

---

# Einführung in Web- und Data-Science

Prof. Dr. Ralf Möller

**Universität zu Lübeck**

**Institut für Informationssysteme**

Tanya Braun (Übungen)

# Übersicht

- Einführung, Klassifikation vs. Regression, parametrisches und nicht-parametrisches überwachtes Lernen
- Netze aus differenzierbaren Modulen („neuronal“ Netze), Support-Vektor-Maschinen
- Häufungsanalysen, Warenkorbanalyse, Empfehlungen
- Statistische Grundlagen: Stichproben, Schätzer, Verteilung, Dichte, kumulative Verteilung, Skalen: Nominal-, Ordinal-, Intervall- und Verhältnisskala, Hypothesentests, Konfidenzintervalle, Reliabilität, Interne Konsistenz, Cronbach Alpha, Trennschärfe
- Bayessche Statistik, Bayessche Netze zur Spezifikation von diskreten Verteilungen, Anfragen, Anfragebeantwortung, Lernverfahren für Bayessche Netze bei vollständigen Daten
- Induktives Lernen: Versionsraum, Informationstheorie, Entscheidungsbäume, Lernen von Regeln
- Ensemble-Methoden, Bagging, Boosting, Random Forests
- Clusterbildung, K-Means, Analyse der Variation (Analysis of Variation, ANOVA), Inter-Cluster-Variation, Intra-Cluster-Variation, F-Statistik, Bonferroni-Korrektur, MANOVA
- Analyse Sozialer Strukturen
- Deep Learning, Einbettungstechniken
- Zusammenfassung

# Clustering

---

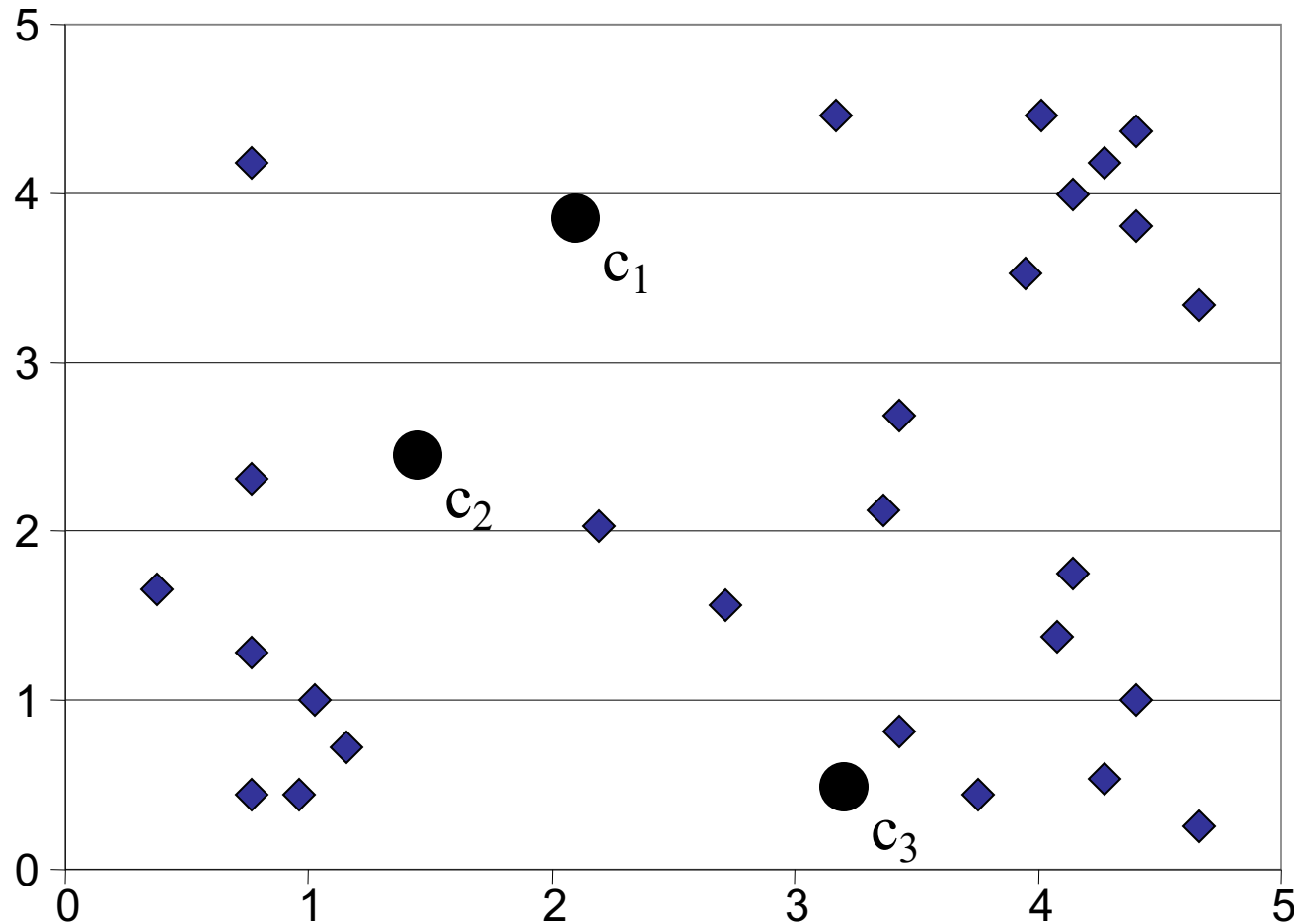
- Form des unüberwachten Lernens
- Suche nach natürlichen Gruppierungen von Objekten
  - Klassen direkt aus Daten bestimmen
    - Hohe Intra-Klassen-Ähnlichkeit
    - Kleine Inter-Klassen-Ähnlichkeit
  - Ggs.: Klassifikation
- Distanzmaße
  - z. B. Minkowski Distanz (im  $\mathbb{R}^n$ ):

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

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

# Partitionierung: K-means Clustering (1)

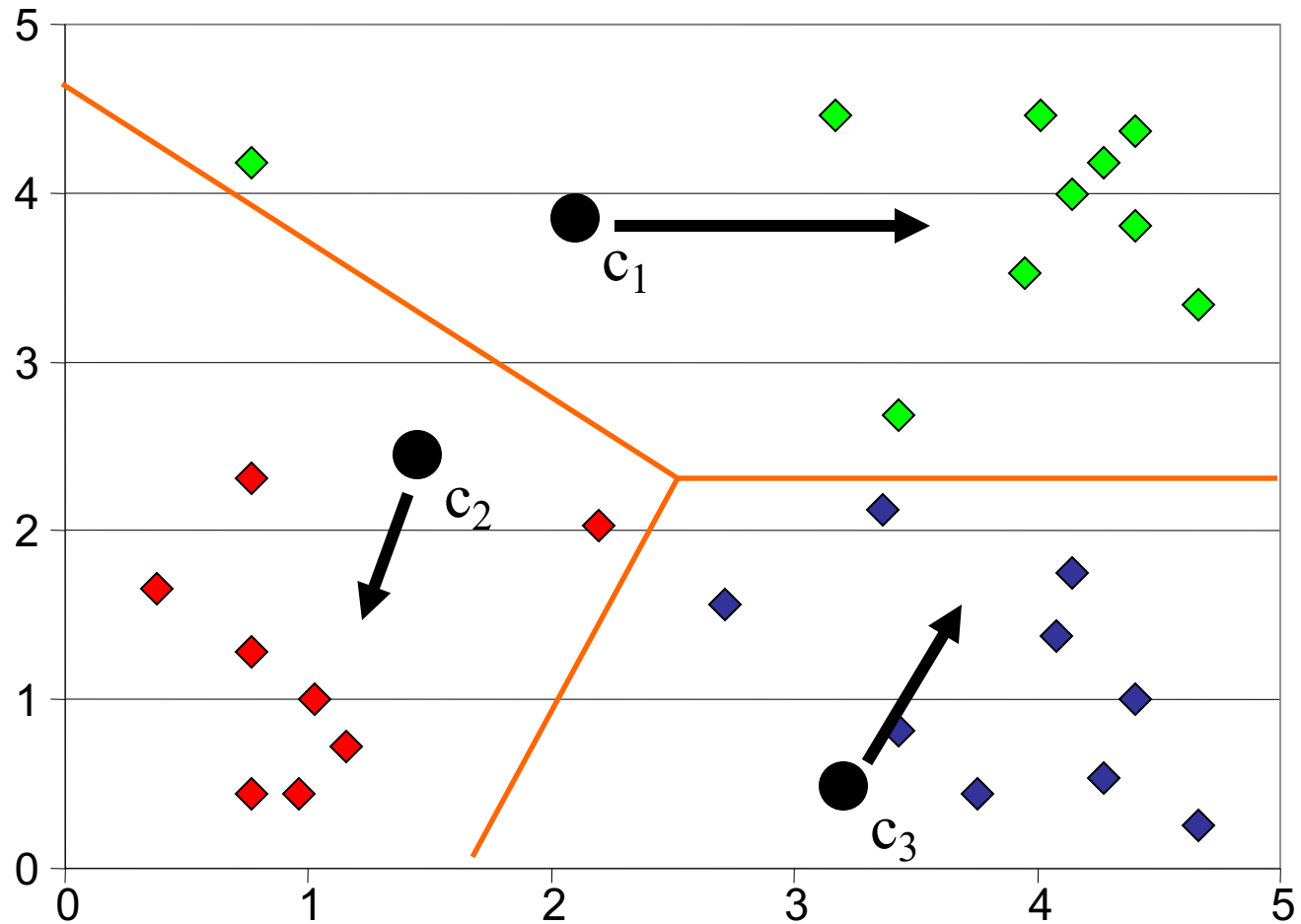
Distanzmaß: Euklidische Distanz



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

# K-means Clustering (2)

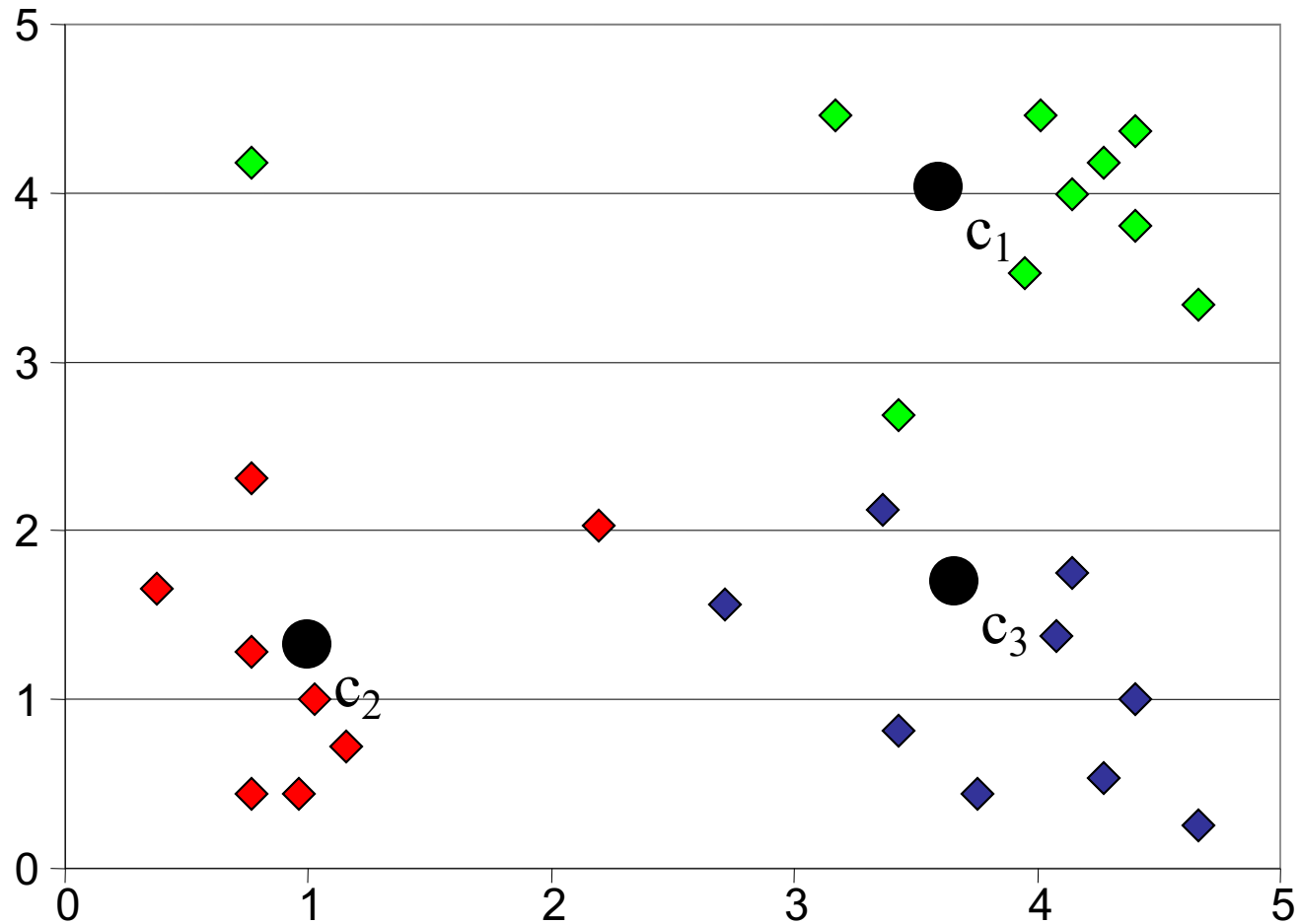
Distanzmaß: Euklidische Distanz



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

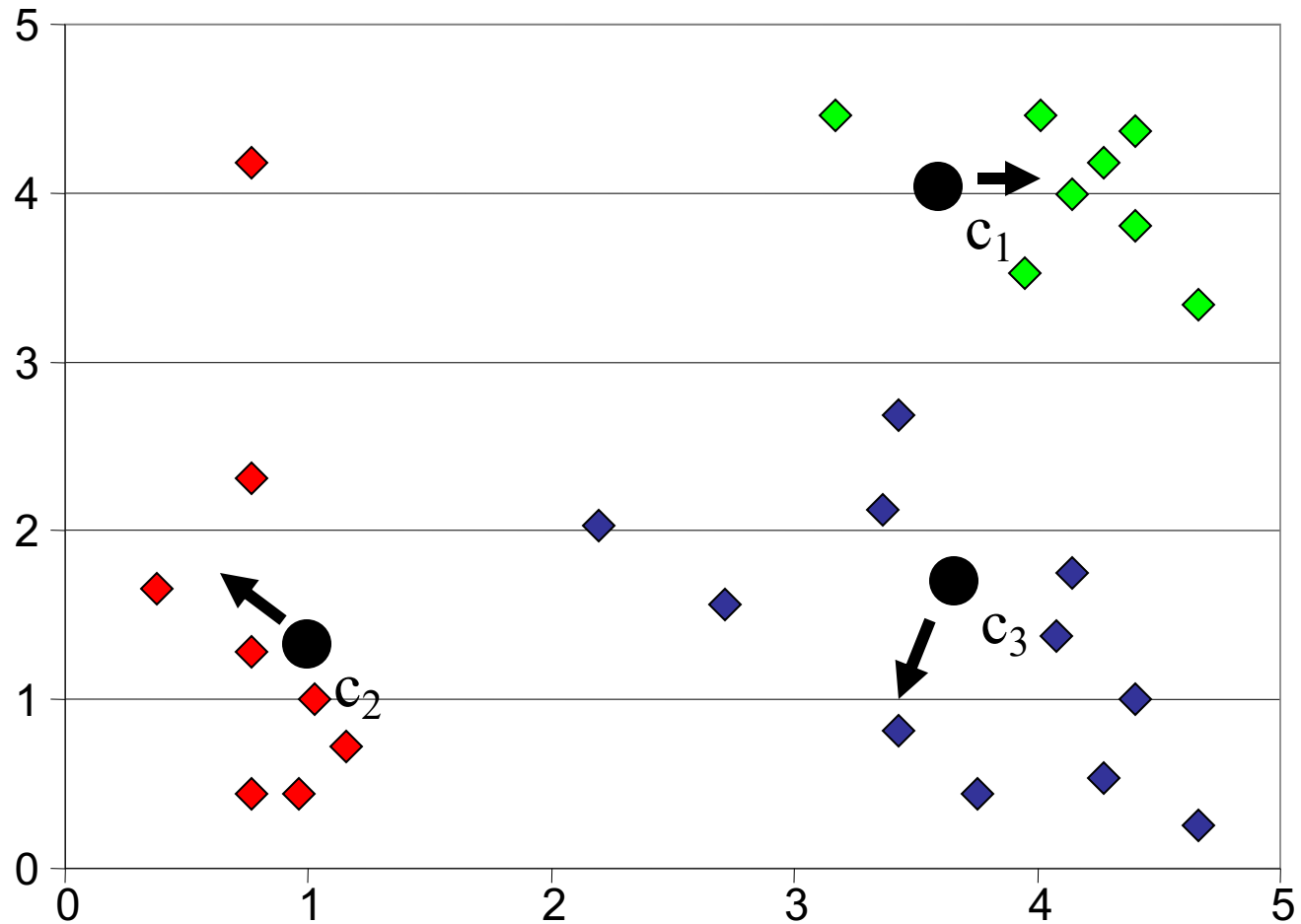
# K-means Clustering (3)

Distanzmaß: Euklidische Distanz



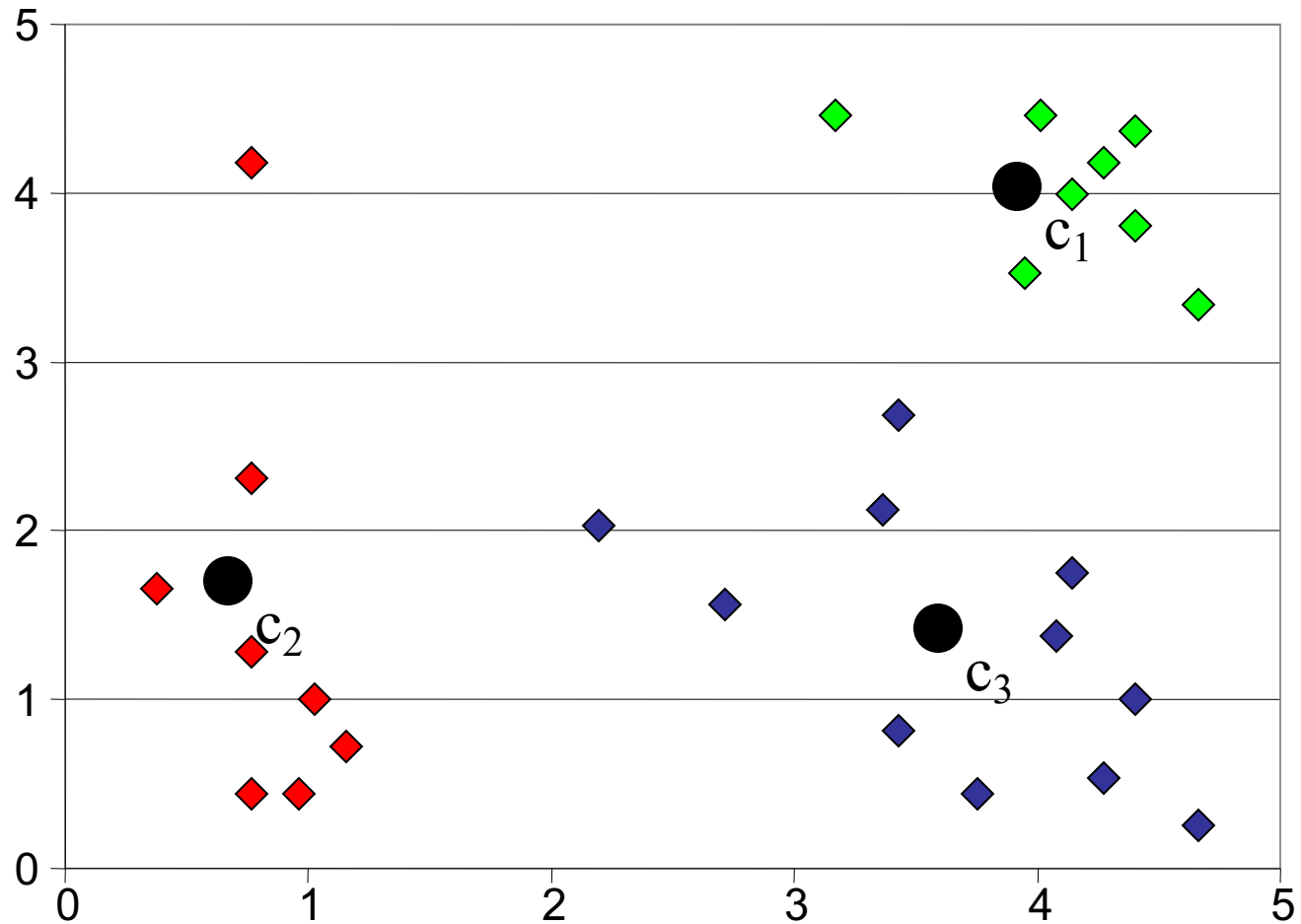
# K-means Clustering (4)

Distanzmaß: Euklidische Distanz



# K-means Clustering (5)

Distanzmaß: Euklidische Distanz



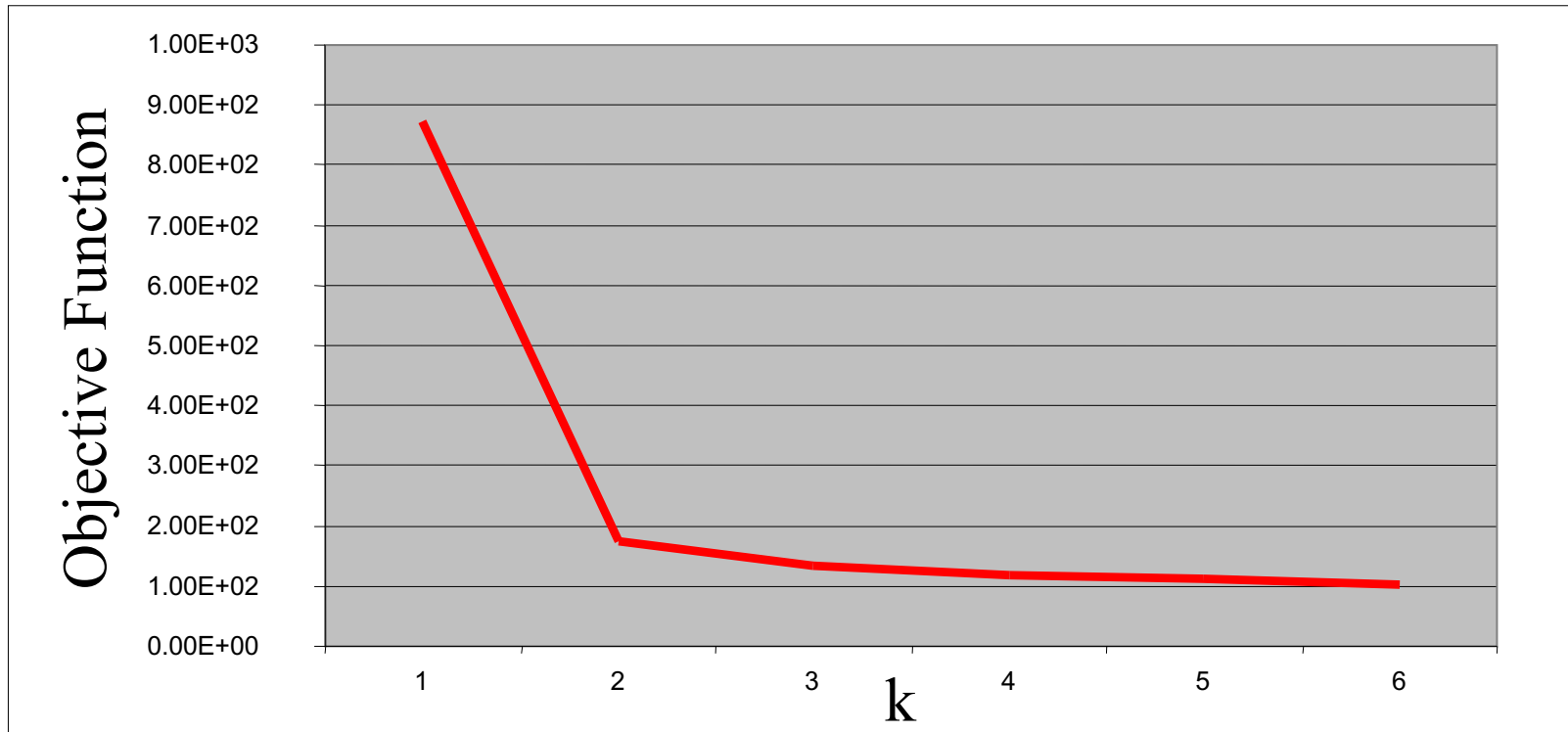


# Wann ist eine Gruppierung gut?

---

- Ideen für Bewertungsmaß (objective function)
  - Hohe Intra-Klassen-Ähnlichkeit
  - Kleine Inter-Klassen-Ähnlichkeit
- Formalisierung
  - Intra-Cluster-Varianz kleiner als Inter-Cluster-Varianz

# Was ist die richtige Clusteranzahl?



Variiere  $k$  und finde Knick in Graph der Bewertungsfunktion (Ellenbogen)

# Und wenn die Cluster schon gegeben sind?

Beispiel: 25 Patienten mit Blasen auf der Haut

Behandlung: Methode A, Methode B, Placebo

Messwerte: # der Tage bis zur Abheilung der Blasen

Daten aus Studie [und Mittelwerte]:

- 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]

Können wir sagen, dass Methode A die beste ist?

Sind die Differenzen der Mittelwerte signifikant?

Variation ZWISCHEN Gruppen vs. Variation IN Gruppen (clusters)

Analysis of variation notwendig: ANOVA

# Was macht ANOVA?

---

In der einfachen Form (es gibt viele Erweiterungen)  
testet ANOVA folgende Hypothese:

$H_0$ : Die Mittelwerte sind gleich (unterscheiden sich nicht)

$H_a$ : Nicht alle Mittelwerte sind gleich,  
der **Unterschied** ist signifikant

- Sagt nichts darüber, welche sich unterscheiden
- Muss durch multiple Vergleiche später herausgefunden werden

**Unterschiedshypothese**

# Unterschiedshypothesen

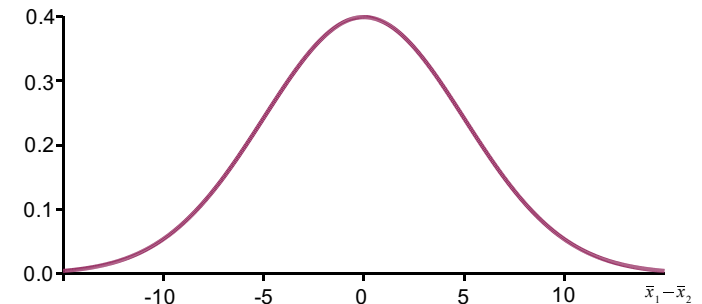
---

- Sind Frauen ängstlicher als Männer?
  - Unterscheiden sich die Mittelwerte von zwei Gruppen?
  - Unabhängige Stichproben
- Ist der Mittelwert der Ängstlichkeit nach einer Therapie größer als vor der Therapie?
  - Unterscheidet sich der Mittelwert einer Stichprobe zu zwei Messzeitpunkten?
  - Abhängige Stichproben
- Liegt der mittlere IQ einer Gruppe über 100?
  - Unterscheidet sich der Mittelwert einer Gruppe von einem vorgegeben Wert?
  - Test bzgl. Gruppe

# Unterschiedshypothesen: Unabhängige Stichproben

Unterscheiden sich die Mittelwerte von zwei Gruppen?

- Differenz der Mittelwerte zweier Stichproben:  $\Delta_x = \bar{x}_1 - \bar{x}_2$
- Schätze die Dichtefunktion für  $\Delta_x$  wenn  $H_0$  wahr ist
- Wenn  $p < \alpha$ , wird  $H_0$  verworfen und  $H_1$  angenommen
- Stichprobenkennwerteverteilung:  
Verteilung der Mittelwertsdifferenzen unter  $H_0$
- Wie verteilen sich empirische Mittelwertsdifferenzen, wenn man sehr oft Stichproben zieht?
- Verteilung von Mittelwertsdifferenzen bei großen Stichproben normalverteilt



# Wiederholung: Standardfehler

---

Der Standardfehler des arithmetischen Mittels ist gleich

$$\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}},$$

wobei  $\sigma$  die Standardabweichung einer einzelnen Messung bezeichnet.

# Hier: Standardfehler der Mittelwertsdifferenzen

---

- Hängt von den Varianzen und den Größen der beiden Teilstichproben ab:

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}$$

- Benötigt, um gefundene Mittelwertsdifferenz interpretieren zu können



# t-Verteilung

- Empirische (gefundene) Mittelwertsdifferenz durch Standardfehler dividiert ergibt sog. **t-Verteilung**

$$t_{df} = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}}$$

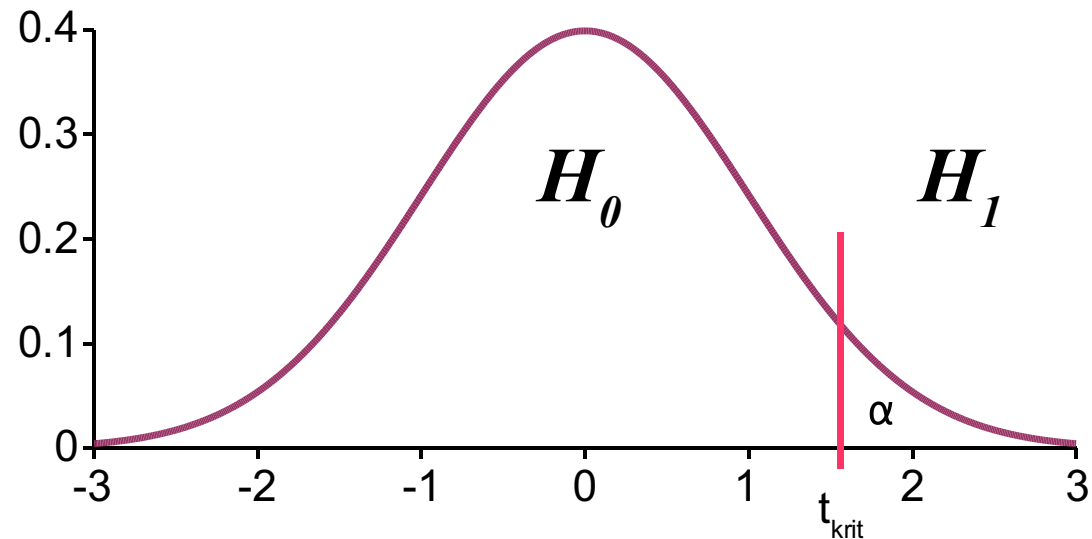
- Die genaue Form der t-Verteilung hängt von deren Freiheitsgraden ( $df = \text{degree of freedom}$ ) ab

$$df = N_1 + N_2 - 2$$

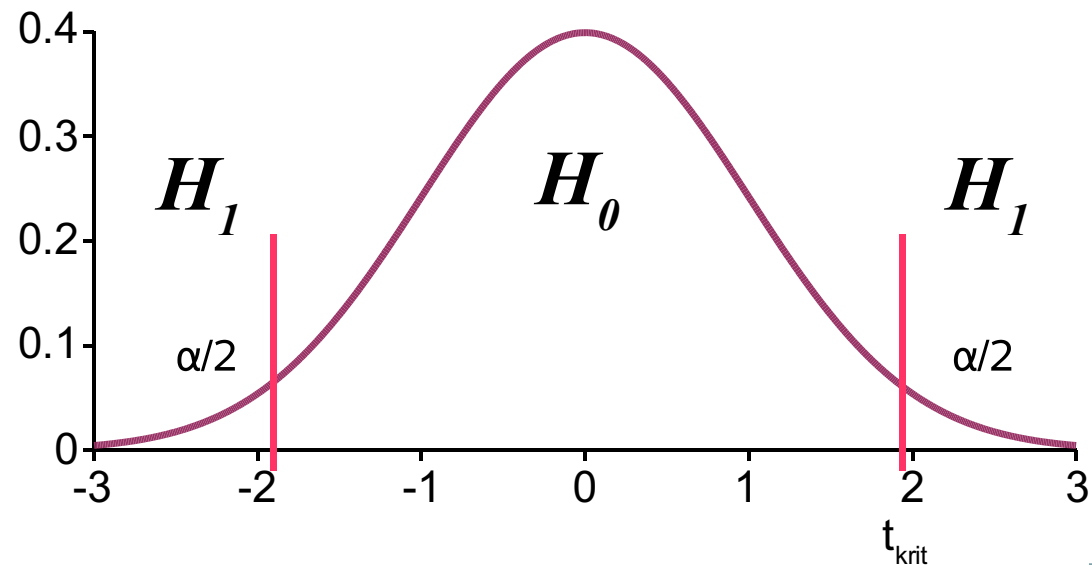
- Bei  $df > 120$  nahezu identisch mit z-Verteilung (St.Norm.V.)
  - Je kleiner  $df$ , desto schmalgipfliger die t-Verteilung
- Die Herleitung der Dichtefunktion und der kumulativen Funktion erfolgt später

# Der $t$ -Test für unabhängige Stichproben

einseitiger Test  
(gerichtete  $H_0$ )



zweiseitiger Test  
(ungerichtete  $H_0$ )



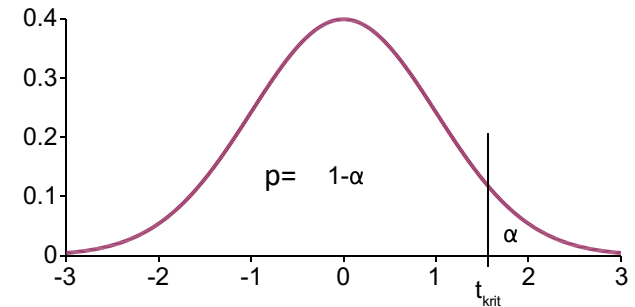
# Entscheidung über die Nullhypothese

---

- Mittels einer  $t$ -Tabelle wird der empirische (gefundene)  $t$ -Wert interpretiert
- Dazu wird ein kritischer Wert aus der  $t$ -Tabelle entnommen
  - Der kritische  $t$ -Wert hängt dabei ab:
    - von den Freiheitsgraden,
    - von dem gewählten  $\alpha$ -Niveau
    - von der Art des Tests (einseitig vs. zweiseitig)
  - Der kritische  $t$ -Wert definiert die Grenze des Bereichs für den empirischen  $t$ -Wert, ab dem  $H_0$  verworfen wird

# Die t-Verteilung

df	p=.800	p=.900	p=.950	p=.975	p=.990	p=.995
1	1,376	3,078	6,314	12,706	31,821	63,657
2	1,061	2,920	2,920	4,303	6,965	9,925
3	0,978	2,353	2,353	3,182	4,541	5,841
4	0,941	2,132	2,132	2,776	3,747	4,604
5	0,920	2,015	2,015	2,571	3,365	4,032
6	0,906	1,943	1,943	2,447	3,143	3,707
7	0,896	1,895	1,895	2,365	2,998	3,499
8	0,889	1,860	1,860	2,306	2,896	3,355
9	0,883	1,833	1,833	2,262	2,821	3,250
10	0,879	1,812	1,812	2,228	2,764	3,169
20	0,860	1,725	1,725	2,086	2,528	2,845
30	0,854	1,697	1,697	2,042	2,457	2,750
40	0,851	1,684	1,684	2,021	2,423	2,704
50	0,849	1,676	1,676	2,009	2,403	2,678
60	0,848	1,671	1,671	2,000	2,390	2,660
70	0,847	1,667	1,667	1,994	2,381	2,648
80	0,846	1,664	1,664	1,990	2,374	2,639
90	0,846	1,662	1,662	1,987	2,368	2,632
100	0,845	1,660	1,660	1,984	2,364	2,626
200	0,843	1,653	1,653	1,972	2,345	2,601
1000	0,842	1,646	1,646	1,962	2,330	2,581



Kritische t-Werte:

$\alpha = .05$ , einseitig,  $df=100$ :

$$t_{krit}(100) = 1.66$$

$\alpha = .05$ , zweiseitig,  $df=100$ :

$$t_{krit}(100) = 1.98$$

$\alpha = .01$ , einseitig,  $df=100$ :

$$t_{krit}(100) = 2.36$$

# Der $t$ -Test für unabhängige Stichproben

## Entscheidungsregeln

- Einseitiger Test:
  - Wenn  $t_{emp} > t_{krit}$  wird  $H_0$  verworfen
- Zweiseitiger Test
  - Wenn  $|t_{emp}| > t_{krit}$  wird  $H_0$  verworfen
- In der  $t$ -Tabelle werden immer Werte für den einseitigen Test angegeben.
- Für einen 2-seitigen Test muss  $t_{krit}$  so gewählt werden, dass ein Bereich von  $\alpha/2$  „von der Verteilung abgeschnitten wird“

# Voraussetzungen t-Test

---

- (1) Variable besitzt Intervallskala (arithm. Mittel ist definiert)
- (2) Normalverteilung des Merkmals in der Grundgesamtheit
- (3) „Varianzhomogenität“
  - „Gleiche“ Varianzen des Merkmals in beiden Populationen
  - „Varianz der Varianz“ klein
- (4) Unabhängigkeit der Stichproben

# Back to ANOVA: The basic 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

# Assumptions of ANOVA

---

- Each group approximately normally distributed
  - Check this by looking at histograms or use assumptions
  - Can sensibly handle some non-normality, but not severe discrepancies
- Standard deviations of each group 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
- $1.458 \times 2 = 2.916 > 1.764$

# 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

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

N: Normalization value  
(corrected: degrees of freedom)

- Variation WITHIN groups (**MSE**)  
for each data value  $x_{ij}$  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:

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

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 = 1)

- 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				
<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>
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:  $I-1$

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

number of data values -  
number of groups:  $n-I$   
(equals df for each group  
added together)

# Computing ANOVA F statistic

			WITHIN		BETWEEN	
			difference:		difference	
		group	data - group mean		group mean - overall mean	
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

$$= \text{MSG} / \text{MSE}$$

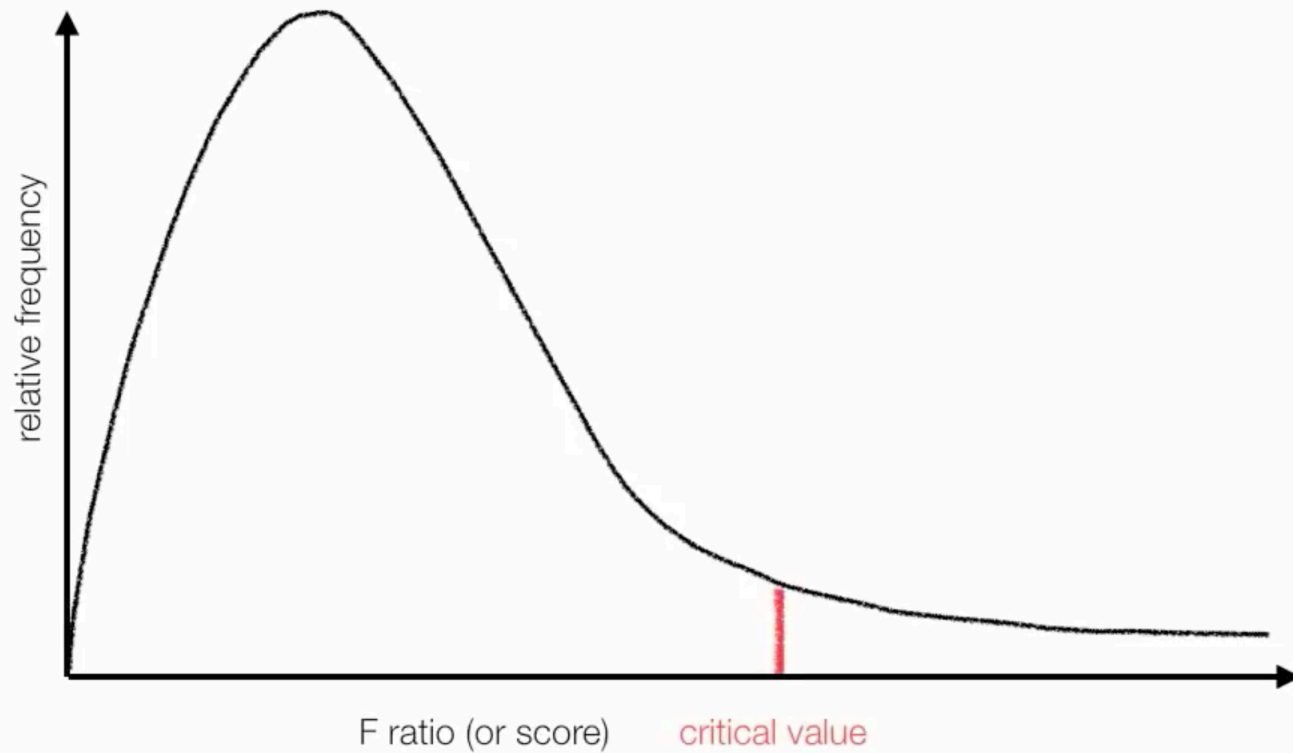
A large value of F indicates relatively more difference between groups than within groups  
(evidence against  $H_0$ )

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

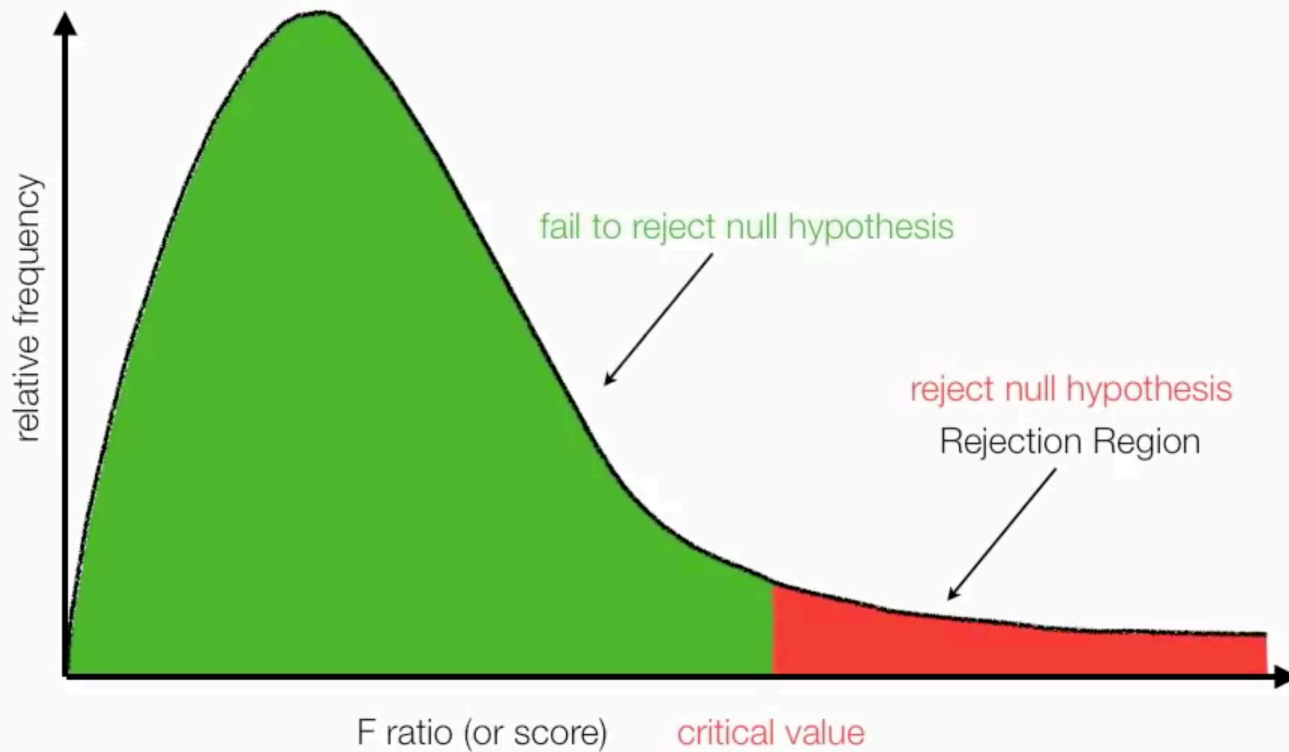
- $l-1$  degrees of freedom in numerator (# groups -1)
- $n - l$  degrees of freedom in denominator (rest of df)



# F-Distribution



# Critical Value



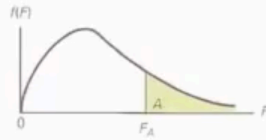
# F-Table

$\alpha = 0.05$  (use another table for different  $\alpha$ )

Computed F-Value = 10.21

Critical Value  $F(2, 9) = 4.26$

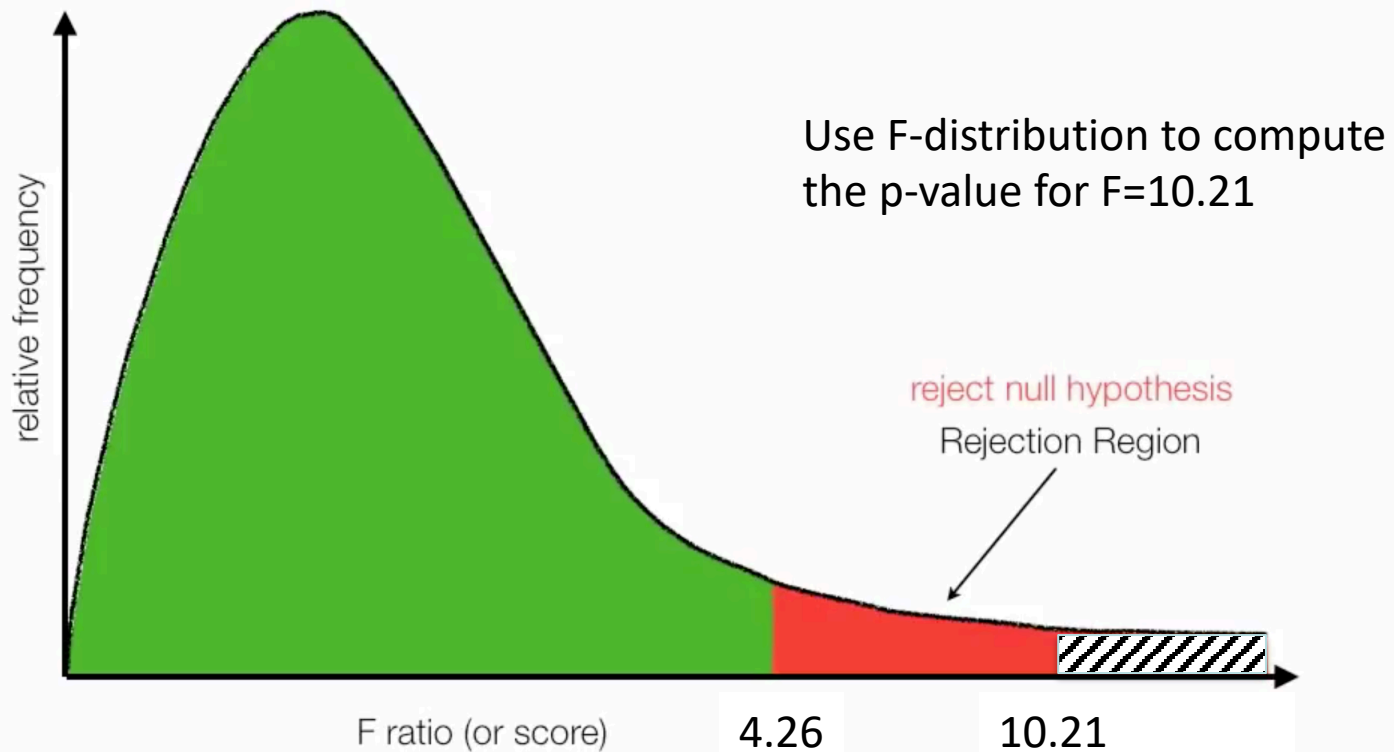
Table 6(a) Critical Values of  $F$ :  $\alpha = .05$



relates to groups or samples

$\nu_2$	$\nu_1$	NUMERATOR DEGREES OF FREEDOM								
		1	2	3	4	5	6	7	8	9
relates to number of observations DENOMINATOR DEGREES OF FREEDOM	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	3.38	2.99	2.76	2.60	2.49	2.40	2.34	2.28

# Rejection of Null Hypothesis



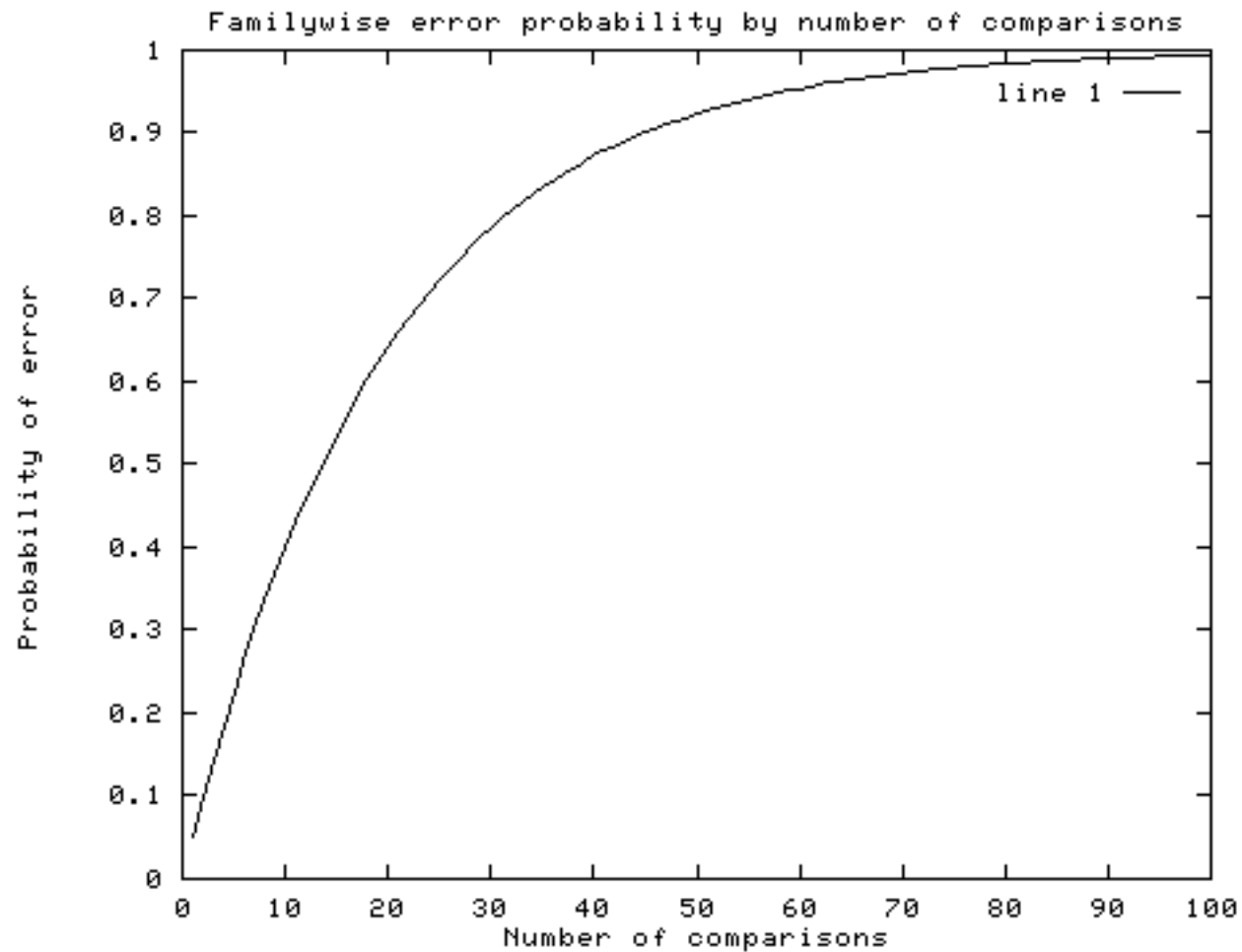
# Why not just do 3 pairwise t-tests?

---

## 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 critical value	Original Alpha	# tests	New Alpha
.001	.05	5	.010
.011	.05	4	.013
.019	.05	3	.017
.032	.05	2	.025
.048	.05	1	.050



# 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