Einführung in Web- und Data-Science

Prof. Dr. Ralf Möller

Universität zu Lübeck

Institut für Informationssysteme

Tanya Braun (Übungen)

Acknowledgements

This lecture is based on
the following presentation:

ANOVA:
Analysis of Variation

Math 243 Lecture

R. Pruim

(but contains additions and modifications)

Example

Subjects: 25 patients with blisters
Treatments: Treatment A, Treatment B, Placebo
Measurement: # of days until blisters heal

Data [and means]:

- A: 5, 6, 6, 7, 7, 8, 9, 10 [7.25]
- B: 7, 7, 8, 9, 9, 10, 10, 11 [8.875]
- P: 7, 9, 9, 10, 10, 10, 11, 12, 13 [10.11]

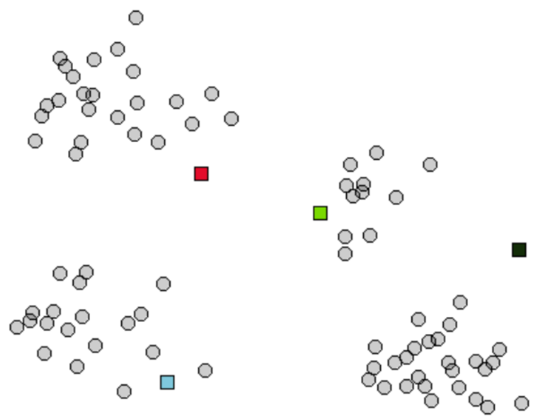
Are these differences significant?

Variation BETWEEN groups vs. variation WITHIN groups

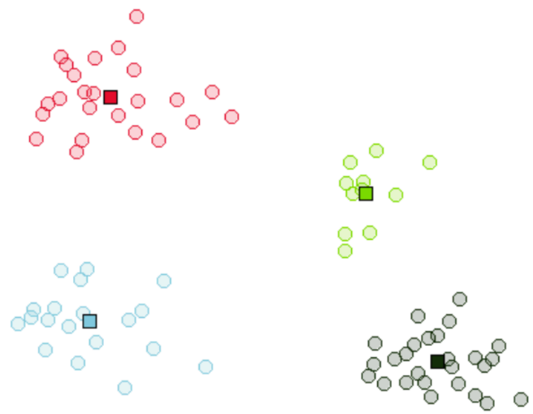
Analysis of variation required: ANOVA

ANOVA and Clustering

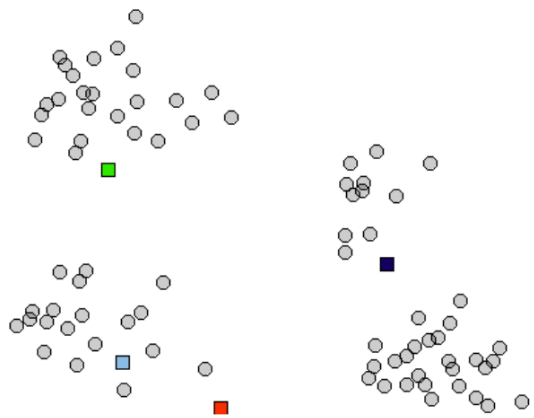
Init values



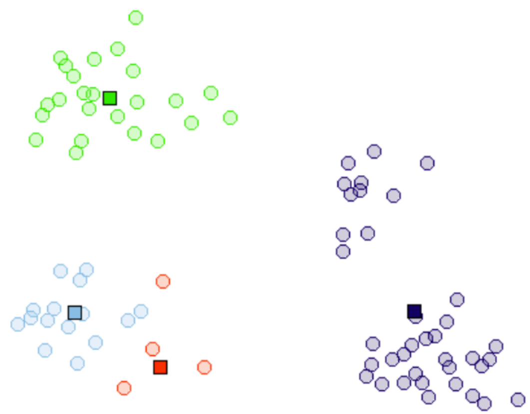
Good result ?



Init values



Bad result ?



The basic ANOVA situation

Two variables: 1 Categorical (type, group), 1 Quantitative (value)

Main Question: Do the (means of) the quantitative variables depend on the group (given by categorical variable) the individual is in?

If categorical variable has only 2 values:

- 2-sample t-test

ANOVA allows for 3 or more groups

Informal Investigation

Graphical investigation:

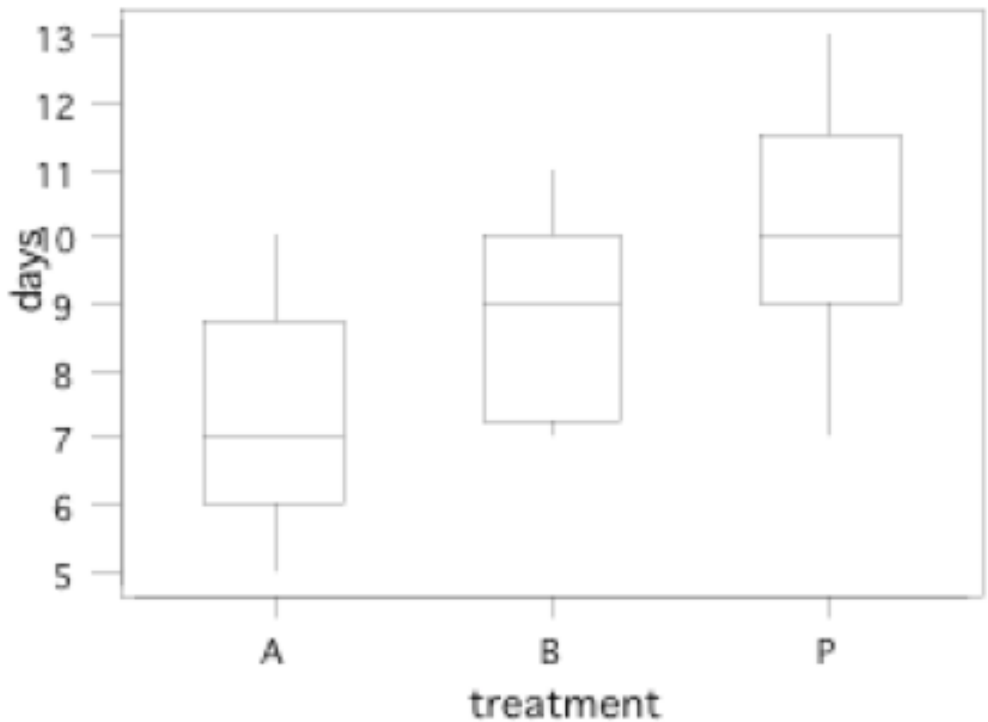
- side-by-side box plots
- multiple histograms

Whether the differences between the groups are significant depends on

- the difference in the means
- the standard deviations of each group
- the sample sizes (aka degrees of freedom df)

Need p-value to make a decision
ANOVA determines p-value from a specific statistic

Side by Side Boxplots



What does ANOVA do?

At its simplest (there are extensions) ANOVA tests the following hypotheses:

H_0: The means of all the groups are equal.

H_a: Not all the means are equal

- doesn't say how or which ones differ.
- Can follow up with "multiple comparisons"

Note: we usually refer to the sub-populations as "groups" when doing ANOVA.

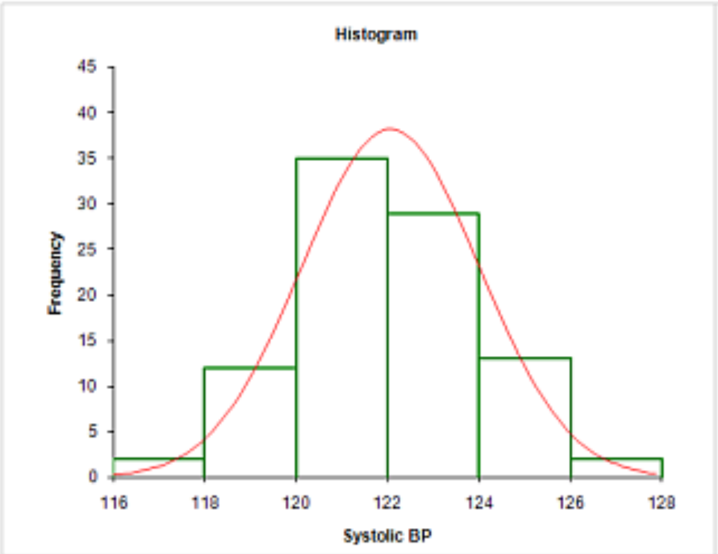
Assumptions of ANOVA

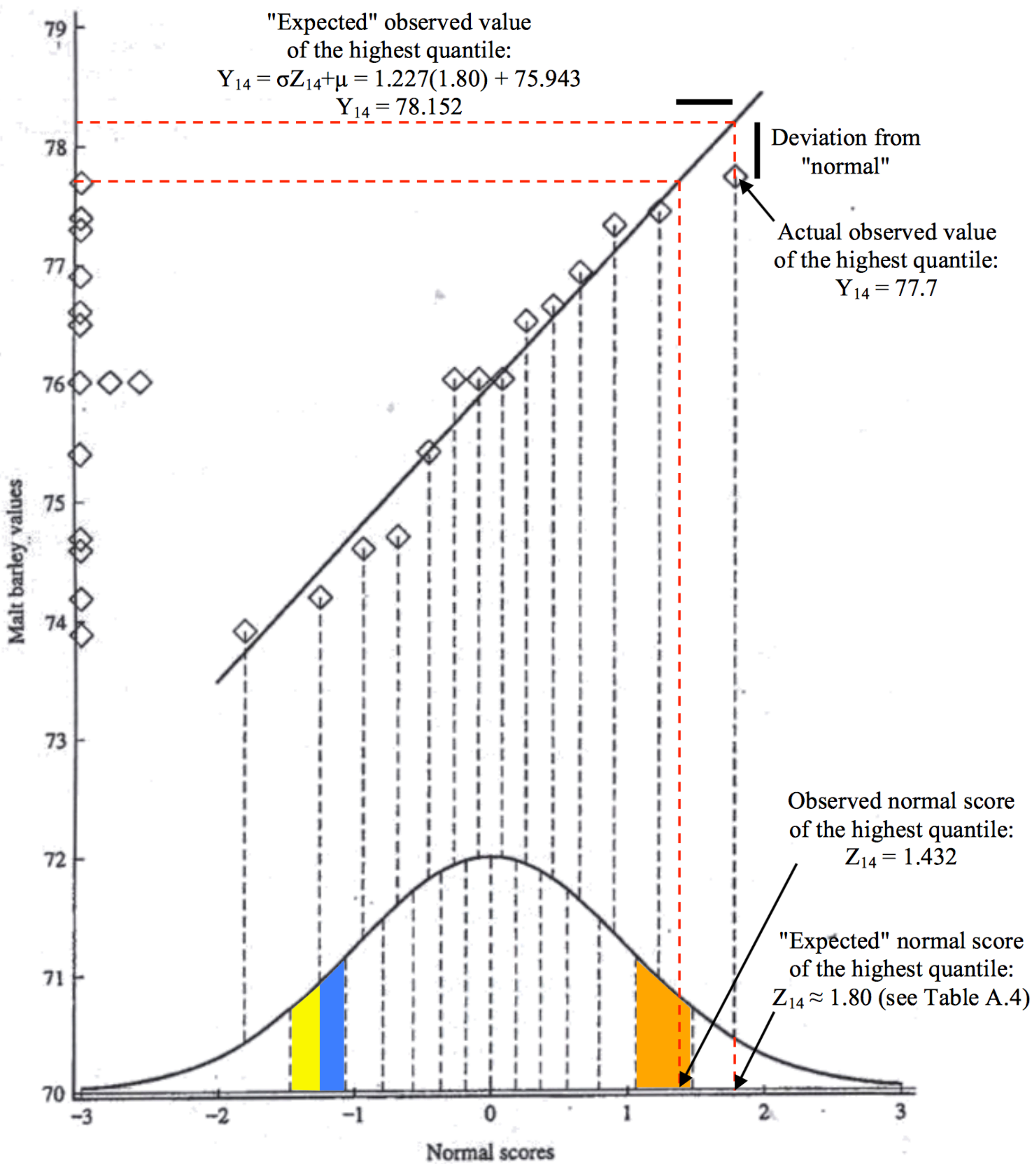
- Each group is approximately normal

Normality Check

We should check for normality using:

- Assumptions about population
- Histograms for each group
- Normal quantile plot for each group



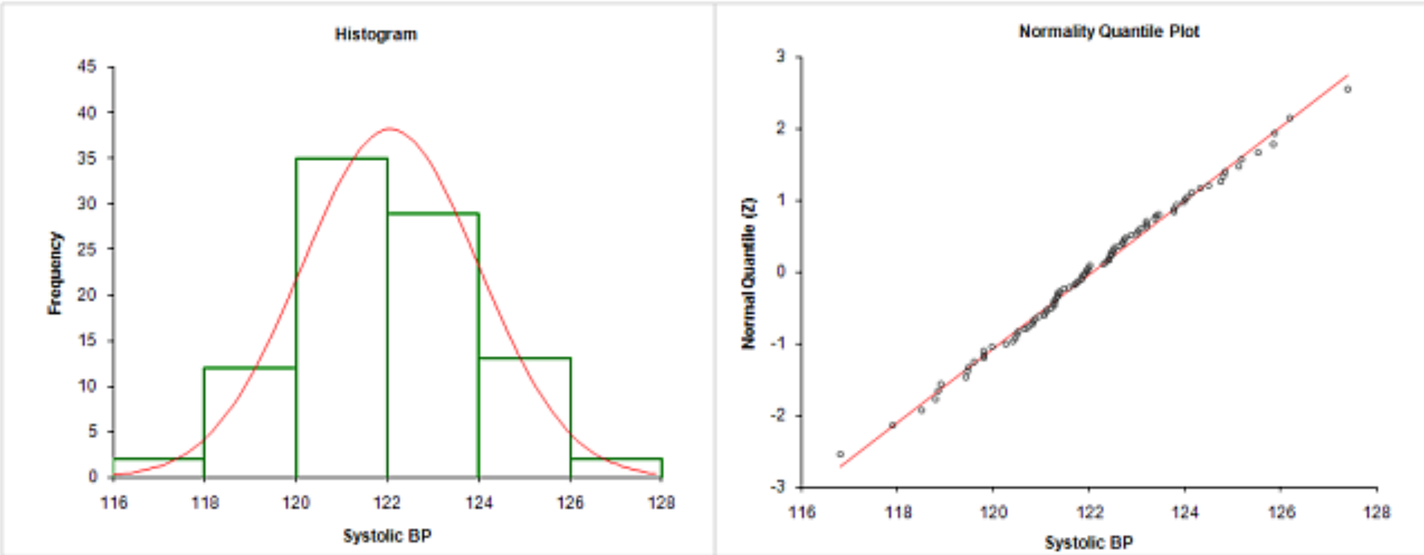

"Expected" observed value
of the highest quantile:
$Y_{14} = \sigma Z_{14} + \mu = 1.227(1.80) + 75.943$
$Y_{14} = 78.152$
Deviation from "normal"
Actual observed value
of the highest quantile:
$Y_{14} = 77.7$
Observed normal score
of the highest quantile:
$Z_{14} = 1.432$
"Expected" normal score
of the highest quantile:
$Z_{14} \approx 1.80$ (see Table A.4)
Malt barley values
Normal scores
79
78
77
76
75
74
73
72
71
70
−3
−2
−1
0
1
2
3

Normality Check

We should check for normality using:

- Assumptions about population
- Histograms for each group
- Normal quantile plot for each group

Useful only for "large" datasets



With small data sets, there really isn't a really good way to check normality from data, but we make the common assumption that physical measurements of people tend to be normally distributed (but see Kolmogorov-Smirnov-Test)

Assumptions of ANOVA

- Each group is approximately normal
  - Check this by looking at histograms and/or normal quantile plots, or use assumptions
  - Can handle some non-normality, but not severe outliers
- Standard deviations of each group are approximately equal
  - Rule of thumb: ratio of largest to smallest sample st. dev. must be less than 2:1

Standard Deviation Check

Variable	treatment	N	Mean	Median	StDev
days	A	8	7.250	7.000	1.669
	B	8	8.875	9.000	1.458
	P	9	10.111	10.000	1.764

Compare largest and smallest standard deviations:

- largest: 1.764
- smallest: 1.458
- 1.458 x 2 = 2.916 > 1.764

Notation for ANOVA

- n = number of individuals all together
- I = number of groups
- $\bar{x}$ = mean for entire data set

Group i has

- n_i = # of individuals in group i
- x_{ij} = value for individual j in group i
- $\bar{x}_i$ = mean for group i
- s_i = standard deviation for group i

How ANOVA works (outline)

ANOVA measures two sources of variation in the data and compares their relative sizes

- Variation BETWEEN groups (MSG)
  for each group look at the difference between its mean and the overall mean

$$N^{-1}\Sigma_i \; (\bar{x}_i - \bar{x})^2$$

- Variation WITHIN groups (MSE)
  for each data value x_j of group i we look at the difference between that value and the mean of its group $M^{-1}\Sigma_{obs_{ij}}(x_{ij} - \bar{x}_i)^2$

F Statistic

The ANOVA F-statistic is a ratio of the Between Group Variaton divided by the Within Group Variation:

$$F = \frac{Between}{Within} = \frac{MSG}{MSE}$$

A large F is evidence *against* H_0, since it indicates that there is more difference between groups than within groups
(hence the means between at least two groups differ).

H_0: The means of all the groups are equal.

Computations

We want to measure the amount of
variation due to BETWEEN group variation and WITHIN group variation

For each data value, we calculate its contribution to:

- BETWEEN group variation: $\left(\bar{x}_i - \bar{\bar{x}}\right)^2$

- WITHIN group variation: $(x_{ij} - \bar{x}_i)^2$

An even smaller example

Suppose we have three groups

- Group 1: 5.3, 6.0, 6.7
- Group 2: 5.5, 6.2, 6.4, 5.7
- Group 3: 7.5, 7.2, 7.9

We get the following statistics:

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	3	18	6	0.49
Column 2	4	23.8	5.95	0.17666
Column 3	3	22.6	7.53333	0.12333

ANOVA Output

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	5.127333	2	2.563667	10.21575	0.008394	4.73741
Within Groups	1.756667	7	0.250952			
Total	6.884	9				

1 less than number of groups

1 less than number of individuals (just like other situations)

number of data values - number of groups (equals df for each group added together)

Computing ANOVA F statistic

			WITHIN		BETWEEN	
			difference:		difference	
		group	data - group mean		group mean - overall me	
data	group	mean	plain	squared	plain	squared
5.3	1	6.00	-0.70	0.490	-0.4	0.194
6.0	1	6.00	0.00	0.000	-0.4	0.194
6.7	1	6.00	0.70	0.490	-0.4	0.194
5.5	2	5.95	-0.45	0.203	-0.5	0.240
6.2	2	5.95	0.25	0.063	-0.5	0.240
6.4	2	5.95	0.45	0.203	-0.5	0.240
5.7	2	5.95	-0.25	0.063	-0.5	0.240
7.5	3	7.53	-0.03	0.001	1.1	1.188
7.2	3	7.53	-0.33	0.109	1.1	1.188
7.9	3	7.53	0.37	0.137	1.1	1.188
TOTAL				1.757		5.106
TOTAL/df				0.25095714		2.5527

overall mean: 6.44 $\qquad$ F = 2.5528/0.25025 = 10.21575

ANOVA Output

```
Analysis of Variance for days
Source       DF        SS        MS        F        P
treatment     2     34.74     17.37     6.45    0.006
Error        22     59.26      2.69
Total        24     94.00
```

1 less than # of groups

1 less than # of individuals (just like other situations)

# of data values - # of groups (equals df for each group added together)

ANOVA Output for Drug Example

Analysis of Variance for days

Source	DF	SS	MS	F	P
treatment	2	34.74	17.37	6.45	0.006
Error	22	59.26	2.69		
Total	24	94.00			

$$\sum_{obs}(x_{ij} - \bar{x}_i)^2$$

$$\sum_{obs}(x_{ij} - \bar{\bar{x}})^2$$

$$\sum_{obs}(\bar{x}_i - \bar{\bar{x}})^2$$

SS stands for sum of squares

- ANOVA splits this into 3 parts

ANOVA Output

```
Analysis of Variance for days
Source      DF        SS        MS       F       P
treatment    2     34.74     17.37    6.45   0.006
Error       22     59.26      2.69
Total       24     94.00
```

MSG = SSG / DFG
MSE = SSE / DFE

F = MSG / MSE

P-value comes from F(DFG,DFE)

(P-values for the F statistic are in table as usual)

So How big is F?

Since F is
Mean Square Between / Mean Square Within

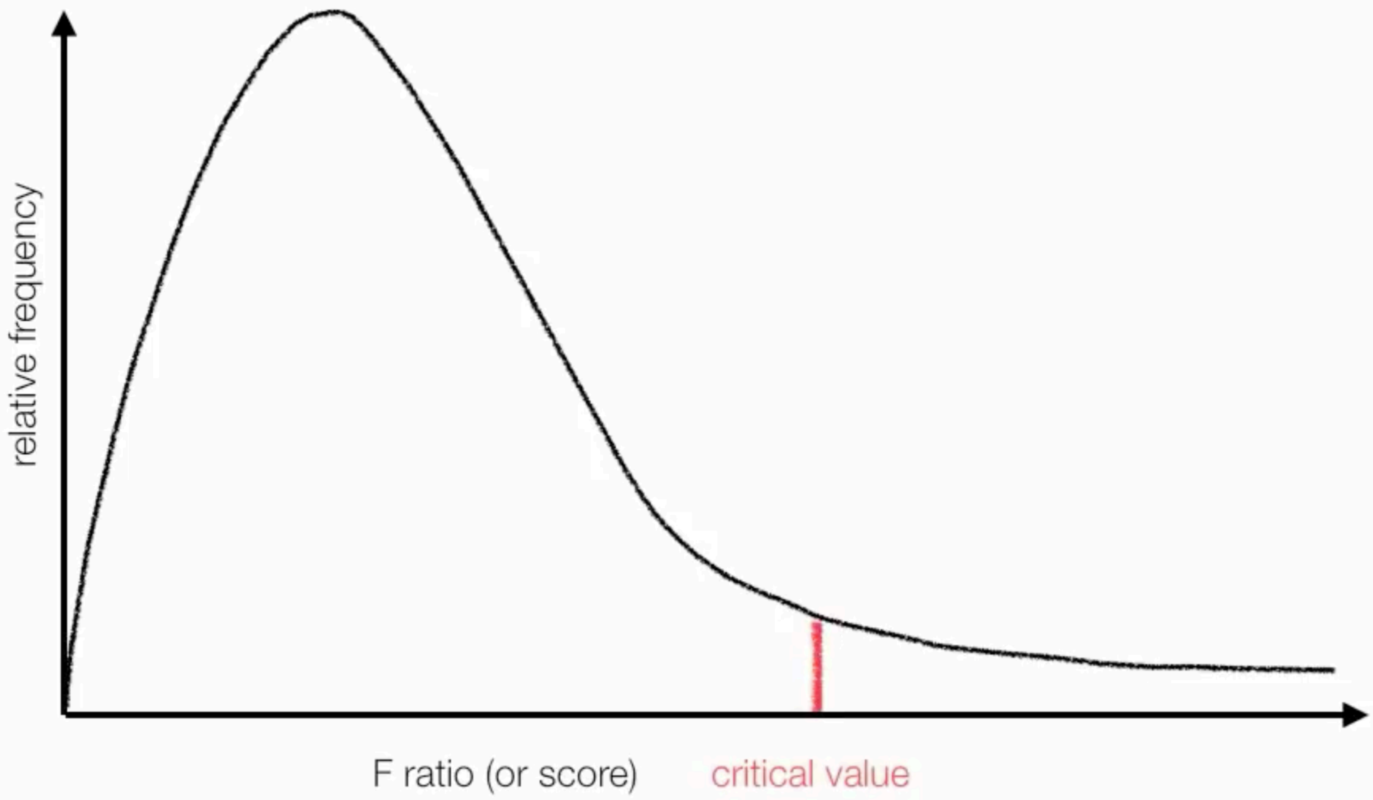
= MSG / MSE

A large value of F indicates relatively more difference between groups than within groups (evidence against H_0)

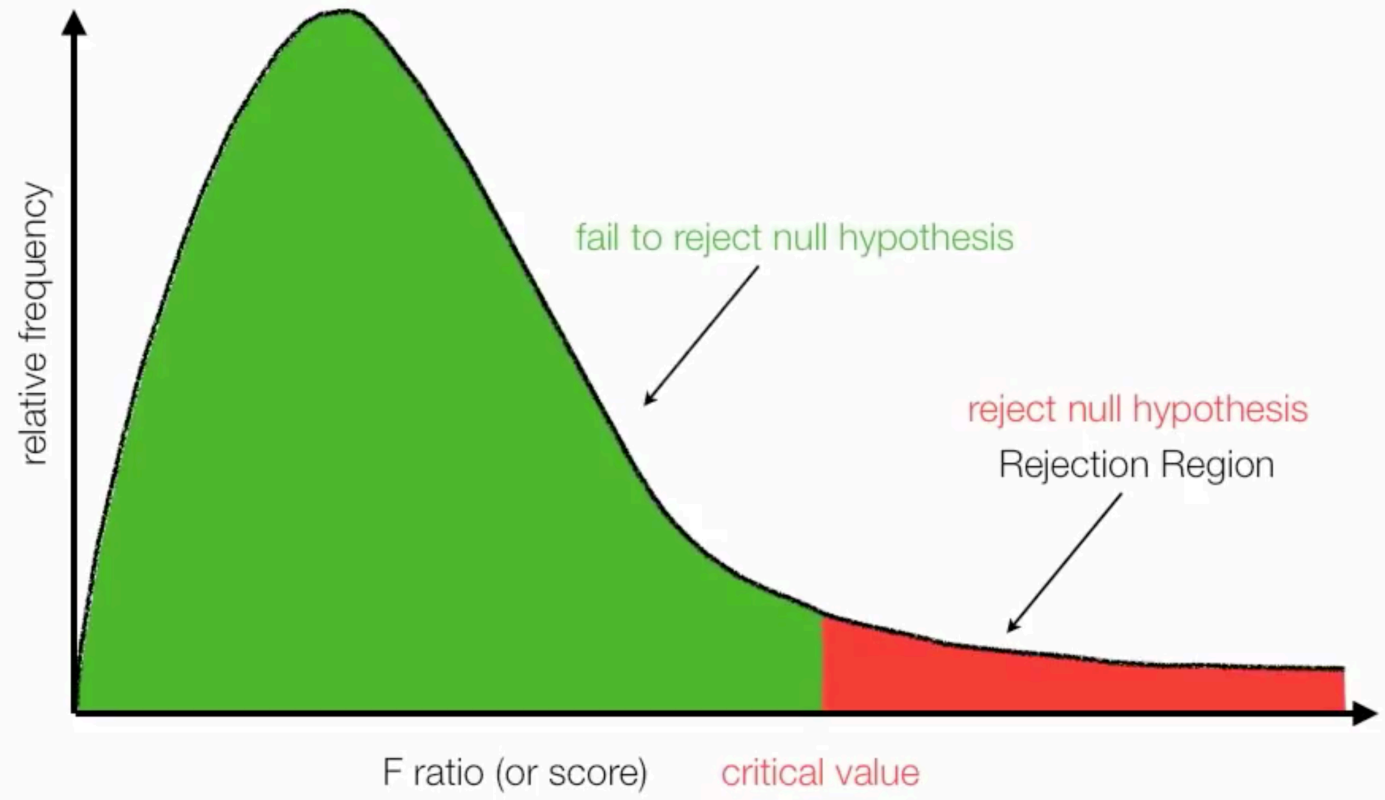
To get the P-value, we compare to *F(I-1,n-I)*-distribution

- *I-1* degrees of freedom in numerator (# groups -1)
- *n - I* degrees of freedom in denominator (rest of df)

F-Distribution



Critical Value



Example: $\alpha = 0.05$

$$F(2, 12) = 22.59,\ p < .05$$

degrees of freedom numerator
relates to groups or samples

Example: $\alpha = 0.05$

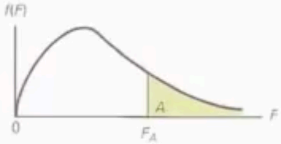
relates total observations
degrees of freedom denominator

$$F(2, 12) = 22.59,\ p < .05$$

degrees of freedom numerator

F-Table

Table 6(a) Critical Values of F: $A = .05$



relates to groups or samples

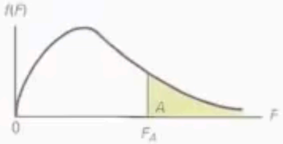
relates to number of observations

ν_2 \ ν_1 (NUMERATOR DEGREES OF FREEDOM)	1	2	3	4	5	6	7	8	9
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30

DENOMINATOR DEGREES OF FREEDOM

Critical Value for $\alpha = 0.05$

Table 6(a) Critical Values of F: $A = .05$



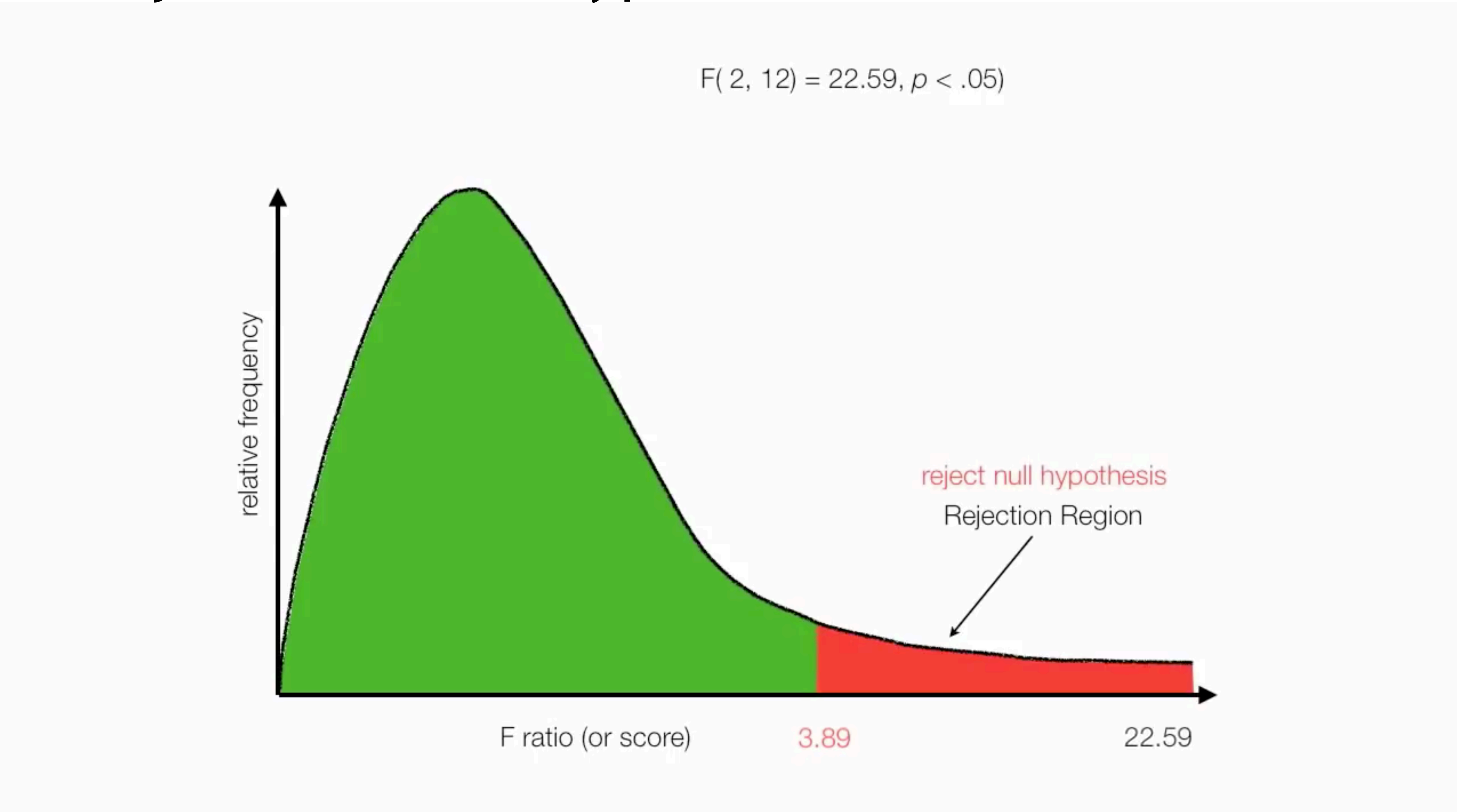
F(2, 12) = 22.59, p < .05)

degrees of freedom denominator

ν_2 \ ν_1	NUMERATOR DEGREES OF FREEDOM 1	2	3	4	5	6	7	8	9
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28

DENOMINATOR DEGREES OF FREEDOM

Rejection of Null Hypothesis



Connections between SST, MST, and standard deviation

If ignore the groups for a moment and just compute the standard deviation of the entire data set, we see

$$s^2 = \frac{\sum\left(x_{ij} - \overline{\overline{x}}\right)^2}{n-1} = \frac{SST}{DFT} = MST$$

So $SST = (n-1)\, s^2$, and $MST = s^2$. That is, *SST* and *MST* measure the TOTAL variation in the data set.

SST: Sum of Squares Total

DFT: Degrees of Freedom Total

MST: Mean Sum of Squares Total

Connections between SSE, MSE, and standard deviation

Remember:

$$s_i^{\ 2} = \frac{\sum \left(x_{ij} - \overline{x}_i\right)^2}{n_i - 1} = \frac{SS[\text{WithinGroup } i]}{df_i}$$

So *SS[Within Group i] =* $(s_i^2)\ (df_i)$

This means that we can compute SSE from the standard deviations and sizes (df) of each group:

$$SSE = SS[Within] = \sum SS[WithinGroup\ i]$$
$$= \sum s_i^2 (n_i - 1) = \sum s_i^2 (df_i)$$

Pooled estimate for st. dev

One of the ANOVA assumptions is that all groups have the same standard deviation. We can estimate this with a weighted average:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots + (n_I - 1)s_I^2}{n - I}$$

$$s_p^2 = \frac{(df_1)s_1^2 + (df_2)s_2^2 + \ldots + (df_I)s_I^2}{df_1 + df_2 + \ldots + df_I}$$

$$s_p^2 = \frac{SSE}{DFE} = MSE$$

so MSE is the pooled estimate of variance

In Summary

$$SST = \sum_{obs} (x_{ij} - \overline{\overline{x}})^2 = s^2(DFT)$$

$$SSE = \sum_{obs} (x_{ij} - \overline{x_i})^2 = \sum_{groups} s_i^{\,2}(df_i)$$

$$SSG = \sum_{obs} (\overline{x_i} - \overline{\overline{x}})^2 = \sum_{groups} n_i(\overline{x_i} - \overline{\overline{x}})^2$$

$$SSE + SSG = SST; \quad MS = \frac{SS}{DF}; \quad F = \frac{MSG}{MSE}$$

R^2 Statistic

R^2 gives the percent of variance due to between group variation

$$R^2 = \frac{SS[Between]}{SS[Total]} = \frac{SSG}{SST}$$

Where's the Difference?

Once ANOVA indicates that the groups do not all appear to have the same means, what do we do?

```
Analysis of Variance for days
Source      DF        SS        MS        F        P
treatmen     2     34.74     17.37     6.45    0.006
Error       22     59.26      2.69
Total       24     94.00
                                   Individual 95% CIs For Mean
                                   Based on Pooled StDev
Level        N      Mean     StDev ----------+---------+---------+------
A            8     7.250     1.669 (-------*-------)
B            8     8.875     1.458            (-------*-------)
P            9    10.111     1.764                     (------*-------)
                                   ----------+---------+---------+------
Pooled StDev =     1.641                    7.5       9.0      10.5
```

Clearest difference: P is worse than A (CI's don't overlap)

Multiple Comparisons

Once ANOVA indicates that the groups do not all have the same means, we can compare them two by two using the 2-sample t test

- We need to adjust our p-value threshold because we are doing multiple tests with the same data.

- There are several methods for doing this.

- If we really just want to test the difference between one pair of treatments, we should set the study up that way.

Tuckey's Pairwise Comparisons

```
Tukey's pairwise comparisons

    Family error rate = 0.0500
Individual error rate = 0.0199

Critical value = 3.55

Intervals for (column level mean) - (row level mean)

                A             B

       B     -3.685
              0.435

       P     -4.863        -3.238
             -0.859         0.766
```

95% confidence

Use alpha = 0.0199 for each test.

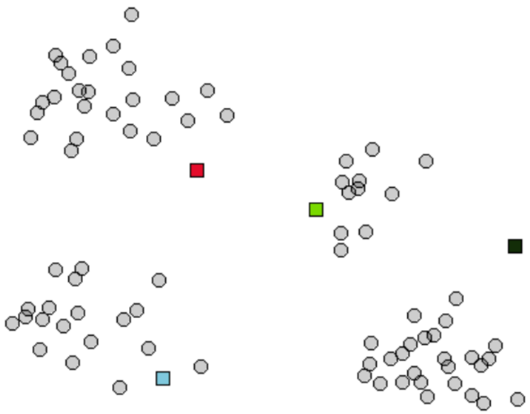
These give 98.01% CI's for each pairwise difference.

Only P vs A is significant
(both values have same sign)

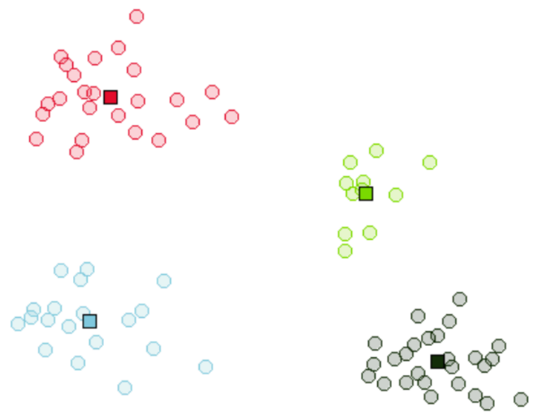
98% CI for A-P is (-0.86,-4.86)

ANOVA and Clustering

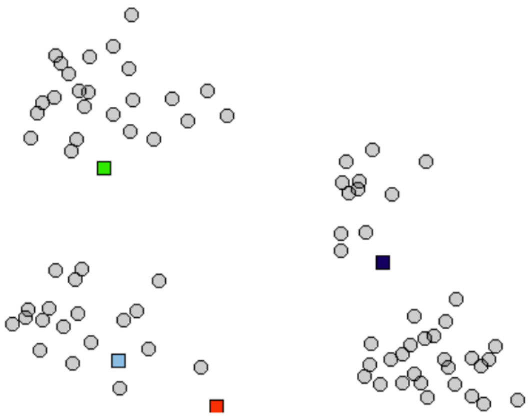
Init values



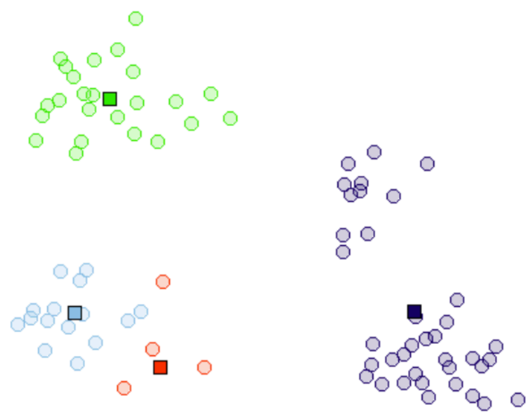
Good result !



Init values



Bad result !



Tukey's Method in R

```
Tukey multiple comparisons of means
    95% family-wise confidence level
```

	diff	lwr	upr
B-A	1.6250	-0.43650	3.6865
P-A	2.8611	0.85769	4.8645
P-B	1.2361	-0.76731	3.2395