CENTRE DE RECHERCHE PUBLIC

Microarray Center

the Luxembourg Bioinformatics Network

UNIVERSITÉ DU LUXEMBOURG

STATISTICAL DATA ANALYSIS IN EXCEL

Lecture 4

Analysis of Variance (ANOVA)

dr. Petr Nazarov

petr.nazarov@crp-sante.lu

14-01-2013

Part I

Inference about Population Variance

INTERVAL ESTIMATION FOR VARIANCE

Variance Sampling Distribution

Variance
A measure of variability based on the squared deviations of the data values about the mean.

population
$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

sample
$$s^2 = \frac{\sum (x_i - m)^2}{n-1}$$

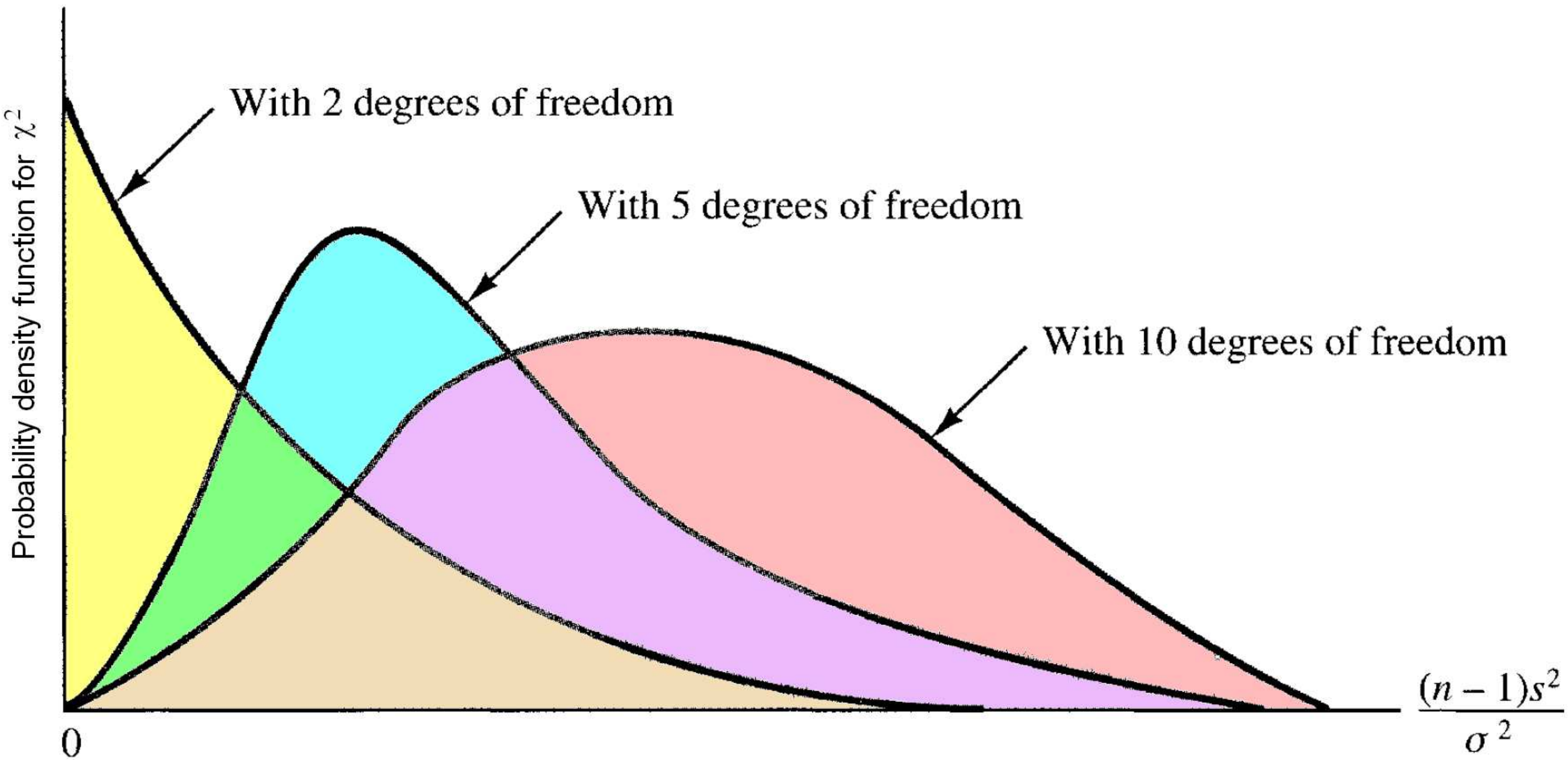
The interval estimation for variance is build using the following measure:

$$(n-1)\frac{s^2}{\sigma^2}$$

Sampling distribution of $(n\text{-}1)s^2/\sigma^2$
Whenever a simple random sample of size n is selected from a normal population, the sampling distribution of **$(n\text{-}1)s^2/\sigma^2$** has a **chi-square distribution** (χ^2) with n-1 degrees of freedom.

$$(n-1)\frac{s^2}{\sigma^2} = \chi^2_{df=n-1}$$

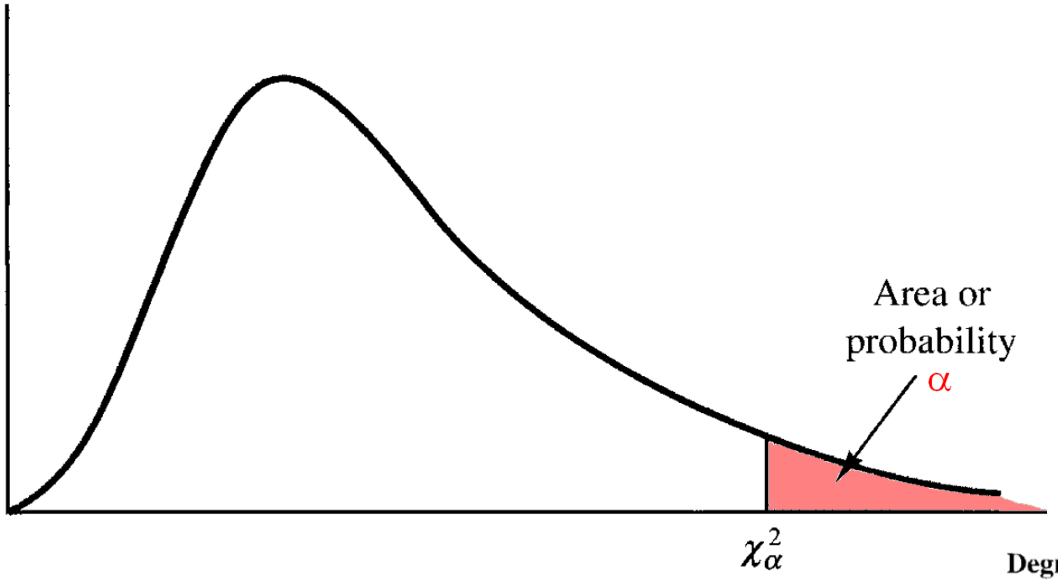
χ^2 Distribution



χ^2 distribution works only for sampling from normal population

$$\chi^2_{df=k} = \sum_{i=1}^{k} x_i^2 \quad where\ x_i - normal$$

χ^2 Probabilities in Table and Excel



Degrees of Freedom	Area in Upper Tail .99	.975	.95	.90	.10	.05	.025	.01
1	.000	.001	.004	.016	2.706	3.841	5.024	6.635
2	.020	.051	.103	.211	4.605	5.991	7.378	9.210
3	.115	.216	.352	.584	6.251	7.815	9.348	11.345
4	.297	.484	.711	1.064	7.779	9.488	11.143	13.277
5	.554	.831	1.145	1.610	9.236	11.070	12.832	15.086
6	.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.041	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191

In Excel ≤2007 use functions:

- **= CHIDIST(χ^2, n-1)**
- **= CHIINV(α, n-1)**

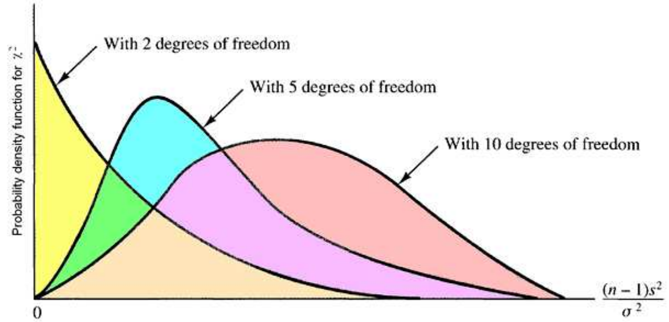
In Excel 2010 use functions:

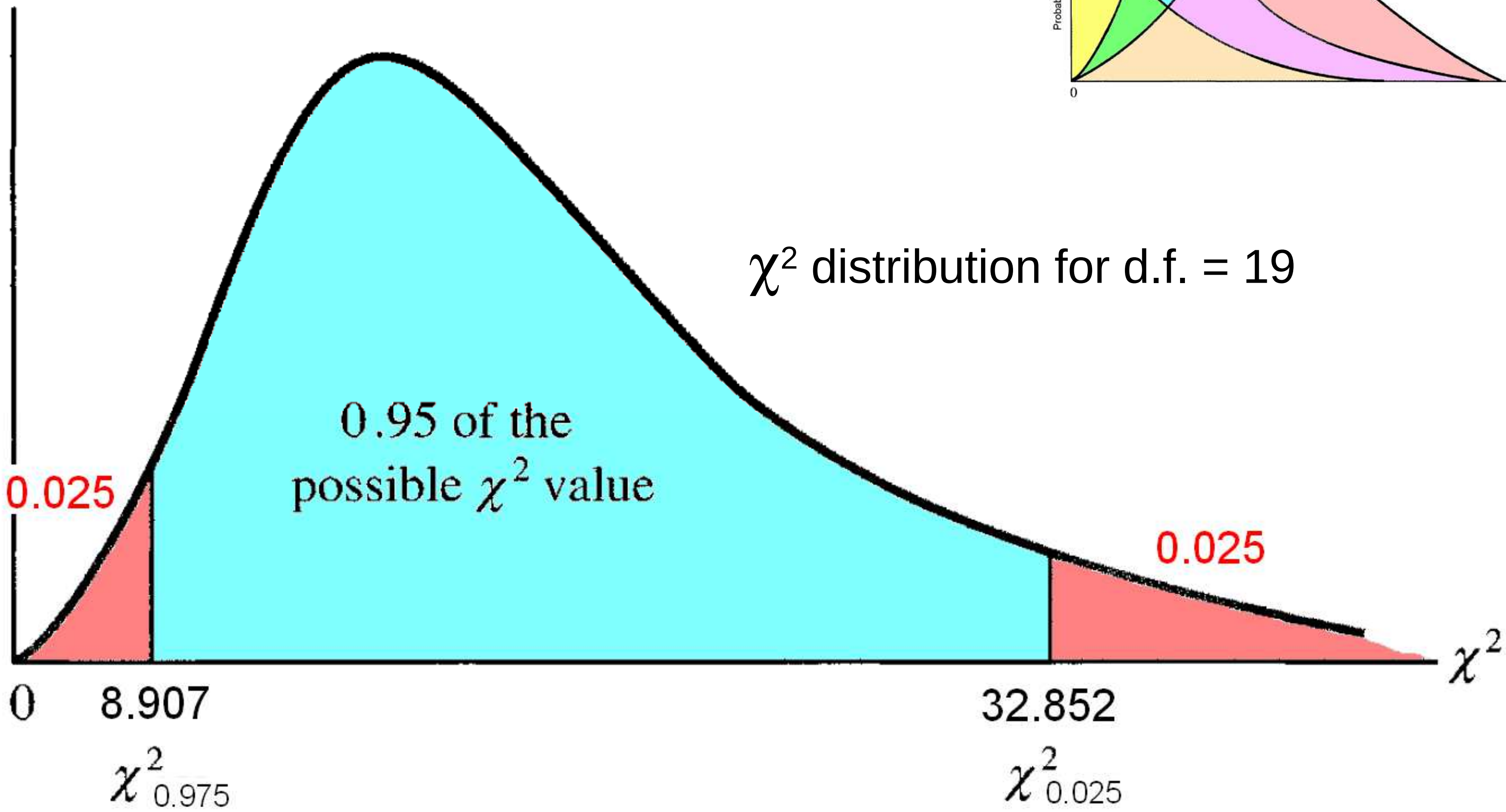
- **= CHISQ.DIST(χ^2, n-1)**
- **= CHISQ.INV(α, n-1)**

INTERVAL ESTIMATION FOR VARIANCE

χ^2 Distribution for Interval Estimation

$$\chi^2 = (n-1)\frac{s^2}{\sigma^2}$$

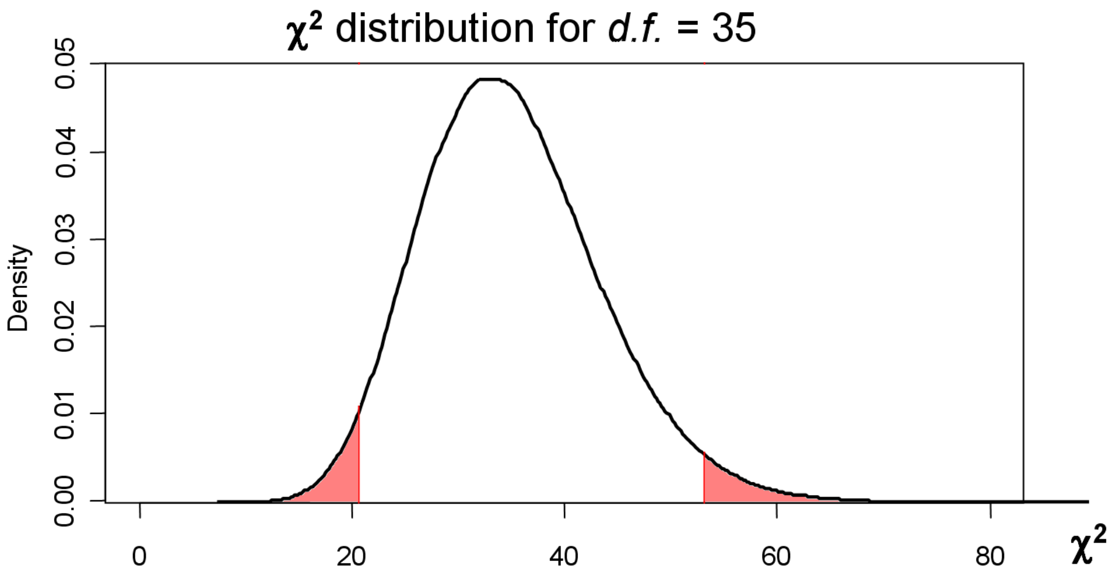




$$\chi^2_{1-\alpha/2} \le (n-1)\frac{s^2}{\sigma^2} \le \chi^2_{\alpha/2}$$

⇓

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$



Suppose sample of *n* = 36 coffee cans is selected and *m* = 2.92 and *s* = 0.18 lbm is observed. Provide 95% confidence interval for the standard deviation

$$\frac{(36-1)0.18^2}{53.203} \le \sigma^2 \le \frac{(36-1)0.18^2}{20.569}$$

```
= CHISQ.INV(α/2, n-1)
           (1-α/2, n-1)
```

$$0.0213 \le \sigma^2 \le 0.0551$$

$$0.146 \le \sigma \le 0.235$$

Hypotheses about Population Variance

H_0: $\sigma^2 \leq$ const
H_a: $\sigma^2 >$ const

H_0: $\sigma^2 \geq$ const
H_a: $\sigma^2 <$ const

H_0: $\sigma^2 =$ const
H_a: $\sigma^2 \neq$ const

	Lower Tail Test	Upper Tail Test	Two-Tailed Test
Hypotheses	$H_0: \sigma^2 \geq \sigma_0^2$ $H_a: \sigma^2 < \sigma_0^2$	$H_0: \sigma^2 \leq \sigma_0^2$ $H_a: \sigma^2 > \sigma_0^2$	$H_0: \sigma^2 = \sigma_0^2$ $H_a: \sigma^2 \neq \sigma_0^2$
Test Statistic	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$
Rejection Rule: p-Value Approach	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$
Rejection Rule: Critical Value Approach	Reject H_0 if $\chi^2 \leq \chi^2_{(1-\alpha)}$	Reject H_0 if $\chi^2 \geq \chi^2_{\alpha}$	Reject H_0 if $\chi^2 \leq \chi^2_{(1-\alpha/2)}$ or if $\chi^2 \geq \chi^2_{\alpha/2}$

VARIANCES OF TWO POPULATIONS

Sampling Distribution

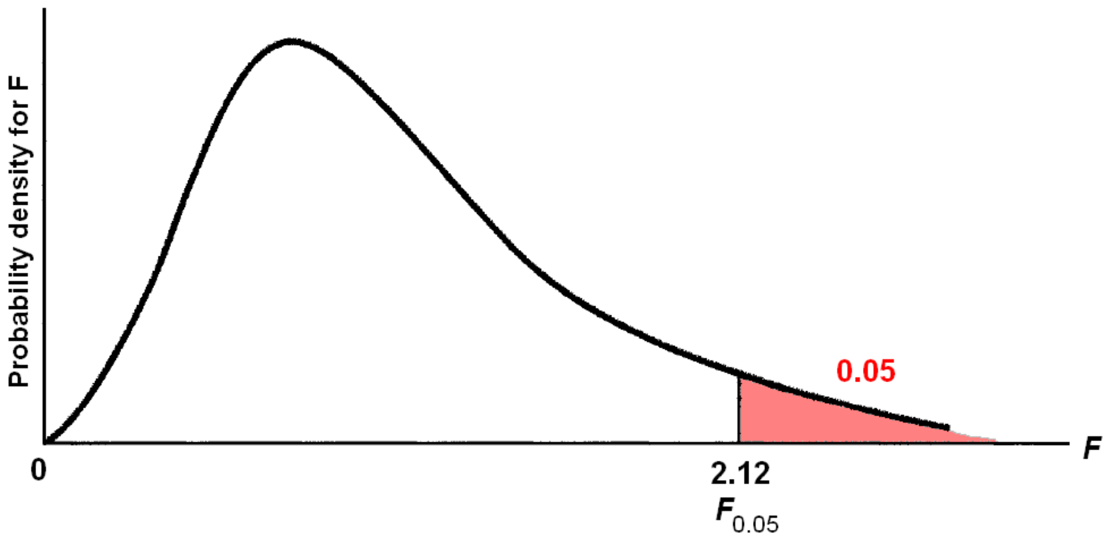
In many statistical applications we need a comparison between variances of two populations. In fact well-known ANOVA-method is base on this comparison.

The statistics is build for the following measure:

$$F = \frac{s_1^2}{s_2^2}$$

Sampling distribution of s_1^2/s_2^2 when $\sigma_1^2 = \sigma_2^2$
Whenever a independent simple random samples of size n_1 and n_2 are selected from two normal populations with equal variances, the sampling of s_1^2/s_2^2 has **F-distribution** with n_1-1 degree of freedom for numerator and n_2-1 for denominator.

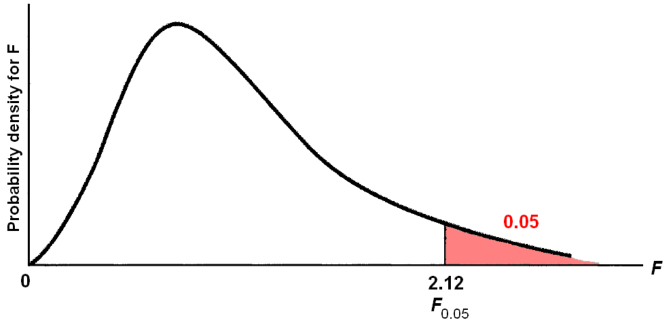
F-distribution for 20 d.f. in numerator and 20 d.f. in denominator



In Excel use functions:

- = **F.TEST(**data1,data2**)**

Hypotheses about Variances of Two Populations



H_0: $\sigma_1^2 \le \sigma_2^2$

H_a: $\sigma_1^2 > \sigma_2^2$

H_0: $\sigma_1^2 = \sigma_2^2$

H_a: $\sigma_1^2 \neq \sigma_2^2$

	Upper Tail Test	Two-Tailed Test
Hypotheses	$H_0 : \sigma_1^2 \le \sigma_2^2$ $H_a : \sigma_1^2 > \sigma_2^2$	$H_0 : \sigma_1^2 = \sigma_2^2$ $H_a : \sigma_1^2 \neq \sigma_2^2$ *Note: Population 1 has the lager sample variance*
Test Statistic	$F = \frac{s_1^2}{s_2^2}$	$F = \frac{s_1^2}{s_2^2}$
Rejection Rule: p-Value Approach	Reject H_0 if p-value $\le \alpha$	Reject H_0 if p-value $\le \alpha$
Rejection Rule: Critical Value Approach	Reject H_0 if $F \ge F_\alpha$	Reject H_0 if $F \ge F_\alpha$

Example

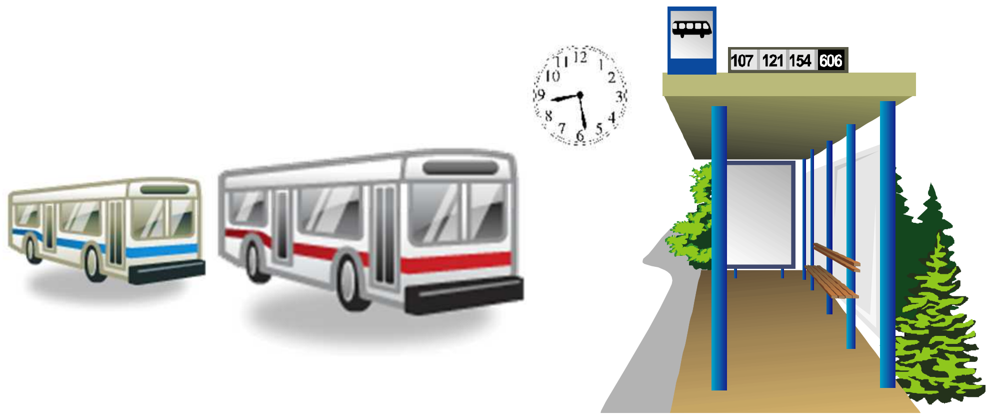
schoolbus.xls

#	Milbank	Gulf Park
1	35.9	21.6
2	29.9	20.5
3	31.2	23.3
4	16.2	18.8
5	19.0	17.2
6	15.9	7.7
7	18.8	18.6
8	22.2	18.7
9	19.9	20.4
10	16.4	22.4
11	5.0	23.1
12	25.4	19.8
13	14.7	26.0
14	22.7	17.1
15	18.0	27.9
16	28.1	20.8
17	12.1	
18	21.4	
19	13.4	
20	22.9	
21	21.0	
22	10.1	
23	23.0	
24	19.4	
25	15.2	
26	28.2	

Dullus County Schools is renewing its school bus service contract for the coming year and must select one of two bus companies, the Milbank Company or the Gulf Park Company. We will use the variance of the arrival or pickup/delivery times as a primary measure of the quality of the bus service. Low variance values indicate the more consistent and higher-quality service. If the variances of arrival times associated with the two services are equal, Dullus School administrators will select the company offering the better financial terms. However, if the sample data on bus arrival times for the two companies indicate a significant difference between the variances, the administrators may want to give special consideration to the company with the better or lower variance service. The appropriate hypotheses follow.

$$H_0\colon \sigma_1^2 = \sigma_2^2$$
$$H_a\colon \sigma_1^2 \neq \sigma_2^2$$

If H_0 can be rejected, the conclusion of unequal service quality is appropriate. We will use a level of significance of $\alpha = .10$ to conduct the hypothesis test.



schoolbus.xls

#	Milbank	Gulf Park
1	35.9	21.6
2	29.9	20.5
3	31.2	23.3
4	16.2	18.8
5	19.0	17.2
6	15.9	7.7
7	18.8	18.6
8	22.2	18.7
9	19.9	20.4
10	16.4	22.4
11	5.0	23.1
12	25.4	19.8
13	14.7	26.0
14	22.7	17.1
15	18.0	27.9
16	28.1	20.8
17	12.1	
18	21.4	
19	13.4	
20	22.9	
21	21.0	
22	10.1	
23	23.0	
24	19.4	
25	15.2	
26	28.2	

1. Let us start from estimation of the variances for 2 data sets

Milbank: $s_1^2 = 48$

Gulf Park: $s_2^2 = 20$

interval estimation (optionally)

Milbank: $\sigma_1^2 \approx 48$ $(29.5 \div 91.5)$

Gulf Park: $\sigma_2^2 \approx 20$ $(10.9 \div 47.9)$

2. Let us calculate the *F*-statistics

$$F = \frac{s_1^2}{s_2^2} = \frac{48}{20} = 2.40$$

3. … and p-value = 0.08

p-value = 0.08 < α = 0.1

In Excel use:

```
= F.TEST(data1,data2)
```

Part II

Analysis of Variance (ANOVA)

Means for more than 2 populations
We have measurements for 5 conditions. Are the means for these conditions equal?

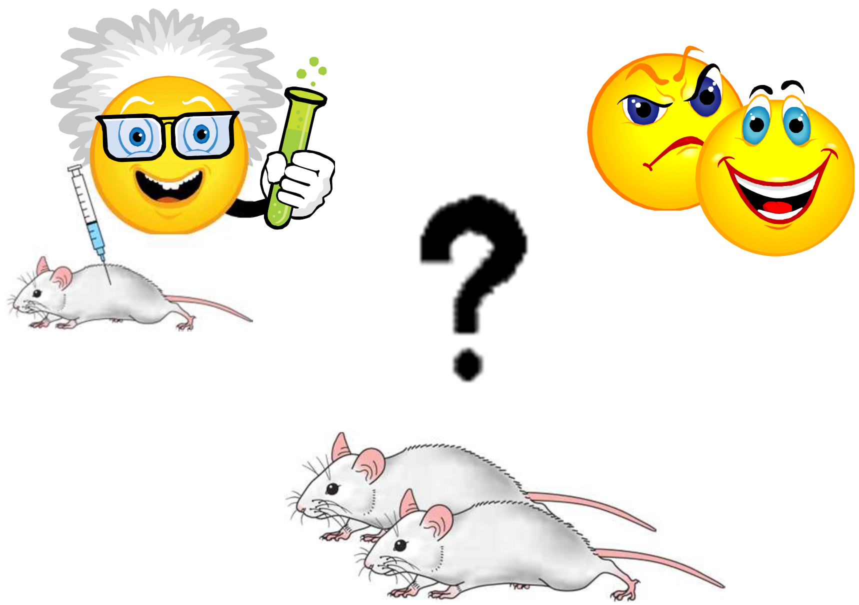
Validation of the effects
We assume that we have several factors affecting our data. Which factors are most significant? Which can be neglected?

ANOVA
example from Partek™

If we would use pairwise comparisons, what will be the probability of getting error?

Number of comparisons: $C_2^5 = \frac{5!}{2!3!} = 10$

Probability of an error: $1-(0.95)^{10} = 0.4$



http://easylink.playstream.com/affymetrix/ambsymposium/partek_08.wvx

Example from Case Problem 3

As part of a long-term study of individuals 65 years of age or older, sociologists and physicians at the Wentworth Medical Center in upstate New York investigated the relationship between geographic location and depression. A sample of 60 individuals, all in reasonably good health, was selected; 20 individuals were residents of Florida, 20 were residents of New York, and 20 were residents of North Carolina. Each of the individuals sampled was given a standardized test to measure depression. The data collected follow; higher test scores indicate higher levels of depression.

Q: Is the depression level same in all 3 locations?

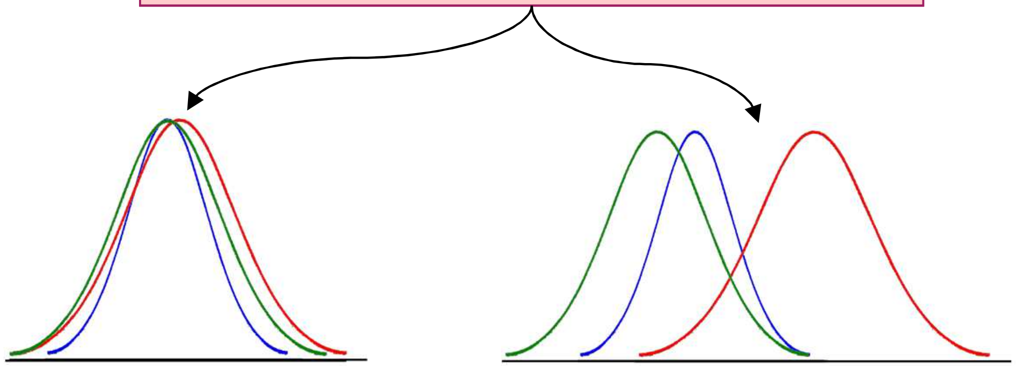
`depression.xls`

1. Good health respondents

Florida	New York	N. Carolina
3	8	10
7	11	7
7	9	3
3	7	5
8	8	11
8	7	8
...	...	...

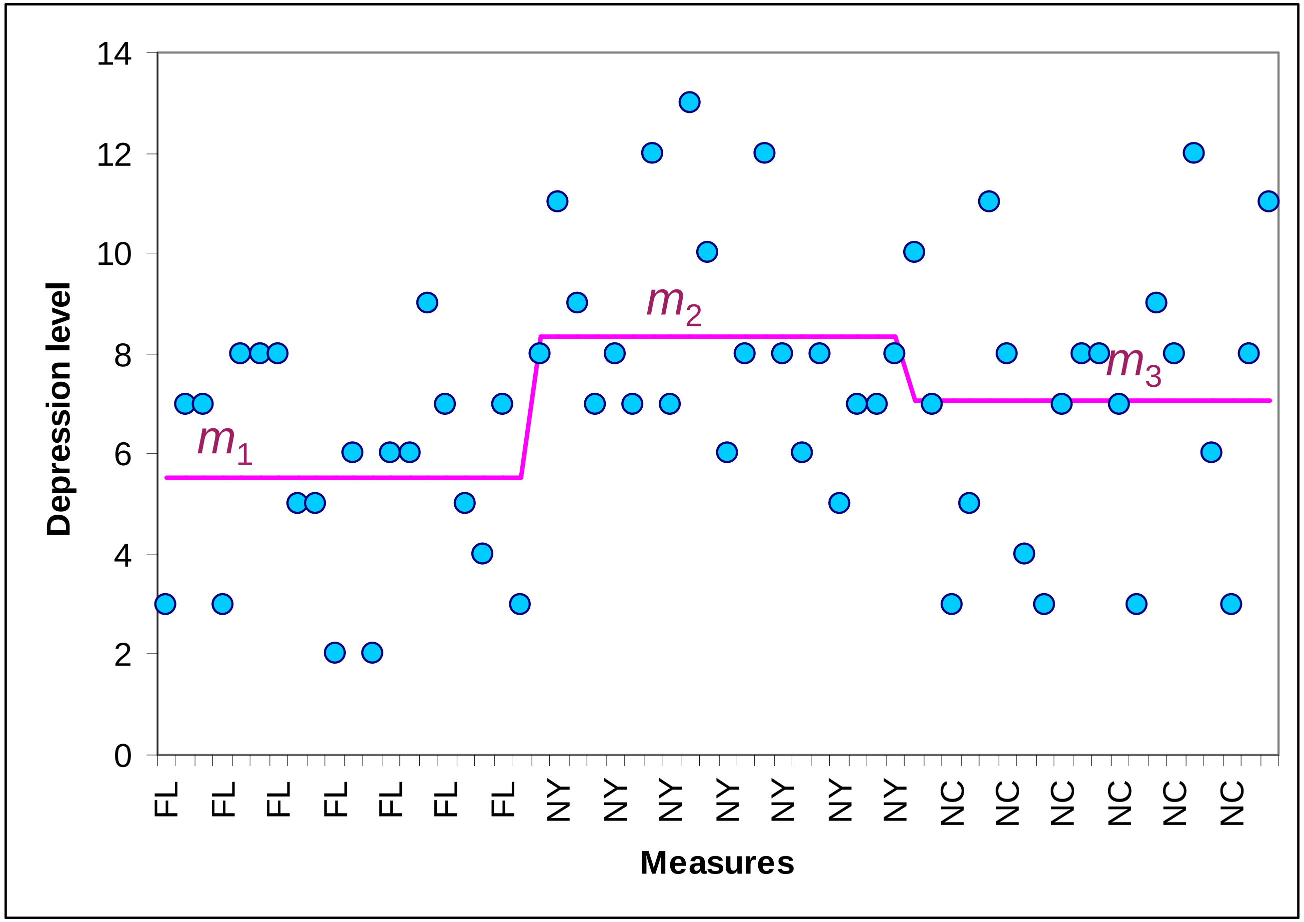
$H_0: \mu_1 = \mu_2 = \mu_3$

H_a: not all 3 means are equal

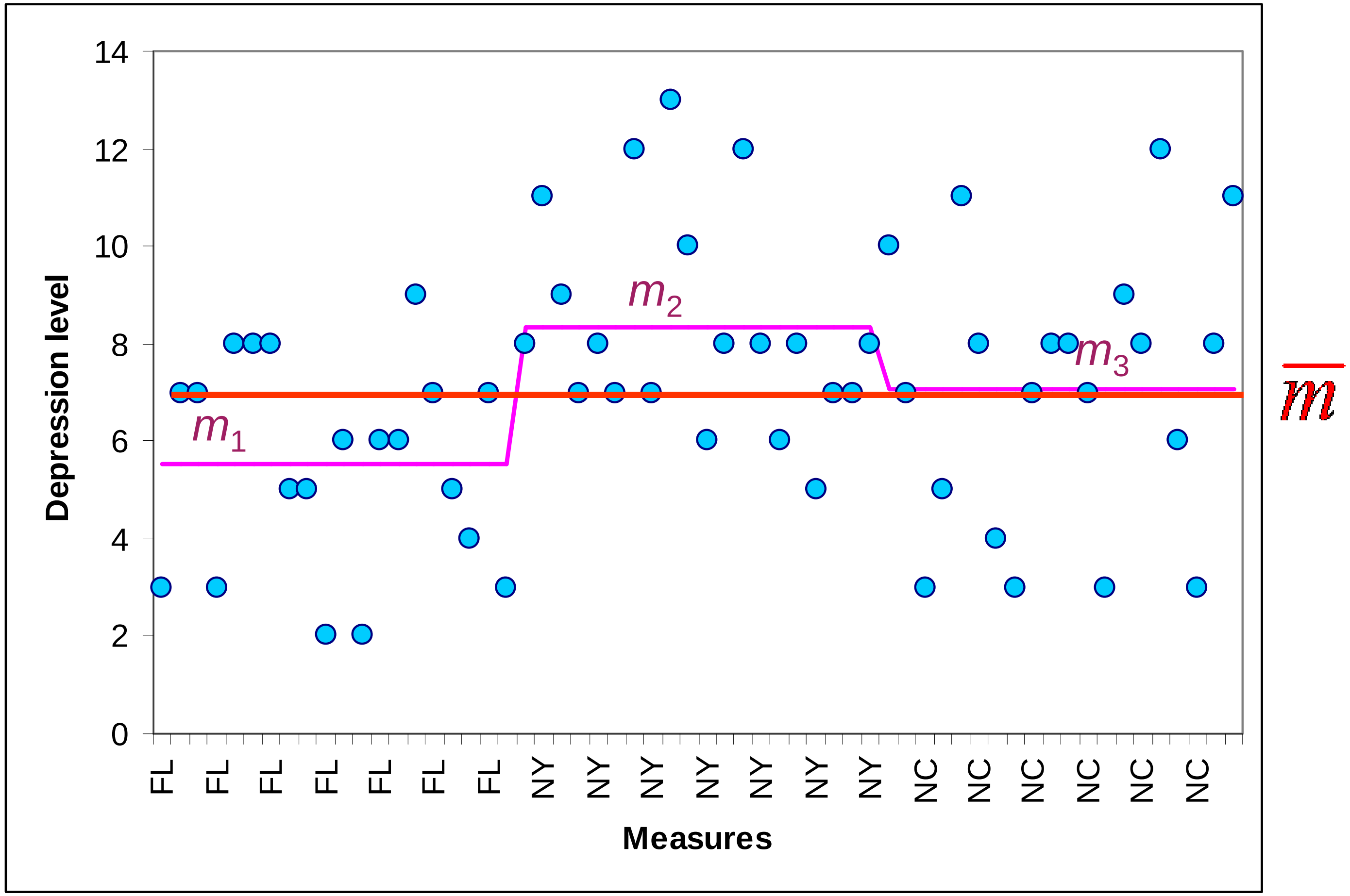


H_0: $\mu_1 = \mu_2 = \mu_3$

H_a: not all 3 means are equal



Example



$$SST = SSTR + SSE$$

ANOVA table
A table used to summarize the analysis of variance computations and results. It contains columns showing the source of variation, the sum of squares, the degrees of freedom, the mean square, and the *F* value(s).

In Excel use:

- Tools → Data Analysis → ANOVA Single Factor

depression.xls

Let's perform for dataset 1: "good health"

SSTR

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	78.53333	2	39.26667	6.773188	0.002296	3.158843
Within Groups	330.45	57	5.797368			
Total	408.9833	59				

SSE

Factors and Treatments

Factor
Another word for the independent variable of interest.

Factorial experiment
An experimental design that allows statistical conclusions about two or more factors.

Treatments
Different levels of a factor.

depression.xls

Factor 1: Health → good health, bad health

Factor 2: Location → Florida, New York, North Carolina

$$\text{Depression} = \mu + \text{Health} + \text{Location} + \text{Health}\times\text{Location} + \varepsilon$$

Interaction
The effect produced when the levels of one factor interact with the levels of another factor in influencing the response variable.

2-factor ANOVA with *r* Replicates: Example

depression.xls → **Factor 1:** Health; **Factor 2:** Location

1. Reorder the data into format understandable for Excel

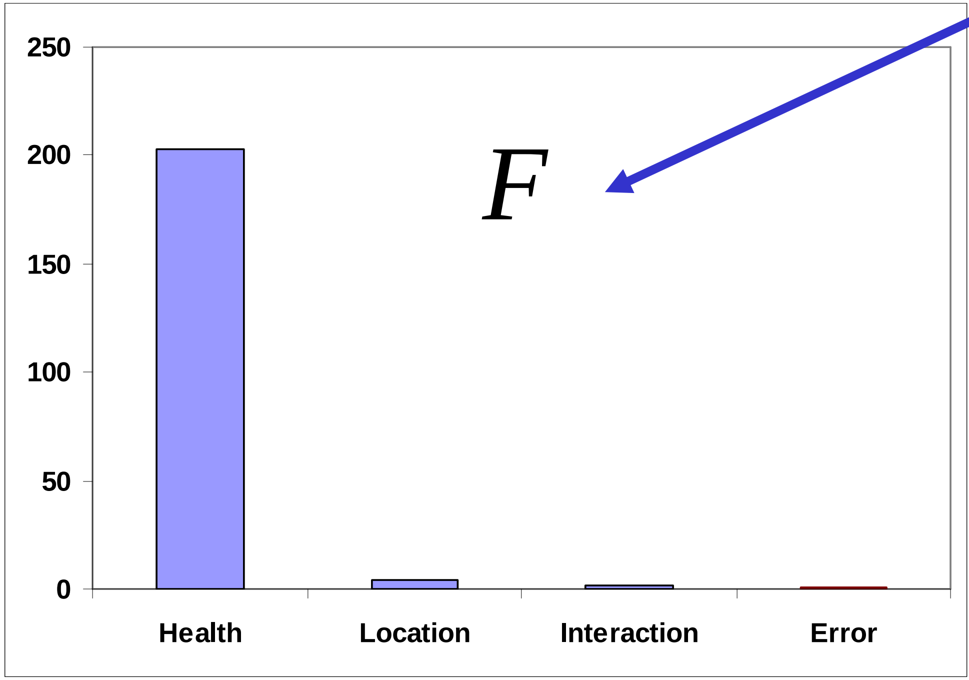
	Florida	New York	North Carolina
Good health	3	8	10
	7	11	7
	7	9	3
	3	7	5
	...	...	...
	7	7	8
	3	8	11
bad health	**13**	**14**	**10**
	12	**9**	**12**
	17	**15**	**15**
	17	**12**	**18**
	...	...	...
	11	**13**	**13**
	17	**11**	**11**

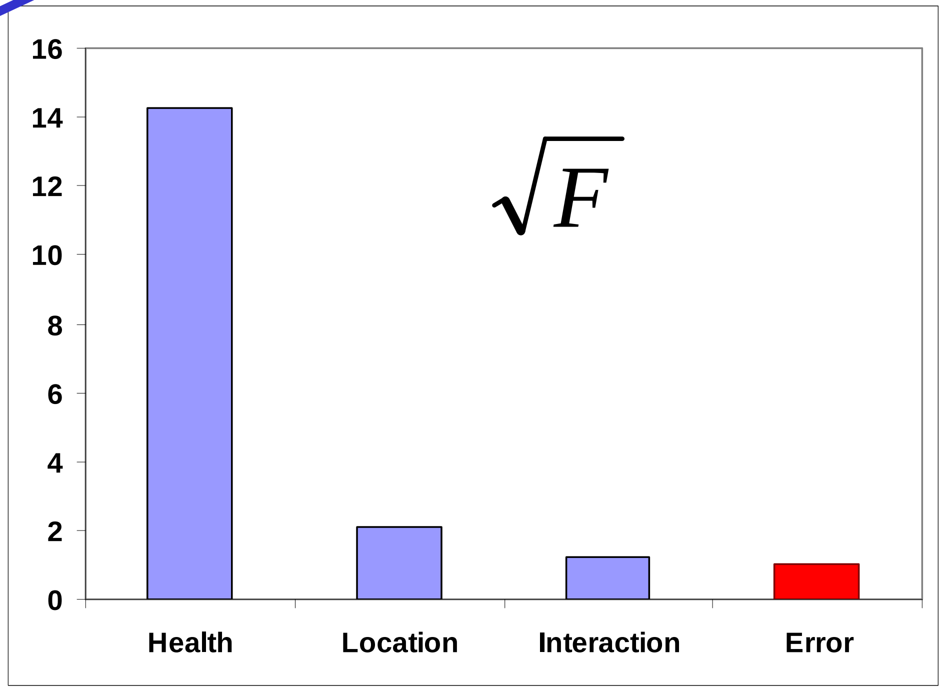
2. Use Tools → Data Analysis → ANOVA: Two-factor with replicates

2-factor ANOVA with *r* Replicates: Example

ANOVA

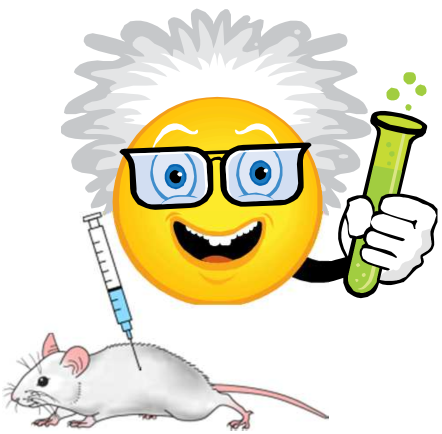
	Source of Variation	SS	df	MS	F	P-value	F crit
Health	Sample	1748.033	1	1748.033	203.094	4.4E-27	3.92433
Location	Columns	73.85	2	36.925	4.290104	0.015981	3.075853
Interaction	Interaction	26.11667	2	13.05833	1.517173	0.223726	3.075853
Error	Within	981.2	114	8.607018			
	Total	2829.2	119				





QUESTIONS ?

Thank you for your attention



to be continued…