# Einführung in Web- und Data-Science

Prof. Dr. Ralf Möller

**Universität zu Lübeck**

**Institut für Informationssysteme**

Tanya Braun (Übungen)

# Clustering

- Form des unüberwachten Lernens
- Suche nach natürlichen Gruppierungen von Objekten
  - Klassen direkt aus Daten bestimmen
    - Hohe Intra-Klassen-Ähnlichkeit
    - Kleine Inter-Klassen-Ähnlichkeit
  - Ggs.: Klassifikation
- Distanzmaße

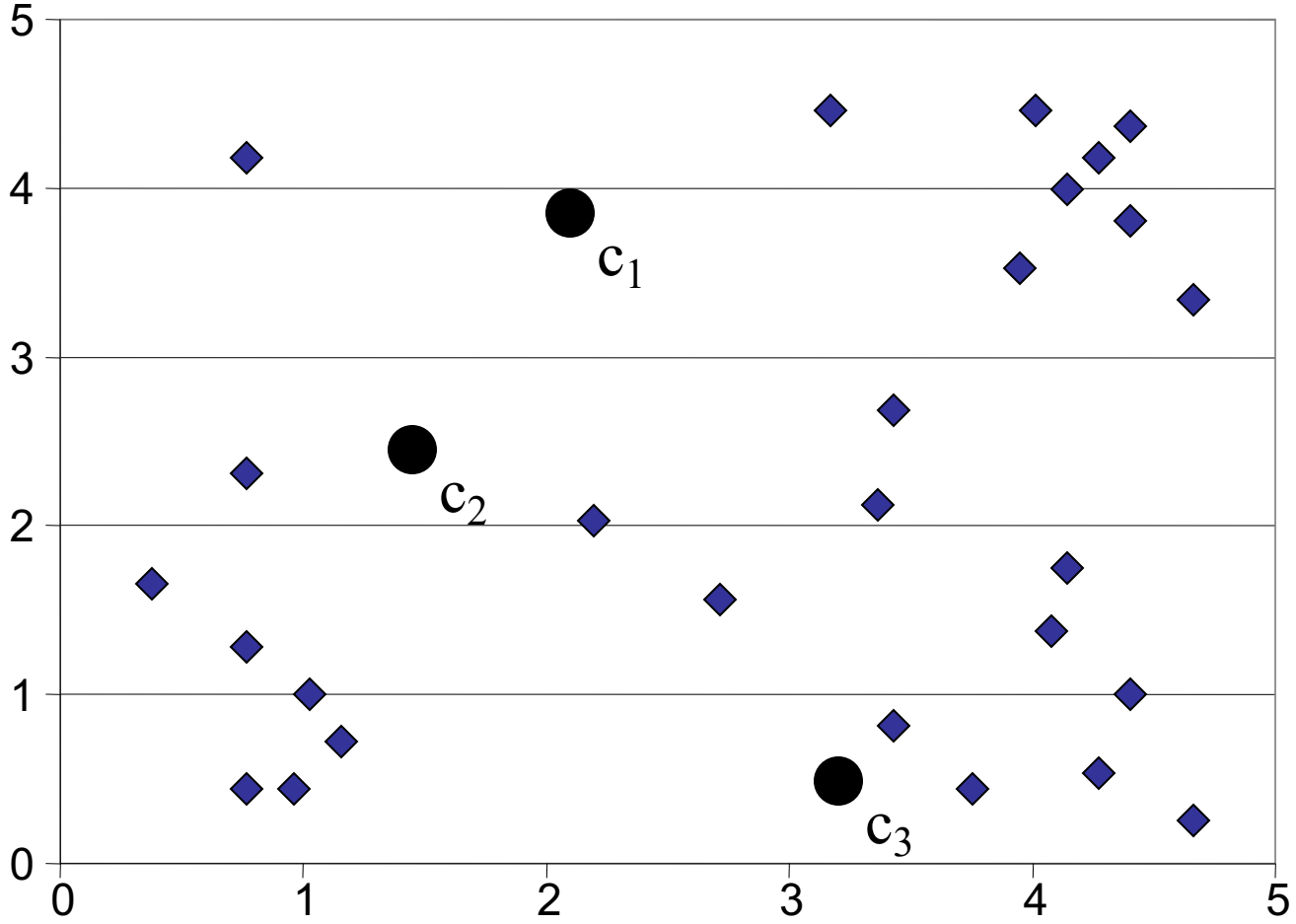
- z. B. Minkowski Distanz (im $\mathbb{R}^n$):

$$d_p(\mathbf{x}, \mathbf{y}) = \left( \sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} = ||\mathbf{x} - \mathbf{y}||_p$$

- für $p = 1$: Manhattan Distanz
- für $p = 2$: Euklidische Distanz

# Partitionierung: K-means Clustering (1)

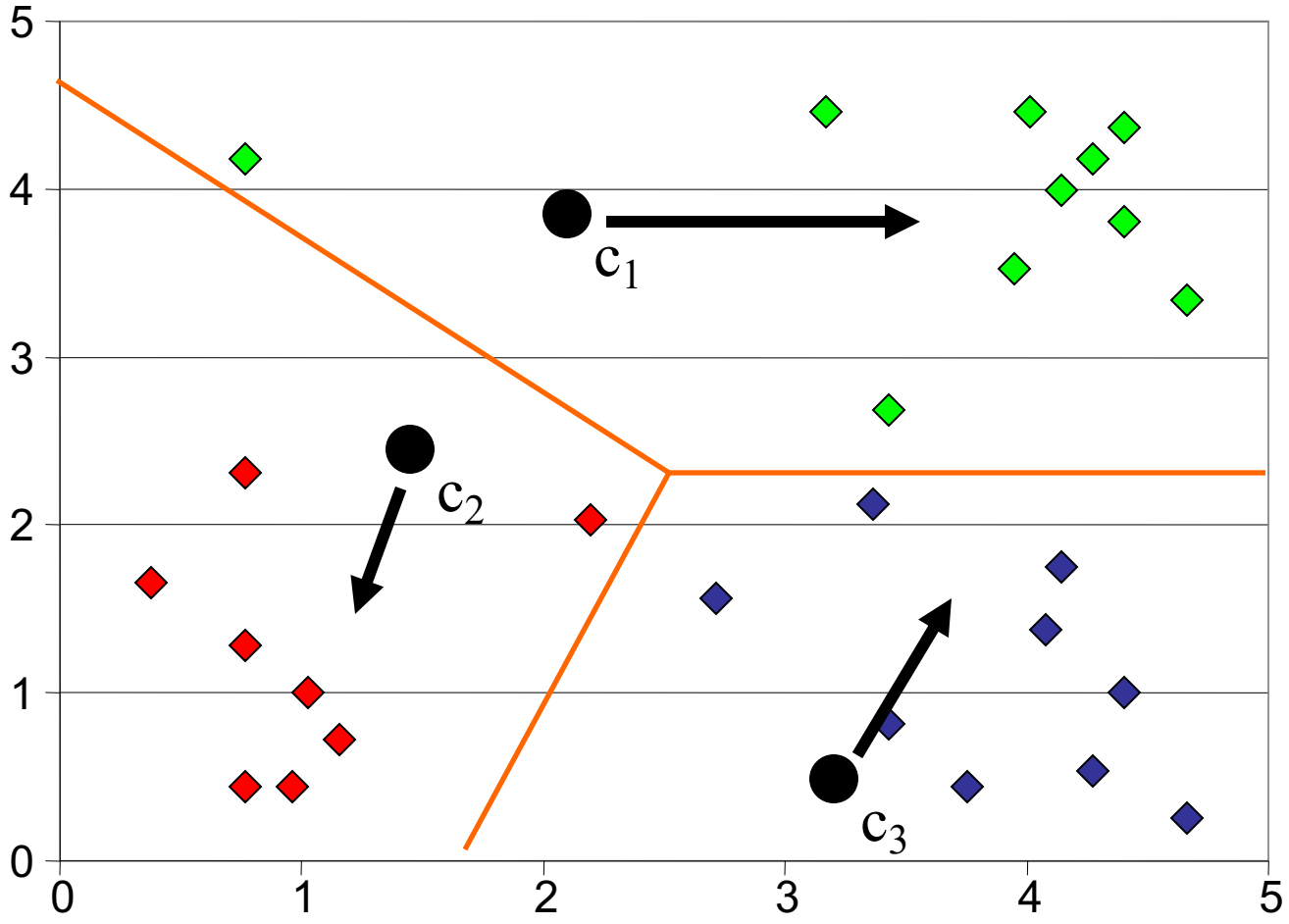
Distanzmaß: Euklidische Distanz




$$C_i^t = \left\{ x_j : \left\| x_j - c_i^t \right\|_2 \leq \left\| x_j - c_r^t \right\|_2 \text{ for all } r = 1 \ldots k, r \neq i \right\}$$

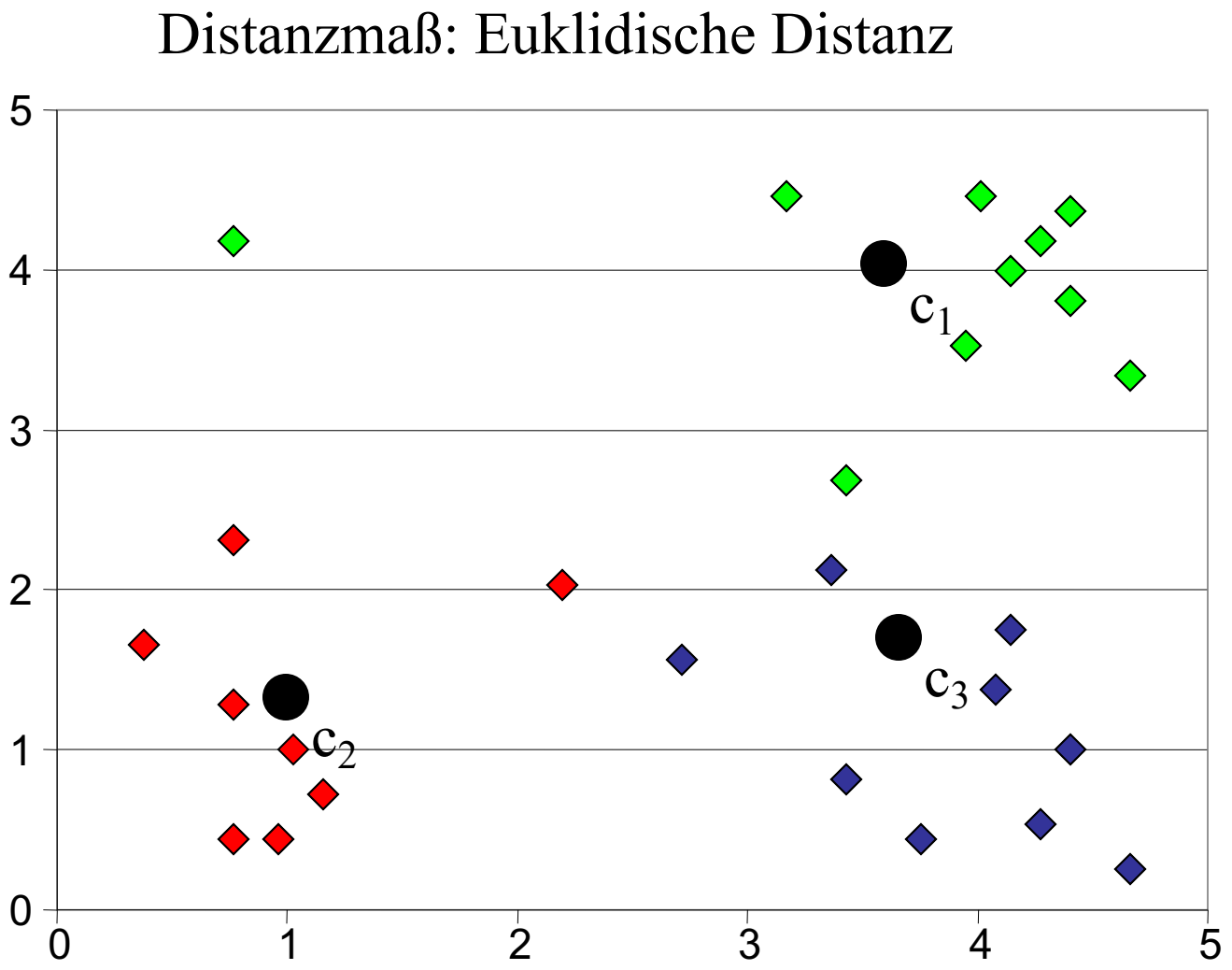
# K-means Clustering (2)

Distanzmaß: Euklidische Distanz

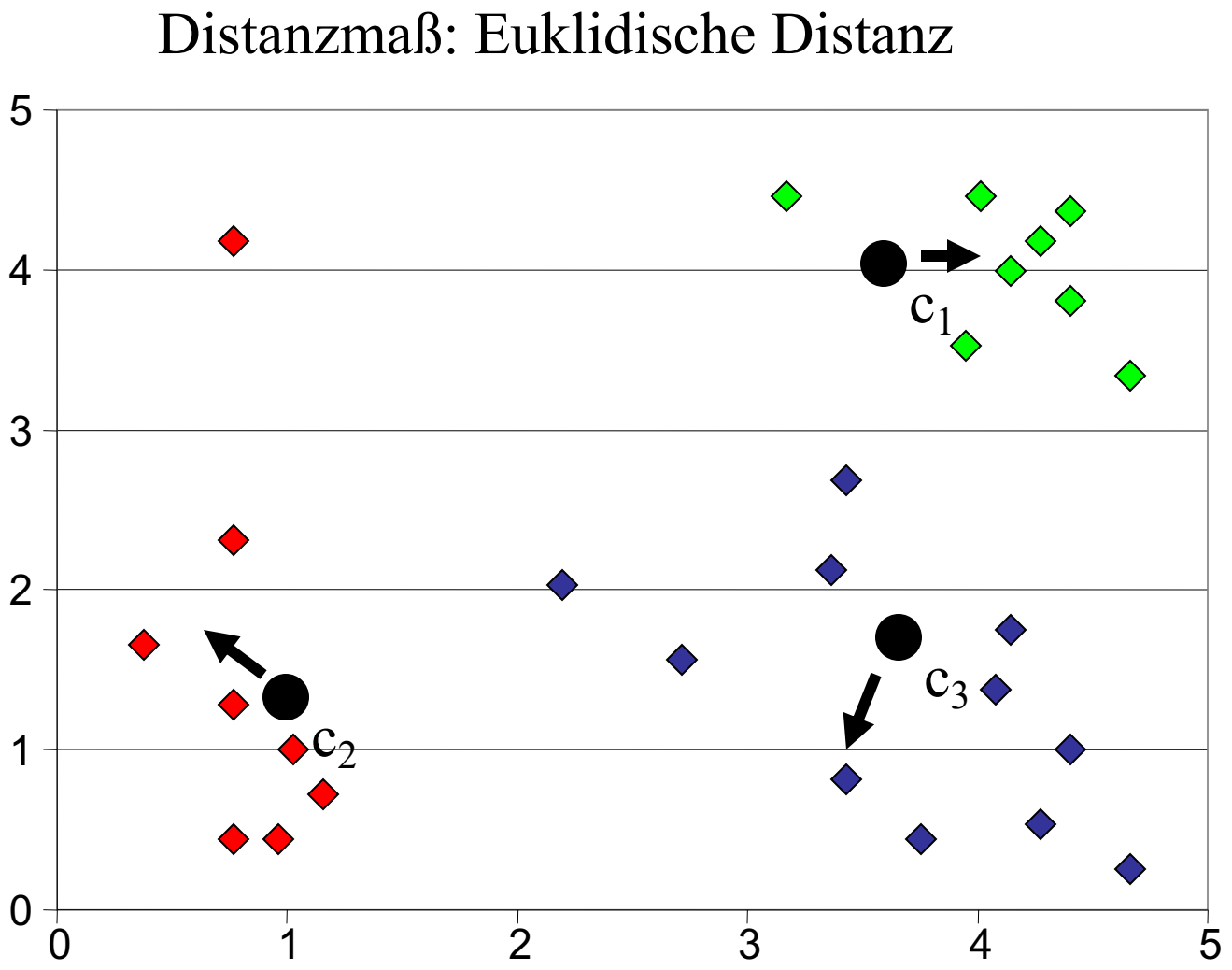



$$c_i^{t+1} = \frac{1}{|C_i^t|} \sum_{x_j \in C_i^t} x_j$$

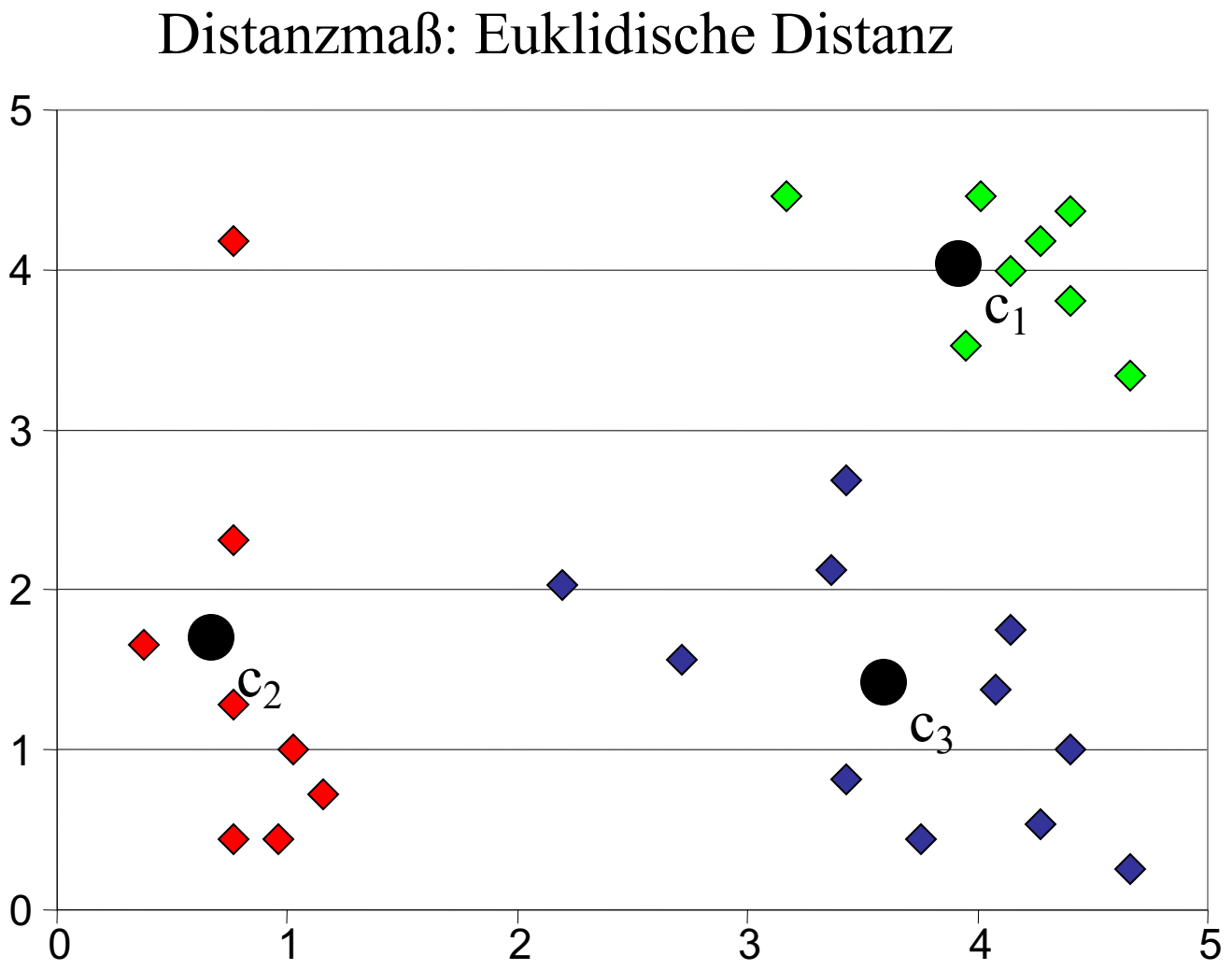
# K-means Clustering (3)



# K-means Clustering (4)



# K-means Clustering (5)



# Acknowledgements

This part is based on
the following presentation:

## ANOVA:
## Analysis of Variation

Math 243 Lecture

R. Pruim

(but contains changes 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


Difference between clusters less than difference within clusters

# 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

# 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
  - Check this by looking at histograms 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
- $\overline{X}$ = mean for entire data set

Group $i$ has

- $n_i$ = # of individuals in group $i$
- $x_{ij}$ = value for individual $j$ in group $i$
- $\overline{x}_i$ = 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

$$N^{-1} \sum_{obs_i} (\bar{x}_i - \bar{x})^2$$

N: Normalization value
(corrected: degrees of freedom)

- 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

$$M^{-1} \sum_{obs_{ij}} (x_{ij} - \bar{x}_i)^2$$

M: Normalization value
(corrected: degrees of freedom)

# 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.

$H_0$ in terms of clusters:
Clusters are bad
(centroids are equal)

# An even smaller example

Suppose we have three groups (#groups = l)

- 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* |
| Group 1 | 3 | 18 | 6 | 0.49 |
| Group 2 | 4 | 23.8 | 5.95 | 0.17666 |
| Group 3 | 3 | 22.6 | 7.53333 | 0.12333 |

# ANOVA Output

| ANOVA | | | | | | |
|---|---|---|---|---|---|---|
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| Between Groups | 5.12733 | 2 | 2.56366 | 10.2157 | 0.00839 | 4.73741 |
| Within Groups | 1.75666 | 7 | 0.25095 | | | |
| | | | | | | |
| Total | 6.884 | 9 | | | | |

1 less than number of groups: l-1

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

number of data values - number of groups: n-l (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

F = 2.5528/0.25025 = 10.21575

# So How big is F?

Since F is

Mean Square Between / Mean Square Within

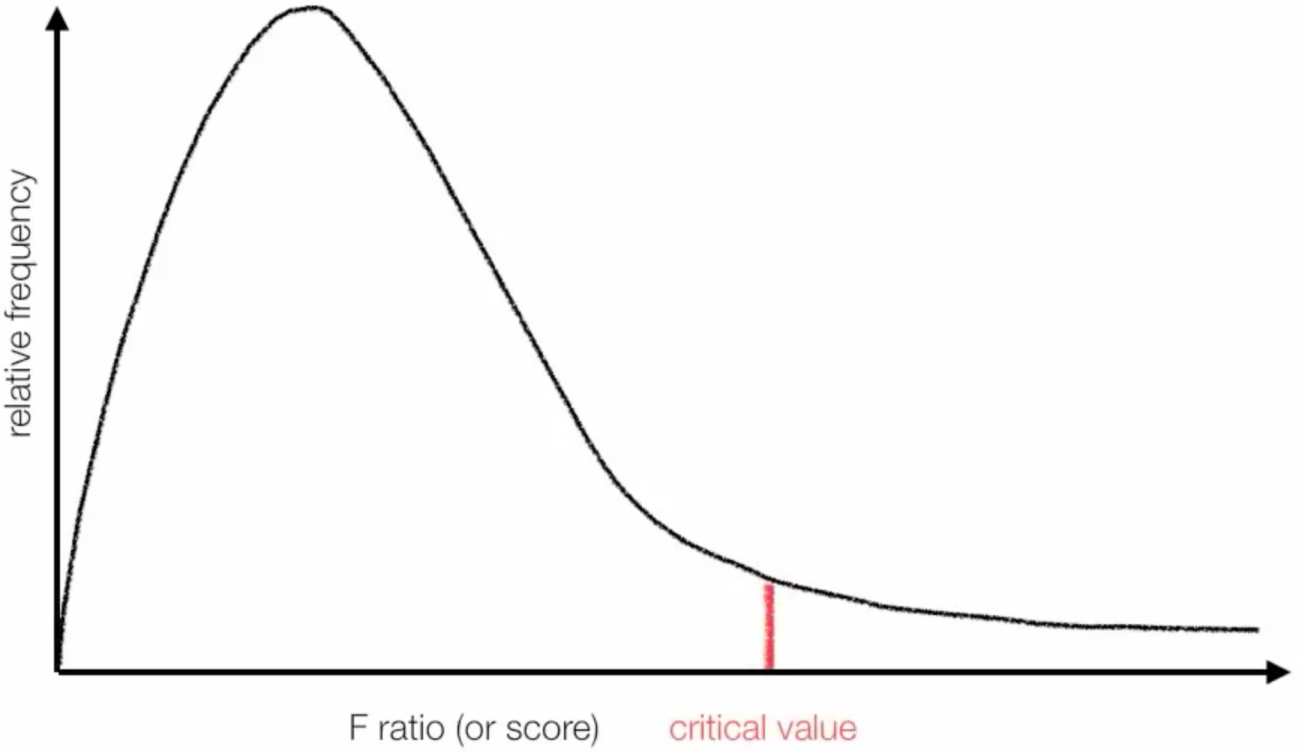
= MSG / MSE

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

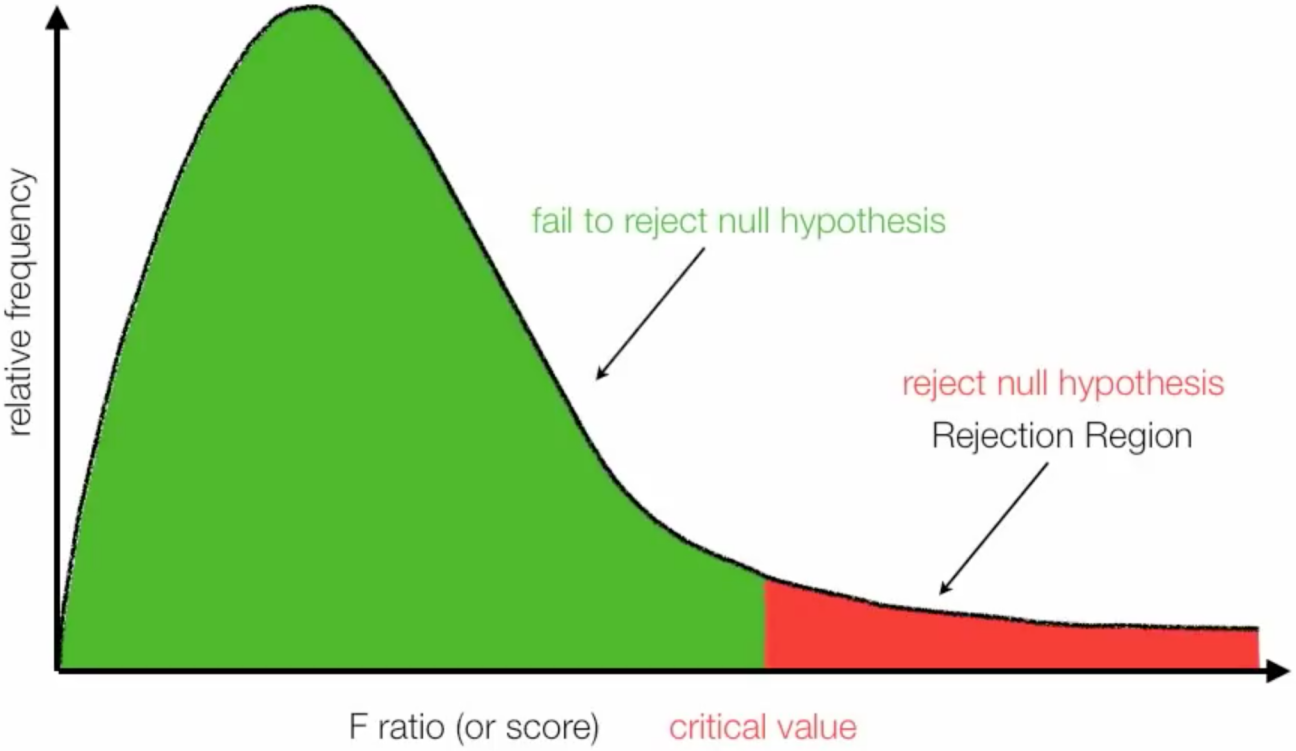
To get the P-value, we compare to *F(I-1,n-I)*-distribution

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

# F-Distribution



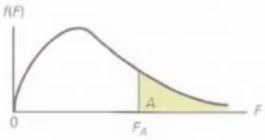
# Critical Value



# F-Table

$\alpha$ = 0.05 (use another table for different $\alpha$)
Computed F-Value = 10.21
Critical Value F(2, 9) = 4.26

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



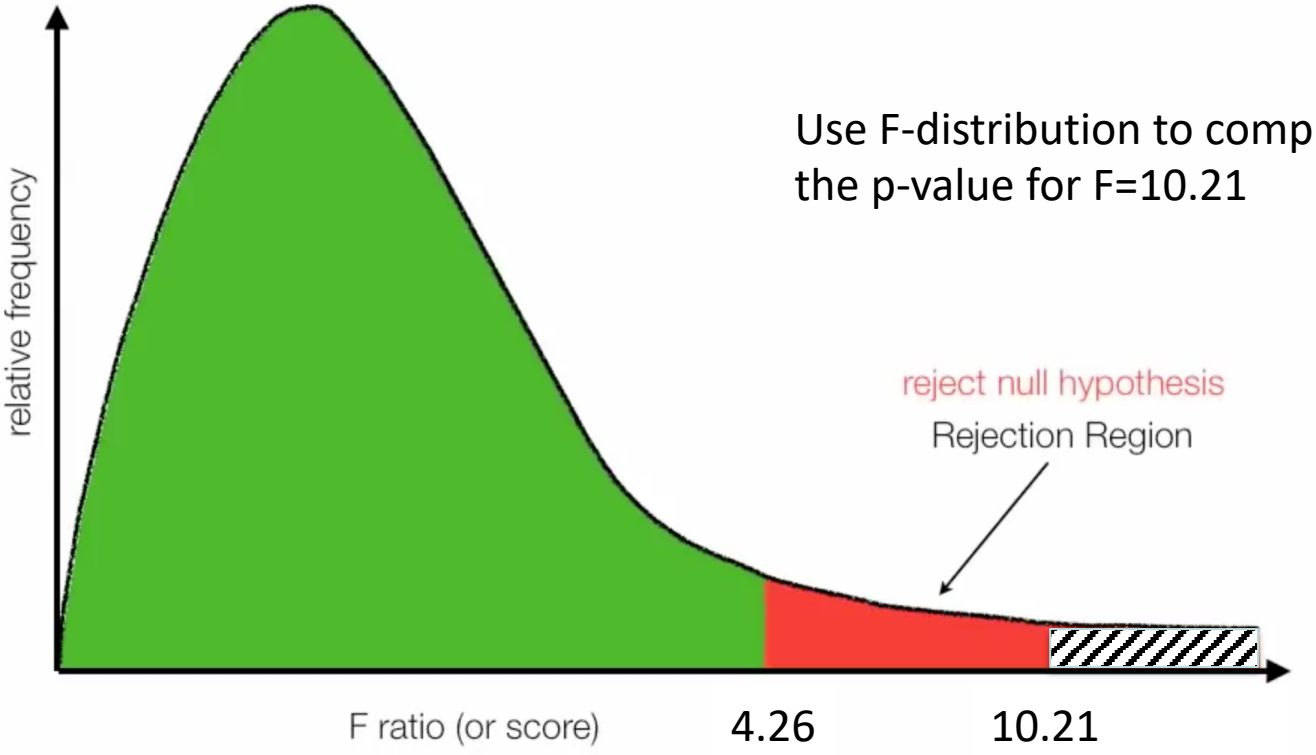

relates to groups or samples

relates to number of observations

| $\nu_2$ \ $\nu_1$ | 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



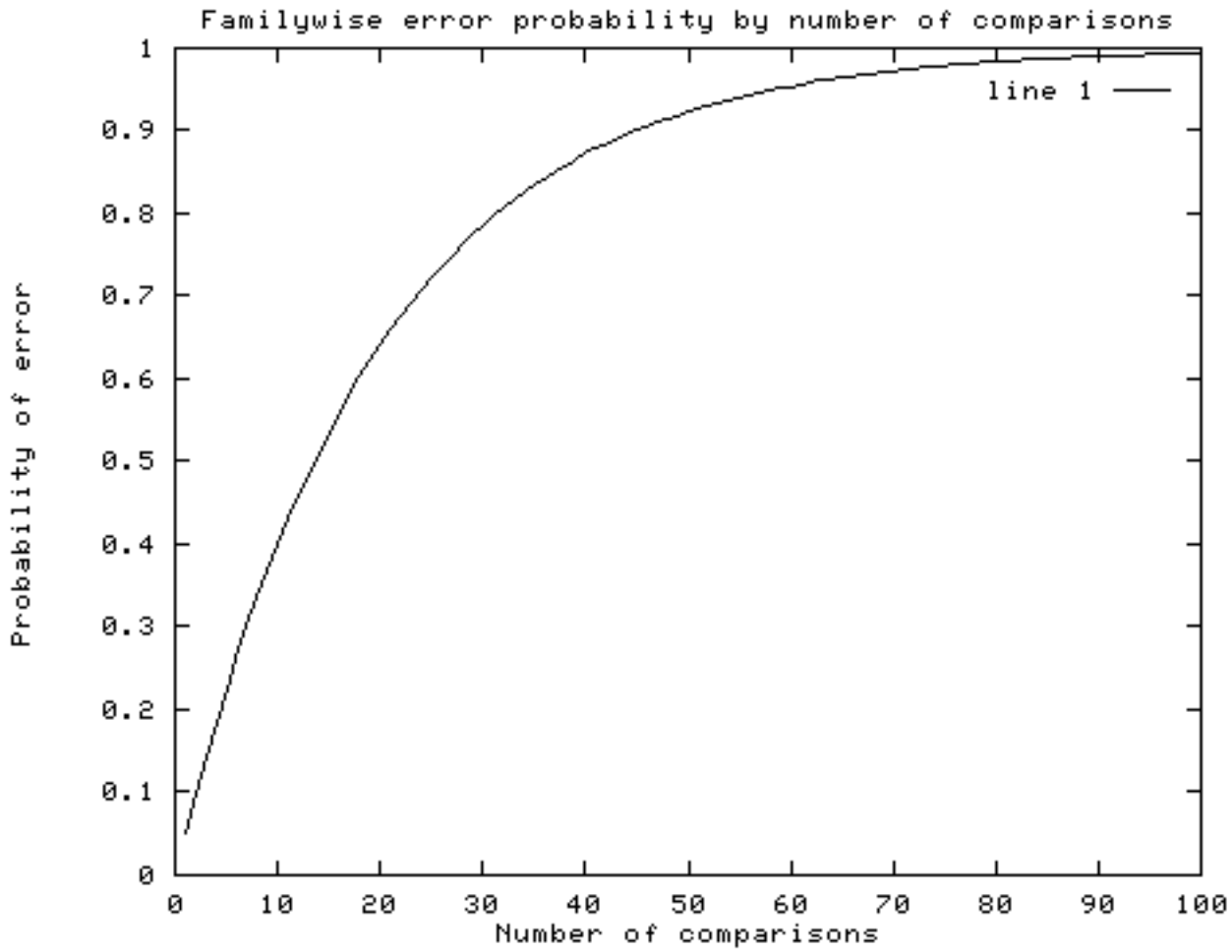

Use F-distribution to compute the p-value for F=10.21

# Why not just do 3 pairwise t-tests?

Answer:

- At an error rate of 5% for each test (max p-value), overall chance of type-I error is up to $1-(.95)^3 = 14\%$
  - If all 3 comparisons independent
- For 6 groups: ${}_6C_2 = 15$ pairwise t-tests;
  - High chance of finding something significant just by chance (if all tests were independent with a type-I error rate of 5% each)
  - Probability of at least one type-I error = $1-(.95)^{15}=54\%$.

# Recall: Multiple comparisons



# Correction for multiple comparisons

How to correct for multiple comparisons *post-hoc*…

- Bonferroni correction (adjust $\alpha$ by most conservative amount; assuming all tests independent, divide $\alpha$ by the number of tests)
- …

# Bonferroni

For example, to make a Bonferroni correction, divide your desired alpha cut-off level (usually .05) by the number of comparisons you are making. Assumes complete independence between comparisons, which is way too conservative.

| Obtained P-value | Original Alpha | # tests | New Alpha | Significant? |
|---|---|---|---|---|
| .001 | .05 | 5 | .010 | Yes |
| .011 | .05 | 4 | .013 | Yes |
| .019 | .05 | 3 | .017 | No |
| .032 | .05 | 2 | .025 | No |
| .048 | .05 | 1 | .050 | Yes |

# Multivariate Analysis of Variance: MANOVA

- An extension of ANOVA in which main effects and interactions are assessed on a combination of DVs
  - IV = independent variable, manipulated variable (e.g., Treatment)
  - DV = dependent variable, measured variable (e.g., Mean)
- MANOVA tests whether mean differences among groups on a combination of DVs is likely to occur by chance
- New DVs are created that are linear combinations of the individual DVs such that the difference between groups is maximized
- The questions are mostly the same as ANOVA just on the linearly combined DVs instead just one DV

# Are there any interactions among the IVs?

- Does change in the linearly combined DV for one IV depend on the levels of another IV?

- For example: Given three types of treatment, does one treatment work better for men and another work better for women?

# Basic Requirements

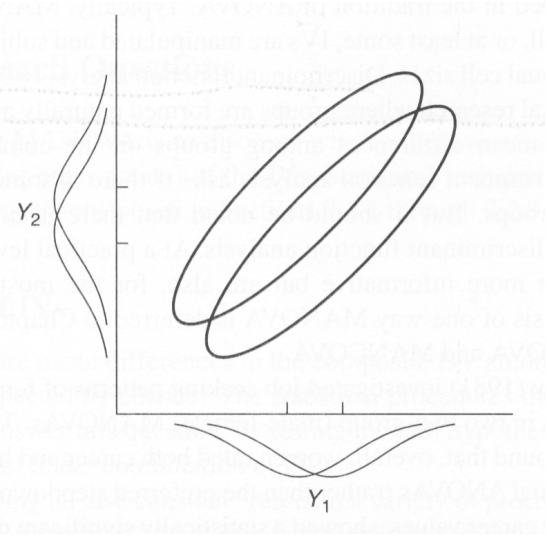
- 2 or more DVs (Interval, Ratio)
- 1 or more categorical IVs (Nominal, Ordinal)

# MANOVA advantages over ANOVA

- By measuring multiple DVs you increase your chances for finding a group difference
- With a single DV you “put all of your eggs in one basket”
- Multiple measures usually do not “cost” a great deal more and you are more likely to find a difference on at least one

# MANOVA advantages over ANOVA

- Using multiple ANOVAs inflates type 1 error rates and MANOVA helps control for the inflation
- Under certain (rare) conditions MANOVA may find differences that do not show up under ANOVA
- The more complex an analysis becomes the less power there is



# MANOVA

- A new DV is created that is a linear combination of the individual DVs that maximizes the difference between groups.
- In factorial designs a different linear combination of the DVs is created for each main effect and interaction that maximizes the group difference separately.
- Also when the IVs have more than one level the DVs can be recombined to maximize paired comparisons

# Discriminant Function Analysis

- Used to predict group membership from a set of continuous predictors
- Think of it as MANOVA in reverse – in MANOVA we asked if groups are significantly different on a set of linearly combined DVs. If this is true, than those same "DVs" can be used to predict group membership.
- How can continuous variables be linearly combined to best classify a subject into a group?

# Basics

- MANOVA and disriminant function analysis are mathematically identical but are different in terms of emphasis
  - discrim is usually concerned with actually putting people into groups (classification) and testing how well (or how poorly) subjects are classified
  - Essentially, discrim is interested in exactly how the groups are differentiated not just that they are significantly different (as in MANOVA)

# Questions

- the primary goal is to find a dimension(s) that groups differ on and create classification functions
- Can group membership be accurately predicted by a set of predictors?
  - Essentially the same question as MANOVA