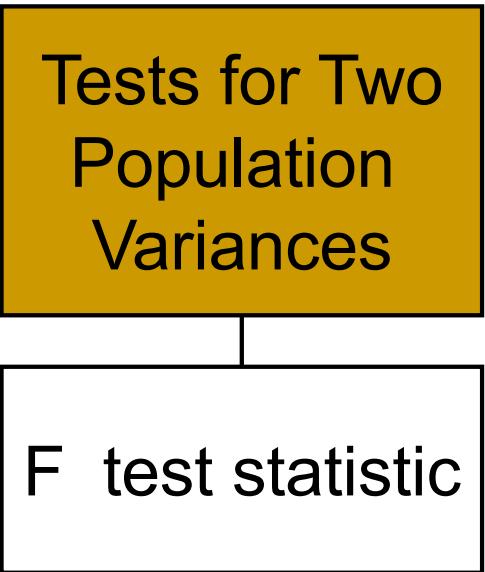


Analysis of Variance

Hypothesis Tests for Variances



$$\begin{aligned} H_0: \sigma_1^2 &= \sigma_2^2 \\ H_1: \sigma_1^2 &\neq \sigma_2^2 \end{aligned}$$

Two-tail test

$$\begin{aligned} H_0: \sigma_1^2 &\geq \sigma_2^2 \\ H_1: \sigma_1^2 &< \sigma_2^2 \end{aligned}$$

Lower-tail test

$$\begin{aligned} H_0: \sigma_1^2 &\leq \sigma_2^2 \\ H_1: \sigma_1^2 &> \sigma_2^2 \end{aligned}$$

Upper-tail test

Test for Equal Variances of Two normally distributed Populations

Hypothesis Tests for Variances

The F test statistic is:

Tests for Two
Population
Variances

$$F = \frac{S_1^2}{S_2^2}$$

F test statistic



S_1^2 = Variance of Sample 1

$n_1 - 1$ = numerator degrees of freedom

S_2^2 = Variance of Sample 2

$n_2 - 1$ = denominator degrees of freedom

Hypothesis Tests for Variances

Test $H_0: \sigma_1^2/\sigma_2^2 \leq 1$ vs. $\sigma_1^2/\sigma_2^2 > 1$

- Reject H_0 in favor of H_a if s_1^2/s_2^2 is significantly greater than 1
- s_1^2 is the variance of a random sample of size n_1 from a population with variance σ_1^2
- s_2^2 is the variance of a random sample of size n_2 from a population with variance σ_2^2
- To decide how large s_1^2/s_2^2 must be to reject H_0 , describe the sampling distribution of s_1^2/s_2^2
- The sampling distribution of s_1^2/s_2^2 is the F distribution

F Distribution

The curve of the F distribution depends on degrees of freedom of numerator and degrees of freedom of denominator .

$$F = \frac{S_1^2}{S_2^2} \quad \begin{aligned} n_1 - 1 &= \text{numerator degrees of freedom,} \\ n_2 - 1 &= \text{denominator degrees of freedom} \end{aligned}$$

```
x=seq(0,10,by=0.01)
y=df(x,10,15)
plot(x,y,type='l',col='blue')
qf(0.95,10,15)
[1] 2.543719
```

```
pf(2.5438,10,15)
[1] 0.95
```

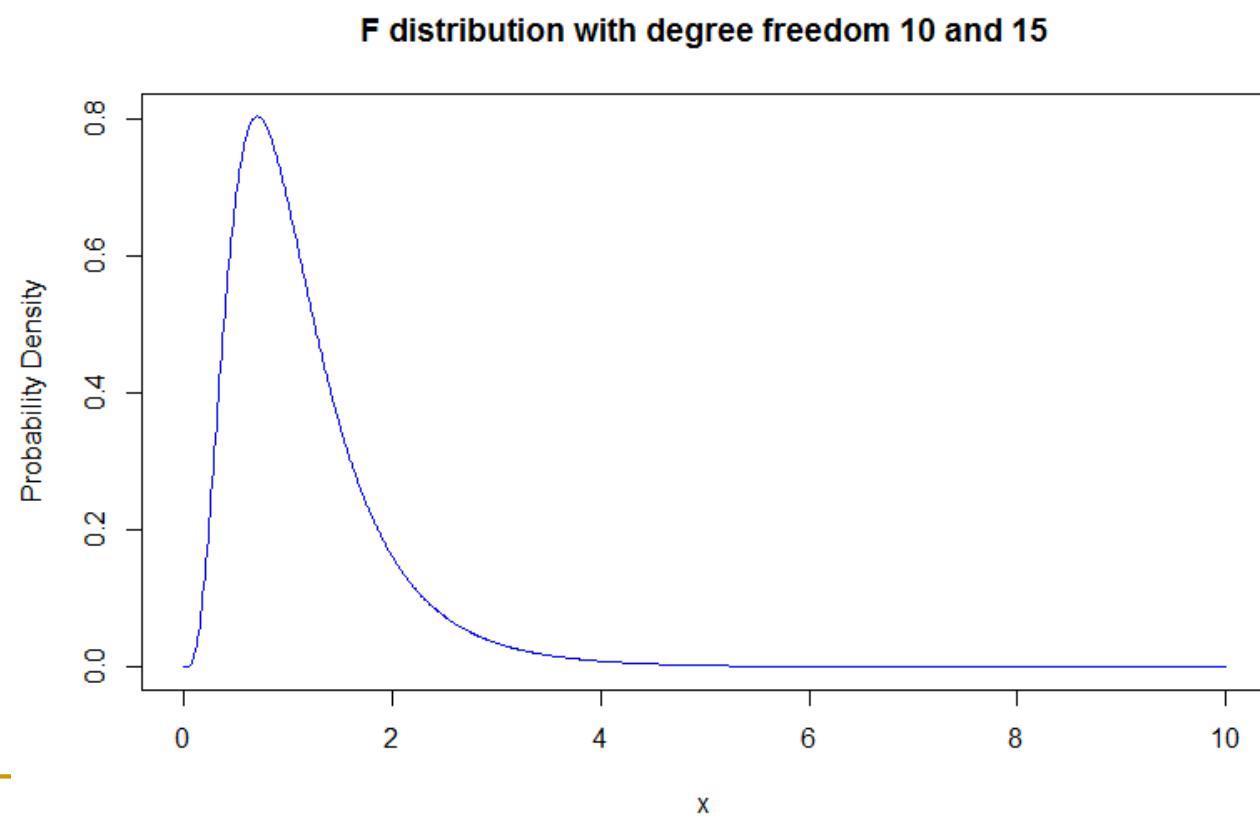


Figure 11.5

Example

Sample 1: Mean rate of return on a sample of 10 internet stocks was 12.6 percent with a standard deviation of 3.9 percent.

Sample 2: Mean rate of return on a sample of 8 utility stocks was 10.9 percent with a standard deviation of 3.5 percent.

Suppose the rate of return on two types of stocks follow normal distribution, at the .05 significance level, can Colin conclude that there is more variation in the software stocks?

Step 1: The hypotheses are $H_0 : \sigma_I^2 \leq \sigma_U^2$

$$H_1 : \sigma_I^2 > \sigma_U^2$$

Step 2: The significance level is .05.

Step 3: The test statistic is the *F* distribution.

Step 4: The value of *F* is computed as follows.

$$F = \frac{(3.9)^2}{(3.5)^2} = 1.2416$$

P-value: $P(F > 1.2416)$ is 0.3965. $1 - pf(1.2416, 9, 7)$ (Use R)

Step 5:

H_0 is not rejected. There is insufficient evidence to show more variation in the internet stocks.

var.test() in R

R provide the var.test function to test if the two variance are the same.

```
Var.test(x,y,alternative="", conf.level= )
```

Example

Refer to the country.txt data which contains numerous population indicators for a sample of 121 countries. Conduct a test of hypothesis to determine if there is a difference in the variance of birthrate for developed and developing country. At the significance level of 0.05.

```
>country<-read.table(file.choose(),header=T)  
>attach(country)
```

```
var.test (birthrate[develop=='developing' ],  
birthrate[develop==developed])
```

Or

```
var.test(birthrate~develop)
```

```
var.test(birthrate~develop)
```

F test to compare two variances

```
data: birthrate by develop
F = 0.5252, num df = 27, denom df = 92, p-value = 0.05882
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.2985616 1.0252946
sample estimates:
ratio of variances
0.525167
```

H₀ is not rejected. There is no sufficient evidence to show a difference in variance of birthrate for developed country and developing country.

Testing for Significant Differences among the means of several populations

- Are there any statistically significant differences between the sample (treatment) means?
- The null hypothesis is that the mean of all p groups are the same
 - $H_0: \mu_1 = \mu_2 = \dots = \mu_p$
- The alternative is that some (or all, but at least two) of the p treatments have different effects on the mean response
 - $H_a: \text{at least two of } \mu_1, \mu_2, \dots, \mu_p \text{ differ}$

Comparison of Several Means

- Evaluate the difference among the means of three or more groups

Examples: Expected mileage for five brands of tires

- Assumptions
 - Populations are normally distributed
 - Populations have equal variances
 - Samples are randomly and independently drawn

Comparison of Several Means

Example Fuel consumption comparisons

- Suppose we want to compare the fuel consumption recorded for three types of automobile, A-cars, B-cars and C-cars.
- The data are listed in the table

	A-CARS	B-CARS	C-CARS
	22.2	24.6	22.7
	19.9	23.1	21.9
	20.3	22.0	23.2
	21.4	23.5	24.1
	21.2	23.6	22.1
	21.0	22.1	23.4
	20.3	23.5	—
Sums	146.3	162.4	137.4

Comparison of Several Means

POPULATION			
1	2	...	K
x_{11}	x_{21}	...	x_{K1}
x_{12}	x_{22}	...	x_{K2}
.	.	.	.
.	.	.	.
.	.	.	.
x_{1n}	x_{2n}	...	x_{Kn}

Suppose that we have independent random samples of n_1, n_2, \dots, n_K observations from K populations. Denote the K group sample means by $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_K$ and the overall sample mean by \bar{x}

One-way analysis of variance (ANOVA)

- Analyzing the Variability
- Variations of observations will be decomposed in to
 - Within-Groups variability
 - Between-Groups variability
- Compare the **between-groups variability** to the **within-groups variability**

Partitioning the Variation

The framework for one-way ANOVA

Within – Groups :

$$SSW = \sum_{i=1}^K \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

Between – Groups :

$$SSG = \sum_{i=1}^K n_i (\bar{x}_i - \bar{x})^2$$

Total :

$$SST = \sum_{i=1}^K \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$$

where x_{ij} denotes the jth sample observation in the ith group.

Then

$$SST = SSW + SSG$$

Partitioning the Variation

$$SST = SSG + SSW$$

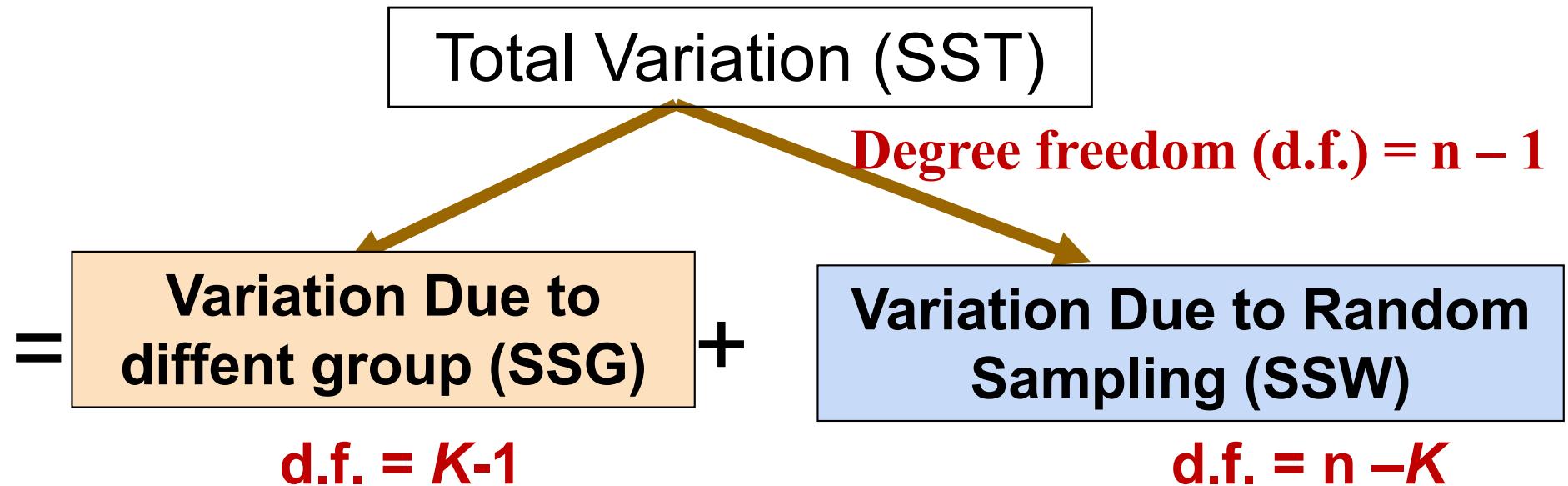
Total Variation = the aggregate dispersion of the individual data values across the various factor levels (SST)

Between-Group Variation = dispersion between the factor sample means (SSG)

Within-Group Variation = dispersion that exists among the data values within a particular factor level (SSW)

Comparison of Several Means

Denote by n the total sample size $n = n_1 + n_2 + \dots + n_K$



We define the mean squares as follows:

$$\text{Within - Groups :} \quad \text{MSW} = \frac{\text{SSW}}{n - K}$$

$$\text{Between - Groups :} \quad \text{MSG} = \frac{\text{SSG}}{K - 1}$$

One-Way ANOVA Table

Source of Variation	SS	df	MS (Variance)	F ratio
Among Groups	SSG	K - 1	$MSG = \frac{SSG}{K - 1}$	$F = \frac{MSG}{MSW}$
Within Groups	SSW	n - K	$MSW = \frac{SSW}{n - K}$	
Total	$SST = SSA + SSW$	n - 1		

K = number of groups

n = sum of the sample sizes from all groups

df = degrees of freedom

One-Way ANOVA

F Test Statistic

$$H_0: \mu_1 = \mu_2 = \dots = \mu_c$$

H_1 : At least two population means are different

- Test statistic

$$F = \frac{MSG}{MSW}$$

MSG is mean squares **among** groups

MSW is mean squares **within** groups

- Degrees of freedom

- $df_1 = K - 1$ (K = number of groups)

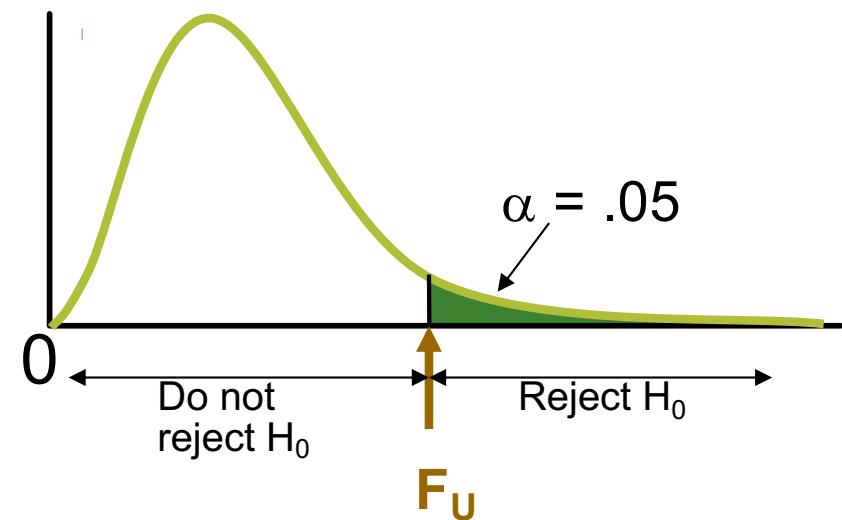
- $df_2 = n - K$ (n = sum of sample sizes from all populations)

F Statistic

- The F statistic is the ratio of mean squares **among** groups and mean squares **within** groups

Decision Rule:

- Reject H_0 if $F > F_U$, otherwise do not reject H_0 .
- P-value is smaller than α .



Comparison of Several Means

	A-CARS	B-CARS	C-CARS
	22.2	24.6	22.7
	19.9	23.1	21.9
	20.3	22.0	23.2
	21.4	23.5	24.1
	21.2	23.6	22.1
	21.0	22.1	23.4
	20.3	23.5	—
Sums	146.3	162.4	137.4

ANOVA Table for Fuel Consumption Data

ANOVA

Between Groups	21.550	2	10.775	15.039	.000
Within Groups	12.180	17	.716		
Total	33.730	19			

- Write an R program to test if the fuel consumption for three types of automobile, A-cars, B-cars and C-cars are equal.

One way ANONA in R

oneway.test(formula, data, var.equal =TRUE)

Or

anova(lm(formula))

Example: The data set red.cell.folate (Ispace) contains red cell folate levels in patients receiving three different methods of ventilation during anesthesia.

red.cell.folate

	folate	ventilation
1	243	N2O+O2, 24h
2	251	N2O+O2, 24h
3	275	N2O+O2, 24h
4	291	N2O+O2, 24h
5	347	N2O+O2, 24h
6	354	N2O+O2, 24h
7	380	N2O+O2, 24h
8	392	N2O+O2, 24h
9	206	N2O+O2, op
10	210	N2O+O2, op
11	226	N2O+O2, op
12	249	N2O+O2, op
13	255	N2O+O2, op
14	273	N2O+O2, op
15	285	N2O+O2, op
16	295	N2O+O2, op
17	309	N2O+O2, op
18	241	O2, 24h
19	258	O2, 24h
20	270	O2, 24h
21	293	O2, 24h
22	328	O2, 24h

One way ANONA in R

```
> attach(red.cell.folate)  
> anova(lm(folate~ventilation))
```

Analysis of Variance Table

Response: folate

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ventilation	2	15516	7757.9	3.7113	0.04359 *
Residuals	19	39716	2090.3		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

➤`oneway.test(folate~ventilation,var.equal = TRUE)`

One-way analysis of means

data: folate and ventilation

F = 3.7113, num df = 2, denom df = 19, p-value = 0.04359

At 0.05 significance level, reject H0. There is a difference in the mean of red cell folate levels among the three different methods of ventilation

One way ANONA in R

Example Fuel Consumption Data

	A-CARS	B-CARS	C-CARS
	22.2	24.6	22.7
	19.9	23.1	21.9
	20.3	22.0	23.2
	21.4	23.5	24.1
	21.2	23.6	22.1
	21.0	22.1	23.4
	20.3	23.5	—
Sums	146.3	162.4	137.4

One way ANONA in R

➤ **oneway.test(car~type.car,data=car.data,var.equal = TRUE)**

One-way analysis of means

data: car and typr.car

F = 15.0386, num df = 2, denom df = 17, p-value = 0.0001737

➤ **oneway.test(car~type.car,data=car.data)**

One-way analysis of means (not assuming equal variances)

data: car and typr.car

F = 14.9056, num df = 2.000, denom df = 11.148, p-value = 0.0007077

➤ **anova(lm(car~type.car,data=car.data))**

Analysis of Variance Table

Response: car

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
typr.car	2	21.5495	10.7748	15.039	0.0001737	***
Residuals	17	12.1800	0.7165			

Pairwise comparisons and multiple testing

- The F test gives an answer about equality of several means. When the test is rejected, we want to know more detailed comparisons among those means.

**>`pairwise.t.test()` Conduct pairwise comparisons
between group levels**

Pairwise comparisons and multiple testing

Example: Data set red.cell.folate

```
> pairwise.t.test(folate,ventilation)
```

Pairwise comparisons using t tests with pooled SD

data: folate and ventilation

	N2O+O2,24h	N2O+O2,op
N2O+O2, op	0.042	-
O2,24h	0.310	0.408

The output is a table of p-value for the pairwise comparisons.

From this table, only the difference in the mean of red cell folate levels between N2o+o2,24h and N2o+o2,op is significant at 5% significance level.

Two-way ANOVA

Two-way ANOVA

- Allows two different factors to be examined simultaneously
- Examines the effects of two different factors and their interactions.
- Each of the **main effects** is one-way test. The **interaction effect** is simply asking "is there any significant difference in performance when you consider two factors acting together".

Two-way ANOVA

Example: The data set *gssft.txt* is used

Let us consider the average hours worked based on degree and gender.

id	age	degree	hrs1	sex
1	26	3	40	1
2	48	1	38	2
4	39	1	60	2
5	25	2	44	2
6	25	1	40	2
7	36	1	55	1
8	44	2	42	2
9	44	4	50	1
10	47	3	60	1
11	53	4	50	2
12	52	4	40	1
13	52	1	45	1

```
> gssft<-read.table("D:\\gssft.txt",header=T)
```

Download gssft.txt from website www.uic.edu.hk/~HepingST
Read the data into R

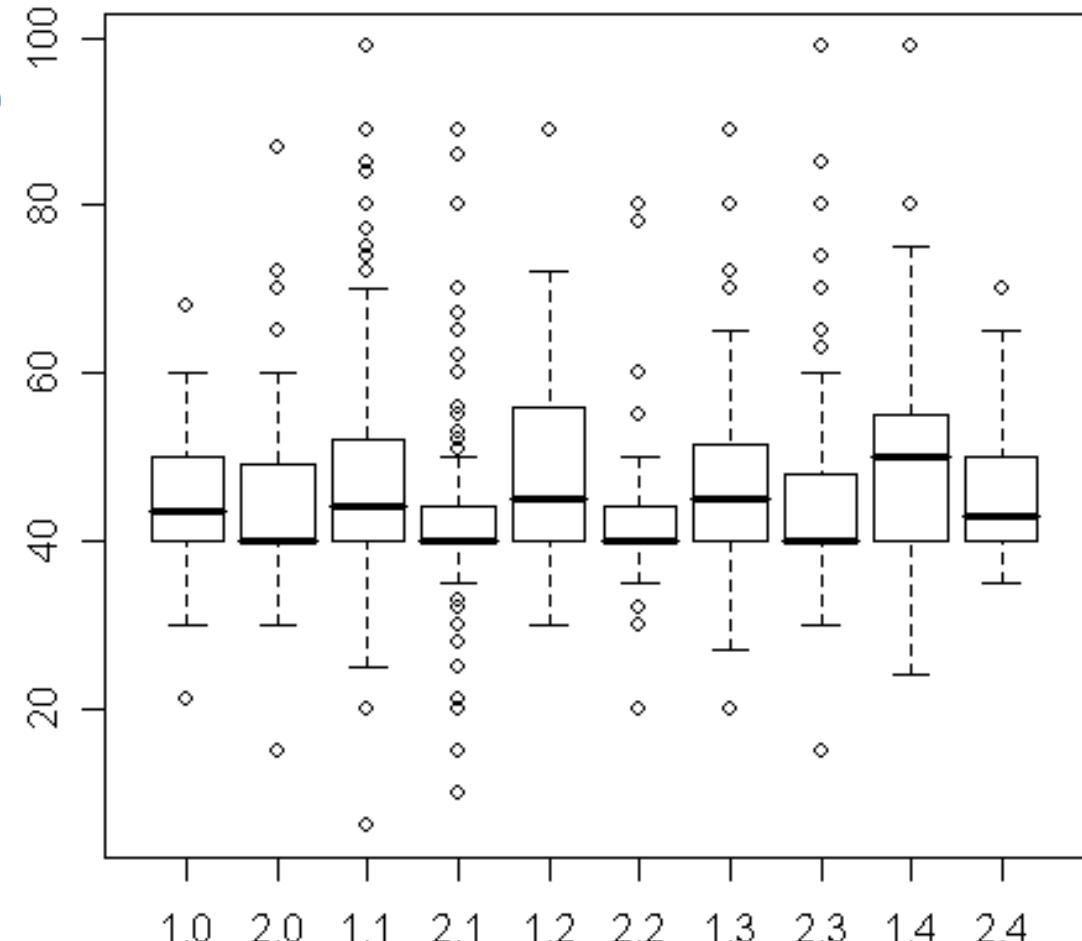
Example: The data set *gssft.txt* is used

id	age	degree	hrs1	sex	
1	26	3	40	1	Degree: highest degree
2	48	1	38	2	0=less than high school
4	39	1	60	2	1=high school, 2=Junior college
5	25	2	44	2	3=bachelor, 4=graduate
6	25	1	40	2	
7	36	1	55	1	
8	44	2	42	2	
9	44	4	50	1	
10	47	3	60	1	
11	53	4	50	2	
12	52	4	40	1	Sex: 1=male, 2=female
13	52	1	45	1	
14	51	3	50	1	Hrs1: average hours work
16	40	1	48	1	
18	44	1	40	1	
19	40	0	50	1	
21	48	3	50	1	
22	49	3	40	2	
32	34	3	40	2	
34	37	1	60	1	
35	22	1	47	1	
36	33	0	52	2	Let us consider the average hours
38	43	1	45	1	worked based on degree and gender .
39	29	3	50	1	
41	31	3	47	2	
42	45	2	45	2	
45	26	1	40	1	
46	46	0	30	2	
--	--	-	--	-	

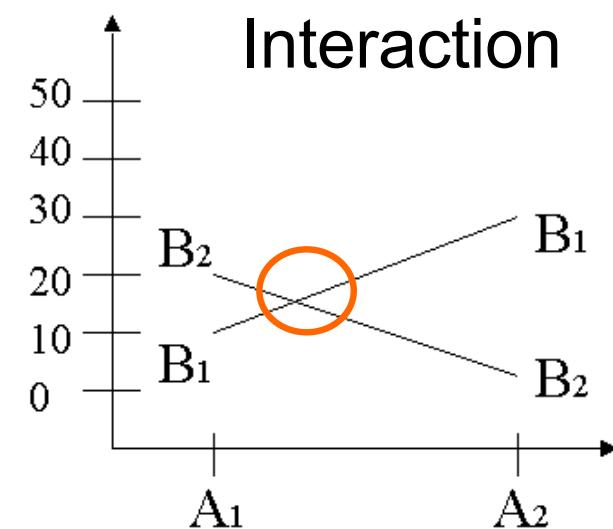
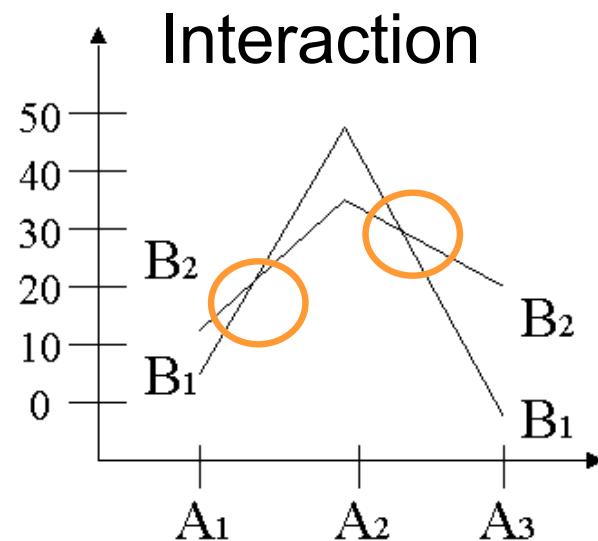
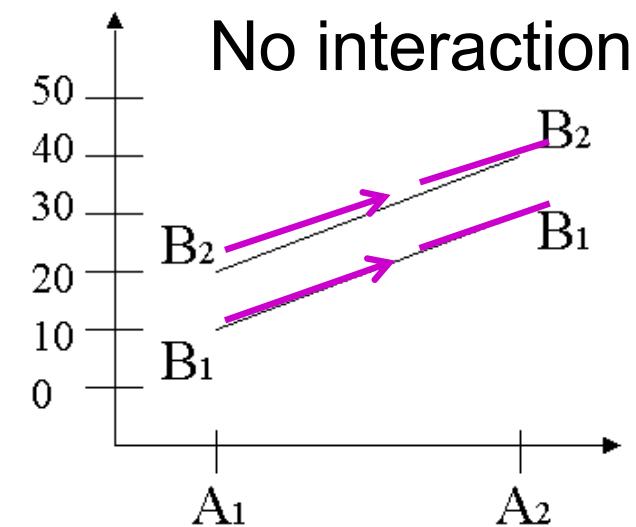
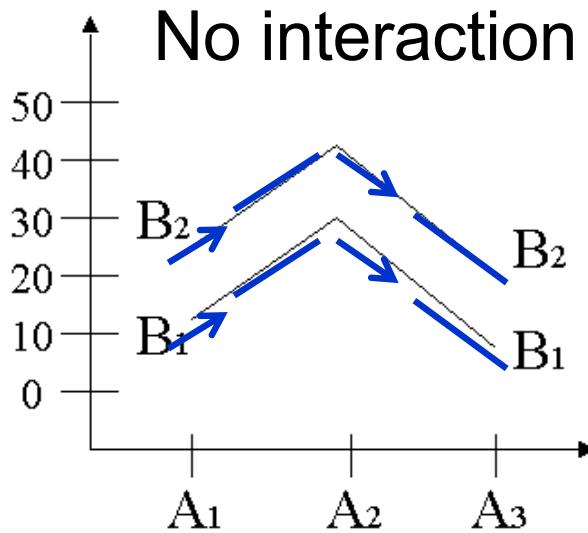
The box plots provide much more information than the bar charts. There are some outliers and extreme points.

`boxplot(hrs1~degree*sex)`

- For each of the five degree categories, the median hours worked is higher for men than for women.
- The length of the boxes for the men is somewhat larger than for the women. That tells us that the interquartile ranges for hours worked are larger for the men than for the women.

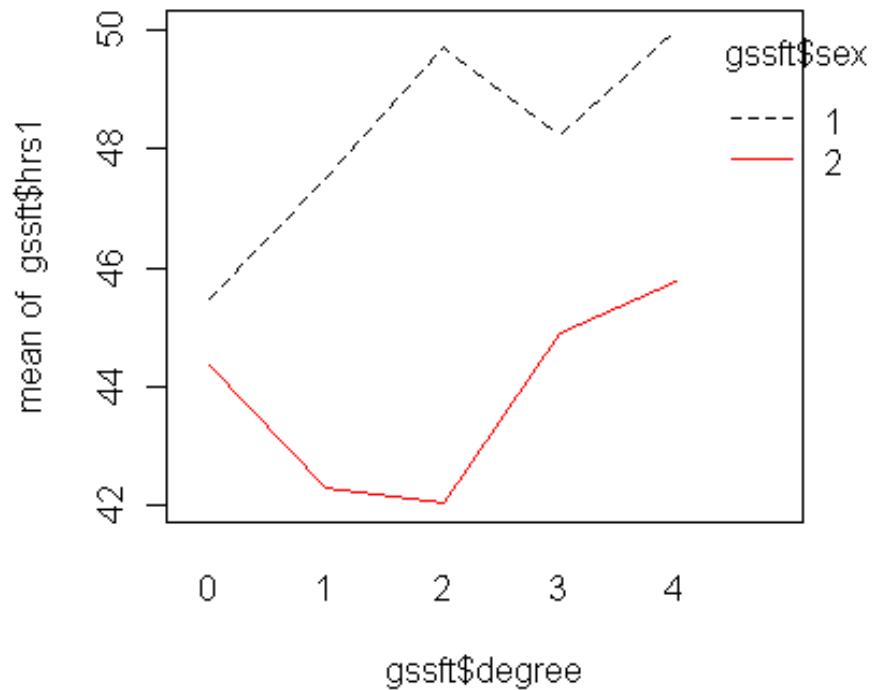


Interaction



Example: The data set *gssft.txt*

```
interaction.plot(gssft$degree,  
gssft$sex,gssft$hrs1,col=1:2)
```



Two-Way ANOVA Null Hypotheses

- 1. No Difference in Means Due to Factor A
 - $H_0: \mu_{1..} = \mu_{2..} = \dots = \mu_{a..}$
- 2. No Difference in Means Due to Factor B
 - $H_0: \mu_{.1} = \mu_{.2} = \dots = \mu_{.b}$
- 3. No Interaction of Factors A & B
 - $H_0: AB_{ij} = 0$

Two-way ANOVA (Example data *gssft*)

■ Testing Hypotheses

- Are the average hours worked the same for the five degree categories?
- Are the average hours worked the same for men and women?
- Is the relationship between average hours worked and degree the same for men and for women?

Two-Way ANOVA Data Table

Factor A	Factor B			
	1	2	...	b
1	X_{111}	X_{121}	...	X_{1b1}
	X_{112}	X_{122}	...	X_{1b2}
2	X_{211}	X_{221}	...	X_{2b1}
	X_{212}	X_{222}	...	X_{2b2}
:	:	:	:	:
a	X_{a11}	X_{a21}	...	X_{ab1}
	X_{a12}	X_{a22}	...	X_{ab2}

Observation k

X_{ijk}

Level i Level j
 Factor Factor

A B

Partitioning the Variation

- Total variation can now be split into three parts:

$$SST = SSA + SSB + SSAB + SSE$$

SST = Total variation

SSA = Variation due to factor A

SSB = Variation due to factor B

SSAB= Variation due to interaction

SSE = Random variation

$$SST = \sum_i \sum_j \sum_k (x_{ijk} - \bar{x})^2$$

$$SSA = bn \sum_{i=1}^a (\bar{x}_{i..} - \bar{x})^2$$

$$SSB = an \sum_{j=1}^b (\bar{x}_{.j.} - \bar{x})^2$$

$$SSAB = n \sum_{i=1}^a \sum_{j=1}^b (\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x})^2$$

Suppose that we have a sample of observations on a levels for factor A and b levels for factor B, with n observations per cell.

Two-way ANOVA

Source	SS	d.f.	MS	F
Factor A	SSA	$a-1$	$\frac{SSA}{(a-1)}$	$\frac{MSA}{s^2}$
Factor B	SSB	$b-1$	$\frac{SSB}{(b-1)}$	$\frac{MSB}{s^2}$
Interaction A:B	SSAB	$(a-1)(b-1)$	$\frac{SSAB}{(a-1)(b-1)}$	$\frac{MSAB}{s^2}$
Error	SSE	$ab(n-1)$	$s^2 = \frac{SSE}{ab(n-1)}$	
Total	SST	$abn-1$		

Main Factor Test

$H_0 : \mu_{1.} = \mu_{2.} = \mu_{3.} = \dots$

$$F = \frac{MSA}{MSE}$$

- $df_1 = a - 1$
- $df_2 = ab(n-1)$

Reject H_0 if $F > F_U$

$H_0 : \mu_{.1} = \mu_{.2} = \mu_{.3} = \dots = \mu_{.b}$

$$F = \frac{MSB}{MSE}$$

- $df_1 = b - 1$
- $df_2 = ab(n-1)$

Reject H_0 if $F > F_U$

Interaction effect test

□ $H_0: AB_{ij} = 0$

$$F = \frac{MSAB}{MSE}$$

- $df_1 = (a - 1)(b - 1)$
- $df_2 = ab(n - 1)$

Reject H_0 if $F > F_U$

Two-way ANOVA

`aov(hrs1~degree*sex)`

Call:

```
aov(formula = hrs1 ~ degree * sex)
```

Terms:

	degree	sex	degree:sex	Residuals
Sum of Squares	1797.78	8102.84	824.49	166915.76
Deg. of Freedom	4	1	4	1490

Residual standard error: 10.58414

Estimated effects may be unbalanced

Remind: `sex<-as.factor(sex)` `degree<-as.factor(degree)`

Alternative

`a<-lm(hrs1~degree*sex) # a<-lm(hrs1~degree+sex+degree:sex)`
`anova(a)`

Two-way ANOVA

```
> a<-aov(hrs1~degree*sex)
> summary(a)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
degree	4	1798	449	4.012	0.003056 **
sex	1	8103	8103	72.331	< 2.2e-16 ***
degree:sex	4	824	206	1.840	0.118661
Residuals	1490	166916	112		

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

Two-way ANOVA

- The null hypotheses for the **interaction** is that the effect of type of degree on average hours worked is the same for males and females in the population.
- The F test shows the p -value for testing no-interaction to be 0.12.
- The absence of interaction tells that it's reasonable to believe that the difference in average hours worked between males and females is the same for all degree categories.
- The F test for the degree is significant with p-value 0.0030.
- The observed significance level for the sex main effect is very small, less than 0.005. The average hours worked for males and females are significant different.

Two-way ANOVA

■ Removing the interaction effect

```
> a<-lm(hrs1~degree+sex)  
> anova(a)
```

Analysis of Variance Table

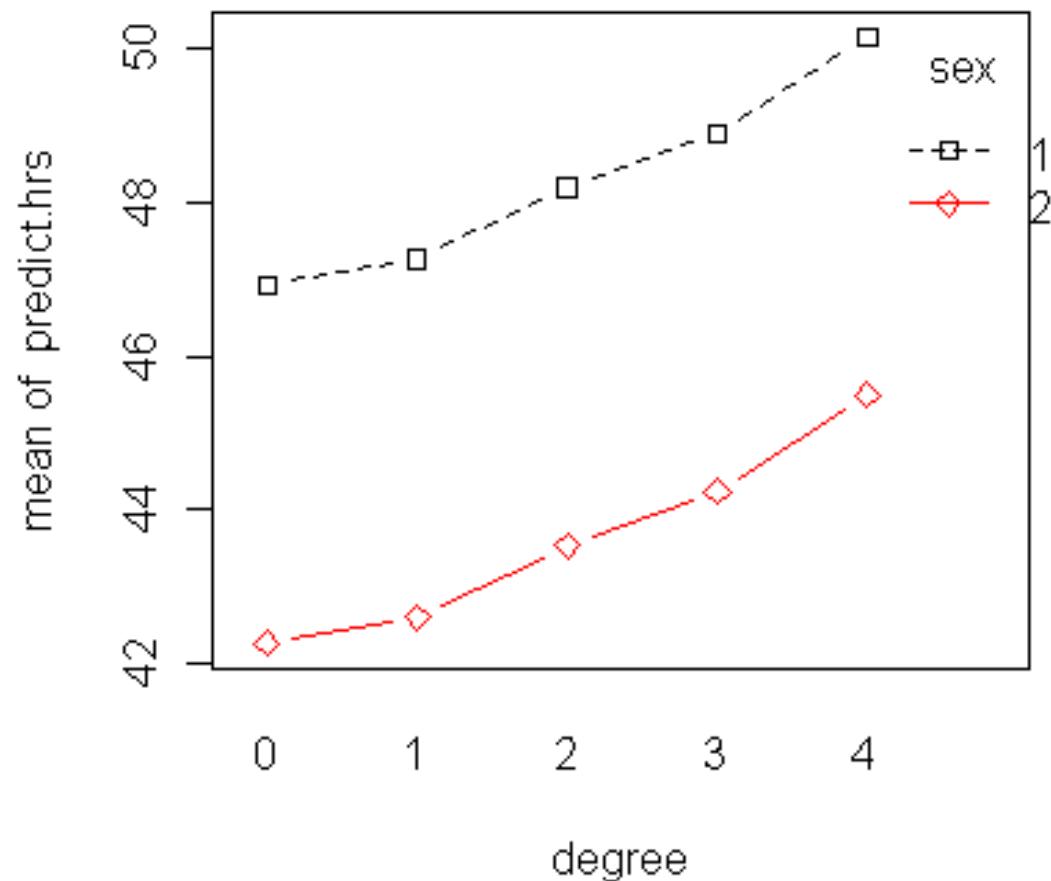
Response: hrs1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
degree	4	1845	461	4.0948	0.002641 **
sex	1	7970	7970	70.7491	< 2.2e-16 ***
Residuals	1484	167166	113		

Remind: sex<-as.factor(sex) degree<-as.factor(degree)

The line plots of the **predicted means** for the 10 gender-and-degree combinations under the model without the interaction effect.

```
sex<-as.factor(sex)
degree<-as.factor(degree)
a<-lm(hrs1~degree+sex)
predict.hrs<-predict(a)
```



```
interaction.plot(degree,sex,predict.hrs,col=1:2,type='b',pch=2
2:23)
```

Example: data *weightgain*

The data ***weightgain*** (ISapce) is collected from an experiment to study the gain in weight of rats fed on four different diets, distinguished by amount of protein (low and high) and by source of protein (beef and cereal). Ten rats are randomized to each of the four treatments and the weight gain in grams recorded. The question of interest is how diet affects weight gain.

```
weightgain1<-read.csv('d:weightgain.csv',header=T)
```

```
> weightgain1
   source type weightgain
1   Beef  Low     90
2   Beef  Low     76
3   Beef  Low     90
4   Beef  Low     64
5   Beef  Low     86
6   Beef  Low     51
7   Beef  Low     72
8   Beef  Low     90
9   Beef  Low     95
10  Beef  Low     78
11  Beef  High    73
```

Useful command: tapply()

- “**Tapply**” function groups the observations along the value of one (or several) factors and applies a function (mean, etc.) to the resulting groups.

```
> gl(2,2,20)
[1] 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2
Levels: 1 2
```

```
> tapply(1:20, gl(2,2,20) ,sum)
 1   2
95 115
```

Example:

```
attach(weightgain1)
```

Show the mean and sd of weight gain for the different sources and different types, respectively

```
tapply(weightgain,source,mean)
```

Beef Cereal
89.6 84.9

```
tapply(weightgain,type,mean)
```

High	Low
92.95	81.55

Example:

Show the mean and sd for the 4 combinations of source and type.

```
tapply(weightgain, list(source, type),mean) #calculate the  
mean for each combination of
```

	High	Low
Beef	100.0	79.2
Cereal	85.9	83.9

```
tapply(weightgain, list(source, type),sd)
```

	High	Low
Beef	15.13642	13.88684
Cereal	15.02184	15.70881

Example: data *weightgain*

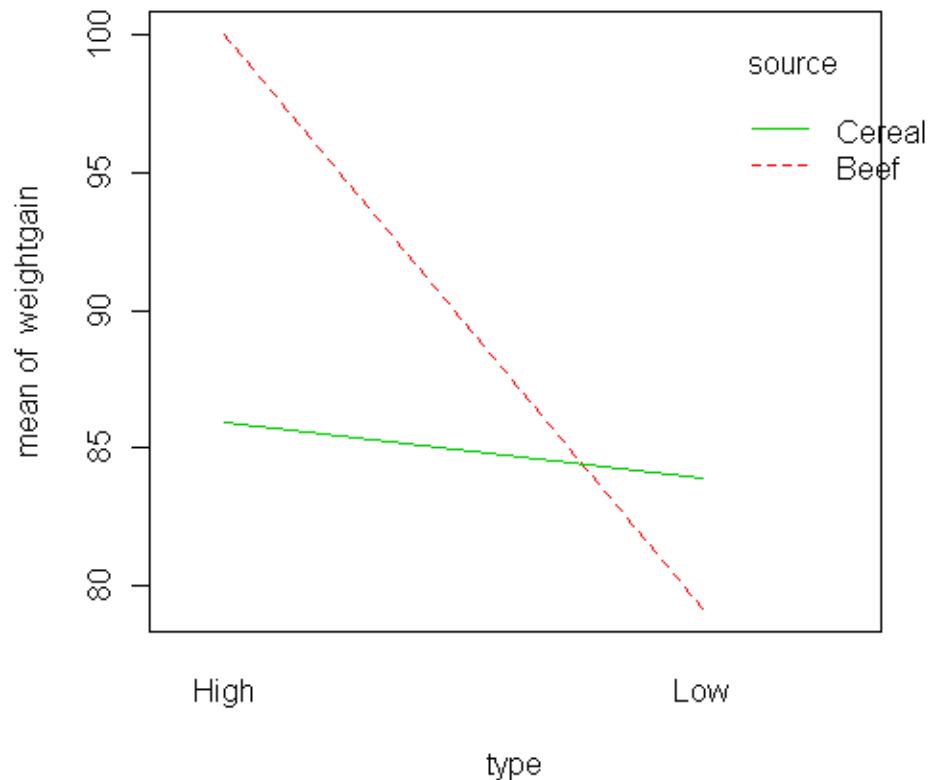
Interaction plot:

```
data1<-weightgain
```

```
attach(data1)
```

```
interaction.plot(type,source,weightgain,col=2:3)
```

```
detach(data1)
```



Anova: data *weightgain*

```
wg.aov<-aov(weightgain~source?type,data=weightgain)
summary(wg.aov)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
source	1	220.9	220.9	0.9879	0.32688
type	1	1299.6	1299.6	5.8123	0.02114 *
source:type	1	883.6	883.6	3.9518	0.05447 .
Residuals	36	8049.4	223.6		

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