ORIGIN ≡ 1

One-Way Analysis of Variance with Fixed Effects Model:Generating the ANOVA Table & F-Test

Analysis of Variance (ANOVA) approaches comprise a broad class of statistical methods that also fall under the General Linear Models framework. Independent variable(s) in ANOVA involve membership in classes. Since more than two classes may be present, this approach allows extension of the t-test strategy of compairing means between two populations to comparisions of means between multiple populations. Since ANOVA is ubiquitous in many settings in biology, proficient use is often viewed as evidence of good experimental design.

Data Structure:

k groups with not necessarily the same numbers of observations and different means.

Let index i,j indicate the ith column (treatment class) and jth row (object).

One-Way ANOVA					
	Treatment Classes:				
Objects (Replicates)	#1	#2	#3	...	#k
1					
2					
3					
...					
n	n1	n2	n3		nk
means:	X1bar	X2bar	X3bar	...	Xkbar

ZAR := READPRN("c:/DATA/Biostatistics/ZarEX10.1R.txt")

Zar's Example 10.1

$j := 1..5$

$X1_j := \left(ZAR^{\langle 2\rangle}\right)_j$

$X2_j := \left(ZAR^{\langle 2\rangle}\right)_{j+5}$

$j := 1..4$

$X3_j := \left(ZAR^{\langle 2\rangle}\right)_{j+10}$

$j := 1..5$

$X4_j := \left(ZAR^{\langle 2\rangle}\right)_{j+14}$

$X := ZAR^{\langle 2\rangle}$

ZAR =

	1	2	3
1	1	60.8	1
2	2	67	1
3	3	65	1
4	4	68.6	1
5	5	61.7	1
6	6	68.7	2
7	7	67.7	2
8	8	75	2
9	9	73.3	2
10	10	71.8	2
11	11	69.6	3
12	12	77.1	3
13	13	75.2	3
14	14	71.5	3
15	15	61.9	4
16	16	64.2	4
17	17	63.1	4
18	18	66.7	4
19	19	60.3	4

$X1 = \begin{pmatrix} 60.8 \\ 67 \\ 65 \\ 68.6 \\ 61.7 \end{pmatrix}$ $X1_{bar} := mean(X1)$ $X1_{bar} = 64.62$; $n_1 := length(X1)$ $n_1 = 5$; $s_1 := \sqrt{Var(X1)}$ $s_1 = 3.3454$

$X2 = \begin{pmatrix} 68.7 \\ 67.7 \\ 75 \\ 73.3 \\ 71.8 \end{pmatrix}$ $X2_{bar} := mean(X2)$ $X2_{bar} = 71.3$; $n_2 := length(X2)$ $n_2 = 5$; $s_2 := \sqrt{Var(X2)}$ $s_2 = 3.0684$

$X3 = \begin{pmatrix} 69.6 \\ 77.1 \\ 75.2 \\ 71.5 \end{pmatrix}$ $X3_{bar} := mean(X3)$ $X3_{bar} = 73.35$; $n_3 := length(X3)$ $n_3 = 4$; $s_3 := \sqrt{Var(X3)}$ $s_3 = 3.4142$

$X4 = \begin{pmatrix} 61.9 \\ 64.2 \\ 63.1 \\ 66.7 \\ 60.3 \end{pmatrix}$ $X4_{bar} := mean(X4)$ $X4_{bar} = 63.24$; $n_4 := length(X4)$ $n_4 = 5$; $s_4 := \sqrt{Var(X1)}$ $s_4 = 3.3454$

index variables:
j indexes individual cases within classes and ranges from 1 to n_j for each group.
i indexes classes from 1 to k, k is the total number of classes.

$k := 4$

$i := 1..k$

$X_{bar} := \begin{pmatrix} X1_{bar} \\ X2_{bar} \\ X3_{bar} \\ X4_{bar} \end{pmatrix}$ $X_{bar} = \begin{pmatrix} 64.62 \\ 71.3 \\ 73.35 \\ 63.24 \end{pmatrix}$ $n := \begin{pmatrix} n_1 \\ n_2 \\ n_3 \\ n_4 \end{pmatrix}$ $n = \begin{pmatrix} 5 \\ 5 \\ 4 \\ 5 \end{pmatrix}$ $s := \begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \end{pmatrix}$ $s = \begin{pmatrix} 3.3454 \\ 3.0684 \\ 3.4142 \\ 3.3454 \end{pmatrix}$ **< summary statistics**

Treatments Model:

$X_{i,j} = \mu + \alpha_i + \varepsilon_{i,j}$ < where:

μ is the grand mean of all objects.
α_i where $\mu + \alpha_i$ is the average for each class i.
$\varepsilon_{i,j}$ is the error term specific to each object i,j

With Restriction:

$\sum_i n_i \cdot \alpha_i := 0$ or $\sum_i \alpha_i := 0$ or $\alpha_k := 0$ < allows estimation of k parameters. See Rosner 2006 p. 558

Cell Means Model:

$X_{i,j} = \mu_i + \varepsilon_{i,j}$ < where:

μ_i is mean for each treatment class.
$\varepsilon_{i,j}$ is the error term specific to each object i,j

Assumptions:

ε_{ij} are a random sample ~ $N(0,\sigma^2)$

Number & Means:

$N := \sum_i n_i$ $N = 19$ < total number of observations

$GM := \text{mean}(X)$ $GM = 67.8526$ < grand mean - sample estimate of μ

$X_{bar} = \begin{pmatrix} 64.62 \\ 71.3 \\ 73.35 \\ 63.24 \end{pmatrix}$ < sample means for each class

$X = \begin{pmatrix} 60.8 \\ 67 \\ 65 \\ 68.6 \\ 61.7 \\ 68.7 \\ 67.7 \\ 75 \\ 73.3 \\ 71.8 \\ 69.6 \\ 77.1 \\ 75.2 \\ 71.5 \\ 61.9 \\ 64.2 \\ 63.1 \\ 66.7 \\ 60.3 \end{pmatrix}$

Sums of Squares:

$n = \begin{pmatrix} 5 \\ 5 \\ 4 \\ 5 \end{pmatrix}$ $(X_{bar} - GM)^2 = \begin{pmatrix} 10.4499 \\ 11.8843 \\ 30.2211 \\ 21.2764 \end{pmatrix}$ < square of difference of each class mean from the grand mean. This is the sample estimate of α^2.

Between:

$SS_B := \sum_i n_i \cdot \left[(X_{bar} - GM)^2\right]_i$ $SS_B = 338.9374$ < Between (Treatment) Sum of Squares

^ sum of the above squared differences times the number in each class.

Within:

$(X1 - X_{bar_1})^2 = \begin{pmatrix} 14.5924 \\ 5.6644 \\ 0.1444 \\ 15.8404 \\ 8.5264 \end{pmatrix}$ $(X2 - X_{bar_2})^2 = \begin{pmatrix} 6.76 \\ 12.96 \\ 13.69 \\ 4 \\ 0.25 \end{pmatrix}$ $(X3 - X_{bar_3})^2 = \begin{pmatrix} 14.0625 \\ 14.0625 \\ 3.4225 \\ 3.4225 \end{pmatrix}$ $(X4 - X_{bar_4})^2 = \begin{pmatrix} 1.7956 \\ 0.9216 \\ 0.0196 \\ 11.9716 \\ 8.6436 \end{pmatrix}$

^ square of the difference of each item in a class from its respective class mean

$SS_W := \sum (X1 - X_{bar_1})^2 + \sum (X2 - X_{bar_2})^2 + \sum (X3 - X_{bar_3})^2 + \sum (X4 - X_{bar_4})^2$ $SS_W = 140.75$

^ sum over all individuals and all classes

^ Within Sum of Squares

Total:

$jj := 1 .. N$ **< jj is index for all cases regardless of class**

$SS_T := \sum_{jj} (X_{jj} - GM)^2$ $SS_T = 479.6874$ **< Total Sum of Squares**

$(N-1) \cdot Var(X) = 479.6874$

$SS_B + SS_W = 479.6874$

One-Way ANOVA Table:

Source:	SS	df		MS	
Between	$SS_B := 338.9373684$	$df_B := k - 1$	$df_B = 3$	$MS_B := \frac{SS_B}{df_B}$	$MS_B = 112.9791$
Within	$SS_W := 140.7500000$	$df_W := N - k$	$df_W = 15$	$MS_W := \frac{SS_W}{df_W}$	$MS_W = 9.3833$
TOTAL	$SS_T := SS_B + SS_W$				
	$SS_T = 479.6874$				

Prototype in R:

NOTE that there is a RIGHT WAY and a WRONG WAY to do ANOVA in R:

```
#CALCULATING A ONE-WAY ANOVA TABLE

ZAR=read.table("c:/DATA/Biostatistics/ZarEX10.1.txt")
ZAR
attach(ZAR)
Y=weights
```

WRONG WAY:

```
#WRONG WAY:
X=feed
anova(lm(Y~X))
```

```
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value Pr(>F)
X          1   2.57    2.57  0.0916 0.7659
Residuals 17 477.12   28.07
```

What this does is produce an ANOVA on the Linear Regression of the dependent variable (Y) with the *numerical values* reported for the independent variable (X). Since these values are class indicators (1,2,3) they are meaningless. The fact that this is an incorect Linear Regression ANOVA can be seen in the report of 1 DF for variable X in the ANOVA chart.

RIGHT WAY:

< It is important to tell R that the class variable is a "factor" using the factor() function...

```
#RIGHT WAY:
X=factor(feed)
anova(lm(Y~X))
```

```
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value   Pr(>F)
X          3 338.94  112.98  12.040 0.000283 ***
Residuals 15 140.75    9.38
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

^ These values match above.

F-Test for H_0: All $\alpha_i = 0$ in One-Way ANOVA with Fixed Effects Model

Inferences on the means of the multiple populations indicated by the class ("factor" or "group") variable follow directly from the ANOVA table. The following test is often referred to as *omnibus test* given the generality of the alternate hypothesis H_1.

Hypotheses for Treatments Model:

H_0: $\alpha_i = 0$ *for all* i

H_1: *At least one* $\alpha_i \neq 0$ **< One sided test**

^ here α_i is NOT the same thing as Type I error rate α below

Hypotheses for Cell Means Model:

Note that hypotheses here are fully equivalent!

H_0: μ_i are *all* equal *for all* i

H_1: *At least one* μ_i not the same as the others **< One sided test**

Test Statistic:

$F := \frac{MS_B}{MS_W}$ $F = 12.0404$ **< Ratio of "between" versus "within" Mean Squares**

Distribution of the test Statistic F:

If H_0 is true then $F \sim F_{((k-1),(N-k))}$ **where: k = number of classes**
N = total number of observations

Critical Value of the Test:

$\alpha := 0.05$ **< Probability of Type I error must be explicitly set**

$CV := qF[1-\alpha,(k-1),(N-k)]$ $CV = 3.2874$

Decision Rule:

IF F > CV, THEN REJECT H_0, OTHERWISE ACCEPT H_0

Probability Value:

$P := 1 - pF[F,(k-1),(N-k)]$ $P = 0.00028301$

Prototype in R:

F-statistic & Probability value are given in ANOVA report above.

Note: to obtain more complete values in the ANOVA report, I set number of significant digits to 12 using the options switch within the lm() function, within anova().

#TO INCREASE NUMBER OF SIGNIFICANT DIGITS DO THIS:
anova(lm(Y~X),options(digits=12))

```
Analysis of Variance Table

Response: Y
          Df       Sum Sq      Mean Sq F value      Pr(>F)
X          3 338.9373684 112.9791228 12.0404 0.00028301 ***
Residuals 15 140.7500000   9.3833333
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```