

# Analysis of Variance

1. One-Way ANOVA
2. Two-Way ANOVA
3. Multivariate Analysis of Variance

Variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

**Analysis of variance (ANOVA)** is used

● to investigate the relationship between categorical independent variables and continuous dependent variables

```
InsectSprays[1:3,]
```

count	spray
10	A
7	A
20	A

Continuous Dependent Variable

Categorical Independent Variable

```
table(InsectSprays$spray)
```

A	B	C	D	E	F
12	12	12	12	12	12

● to analyze the differences among group means and their associated procedures

## 1. One-Way ANOVA

One-way ANOVA is used to determine whether there are any statistically significant differences between the means of two or more independent groups.

Data Structure

Group	1	2	...	$k$
	$x_{11}$	$x_{21}$	...	$x_{k1}$
	$x_{12}$	$x_{22}$	...	$x_{k2}$
	$\vdots$	$\vdots$		$\vdots$
	$x_{1n_1}$	$x_{2n_2}$	...	$x_{kn_k}$
Group mean	$\mu_1$	$\mu_2$	...	$\bar{\mu}_k$

### Hypothesis

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

$$H_A: \exists i, j: \mu_i \neq \mu_j$$



## Sample Data : InsectSprays

### **InsectSprays** {datasets} Effectiveness of Insect Sprays

The counts of insects in agricultural experimental units treated with different insecticides.

```
> head(InsectSprays,4)
```

```
count spray
1    10    A
2     7    A
3    20    A
4    14    A
```

```
> table(InsectSprays$spray)
```

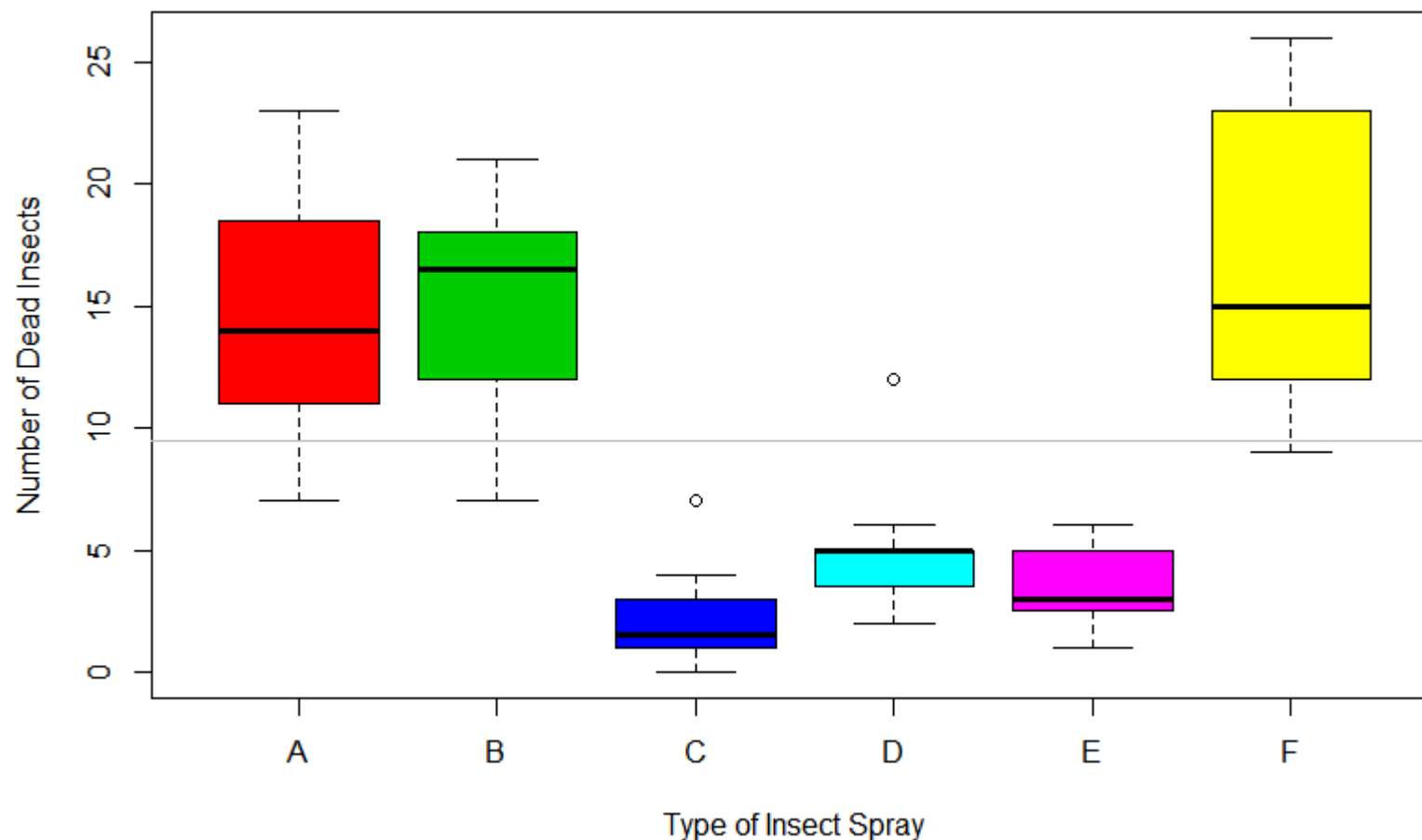
```
 A  B  C  D  E  F
12 12 12 12 12 12
```

spray	A	B	C	D	E	F
	10	11	0	3	3	11
	7	17	1	5	5	9
	20	21	7	12	3	15
	14	11	2	6	5	22
	14	16	3	4	3	15
	12	14	1	3	6	16
	10	17	2	5	1	13
	23	17	1	5	1	10
	17	19	3	5	3	26
	20	21	0	5	2	26
	14	7	1	2	6	24
	13	13	4	4	4	13
mean	14.5	15.3	2.1	4.9	3.5	16.7



## Step 1. Examine the mean differences

```
boxplot(count~spray, data=InsectSprays,  
        xlab='Type of Insect Spray',  
        ylab='Number of Dead Insects', col=2:7)  
abline(h=mean(InsectSprays$count),col='gray')
```



## Step 2. Compute ANOVA table for a fitted model

```
aov(formula, data, ... )
```

Fit an analysis of variance model by a call to `lm` for each stratum.

formula:  $y(\text{continuous variable}) \sim X$  (categorical variable with two or more groups)

```
#Conducting One-Way ANOVA
```

```
aov.out <- aov(count~spray,data=InsectSprays)
```

```
summary(aov.out)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
spray	5	2669	533.8	34.7	<2e-16 ***
Residuals	66	1015	15.4		

#mean square  
=sample variance  
=sum of squares/Df

#SST = sum of the squares of the deviations of all observations

$$SS_{Total} = \text{Total Sums of Squares} = \sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2 \quad \bar{y} = \text{grand mean}$$

$$SS_A = \text{Explanatory Variable A's Sums of Squares} = \sum_{j=1}^J \sum_{i=1}^{n_j} (\bar{y}_j - \bar{y})^2$$

$$SS_E = \text{Error (Residual) Sums of Squares} = \sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

$$SS_{Total} = SS_A + SS_E$$

**anova**(object, ...)

Compute analysis of variance tables for one or more fitted model objects.

```
> an3 <- anova(lm(count ~ spray))
```

```
> an3
```

Analysis of Variance Table

Response: count

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
spray	5	2668.8	533.77	34.702	< 2.2e-16 ***
Residuals	66	1015.2	15.38		

Anova Analysis Result

	Sum Sq	Df	Mean Sq	F	p-value
spray	2668.8	5	533.8	34.702	0.000
Residuals	1015.2	66	15.38		
Sum	3684.0	71			

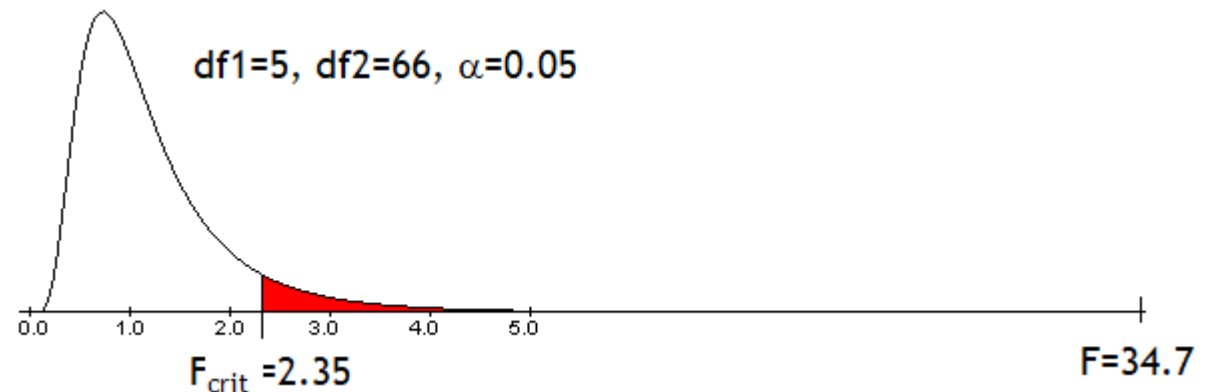
The difference between lm and aov is mainly in the form of the output. If you have multiple error terms then you must use aov because lm does not support the Error term.

### Step 3. Conclude with the F test result

```
qf(p, df1, df2, ncp, lower.tail=TRUE)
```

Quantile function for the F distribution with df1 and df2 degrees of freedom

```
>  
> df1 = 5; df2 = 66  
> a = 0.05  
> Fc = qf(1-a, df1, df2) #critical F  
> Fc  
[1] 2.353809
```



Interpretation:

- (1)  $p < 0.05$
- (2)  $F = 34.7$  (Large F-value indicates that the model is significant.)
- (3) We reject the null hypothesis.
- (4) We conclude that count of insects varies with respect to spray type.

**model.tables**(x, ...)

Computes summary tables for model fits, especially complex aov fits.

```
> print(model.tables(aov.out, "means"), digits=3)
```

Tables of means

Grand mean

9.5

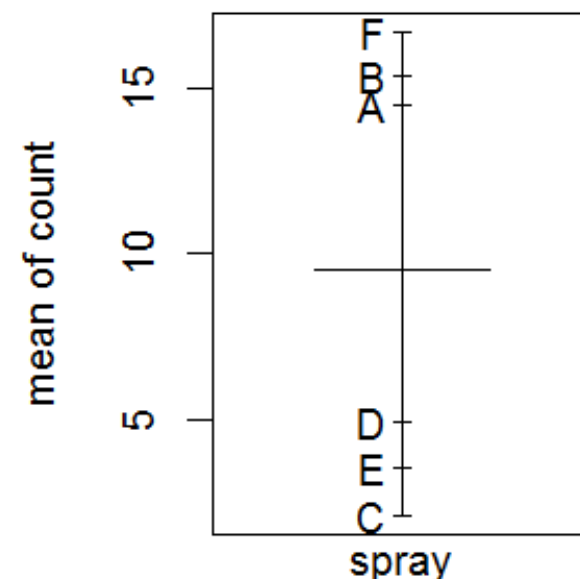
spray

	A	B	C	D	E	F
	14.50	15.33	2.08	4.92	3.50	16.67

**plot.design**(x, fun, data=NULL, ... )

Plot univariate effects of one or more factors, typically for a designed experiment as analyzed by aov().

```
plot.design(InsectSprays)
```





## Step 4. Perform Tukey HSD (Honestly Significant Difference) Post Hoc Test

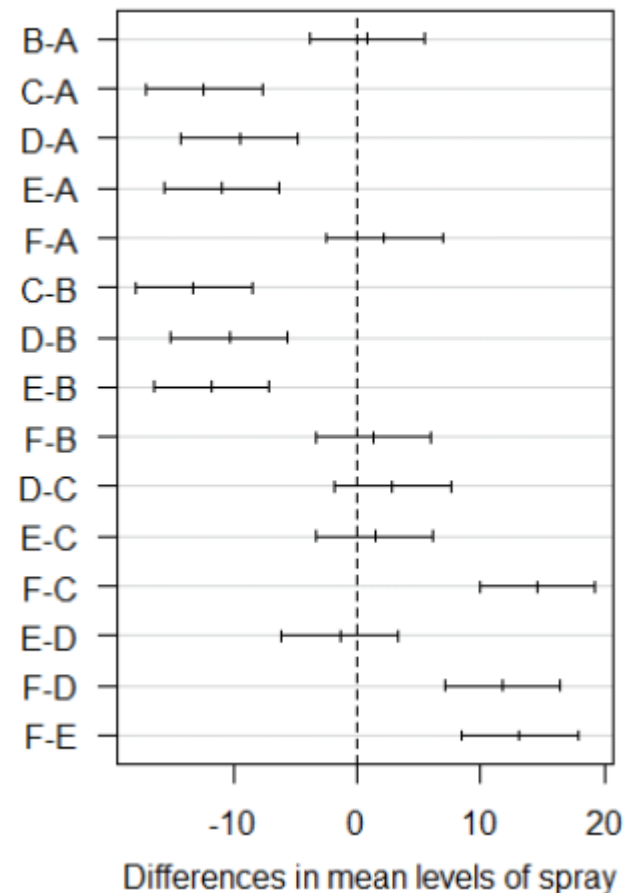
```
> tkh <- TukeyHSD(aov.out, conf.level=0.95)
> tkh
Tukey multiple comparisons of means
 95% family-wise confidence level

$spray
```

	diff	lwr	upr	p adj
B-A	0.8333333	-3.866075	5.532742	0.9951810
C-A	-12.4166667	-17.116075	-7.717258	0.0000000
D-A	-9.5833333	-14.282742	-4.883925	0.0000014
E-A	-11.0000000	-15.699409	-6.300591	0.0000000
F-A	2.1666667	-2.532742	6.866075	0.7542147
C-B	-13.2500000	-17.949409	-8.550591	0.0000000
D-B	-10.4166667	-15.116075	-5.717258	0.0000002
E-B	-11.8333333	-16.532742	-7.133925	0.0000000
F-B	1.3333333	-3.366075	6.032742	0.9603075
D-C	2.8333333	-1.866075	7.532742	0.4920707
E-C	1.4166667	-3.282742	6.116075	0.9488669
F-C	14.5833333	9.883925	19.282742	0.0000000
E-D	-1.4166667	-6.116075	3.282742	0.9488669
F-D	11.7500000	7.050591	16.449409	0.0000000
F-E	13.1666667	8.467258	17.866075	0.0000000

```
plot(tkh, las=1)
```

95% family-wise confidence level



This output indicates that the differences B-A, F-A, F-B, D-C, E-C, and E-D are not significant. Meanwhile, C-A, D-A, E-A, C-B, D-B, E-B, F-C, F-D and F-E are significant.

## 2. Two-Way ANOVA

The **two-way analysis of variance (ANOVA)** examines the influence of **two different categorical independent variables (called factors)** on **one continuous dependent variable**.

Understand if there is an **interaction** between the two variables. Two-Way ANOVA is referred to a Factorial ANOVA.

Source of variation	DF	Sum of squares SS	Mean square MS	F	P-value
Factor A	$DFA = I - 1$	SSA	$SSA/DFA$	$MSA/MSE$	for $F_A$
Factor B	$DFB = J - 1$	SSB	$SSB/DFB$	$MSB/MSE$	for $F_B$
Interaction	$DFAB = (I-1)(J-1)$	SSAB	$SSAB/DFAB$	$MSAB/MSE$	for $F_{AB}$
Error	$DFE = N - IJ$	SSE	$SSE/DFE$		
Total	$DFT = N - 1$ $= DFA + DFB + DFAB + DFE$	$SST$ $= SSA + SSB + SSAB + SSE$	$SST/DFT$		

## Examples:

**One-way ANOVA** : Do mice weigh more in early or late mating season?

**Two-way ANOVA** : Are mice heavier in early or late mating season and does that depend on the gender of the mice?

**Hypothesis** for two factors ( $A$ ,  $B$ ) and a dependent variable ( $D$ )

**Interaction Effect:** Two variables interact if a combination of variables affects results that would not be anticipated from their main effects.

$H_0$ :  $A$  and  $B$  interaction will have no significant effect on  $D$

### Main effect :

$A$ :  $A$  will have no significant effect on  $D$  ( $H_0$ )

$B$ :  $B$  will have no significant effect on  $D$  ( $H_0$ )

## Sample Data : schooldays

```
data("schooldays") {HSAUR3}
```

Data from a sociological study, the number of days absent from school is the response variable.

race : race of the child, a factor with levels aboriginal and non-aboriginal.

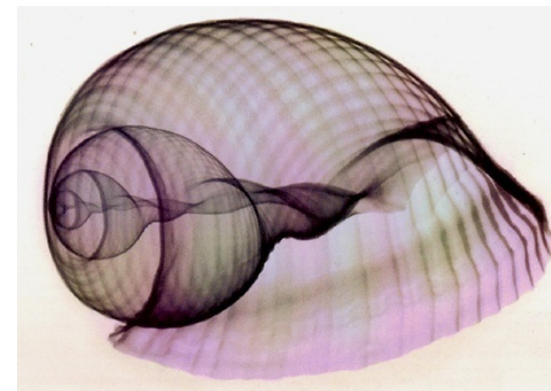
school : the school type, a factor with levels F0 (primary), F1 (first), F2 (second) and F3 (third form).

absent : number of days absent from school.

```
> library(HSAUR3)
```

```
> head(schooldays,2)
```

	race	gender	school	learner	absent
1	aboriginal	male	F0	slow	2
2	aboriginal	male	F0	slow	11



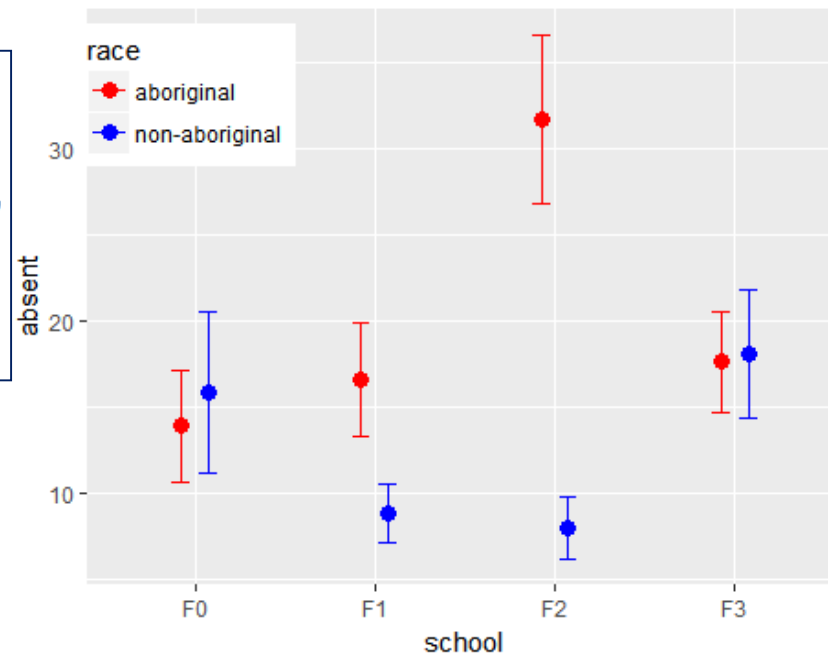
# Means and Summary Statistics by Group

```
> library(Rmisc)
> sum = summarySE(schooldays,measurevar="absent",
+                 groupvars=c("race","school"))
> sum
```

	race	school	N	absent	sd	se	ci
1	aboriginal	F0	13	13.846154	11.859217	3.289155	7.166453
2	aboriginal	F1	20	16.550000	14.752252	3.298704	6.904267
3	aboriginal	F2	20	31.650000	21.947965	4.907714	10.271964
4	aboriginal	F3	21	17.571429	13.253571	2.892166	6.032953
5	non-aboriginal	F0	14	15.785714	17.493641	4.675372	10.100528
6	non-aboriginal	F1	28	8.821429	9.059693	1.712121	3.512982
7	non-aboriginal	F2	18	7.944444	7.741958	1.824797	3.849985
8	non-aboriginal	F3	20	18.050000	16.693759	3.732838	7.812920

## Standard Error Plot Using Summary Statistics

```
library(ggplot2)
pd = position_dodge(0.3)
ggplot(sum, aes(x=school,y=absent,color=race)) +
  geom_errorbar(aes(ymin=absent-se,ymax=absent+se),
               width=0.2,size=0.7,position=pd) +
  geom_point(shape=16, size=3, position=pd) +
  scale_color_manual(values=c("red","blue")) +
  theme(legend.position=c(0.13,0.85))
```



# Conducting Two-Way ANOVA

```
> aov2 <- aov(absent ~ race*school, data=schooldays)
> summary(aov2)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
race	1	2646	2645.7	12.461	0.000556	***
school	3	1326	441.8	2.081	0.105245	
race:school	3	3738	1246.1	5.869	0.000822	***
Residuals	146	30997	212.3			

#race\*school = race + school + race:school

## Main effect 1:

P-value for race is 0.001. We reject the null hypothesis that the means of absent evaluated according to the race are equal.

## Main effect 2:

P-value for school is 0.105. We retain the null hypothesis that the means of absent evaluated according to the school are equal.

## Interaction effect:

P-value for race:school is 0.001. The interaction between race and school is statistically significant and we reject the null hypothesis.



# Summary Tables for ANOVA Model Fits

```
> print(model.tables(aov2,"means"),digits=4)
```

Tables of means

Grand mean

16.13636

race

aboriginal	non-aboriginal
20.45	12.15

rep 74.00 80.00 ← n

school

F0	F1	F2	F3
14.84	12.57	20.04	17.54

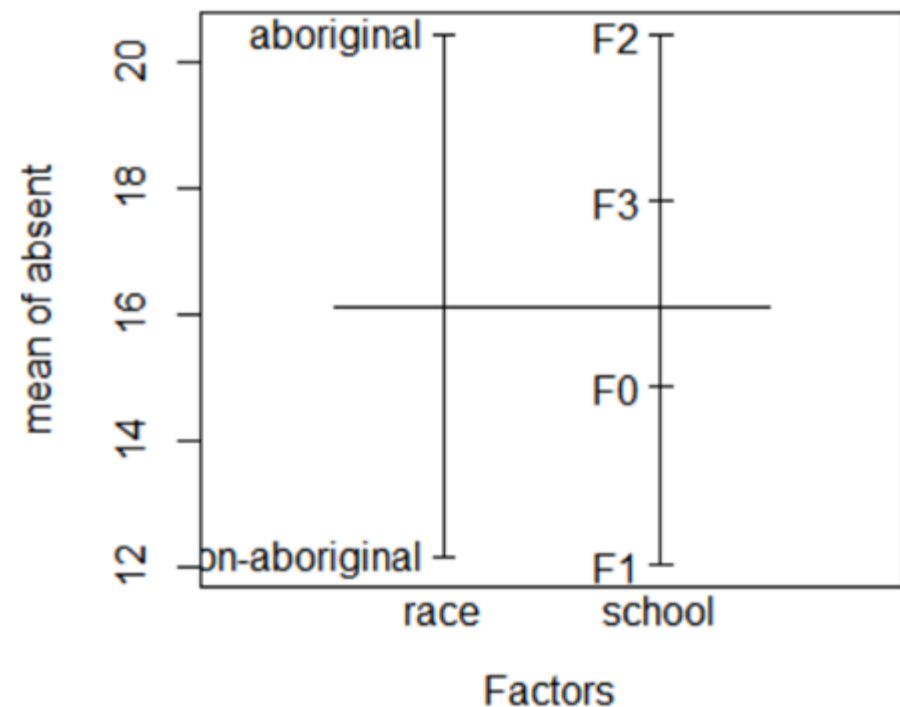
rep 27.00 48.00 38.00 41.00

race:school

race	school			
	F0	F1	F2	F3
aboriginal	13.85	16.55	31.65	17.57
rep	13.00	20.00	20.00	21.00
non-aboriginal	15.79	8.82	7.94	18.05
rep	14.00	28.00	18.00	20.00

## Univariate Effects of race and school Factors

```
plot.design(absent ~ race+school,
             data=schooldays)
```





## Interaction Plot

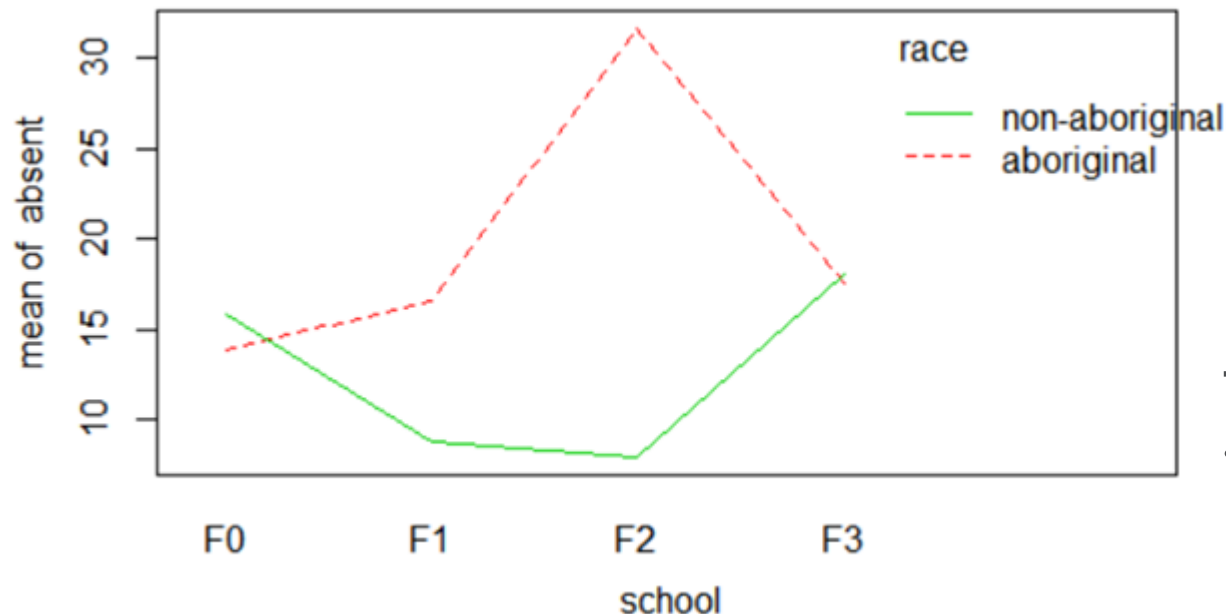
An interaction plot is used to characterize the relationship between variables.

```
interaction.plot(x.factor, trace.factor, response, fun=mean, ... ) {stats}
```

Two-way Interaction Plot

Plots the mean (or other summary) of the response for two-way combinations of factors, thereby illustrating possible interactions.

```
with(schooldays, interaction.plot(x.factor=school,  
                                trace.factor=race, response=absent, col=2:3))
```



In this interaction plot, the lines are not parallel. This interaction effect indicates the relationship between school and race.



# Conducting Two-Way ANOVA

## Finding the minimal adequate model:

In the previous slide, factor “school” alone was not significant. So here we remove this factor to arrive at a minimal adequate model.

Step	Procedure	Explanation
1	Fit the maximal model	Fit all the factors, interactions and covariates of interest. Note the residual deviance. If you are using Poisson or binomial errors, check for overdispersion and rescale if necessary
2	Begin model simplification	Inspect the parameter estimates using <b>summary</b> . Remove the least significant terms first, using <b>update -</b> , starting with the highest order interactions
3	If the deletion causes an insignificant increase in deviance	Leave that term out of the model Inspect the parameter values again Remove the least significant term remaining
4	If the deletion causes a significant increase in deviance	Put the term back in the model using <b>update +</b> . These are the statistically significant terms as assessed by deletion from the maximal model
5	Keep removing terms from the model	Repeat steps 3 or 4 until the model contains nothing but significant terms This is the minimal adequate model If none of the parameters is significant, then the minimal adequate model is the null model

### 3. Multivariate Analysis of Variance

The multivariate analysis of variance (MANOVA) is a type of multivariate analysis used to analyze data that involves more than one dependent variable at a time.

MANOVA allows us to test hypotheses regarding the effect of one or more independent variables on two or more dependent variables.

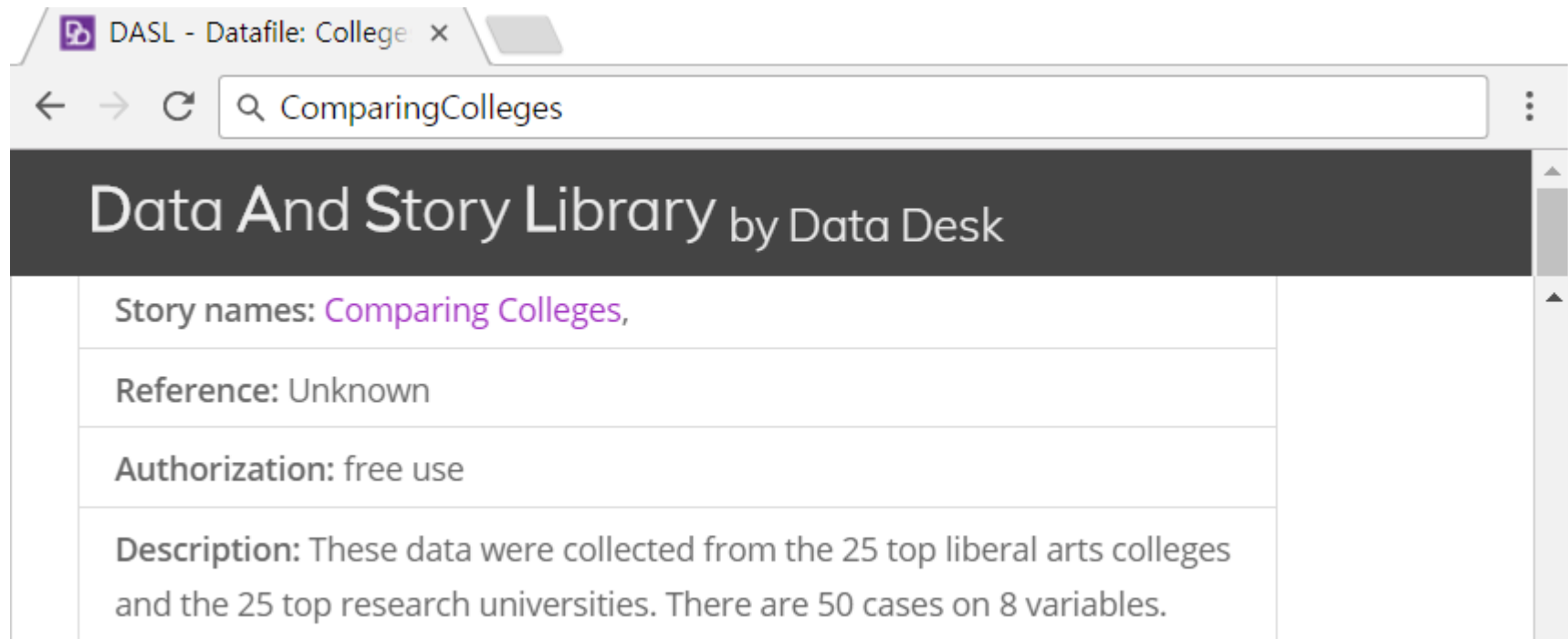
#### **One-Way MANOVA**

The one-way MANOVA is presented using the data set that contains one independent variable and two or more dependent variables.



## Example : ComparingColleges

<http://dasl.datadesk.com/datafiles>



The screenshot shows a web browser window with the title 'DASL - Datafile: College'. The address bar contains 'ComparingColleges'. The main heading is 'Data And Story Library by Data Desk'. Below this, there is a table with the following information:

Story names:	Comparing Colleges,
Reference:	Unknown
Authorization:	free use
Description:	These data were collected from the 25 top liberal arts colleges and the 25 top research universities. There are 50 cases on 8 variables.

The [ComparingColleges](#) dataset contains information on six continuous dependent variables along with a single discrete variable School Type. Twenty-five of each type of school were surveyed.

1. School: Contains the name of each school
2. School\_Type: Coded 'LibArts' for liberal arts and 'Univ' for university
3. SAT: Median combined Math and Verbal SAT score of students
4. Acceptance: % of applicants accepted
5. StudentP: Money spent per student in dollars
6. Top10P: % of students in the top 10% of their h.s. graduating class
7. PhDP: % of faculty at the institution that have PhD degrees
8. GradP: % of students at institution who eventually graduate

```
> cdt <- read.csv("ComparingColleges.csv",header=T)
> attach(cdt); dim(cdt)
[1] 50  8
> head(cdt)
```

	School	School_Type	SAT	Acceptance	StudentP	Top10P	PhDP	GradP
1	Amherst	Lib Arts	1315	22	26636	85	81	93
2	Swarthmore	Lib Arts	1310	24	27487	78	93	88
3	Williams	Lib Arts	1336	28	23772	86	90	93
4	Bowdoin	Lib Arts	1300	24	25703	78	95	90
5	Wellesley	Lib Arts	1250	49	27879	76	91	86
6	Pomona	Lib Arts	1320	33	26668	79	98	80

```
> table(cdt$School_Type)
```

Lib Arts	25	Univ	25
----------	----	------	----

## Conducting one-Way MANOVA

**manova**(formula, data, ...) {stats}

A class for the multivariate analysis of variance.

We will use MANOVA to determine if the **school types** differ across all six dependent variables (SAT, Acceptance, StudentP, PhdP, GradP, Top10P) simultaneously.

```
> Y <- cbind(SAT,Acceptance,StudentP,PhdP,GradP,Top10P)
> table(School_Type)
School_Type
Lib Arts      Univ
      25      25
> fit <- manova(Y ~ School_Type)
> summary(fit)
```

	Df	Pillai	approx F	num Df	den Df	Pr(>F)
School_Type	1	0.54081	8.4405	6	43	4.507e-06 ***
Residuals	48					

We reject the null hypothesis at  $\alpha=0.05$ .

MANOVA of the data with School\_Type as the factor reveals that there is a significant difference between research universities and liberal arts colleges at the 5% level.



The `summary.aov()` function will yield the ANOVA univariate statistics for each of the dependent variables.

```
> summary.aov(fit)
```

```
Response SAT :
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
School_Type	1	2679	2679.1	0.6852	0.4119
Residuals	48	187685	3910.1		

```
Response Acceptance :
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
School_Type	1	369.9	369.92	2.1187	0.152
Residuals	48	8380.8	174.60		

```
Response StudentP :
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
School_Type	1	3605397494	3605397494	22.146	2.177e-05
Residuals	48	7814347896	162798914		

```
Response PhDP :
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
School_Type	1	269.12	269.120	4.2034	0.04583 *
Residuals	48	3073.20	64.025		

```
Response GradP :
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
School_Type	1	20.48	20.480	0.3539	0.5547
Residuals	48	2778.00	57.875		

```
Response Top10P :
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
School_Type	1	2592.0	2592.00	19.567	5.559e-05
Residuals	48	6358.3	132.47		



The differences lie in the 3 dependent variables.

- Which school type has the higher mean for the three significant variables?

```
> tapply(StudentP,School_Type,mean)
Lib Arts      Univ
21755.56 38738.84
> tapply(PhDP,School_Type,mean)
Lib Arts      Univ
  88.24    92.88
> tapply(Top10P,School_Type,mean)
Lib Arts      Univ
  67.24    81.64
```

```
par(mai=c(0.8,0.8,0.2,0.2),mfrow=c(1,3))
plot.design(StudentP~School_Type)
plot.design(PhDP ~ School_Type)
plot.design(Top10P ~ School_Type)
```

The liberal arts colleges have more students per \$, which means that the universities spend more per student.

Universities have more PhD's and Top 10% students than liberal arts colleges.

