is the effect of smoking on heart rate. In one study, six nonsmokers, six light smokers, six moderate smokers, and six heavy smokers each engaged in sustained physical exercise. Table 8.1.1 lists their heart rates after they had rested for three minutes.

Table 8.1.1 Heart Rates

	Nonsmokers	Light Smokers	Moderate Smokers	Heavy Smokers
	69	55	66	91
	52	60	81	72
	71	78	70	81
	58	58	77	67
	59	62	57	95
	65	66	79	84
Averages:	62.3	63.2	71.7	81.7

Show whether smoking affects heart rates at $\alpha = 0.05$.

E.g. 2 A certain fraction of antibiotics injected into the bloodstream are "bound" to serum proteins. This phenomenon bears directly on the effectiveness of the medication, because the binding decreases the systemic uptake of the drug. Table below lists the binding percentages in bovine serum measured for five widely prescribed antibiotics. Which antibiotics have similar binding properties, and which are different?

Table 12.3.1

	Penicillin G	Tetra-cycline	Strepto-mycin	Erythro-mycin	Chloram-phenicol
	29.6	27.3	5.8	21.6	29.2
	24.3	32.6	6.2	17.4	32.8
	28.5	30.8	11.0	18.3	25.0
	32.0	34.8	8.3	19.0	24.2
$T_{.j}$	114.4	125.5	31.3	76.3	111.2
$\overline{Y}_{.j}$	28.6	31.4	7.8	19.1	27.8

Table 12.1.1				
	Treatment Level			
	1	2	$\dots$	k
	Y_{11}	Y_{12}		Y_{1k}
	Y_{21}	Y_{22}		
	$\vdots$	$\vdots$	$\dots$	$\vdots$
	$Y_{n_1 1}$	$Y_{n_2 2}$		$Y_{n_k k}$
Sample sizes:	n_1	n_2	$\dots$	n_k
Sample totals:	$T_{.1}$	$T_{.2}$		$T_{.k}$
Sample means:	$\overline{Y}_{.1}$	$\overline{Y}_{.2}$		$\overline{Y}_{.k}$
True means:	μ_1	μ_2		μ_k

- k treatment levels; k independent random sample of size $n_1, \cdots, n_k$
- Total sample size: $n = \sum_{i=1}^{k} n_i$
- Y_{ij}: i-th observation for the j-th level.
- Sample total: $T_{\cdot j} = \sum_{i=1}^{n_j} Y_{ij}$
- Sample mean: $\overline{Y}_{\cdot j} = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij} = \frac{T_{\cdot j}}{n_j}$
- Overall total: $T_{\cdot\cdot} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} Y_{ij} = \sum_{j=1}^{k} T_{\cdot j}$
- Overall mean: $\overline{Y}_{\cdot\cdot} = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} Y_{ij} = \frac{1}{n} \sum_{j=1}^{k} n_j \overline{Y}_{\cdot j} = \frac{1}{n} \sum_{j=1}^{k} T_{\cdot j}$

Assume For $j = 1, \cdots, k$, $Y_{ij} \sim N(\mu_j, \sigma_j^2)$ and $\sigma_1^2 = \cdots = \sigma_k^2 = \sigma^2$ (unknown).

Problem Testing

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

versus

$$H_1 : \text{not all the } \mu_j\text{'s are equal}$$

Or testing *subhypotheses* such as

$$H_0 : \mu_i = \mu_j \quad \text{or} \quad H_0 : \mu_3 = (\mu_1 + \mu_2)/2$$

ANOVA was developed by statistician and evolutionary biologist —

Ronald Fisher

Statistician

Sir Ronald Aylmer Fisher FRS was a British statistician and geneticist. For his work in statistics, he has been described as "a genius who almost single-handedly created the foundations for modern statistical science" and "the single most important figure in 20th century statistics". Wikipedia

Born: February 17, 1890, East Finchley, London, United Kingdom

Died: July 29, 1962, Adelaide, Australia

Known for: Fisher's principle, Fisher information

Residence: United Kingdom, Australia

Education: Gonville & Caius College, University of Cambridge, Harrow School https://www.youtube.com/watch?v=0XsovsSnRuw

Model assumptions

1. Independence of observations
2. Normality
3. Homogeneity of variances

Table 12.1.1

	Treatment Level			
	1	2	...	k
	Y_{11}	Y_{12}		Y_{1k}
	Y_{21}	Y_{22}		
	$\vdots$	$\vdots$	...	$\vdots$
	$Y_{n_1 1}$	$Y_{n_2 2}$		$Y_{n_k k}$
Sample sizes:	n_1	n_2	...	n_k
Sample totals:	$T_{.1}$	$T_{.2}$		$T_{.k}$
Sample means:	$\overline{Y}_{.1}$	$\overline{Y}_{.2}$		$\overline{Y}_{.k}$
True means:	μ_1	μ_2		μ_k

$\Updownarrow$

Assume:
$\forall j = 1, \cdots, k, \forall j = 1, \cdots, n_i,$

1. Y_{ij} are independent.
2. $Y_{ij} \sim N(\mu_j, \sigma^2)$

$\Longleftrightarrow$

Assume:
$\forall j = 1, \cdots, k, \forall j = 1, \cdots, n_i,$

$$Y_{ij} = \mu_j + \epsilon_{ij}$$

1. ϵ_{ij} are independent.
2. $\epsilon_{ij} \sim N(0, \sigma^2)$

Likelihood ratio test

1. The parameter spaces are

$$\Omega = \{(\mu_1, \cdots, \mu_k, \sigma^2) : -\infty < \mu_1, \cdots, \mu_k < \infty, \sigma^2 > 0\}$$

$$\omega = \{(\mu_1, \cdots, \mu_k, \sigma^2) : -\infty < \mu_1 = \cdots = \mu_k < \infty, \sigma^2 > 0\}$$

2. The likelihood functions are

$$L(\omega) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \mu)^2\right\}$$

$$L(\Omega) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \mu_j)^2\right\}$$

3. Now

$$\frac{\partial \ln L(\omega)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \mu)$$

$$\frac{\partial \ln L(\omega)}{\partial (\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \mu)^2$$

Setting the above derivatives to zero, the solutsions for μ and σ^2 are,

$$\frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{ij} = \bar{y}_{..}$$

$$\frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_{..})^2 = v$$

3' Similarly,

$$\frac{\partial \ln L(\Omega)}{\partial \mu_j} = \frac{1}{\sigma^2} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \mu_j), \quad j = 1, \cdots, k$$

$$\frac{\partial \ln L(\Omega)}{\partial (\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \mu_j)^2$$

Setting the above derivatives to zero, the solutsions for μ_j and σ^2 are,

$$\frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij} = \bar{y}_{\cdot j}$$

$$\frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{\cdot j})^2 = w$$

4. Hence,

$$L(\hat{\omega}) = \left(\frac{n}{2\pi \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{..})^2} \right)^{n/2} \exp\left\{ -\frac{n \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{..})^2}{2 \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{..})^2} \right\}$$

$$||$$

$$\left(\frac{n}{2\pi \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{..})^2} \right)^{n/2} e^{-n/2}$$

Similarly,

$$L(\hat{\Omega}) = \left(\frac{n}{2\pi \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{.j})^2} \right)^{n/2} \exp\left\{ -\frac{n \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{.j})^2}{2 \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{.j})^2} \right\}$$

$$||$$

$$\left(\frac{n}{2\pi \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{.j})^2} \right)^{n/2} e^{-n/2}$$

5. Finally,

$$\lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})} = \left(\frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{\cdot j})^2}{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{\cdot\cdot})^2} \right)^{n/2}$$

$\Rightarrow$ Test statistic:

$$\Lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})} = \left(\frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\cdot j})^2}{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{\cdot\cdot})^2} \right)^{n/2}$$

$$
\begin{aligned}
SSTOT &:= \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(Y_{ij}-\overline{Y}_{\cdot\cdot}\right)^2 \\
&= \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left[\left(Y_{ij}-\overline{Y}_{\cdot j}\right)+\left(\overline{Y}_{\cdot j}-\overline{Y}_{\cdot\cdot}\right)\right]^2 \\
&= \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(Y_{ij}-\overline{Y}_{\cdot j}\right)^2 + \text{zero cross term} + \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(\overline{Y}_{\cdot j}-\overline{Y}_{\cdot\cdot}\right)^2 \\
&= \underbrace{\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(Y_{ij}-\overline{Y}_{\cdot j}\right)^2}_{SSE} + \underbrace{\sum_{j=1}^{k} n_j\left(\overline{Y}_{\cdot j}-\overline{Y}_{\cdot\cdot}\right)^2}_{SSTR}
\end{aligned}
$$

$$\Downarrow$$

$$\Lambda = \left(\frac{SSE}{SSTOT}\right)^{n/2} = \left(\frac{SSE}{SSE+SSTR}\right)^{n/2} = \left(\frac{1}{1+SSTR/SSE}\right)^{n/2}$$

6. Critical regions: for some $\lambda_* \in (0,1)$ close to 0,

$$\begin{aligned}
\alpha &= \mathbb{P}\left(\Lambda \le \lambda_*\right) \\
&= \mathbb{P}\left(\frac{1}{1 + SSTR/SSE} \le \lambda_*^{2/n}\right) \\
&= \mathbb{P}\left(\frac{SSTR}{SSE} \le \lambda_*^{-2/n} - 1\right) \\
&= \mathbb{P}\left(\frac{SSTR/(k-1)}{SSE/(n-k)} \le \left(\lambda_*^{-2/n} - 1\right)\frac{n-k}{k-1}\right)
\end{aligned}$$

7. We will prove that under H_0, $\frac{SSTR/(k-1)}{SSE/(n-k)} \sim$ F-distr. $df_1 = k-1, df_2 = n-k$

$$\Rightarrow \quad \left(\lambda_*^{-2/n} - 1\right)\frac{n-k}{k-1} = F_{1-\alpha,k-1,n-k}.$$

$\square$

Treatment sum of squares: SSTR

Sample size: (Weights)	n_1	n_2	$\cdots$	n_k	$n = \sum_{j=1}^{k} n_j$ *Weighted average*
Sample means:	$\overline{Y}_{\cdot 1}$	$\overline{Y}_{\cdot 2}$	$\cdots$	$\overline{Y}_{\cdot k}$	$\overline{Y}_{\cdot\cdot} = \frac{1}{n} \sum_{j=1}^{k} n_j \overline{Y}_{\cdot j}$
True means:	μ_1	μ_2	$\cdots$	μ_k	$\mu = \frac{1}{n} \sum_{j=1}^{k} n_j \mu_j$
Squares:	$\left(\overline{Y}_{\cdot 1} - \overline{Y}_{\cdot\cdot}\right)^2$	$\left(\overline{Y}_{\cdot 2} - \overline{Y}_{\cdot\cdot}\right)^2$	$\cdots$	$\left(\overline{Y}_{\cdot k} - \overline{Y}_{\cdot\cdot}\right)^2$	*SSTR*

$$SSTR := \sum_{j=1}^{k} n_j \left(\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot}\right)^2$$

1. When $k = 1$, $SSTR \equiv 0$.
2. When $k = 2$, say $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_m$:

$$\overline{Y_{\cdot\cdot}} = \frac{1}{m+n}\left(n\overline{X} + m\overline{Y}\right)$$

$$\begin{aligned} SSTR &= n\left[\overline{X} - \frac{1}{n+m}\left(n\overline{X} + m\overline{Y}\right)\right]^2 + m\left[\overline{Y} - \frac{1}{n+m}\left(n\overline{X} + m\overline{Y}\right)\right]^2 \\ &= n\left[\frac{m(\overline{X} - \overline{Y})}{n+m}\right]^2 + m\left[\frac{n(\overline{X} - \overline{Y})}{n+m}\right]^2 \\ &= \left[\frac{nm^2}{(n+m)^2} + \frac{n^2 m}{(n+m)^2}\right]\left(\overline{X} + \overline{Y}\right)^2 \\ &= \frac{nm}{n+m}\left(\overline{X} - \overline{Y}\right)^2 \end{aligned}$$

$$SSTR = \frac{\left(\overline{X} - \overline{Y}\right)^2}{\frac{1}{m} + \frac{1}{n}}$$

$$
\begin{aligned}
SSTR &= \sum_{j=1}^{k} n_j\left(\overline{Y}_{.j} - \overline{Y}_{..}\right)^2 = \sum_{j=1}^{k} n_j\left[\left(\overline{Y}_{.j} - \mu\right) - \left(\overline{Y}_{..} - \mu\right)\right]^2 \\
&= \sum_{j=1}^{k} n_j\left[\left(\overline{Y}_{.j} - \mu\right)^2 + \left(\overline{Y}_{..} - \mu\right)^2 - 2\left(\overline{Y}_{.j} - \mu\right)\left(\overline{Y}_{..} - \mu\right)\right] \\
&= \sum_{j=1}^{k} n_j\left(\overline{Y}_{.j} - \mu\right)^2 + \sum_{j=1}^{k} n_j\left(\overline{Y}_{..} - \mu\right)^2 - 2\left(\overline{Y}_{..} - \mu\right)\sum_{j=1}^{k} n_j\left(\overline{Y}_{.j} - \mu\right) \\
&= \sum_{j=1}^{k} n_j\left(\overline{Y}_{.j} - \mu\right)^2 + n\left(\overline{Y}_{..} - \mu\right)^2 - 2\left(\overline{Y}_{..} - \mu\right)n\left(\overline{Y}_{..} - \mu\right) \\
&= \sum_{j=1}^{k} n_j\left(\overline{Y}_{.j} - \mu\right)^2 - n\left(\overline{Y}_{..} - \mu\right)^2 \qquad (12.2.1)
\end{aligned}
$$

$$\Downarrow$$

$$
SSTR = \sum_{j=1}^{k} n_j\left[\left(\overline{Y}_{.j} - \mu_j\right)^2 - 2\left(\overline{Y}_{.j} - \mu_j\right)(\mu - \mu_j) + (\mu - \mu_j)^2\right] - n\left(\overline{Y}_{..} - \mu\right)^2
$$

Notice that

$$\overline{Y}_{\cdot j} \sim N(\mu_j, \sigma^2/n_j) \qquad \text{and} \qquad \overline{Y}_{\cdot\cdot} \sim N(\mu, \sigma^2/n)$$

$\Longrightarrow$

$$\mathbb{E}[SSTR] = \sum_{j=1}^{k} n_j \left[\frac{\sigma^2}{n_j} - 2 \times 0 + (\mu - \mu_j)^2\right] - n\frac{\sigma^2}{n}$$

$$= (k-1)\sigma^2 + \sum_{j=1}^{k} n_j(\mu - \mu_j)^2$$

Remark When $\mu_1 = \cdots = \mu_j$ then

0.1 $\mathbb{E}[SSTR] = (k-1)\sigma^2$

0.2 $MSTR := \frac{SSTR}{k-1}$ is an unbiased estimator for σ^2.

0.3 $SSTR/\sigma^2 \sim$ Chi square $(df = k - 1)$. (Homework)

Test $H_0 : \mu_1 = \cdots = \mu_k$ v.s. μ_j are not the same.

Case I. when σ^2 is known.

Reject H_0 if $SSTR/\sigma^2 \geq \chi^2_{1-\alpha,k-1}$.

Case II. when σ^2 is unknown.

......

Sum of Squared Errors: SSE

1. Sum of squred error:

$$SSE := \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(Y_{ij} - \overline{Y}_{\cdot j} \right)^2$$

$$= \sum_{j=1}^{k} (n_j - 1) \left[\frac{1}{n_j - 1} \sum_{i=1}^{n_j} \left(Y_{ij} - \overline{Y}_{\cdot j} \right)^2 \right]$$

$$= \sum_{j=1}^{k} (n_j - 1) S_j^2$$

2. Pooled variance S_p^2:

$$S_p^2 = \frac{SSE}{\sum_{j=1}^{k} (n_j - 1)} = \frac{SSE}{n - k}$$

Mean square of error $MSE = S_p^2$

Notice that

1. $(n_j - 1)S_j^2/\sigma^2$ ~Chi square ($df = n_j - 1$)
2. S_j^2's are independent
3. $SSE/\sigma^2 = (n-k)S_p^2/\sigma^2 = \sum_{j=1}^k (n_j - 1)S_j^2/\sigma^2$,

   Sum of independent of Chi squares

$$\Downarrow$$

Thm. No mattter $H_0 : \mu_1 = \cdots = \mu_k$ is true or not

a. $SSE/\sigma^2 = (n-k)S_p^2/\sigma^2 \sim$ Chi square ($df = \sum_{j=1}^k (n_j - 1) = n - k$)

b. $SSTR \perp SSE$.

Proof. We have shown part (a). Part (b) is trickier. Indeed, both parts are a consequence of **Cochran's theorem**[1] ... $\square$

[1] `https://en.wikipedia.org/wiki/Cochran%27s_theorem`

Let's see two special cases of

Thm. No mattter $H_0 : \mu_1 = \cdots = \mu_k$ is true or not

a. $SSE/\sigma^2 = (n-k)S_p^2/\sigma^2 \sim$ Chi square $(df = \sum_{j=1}^{k}(n_j - 1) = n-k)$

b. $SSTR \perp SSE$.

Cases

1. $k = 1$, one sample case, S_p^2 is sample variance — Chapter 7

   a. $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$

   b. $SSTR \equiv 0$

2. $k = 2$, two sample case — Chapter 9

   a. $(n-2)S_p^2/\sigma^2 \sim \chi^2(n-2)$

   b. $\overline{X} - \overline{Y} \perp S_p^2 \quad \Longleftrightarrow \quad SSTR \perp SSE$

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

1. $SSTR/\sigma^2 \sim \chi^2(k-1)$
2. $SSE/\sigma^2 \sim \chi^2(n-k)$
3. $SSTR \perp SSE$

$$\Longrightarrow \qquad F = \frac{SSTR/(k-1)}{SSE/(n-k)} \sim F(df_1 = k-1, df_2 = n-k)$$

Reject H_0 if $F \geq F_{1-\alpha,k-1,n-k}$

Total Sum of Squares: SSTOT

SSTOT=SSE+SSTR

$$SSTOT := \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(Y_{ij} - \overline{Y}_{\cdot\cdot}\right)^2$$

$$||$$

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left[\left(Y_{ij} - \overline{Y}_{j\cdot}\right) + \left(\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot}\right)\right]^2$$

$$||$$

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(Y_{ij} - \overline{Y}_{j\cdot}\right)^2 + 2\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(Y_{ij} - \overline{Y}_{\cdot j}\right)\left(\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot}\right) + \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot}\right)^2$$

$$||$$

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(Y_{ij} - \overline{Y}_{j\cdot}\right)^2 + 2\sum_{j=1}^{k}\left(\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot}\right)\sum_{i=1}^{n_j}\left(Y_{ij} - \overline{Y}_{\cdot j}\right) + \sum_{j=1}^{k} n_j\left(\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot}\right)^2$$

$$||$$

$$SSE + 0 + SSTR$$

$$
\begin{array}{ccccc}
SSTOT & = & SSE & + & SSTR \\
& \Downarrow & & & \\
\dfrac{SSTOT}{\sigma^2} & = & \dfrac{SSE}{\sigma^2} & + & \dfrac{SSTR}{\sigma^2} \\
\wr & & \wr & & \wr \\
\chi^2(n-1) & & \chi^2(n-k) & \perp & \chi^2(k-1) \\
\text{Under } H_0 & & \checkmark & & \text{Under } H_0
\end{array}
$$

One-way ANOVA Table

Source of Variance	Degree of Freedom (df)	Sum Square (SS)	Mean Square (MS)	F-ratio
Between Groups (Treatment)	k-1	$SSB = \sum_{j=1}^{k}\left(\frac{T_j^2}{n_j}\right) - \frac{T^2}{n}$ $\quad SSB = \sum_{j=1}^{k} n_j(\overline{X}_j - \overline{X}_t)^2$	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSW}$
Within Groups (Error)	n-k	$SSW = \sum_{j=1}^{K}\sum_{i=1}^{n} X_{ij}^2 - \sum_{j=1}^{K}\left(\frac{T_j^2}{n_j}\right)$ $\quad SSW = \sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)^2$	$MSW = \frac{SSW}{n-k}$	
Total	n-1	$SST = \sum_{j=1}^{K}\sum_{i=1}^{n} X_{ij}^2 - \frac{T^2}{n}$ $\quad SST = \sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij} - \overline{X}_t)^2$		

- SST = SSB + SSW

k: number of groups n: number of samples

df: degree of freedom

Source	df	SS	MS	F	P
Treatment	$k-1$	$SSTR$	$MSTR$	$\frac{MSTR}{MSE}$	$P(F_{k-1,n-k} \geq \text{observed} F)$
Error	$n-k$	SSE	MSE		
Total	$n-1$	$SSTOT$			

Common notation

d.f.

k-1	Error sum of squares	$SSE = SSW = SS_{within}$
	Mean square of error	$MSE = MSW = MS_{within} = S_p^2$
	(Pooled sample variance)	
n-k	Treatment sum of squares	$SSTR = SSB = SS_{between}$
	Mean square of treatment	$MSTR = MSB = MS_{between}$
n-1	Total sum of squares:	$SST = SSTOT$

One way ANOVA v.s. Two sample t-test

Let $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_m$ be samples from $N(\mu_X, \sigma^2)$ and $N(\mu_Y, \sigma^2)$, respectively.

Recall

1. $SSTR/\sigma^2 = \dfrac{\left(\overline{X} - \overline{Y}\right)^2}{\sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)} \sim \chi^2(1)$
2. $SSE/\sigma^2 = (n+m-2)S_p^2/\sigma^2 \sim \chi^2(n+m-2)$

$$\Longrightarrow F = \frac{SSTR/1}{SSE/(n+m-2)} = \underbrace{\frac{\left(\overline{X} - \overline{Y}\right)^2}{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}}_{\| \atop T^2} \sim F(df_1 = 1, df_2 = n+m-2)$$

$$\Longrightarrow \alpha = \mathbb{P}\left(|T| \geq t_{\alpha/2, n+m-2}\right) = \mathbb{P}\left(T^2 \geq t^2_{\alpha/2, n+m-2}\right) = \mathbb{P}\left(F \geq F_{1-\alpha, 1, n+m-2}\right)$$

Equivalent!

E.g. 1 Study the relation between smoking and heart rates.

Generations of athletes have been cautioned that cigarette smoking impedes performance. One measure of the truth of that warning is the effect of smoking on heart rate. In one study, six nonsmokers, six light smokers, six moderate smokers, and six heavy smokers each engaged in sustained physical exercise. Table 8.1.1 lists their heart rates after they had rested for three minutes.

Table 8.1.1 Heart Rates

	Nonsmokers	Light Smokers	Moderate Smokers	Heavy Smokers
	69	55	66	91
	52	60	81	72
	71	78	70	81
	58	58	77	67
	59	62	57	95
	65	66	79	84
Averages:	62.3	63.2	71.7	81.7

Show whether smoking affects heart rates at $\alpha = 0.05$.

Sol. Let $\mu_1, \cdots, \mu_4$ be the true heart rates.

Test $H_0 : \mu_0 = \cdots = \mu_4$ or not.

Critical region:

Let $\alpha = 0.05$. For these data, $k = 4$ and $n = 24$, so H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ should be rejected if

$$F = \frac{SSTR/(4-1)}{SSE/(24-4)} \geq F_{1-0.05,4-1,24-4} = F_{.95,3,20} = 3.10$$

(see Figure 12.2.2).

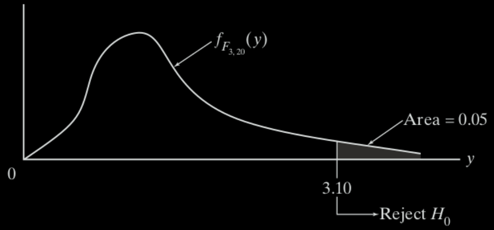


Figure 12.2.2

Computing....

Table 12.2.1

	Nonsmokers	Light Smokers	Moderate Smokers	Heavy Smokers
	69	55	66	91
	52	60	81	72
	71	78	70	81
	58	58	77	67
	59	62	57	95
	65	66	79	84
$T_{.j}$	374	379	430	490
$\overline{Y}_{.j}$	62.3	63.2	71.7	81.7

The overall sample mean, $\overline{Y}_{..}$, is given by

$$\overline{Y}_{..} = \frac{1}{n}\sum_{j=1}^{k} T_{.j} = \frac{374+379+430+490}{24}$$

$$= 69.7$$

Therefore,

$$SSTR = \sum_{j=1}^{4} n_j\left(\overline{Y}_{.j} - \overline{Y}_{..}\right)^2 = 6[(62.3-69.7)^2 + \cdots + (81.7-69.7)^2]$$

$$= 1464.125$$

Similarly,

$$SSE = \sum_{j=1}^{4}\sum_{i=1}^{6}\left(Y_{ij} - \overline{Y}_{.j}\right)^2 = [(69-62.3)^2 + \cdots + (65-62.3)^2] + \cdots + [(91-81.7)^2 + \cdots + (84-81.7)^2]$$

$$= 1594.833$$

The observed test statistic, then, equals *6.12*:

$$F = \frac{1464.125/(4-1)}{1594.833/(24-4)} = 6.12$$

Since $6.12 > F_{.95,3,20} = 3.10$, H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ should be rejected. These data support the contention that smoking influences a person's heart rate.

Figure 12.2.3 shows the analysis of these data summarized in the ANOVA table format. Notice that the small P-value ($= 0.004$) is consistent with the conclusion that H_0 should be rejected.

Source	df	*SS*	*MS*	*F*	*P*
Treatment	3	1464.125	488.04	6.12	0.004
Error	20	1594.833	79.74		
Total	23	3058.958			

Figure 12.2.3

□

http://www.sthda.com/english/wiki/one-way-anova-test-in-r

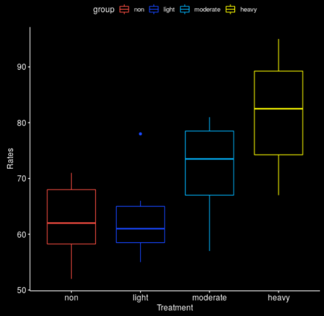
```
> Input <-c("
+ rates group
+ 69  non
+ 52  non
+ 71  non
+ 58  non
+ 59  non
+ 65  non
+ 55  light
+ 60  light
+ 78  light
+ 58  light
+ 62  light
+ 66  light
+ 66  moderate
+ 81  moderate
+ 70  moderate
+ 77  moderate
+ 57  moderate
+ 79  moderate
+ 91  heavy
+ 72  heavy
+ 81  heavy
+ 67  heavy
+ 95  heavy
+ 84  heavy
+ ")
> Data = read.table(textConnection(Input),
+                    header=TRUE)
```

```
> Data
   rates    group
1     69      non
2     52      non
3     71      non
4     58      non
5     59      non
6     65      non
7     55    light
8     60    light
9     78    light
10    58    light
11    62    light
12    66    light
13    66 moderate
14    81 moderate
15    70 moderate
16    77 moderate
17    57 moderate
18    79 moderate
19    91    heavy
20    72    heavy
21    81    heavy
22    67    heavy
23    95    heavy
24    84    heavy
```

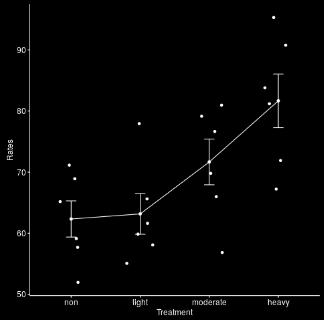
```
> # Check the levels
> levels (Data$group)
[1] "heavy"    " light "    "moderate" "non"
> # Order the groups
> Data$group <- ordered(Data$group, levels = c("non", "light" , "moderate", "heavy"))
> levels (Data$group)
[1] "non"      " light "    "moderate" "heavy"
```

```
> # Compute summary statistics by groups
> # including count, mean, sd:
> library (dplyr) # a grammar of data manipulation
> group_by(Data, group) %>%
+   summarise(
+     count = n() ,
+     mean = mean(rates, na.rm = TRUE),
+     sd = sd(rates, na.rm = TRUE)
+   )
# A tibble : 4 x 4
  group    count  mean    sd
  <ord>    <int> <dbl> <dbl>
1 non          6  62.3  7.26
2 light        6  63.2  8.16
3 moderate     6  71.7  9.16
4 heavy        6  81.7 10.8
```

```
# Box plots
# ++++++++++++++++++++
# Plot rates by group and color by group
 library (ggpubr)
png("Case_12-2-1-ggboxplot.png")
ggboxplot(Data, x = "group", y = "rates",
              color = "group", palette = c("#00AFBB", "#E7B800", "#FC4E07", "blue"),
              order = c("non", " light ", "moderate","heavy"),
              ylab = "Rates", xlab = "Treatment")
dev.off ()
```



```r
# Mean plots
# ++++++++++++++++++++
# Plot rates by group
# Add error bars: mean_se
# (other values include: mean_sd, mean_ci, median_iqr, ....)
png("Case_12-2-1-ggline.png")
 library (ggpubr)
ggline(Data, x = "group", y = "rates",
         add = c("mean_se", " jitter "),
         order =  c("non", " light ",  "moderate","heavy"),
         ylab = "Rates", xlab = "Treatment")
dev.off ()
```



```
> # Compute the analysis of variance
> res.aov <- aov(rates ~ group, data = Data)
> # Summary of the analysis
> summary(res.aov)
            Df Sum Sq Mean Sq F value Pr(>F)
group        3   1464   488.0    6.12 0.00398 **
Residuals   20   1595    79.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> # Tukey multiple  multiple-comparisons
> TukeyHSD(res.aov)
  Tukey multiple  comparisons of means
    95% family-wise confidence level

 Fit :  aov(formula = rates ~ group, data = Data)

$group
                        diff          lwr        upr       p adj
 light -non         0.8333333 -13.596955 15.26362 0.9984448
moderate-non   9.3333333 -5.096955 23.76362 0.2978123
heavy-non       19.3333333  4.903045 33.76362 0.0063659
moderate-light  8.5000000 -5.930289 22.93029 0.3755571
heavy-light      18.5000000  4.069711 32.93029 0.0091463
heavy-moderate 10.0000000 -4.430289 24.43029 0.2438158
```

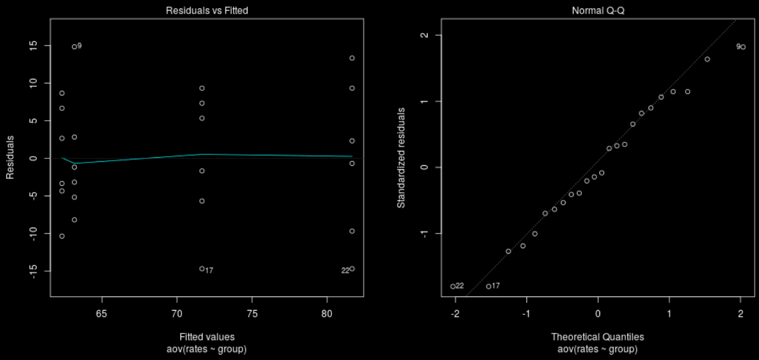
1. diff: difference between means of the two groups
2. lwr, upr: the lower and the upper end point of the C.I. at 95% (default)
3. p adj: p-value after adjustment for the multiple comparisons

Inferences

if p-value $\leq 0.05 \quad \Longleftrightarrow$ if zero is in the C.I.

```
> # Or one may use multcomp package or multiple comparisons
>  library (multcomp)
> summary(glht(res.aov,  linfct  = mcp(group = "Tukey")))

     Simultaneous Tests for General Linear Hypotheses

Multiple  Comparisons of Means: Tukey Contrasts


Fit :  aov(formula = rates ~ group, data = Data)

Linear  Hypotheses:
                              Estimate Std. Error  t  value Pr(>|t |)
 light   – non == 0            0.8333      5.1556   0.162  0.99844
moderate – non == 0      9.3333      5.1556   1.810  0.29776
heavy – non == 0         19.3333      5.1556   3.750  0.00629 **
moderate – light  == 0    8.5000      5.1556   1.649  0.37544
heavy –  light   == 0      18.5000      5.1556   3.588  0.00901 **
heavy – moderate == 0 10.0000      5.1556   1.940  0.24382
–––
 Signif .  codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1
 (Adjusted p values reported –– single–step method)
```

```r
# Check ANOVA assumptions: test validity?
# diagnostic  plots
layout(matrix(c(1,2),1,2)) # optional  1x2 graphs/page
plot(res.aov,c(1,2))
```



1. Residuals vs Fitted: test homogeneity of variances
   One can also use Levene's test for this purpose:

```
> # Use Levene's test to gest homogeneity of variances
> library (car)
> leveneTest(rates ~ group, data = Data)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  3  0.3885 0.7625
      20
```

2. Normal Q-Q plot: Test normality. (It should be close to diagonal line.)
   One can also use Shapiro-Wilk test:

```
# Extract the residuals
> aov_residuals <- residuals(object = res.aov )
> # Run Shapiro-Wilk test
> shapiro.test(x = aov_residuals )

  Shapiro-Wilk normality test

data:  aov_residuals
W = 0.9741, p-value = 0.7677
```

Non-parametric alternative to one-way ANOVA test

```
> # Non–parametric alternative to one–way ANOVA test
> # a non–parametric alternative to one–way ANOVA
> # is Kruskal–Wallis rank sum test, which can be
> # used when ANNOVA assumptions are not met.
> kruskal.test(rates ~ group, data = Data)

  Kruskal–Wallis rank sum test

data:  rates by group
Kruskal–Wallis chi–squared = 10.729, df = 3, p–value = 0.01329
```

See Section 4 of Chapter 14 for more details.

1. John Wilder Tukey (June 16, 1915 – July 26, 2000) was an American mathematician best known for development of the Fast Fourier Transform (FFT) algorithm and box plot.
2. The Tukey range test, the Tukey lambda distribution, the Tukey test of additivity, and the Teichmüller-Tukey lemma all bear his name.
3. He is also credited with coining the term ’bit’.

`https://en.wikipedia.org/wiki/John_Tukey`

$N(\mu_1, \sigma^2)$	$N(\mu_2, \sigma^2)$	$\cdots$	$N(\mu_2, \sigma^2)$
Y_{11}	Y_{12}	$\cdots$	Y_{1k}
Y_{21}	Y_{22}	$\cdots$	Y_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_{r1}	Y_{r2}	$\cdots$	Y_{rk}

Goal For any $i \neq j$, test

$$H_0 : \mu_i = \mu_j \qquad v.s. \qquad H_1 : \mu_i \neq \mu_j$$

at the α level of significance defined as

$$\mathbb{P}\left(\bigcup_{j=1}^{\binom{k}{2}} E_j \right) = \alpha$$

where there are $\binom{k}{2}$ pairs, and E_j is the event of making a type I error for the j-th pair.

Goal' Simultaneous C.I.'s for $\binom{k}{2}$ pairs of means:

Given α, find I_{ij}, the C.I. for $\mu_i - \mu_j$ (with $i, j = 1, \cdots, k$ and $i \neq j$), s.t.

$$\mathbb{P}\left(\mu_i - \mu_j \in I_{ij},\ \forall i, j = 1, \cdots, k, i \neq j\right) = 1 - \alpha.$$

??? Why not the standard pair-wise two-sample t-test?

Suppose $\mathbb{P}(E_j) = \alpha_*$. Then

$$\alpha = \mathbb{P}\left(\bigcup_{j=1}^{\binom{k}{2}} E_j\right) = 1 - \mathbb{P}\left(\bigcap_{j=1}^{\binom{k}{2}} E_j^c\right) \approx 1 - \prod_{j=1}^{\binom{k}{2}} \mathbb{P}(E_j^c) = 1 - (1-\alpha_*)^{\binom{k}{2}}$$

Hence,

$$\alpha_* \approx 1 - (1-\alpha)^{1/\binom{k}{2}}$$

E.g., $\alpha = 0.05$

k	5	8	100
α_*	0.0051162	0.001830	0.00001036

Bonferroni's method

— A straightforward method

$$\mathbb{P}\left(\mu_i - \mu_j \in I_{ij},\ \forall i \neq j\right)$$

$$\|$$

$$\mathbb{P}\left(\bigcap_{i \neq j} \mu_i - \mu_j \in I_{ij}\right)$$

$$\|$$

$$1 - \mathbb{P}\left(\bigcup_{i \neq j} \mu_i - \mu_j \notin I_{ij}\right)$$

$$\geq$$

$$1 - \sum_{i \neq j} \mathbb{P}\left(\mu_i - \mu_j \notin I_{ij}\right)$$

$$\|$$

$$1 - \binom{k}{2}\alpha_*$$

1. If we choose $\alpha_* = \alpha/\binom{k}{2}$,
2. let I_{ij} be the $(1-\alpha_*)100\%$ C.I. $i \neq j$

$$\Downarrow$$

$$\mathbb{P}\left(\mu_i - \mu_j \in I_{ij},\ \forall i \neq j\right)$$

$$\geq$$

$$1 - \binom{k}{2}\alpha_*$$

$$\|$$

$$1 - \alpha.$$

Remark This is an approximation. The resulting C.I. are in general too wide.

The exact, and much more precise, solution is given by J.W. Turkey.

One can also construct simultaneous C.I. for all possible linear combinations of the parameters $\sum_{j=1}^{k} c_j\mu_j$, this can be acchieved by **Scheffé's method**. A simple verson is given in §12.4.

Tukey's HSD (honestly significant difference) test

Let's construct $(1-\alpha)100\%$ C.I.'s simultaneously for all pairs.

$$\mathbb{P}\left(\left|(\overline{Y}_{\cdot i}-\mu_i)-(\overline{Y}_{\cdot j}-\mu_j)\right| \le \mathcal{E}, \quad \forall i \ne j\right) = 1-\alpha$$

$$||$$

$$\mathbb{P}\left(\max_i(\overline{Y}_{\cdot i}-\mu_i)-\min_j(\overline{Y}_{\cdot j}-\mu_j) \le \mathcal{E}\right)$$

$$||$$

$$\mathbb{P}\left(\max_i \overline{Y}_{\cdot i}-\min_j \overline{Y}_{\cdot j} \le \mathcal{E}\right)$$

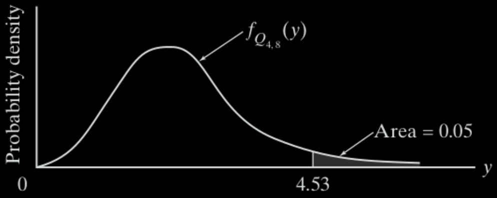
$\Longrightarrow$ Needs to study ...

Def. Let $W_1, \cdots, W_k$ be k i.i.d. r.v.'s from $N(\mu, \sigma^2)$. Let R denote their range:

$$R = \max_i W_i - \min_i W_i.$$

Let S^2 be an unbiased estimator for σ^2 independent of the W_i's and based on ν df. Define the **Studentized range,** $Q_{k,\nu}$, to be the ratio:

$$Q_{k,\nu} := \frac{R}{S}.$$

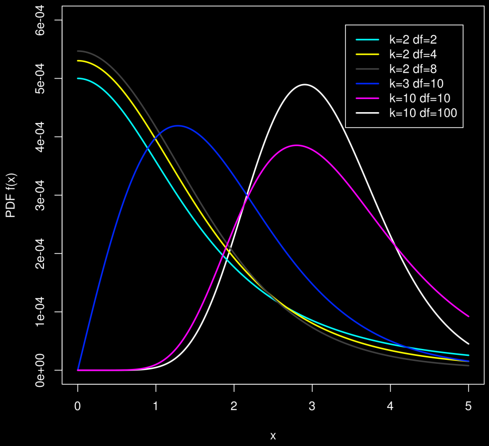


Remark
0.1 We need $R \perp S$ to mimic Student's t-distribution.
0.2 In the following $\nu = n - k = rk - k = r(k-1)$.

$Q_{k,\nu} \sim$ **Studentized range distribution** with parameters k and ν.

k: number of groups.

ν: degrees of freedom.



Let's find one example that all requirements of the $Q_{k,\nu}$ are satisfied.

1. Take $W_j = \overline{Y}_{\cdot j} - \mu_j$, $j = 1, \cdots, k \quad \Longrightarrow \quad W_j \sim N(0, \sigma^2/r)$.

2. MSE or the pooled variance S_p^2 — MSE/r
   is an unbiased estimator for σ^2 — σ^2/r
   is $\perp \{\overline{Y}_{\cdot j}\}_{j=1,\cdots,k}$, hence $\perp \{W_j\}_{j=1,\cdots,k}$

3. df of MSE is equal to $n - k = kr - k = k(r-1)$.

$$\Longrightarrow \qquad \frac{\max_i W_i - \min_j W_j}{\sqrt{MSE/r}} \sim \text{Studentized range distribution}(k,\ rk - k)$$

$$\mathbb{P}\left(\frac{\max_i W_i - \min_j W_j}{\sqrt{MSE/r}} \le Q_{\alpha,k,rk-k}\right) = 1-\alpha$$

$$||$$

$$\mathbb{P}\left(\max_i W_i - \min_j W_j \le \frac{Q_{\alpha,k,rk-k}}{\sqrt{r}}\sqrt{MSE}\right)$$

$$||$$

$$\mathbb{P}\left(|W_i - W_j| \le \frac{Q_{\alpha,k,rk-k}}{\sqrt{r}}\sqrt{MSE},\ \forall i \ne j\right)$$

$$||$$

$$\mathbb{P}\left(\left|\left(\overline{Y}_{\cdot i} - \overline{Y}_{\cdot j}\right) - (\mu_i - \mu_j)\right| \le \frac{Q_{\alpha,k,rk-k}}{\sqrt{r}}\sqrt{MSE},\ \forall i \ne j\right)$$

$$||$$

$$\mathbb{P}\left(\overline{Y}_{\cdot i} - \overline{Y}_{\cdot j} - \frac{Q_{\alpha,k,rk-k}}{\sqrt{r}}\sqrt{MSE} \le \mu_i - \mu_j \le \overline{Y}_{\cdot i} - \overline{Y}_{\cdot j} + \frac{Q_{\alpha,k,rk-k}}{\sqrt{r}}\sqrt{MSE},\ \forall i \ne j\right)$$

Therefore, for all $i \neq j$, the $100(1-\alpha)\%$ C.I. for $\mu_i - \mu_j$ is

$$\overline{Y}_{\cdot i} - \overline{Y}_{\cdot j} \pm \frac{Q_{\alpha,k,rk-k}}{\sqrt{2}} \sqrt{MSE} \sqrt{\frac{2}{r}}$$

To test $H_0 : \mu_i = \mu_j$ for specific $i \neq j$, reject H_0 in favor of $H_1 : \mu_i \neq \mu_j$ if the C.I. does NOT contain 0, at the α level of significance. $\square$

Note: When sample sizes are not equal, use the **Tukey-Kramer method**:

$$\overline{Y}_{\cdot i} - \overline{Y}_{\cdot j} \pm \frac{Q_{\alpha,k,rk-k}}{\sqrt{2}} \sqrt{MSE} \sqrt{\frac{1}{r_i} + \frac{1}{r_j}}$$

E.g. 2 A certain fraction of antibiotics injected into the bloodstream are “bound” to serum proteins. This phenomenon bears directly on the effectiveness of the medication, because the binding decreases the systemic uptake of the drug. Table below lists the binding percentages in bovine serum measured for five widely prescribed antibiotics. Which antibiotics have similar binding properties, and which are different?

Table 12.3.1

	Penicillin G	Tetra-cycline	Strepto-mycin	Erythro-mycin	Chloram-phenicol
	29.6	27.3	5.8	21.6	29.2
	24.3	32.6	6.2	17.4	32.8
	28.5	30.8	11.0	18.3	25.0
	32.0	34.8	8.3	19.0	24.2
$T_{.j}$	114.4	125.5	31.3	76.3	111.2
$\overline{Y}_{.j}$	28.6	31.4	7.8	19.1	27.8

To answer that question requires that we make all $\binom{5}{2} = 10$ pairwise comparisons of μ_i versus μ_j. First, MSE must be computed. From the entries in Table 12.3.1,

$$SSE = \sum_{j=1}^{5} \sum_{i=1}^{4} (Y_{ij} - \overline{Y}_{.j})^2 = 135.83$$

so $MSE = 135.83/(20 - 5) = 9.06$. Let $\alpha = 0.05$. Since $n - k = 20 - 5 = 15$, the appropriate cutoff from the studentized range distribution is $Q_{.05,5,15} = 4.37$. Therefore, $D = 4.37/\sqrt{4} = 2.185$ and $D\sqrt{MSE} = 6.58$.

Table 12.3.2

Pairwise Difference	$\overline{Y}_{.i} - \overline{Y}_{.j}$	Tukey Interval	Conclusion
$\mu_1 - \mu_2$	−2.8	(−9.38, 3.78)	NS
$\mu_1 - \mu_3$	20.8	(14.22, 27.38)	Reject
$\mu_1 - \mu_4$	9.5	(2.92, 16.08)	Reject
$\mu_1 - \mu_5$	0.8	(−5.78, 7.38)	NS
$\mu_2 - \mu_3$	23.6	(17.02, 30.18)	Reject
$\mu_2 - \mu_4$	12.3	(5.72, 18.88)	Reject
$\mu_2 - \mu_5$	3.6	(−2.98, 10.18)	NS
$\mu_3 - \mu_4$	−11.3	(−17.88, −4.72)	Reject
$\mu_3 - \mu_5$	−20.0	(−26.58, −13.42)	Reject
$\mu_4 - \mu_5$	−8.7	(−15.28, −2.12)	Reject

```
> # Case Study 12.3.1
> # Input data first
> Input <- c("
+ rates group
+ 29.6  M1
+ 24.3  M1
+ 28.5  M1
+ 32.0  M1
+ 27.3  M2
+ 32.6  M2
+ 30.8  M2
+ 34.8  M2
+ 5.8   M3
+ 6.2   M3
+ 11.0  M3
+ 8.3   M3
+ 21.6  M4
+ 17.4  M4
+ 18.3  M4
+ 19.0  M4
+ 29.2  M5
+ 32.8  M5
+ 25.0  M5
+ 24.2 M5
+ ")
> Data = read.table(
  textConnection(Input),
+         header=TRUE)
```

```
> # Compute one-way ANOVA test
> res.aov <- aov(rates ~ group, data = Data)
> # Summary of the analysis
> summary(res.aov)
            Df Sum Sq Mean Sq F value Pr(>F)
group        4 1480.8   370.2   40.88 6.74e-08 ***
Residuals   15  135.8     9.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
        1
```

```
> # Tukey multiple pairwise-comparisons
> TukeyHSD(res.aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = rates ~ group, data = Data)

$group
          diff        lwr        upr     p adj
M2-M1   2.775  -3.795401   9.345401 0.6928357
M3-M1 -20.775 -27.345401 -14.204599 0.0000006
M4-M1  -9.525 -16.095401  -2.954599 0.0034588
M5-M1  -0.800  -7.370401   5.770401 0.9952758
M3-M2 -23.550 -30.120401 -16.979599 0.0000001
M4-M2 -12.300 -18.870401  -5.729599 0.0003007
M5-M2  -3.575 -10.145401   2.995401 0.4737713
M4-M3  11.250   4.679599  17.820401 0.0007429
M5-M3  19.975  13.404599  26.545401 0.0000010
M5-M4   8.725   2.154599  15.295401 0.0071611
```

```
> round(TukeyHSD(res.aov)$group,2)
        diff    lwr    upr p adj
M2-M1   2.78  -3.80   9.35  0.69
M3-M1 -20.77 -27.35 -14.20  0.00
M4-M1  -9.52 -16.10  -2.95  0.00
M5-M1  -0.80  -7.37   5.77  1.00
M3-M2 -23.55 -30.12 -16.98  0.00
M4-M2 -12.30 -18.87  -5.73  0.00
M5-M2  -3.58 -10.15   3.00  0.47
M4-M3  11.25   4.68  17.82  0.00
M5-M3  19.97  13.40  26.55  0.00
M5-M4   8.73   2.15  15.30  0.01
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Table 12.3.2

Pairwise Difference	$\overline{Y}_{.i} - \overline{Y}_{.j}$	Tukey Interval	Conclusion
$\mu_1 - \mu_2$	−2.8	(−9.38, 3.78)	NS
$\mu_1 - \mu_3$	20.8	(14.22, 27.38)	Reject
$\mu_1 - \mu_4$	9.5	(2.92, 16.08)	Reject
$\mu_1 - \mu_5$	0.8	(−5.78, 7.38)	NS
$\mu_2 - \mu_3$	23.6	(17.02, 30.18)	Reject
$\mu_2 - \mu_4$	12.3	(5.72, 18.88)	Reject
$\mu_2 - \mu_5$	3.6	(−2.98, 10.18)	NS
$\mu_3 - \mu_4$	−11.3	(−17.88, −4.72)	Reject
$\mu_3 - \mu_5$	−20.0	(−26.58, −13.42)	Reject
$\mu_4 - \mu_5$	−8.7	(−15.28, −2.12)	Reject

```
  > # Or one may use multcomp package or multiple comparisons
> library(multcomp)
> summary(glht(res.aov, linfct = mcp(group = "Tukey")))

   Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = rates ~ group, data = Data)

Linear Hypotheses:
             Estimate Std. Error t value Pr(>|t|)
M2 - M1 == 0    2.775      2.128   1.304  0.69283
M3 - M1 == 0  -20.775      2.128  -9.764  < 0.001 ***
M4 - M1 == 0   -9.525      2.128  -4.477  0.00348 **
M5 - M1 == 0   -0.800      2.128  -0.376  0.99528
M3 - M2 == 0  -23.550      2.128 -11.068  < 0.001 ***
M4 - M2 == 0  -12.300      2.128  -5.781  < 0.001 ***
M5 - M2 == 0   -3.575      2.128  -1.680  0.47374
M4 - M3 == 0   11.250      2.128   5.287  < 0.001 ***
M5 - M3 == 0   19.975      2.128   9.388  < 0.001 ***
M5 - M4 == 0    8.725      2.128   4.101  0.00717 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

```
             Estimate Std. Error t value Pr(>|t|)
M2 - M1 == 0    2.775      2.128   1.304  0.69283
M3 - M1 == 0  -20.775      2.128  -9.764  < 0.001 ***
M4 - M1 == 0   -9.525      2.128  -4.477  0.00348 **
M5 - M1 == 0   -0.800      2.128  -0.376  0.99527
M3 - M2 == 0  -23.550      2.128 -11.068  < 0.001 ***
M4 - M2 == 0  -12.300      2.128  -5.781  < 0.001 ***
M5 - M2 == 0   -3.575      2.128  -1.680  0.47371
M4 - M3 == 0   11.250      2.128   5.287  < 0.001 ***
M5 - M3 == 0   19.975      2.128   9.388  < 0.001 ***
M5 - M4 == 0    8.725      2.128   4.101  0.00719 **
```

Table 12.3.2

Pairwise Difference	$\overline{Y}_{.i} - \overline{Y}_{.j}$	Tukey Interval	Conclusion
$\mu_1 - \mu_2$	−2.8	(−9.38, 3.78)	NS
$\mu_1 - \mu_3$	20.8	(14.22, 27.38)	Reject
$\mu_1 - \mu_4$	9.5	(2.92, 16.08)	Reject
$\mu_1 - \mu_5$	0.8	(−5.78, 7.38)	NS
$\mu_2 - \mu_3$	23.6	(17.02, 30.18)	Reject
$\mu_2 - \mu_4$	12.3	(5.72, 18.88)	Reject
$\mu_2 - \mu_5$	3.6	(−2.98, 10.18)	NS
$\mu_3 - \mu_4$	−11.3	(−17.88, −4.72)	Reject
$\mu_3 - \mu_5$	−20.0	(−26.58, −13.42)	Reject
$\mu_4 - \mu_5$	−8.7	(−15.28, −2.12)	Reject

Two more examples of ANOVA using R

E.g. 1 `http://www.sthda.com/english/wiki/one-way-anova-test-in-r`

E.g. 2 `https://datascienceplus.com/one-way-anova-in-r/`