UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

R course for beginners

Session 2: Statistics

Based on R Lecture by Juan Luis Mateo, COS

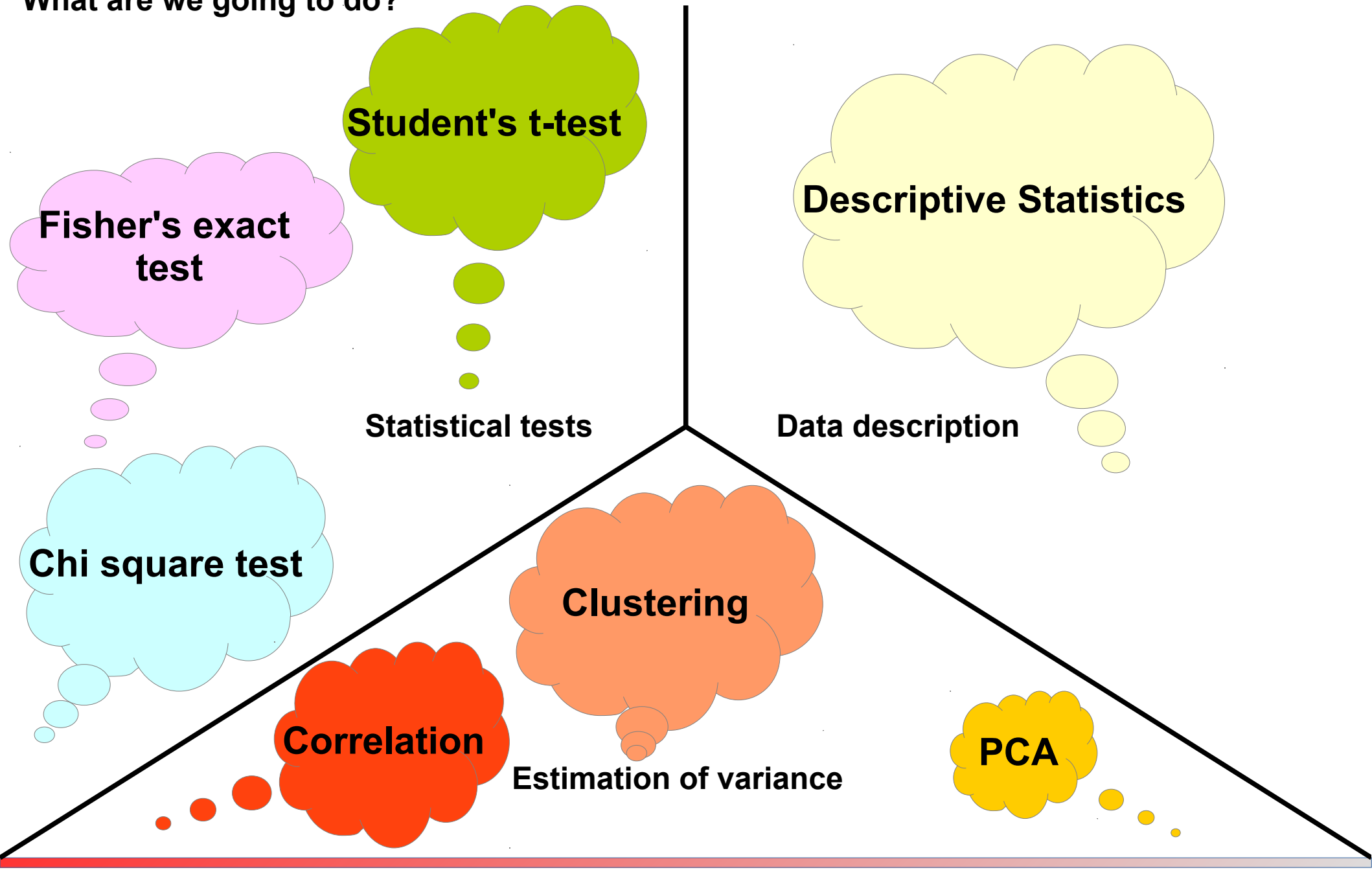
Christine Gläßer ---- ZMBH ---- Room 504 ---- +49(0)6221-54 6824 ---- c.glaesser@zmbh.uni-heidelberg.de

Session 1 – Recap commands

Working directory	I/O	Check data	Create data	Mathematical operations
getwd()	read.delim()	data[row,column]	rbind()	+
setwd()	write.table()	colnames()	<-	-
dir()		rownames()	c()	*
		length()	1:10	/
			seq()	^
			rep()	sum()
			array()	mean()
				sd()

What are we going to do?

What are we going to do?



Data description

Descriptive Statistics

1) Measures of centrality

- Mean: estimate of the mean value of a variable in your sample

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Median: value separating the higher half of your data from the lower half

- Quantiles: value separating x% data from the rest
  ----> the median is also the 2-quantile
  ----> in most cases, 75% and 25% are of interest

Data description

1) Measur

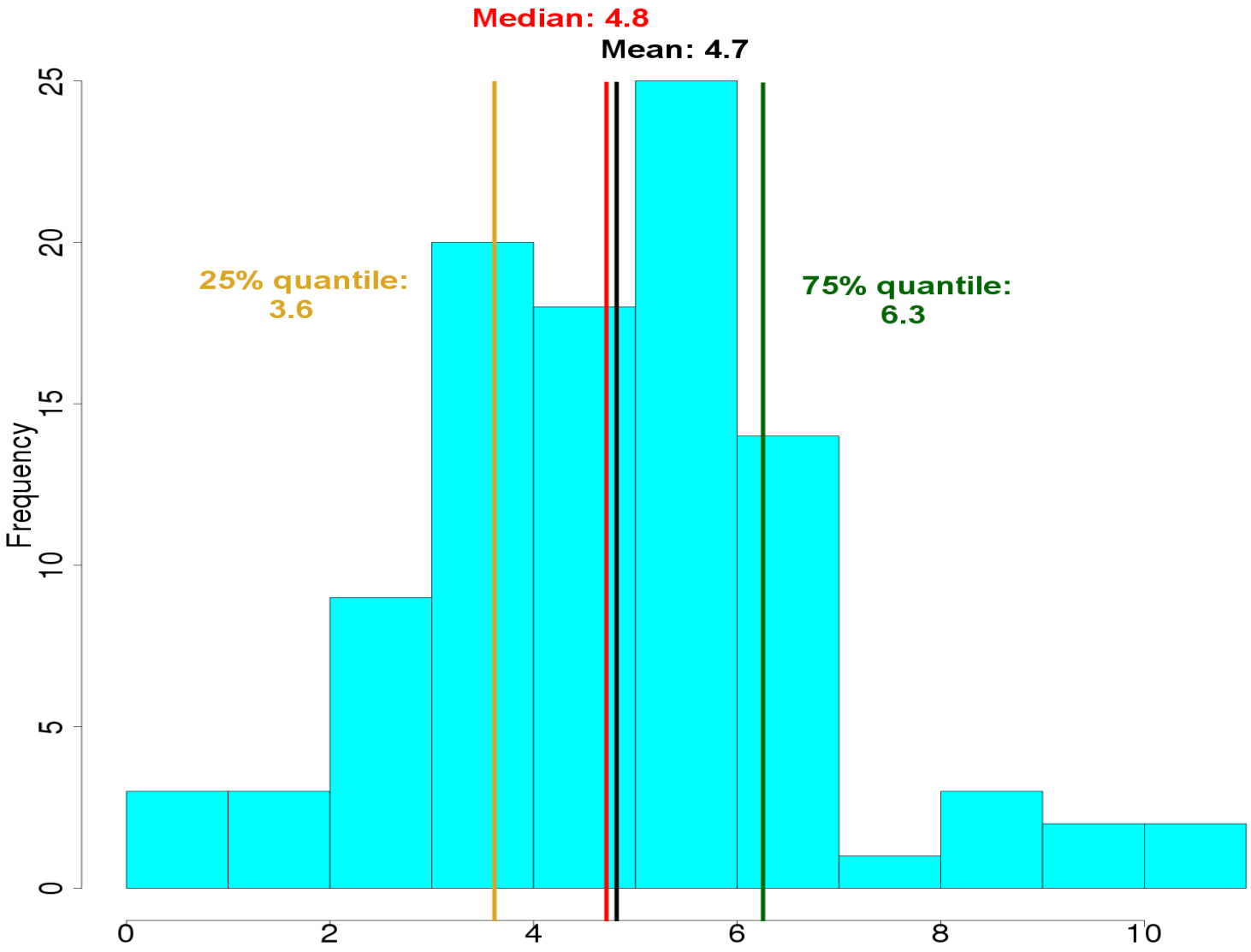
ive Statistics

- Mean: es

- Median:

er half

- Quantiles
  ----> the
  ----> in



Data description

2) Measures of spread

Descriptive Statistics

- Range: difference between minimum and maximum value in your data

- Variance: showing how far values are from the mean value

----> standard deviation as equivalent measure, square root of variance

$$s = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2}{n-1}}$$

Data description – descriptive statistics

Commands: mean, median, min, max, median, quantile, sd, range

Mean/Sd: cp. Session 1
Median: computes the sample median
Min: returns minima of input values
Max: returns maxima of input values
Quantile: calculating the quantiles

OR: use one of various summary commands!

Command: summary

Data description – descriptive statistics

1) Load data „sleep_data_simple.txt“

Remember from session 1: where's your data stored? Direct R to that folder, then load data

2) Describe your data: mean,median,25th and 75th quartiles,min,max

Data description – descriptive statistics

1) Load data „sleep_data_simple.txt“

Remember from session 1: where's your data stored? Direct R to that folder, then load data

2) Describe your data: mean,median,25th and 75th quartiles,min,max

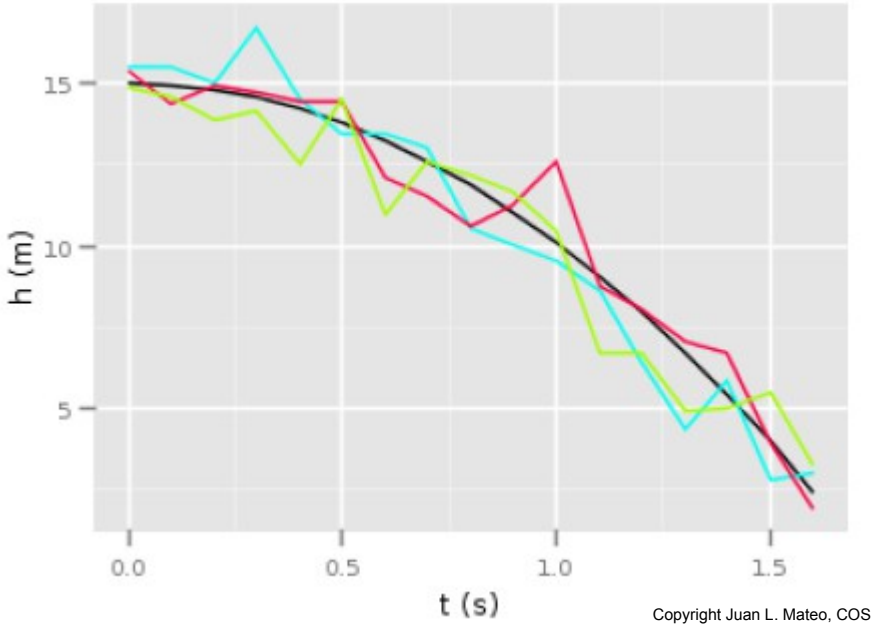
```
> sleep<-read.delim("sleep_data_simple.txt")
> summary(sleep)
 X8.hours.sleep.group..X. X4.hours.sleep.group..Y.
 Min.   :3.0              Min.   :1.00
 1st Qu.:3.0              1st Qu.:1.75
 Median :5.0              Median :4.00
 Mean   :5.0              Mean   :4.00
 3rd Qu.:5.5              3rd Qu.:6.00
 Max.   :9.0              Max.   :8.00
>
```

Statistical tests – Recap

In theory, we don't have noise and events follow a precise law

(e.g. free fall: $h = h_0 - \frac{1}{2} g t^2$)

In reality, measurements are not precise



Averaging to get rid of noise, smooting data

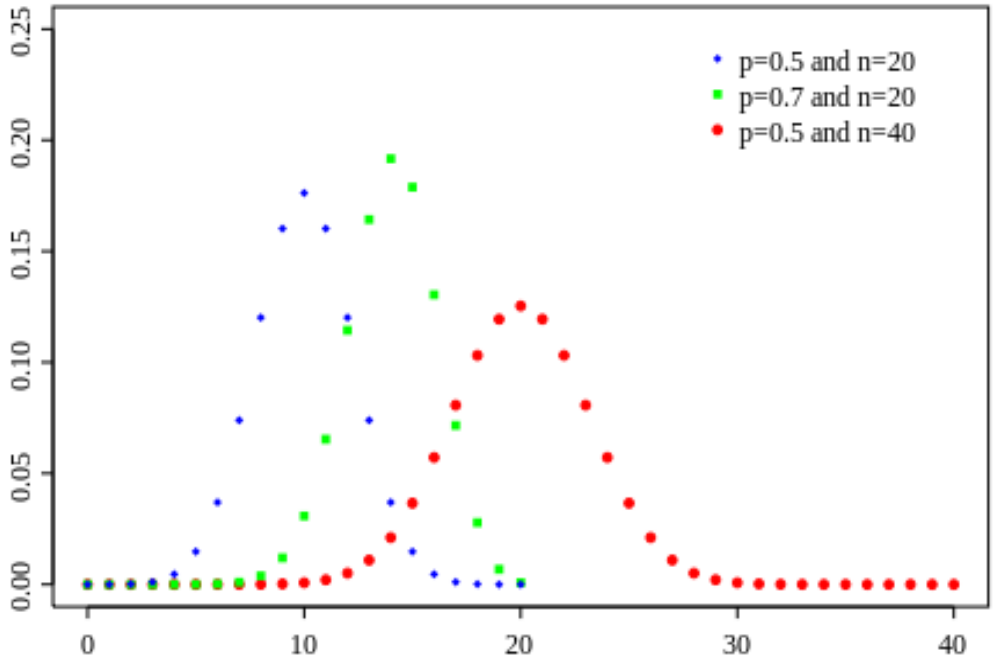
----> idea of statistics

----> the more data, the better

Statistical tests – Recap

Model probability for specific types of events

a) Binomial distribution: repetition with binary outcome



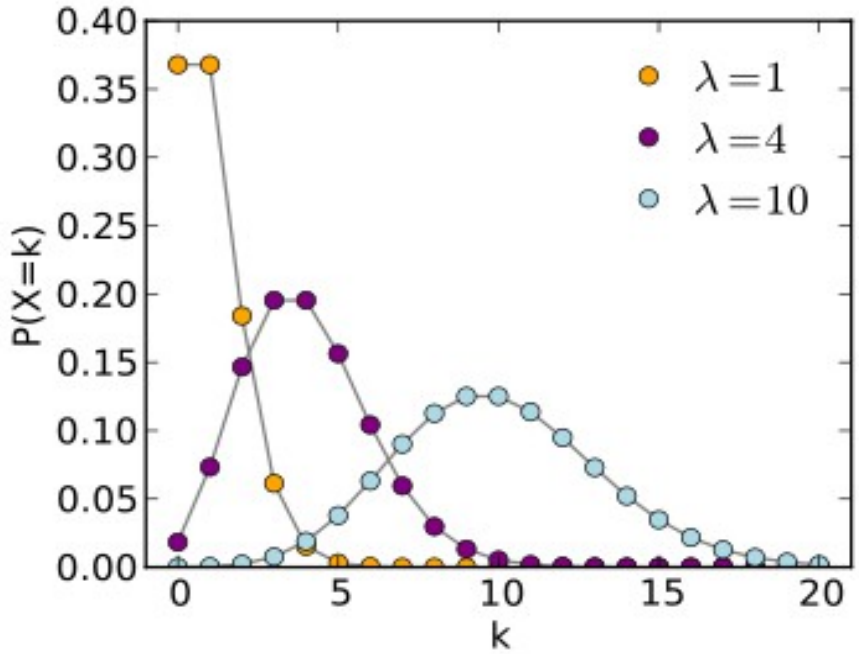
http://en.wikipedia.org/wiki/Binomial_distribution

Statistical tests – Recap

Statistics?

Model probability for specific types of events

b) Poisson distribution: probability of a given number of events in a defined amount of time, we know the average

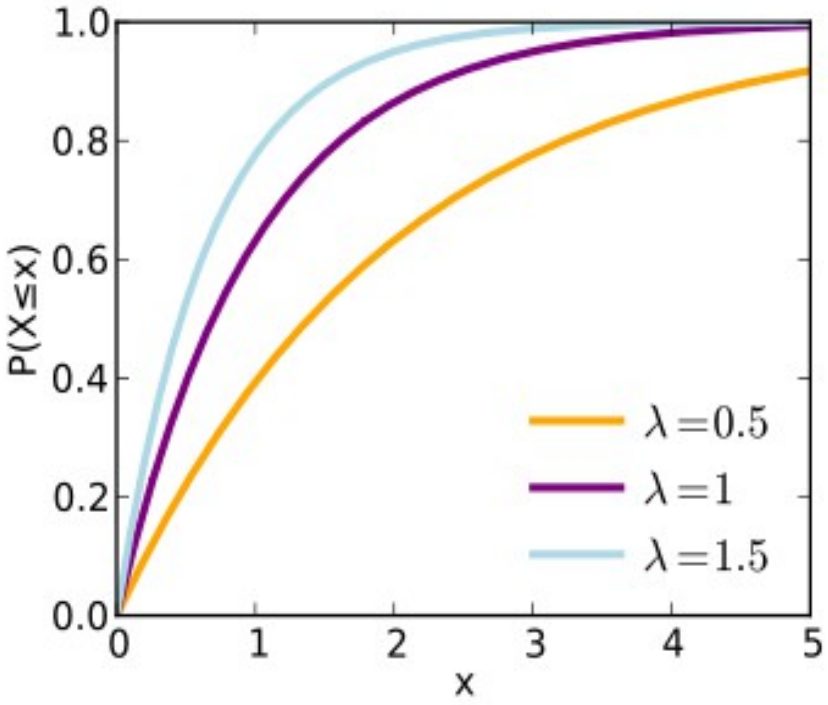


http://en.wikipedia.org/wiki/Poisson_distribution

Statistical tests – Recap

Model probability for specific types of events

c) Exponential distribution: processes with exponential behaviour

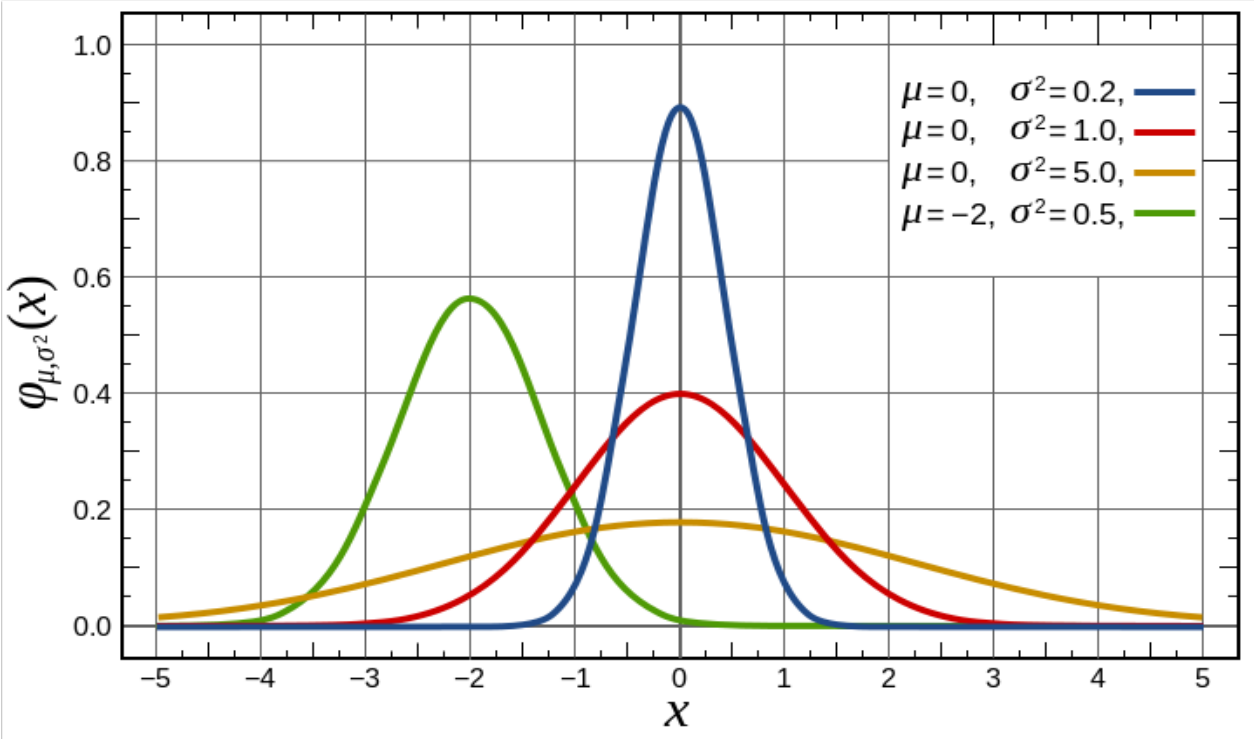


http://en.wikipedia.org/wiki/Exponential_distribution

Statistical tests – Recap

Model probability for specific types of events

d) Normal distribution: probability of an event falling far from the expected value



http://en.wikipedia.org/wiki/Normal_distribution

Statistical tests – Recap

Hypothesis testing: test if our data follows that distribution

- Null hypothesis H0: statement we want to test

---> Standard: two things are comparable

- Alternative hypothesis: H0 is false

- Result: probability

Statistical tests – Recap

Probability: measure of uncertainty

----> p-value

„Gold standard“: p-value of 0.05, meaning 95% confidence that your observation is significant

Statistical tests – Recap

HOWEVER...

False positives!

Imagine: 10,000 genes

Thanks to Simon Anders, EMBL

None is differentially expressed, but you think there are some

Assume a p-value of 0.05

Statistical tests – Recap

HOWEVER...

False positives!

Imagine: 10,000 genes

Thanks to Simon Anders, EMBL

None is differentially expressed, but you think there are some

Assume a p-value of 0.05

P-value definition: result is assigned value p, then probability of seeing a result this strong only due to noise is p-value

Statistical tests – Recap

HOWEVER...

False positives!

Imagine: 10,000 genes

Thanks to Simon Anders, EMBL

None is differentially expressed, but you think there are some

Assume a p-value of 0.05

---> 5% of genes will have p-value <0.05 (500 genes!)

Statistical tests – Recap

HOWEVER...

False positives!

Imagine: 10,000 genes

Thanks to Simon Anders, EMBL

None is differentially expressed, but you think there are some

Assume a p-value of 0.05

---> assume 1000 genes have p-value <0.05; those contain 500 false positives (50%!)

Statistical tests – Recap

HOWEVER...

False positives!

Imagine: 10,000 genes

Thanks to Simon Anders, EMBL

None is differentially expressed, but you think there are some

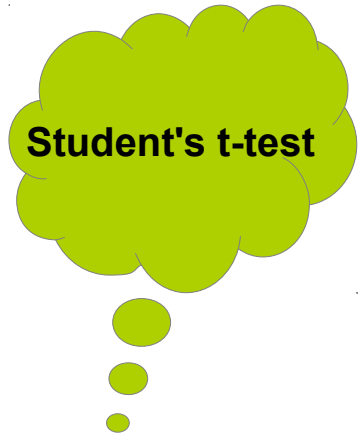
Assume a p-value of 0.05

---> techniques to adjust p-value
---> Benjamini-Hochberg most common, adjusts 0.05 raw to 0.5

Statistical tests – Recap

- Focus on continuous samples
- Parametric tests: tests require assumptions about data distribution
- Non parametric tests: tests do not require assumptions about data distribution

Statistical tests – Student's t-test

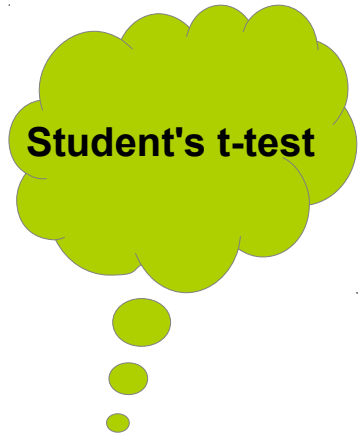


- Parametric test
- Data is normally distributed
- Either one sample: H0: mean has a specific value

  or two samples: H0: samples have equal mean values

  Additional assumption: variances are equal
  ---> if not: Welch's t-test

Statistical tests – Student's t-test

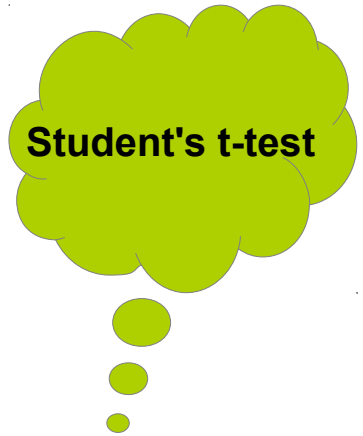


- Parametric test
- Data is normally distributed
- Either one sample: H0: mean has a specific value

  or two samples: H0: samples have equal mean values

  Additional assumption: variances are equal
  ---> if not: Welch's t-test
- Paired tests (e.g. same proband, different arms) give more statistical power; paired t-test possible

Statistical tests – Student's t-test



Q: is there a significant difference between group X and Y?

Do a t-test with sleep data X and Y

Command: t.test

```
> sleep <- read.delim("sleep_data_simple.txt")
> sleep
  X8.hours.sleep.group..X. X4.hours.sleep.group..Y.
1                        5                        8
2                        7                        1
3                        5                        4
4                        3                        6
5                        5                        6
6                        3                        4
7                        3                        1
8                        9                        2
> t.test(sleep[,1], sleep[,2])

        Welch Two Sample t-test

data:  sleep[, 1] and sleep[, 2]
t = 0.8473, df = 13.563, p-value = 0.4115
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.538936  3.538936
sample estimates:
mean of x mean of y
        5         4

>
```

Statistical tests – Chi square test

- Nominal data
- H0: frequencies of values of our samples are independent
- Samples are sufficiently large

Command: chisq.test

```
> chisq.test(sleep)

        Pearson's Chi-squared test

data:  sleep
X-squared = 11.2416, df = 7, p-value = 0.1284

Warnmeldung:
In chisq.test(sleep) : Chi-Quadrat-Approximation kann inkorrekt sein
>
```

Error message refers to small samples!

Statistical tests – Fisher's exact test

- Equivalent to Chi square test, but with ...
- … small samples

Fisher's exact test

Command: fisher.test

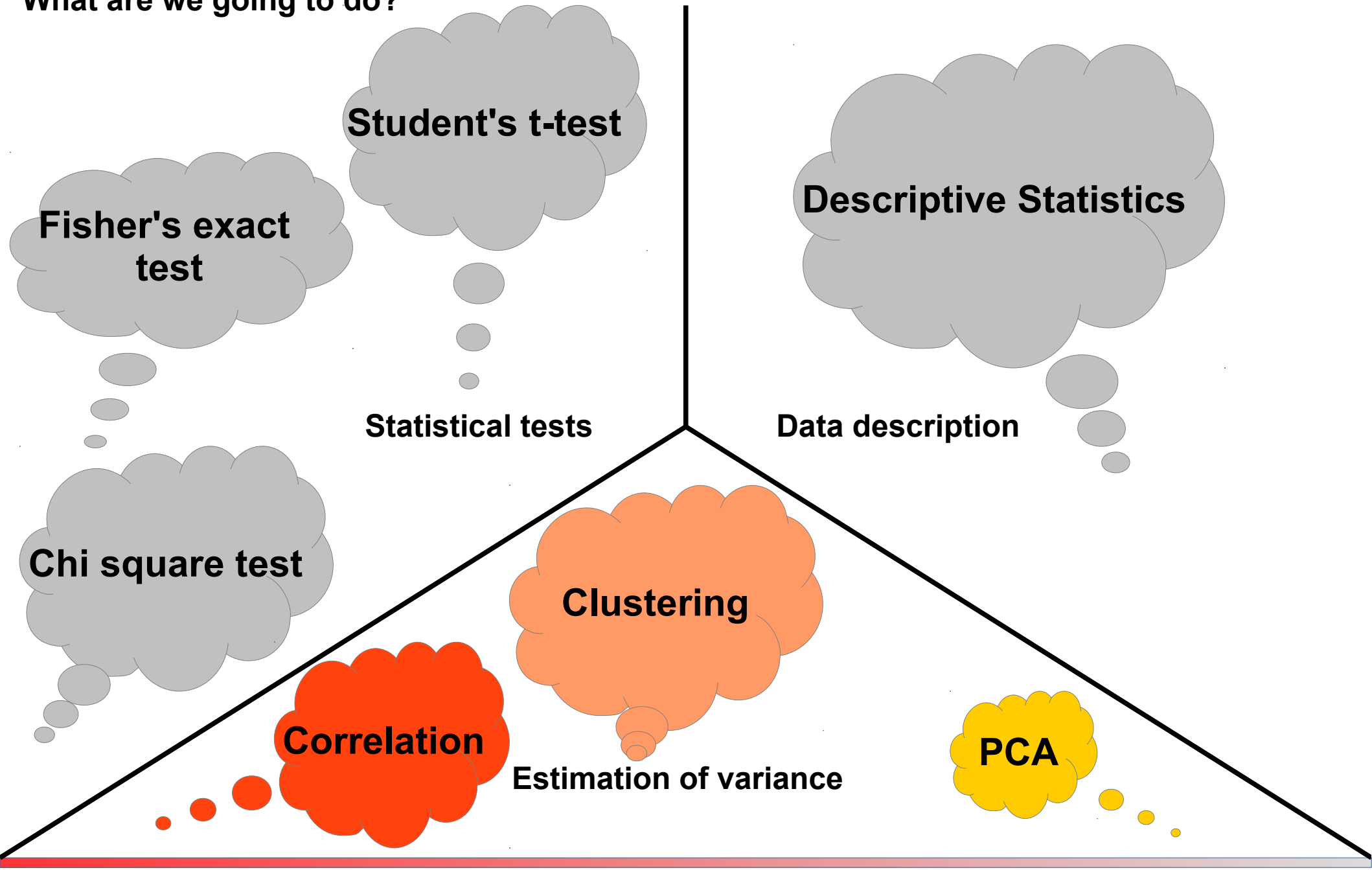
```
> fisher.test(sleep)

        Fisher's Exact Test for Count Data

data:  sleep
p-value = 0.1279
alternative hypothesis: two.sided
```

Two-sided: both directions are considered equally likely

What are we going to do?



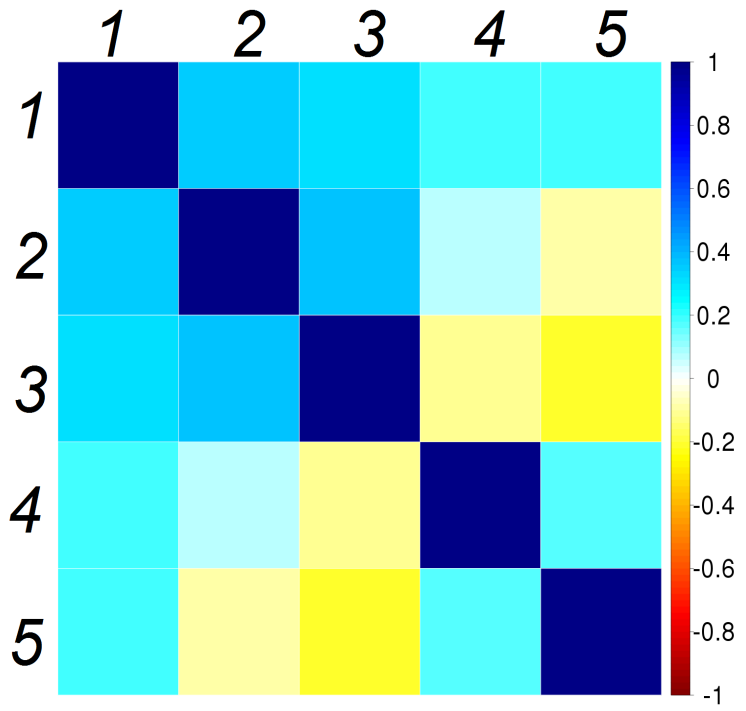
Estimation of variance – Recap

Variance?

Correlation: Statistical relationships involving dependence

- Pearson's correlation coefficient works for linear relationships
- Spearmans' rank correlation coefficient for non-linear relationships

- anti-correlation: negative values
- correlation: positive values

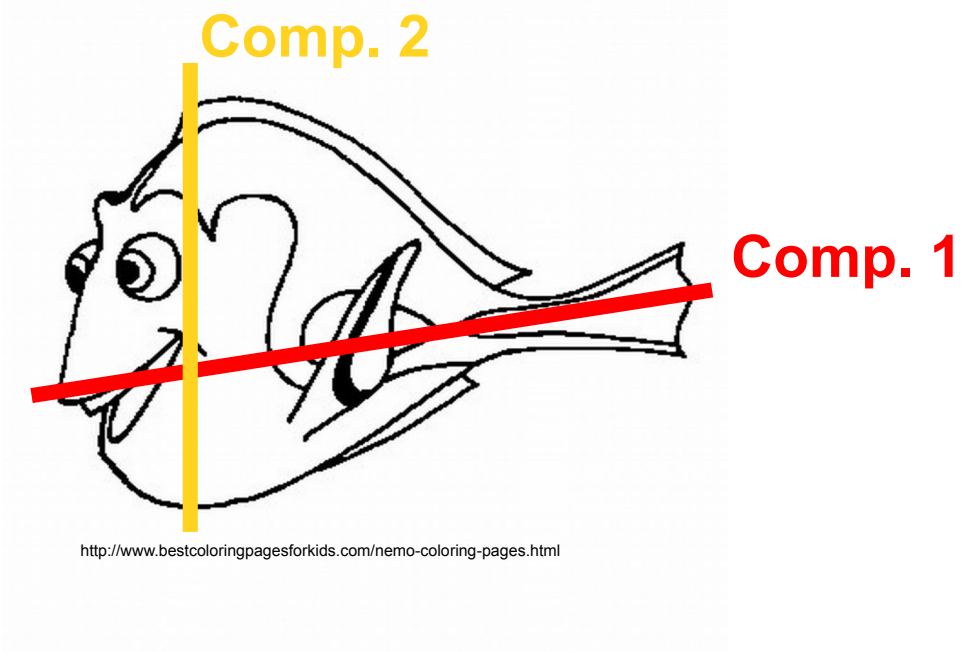


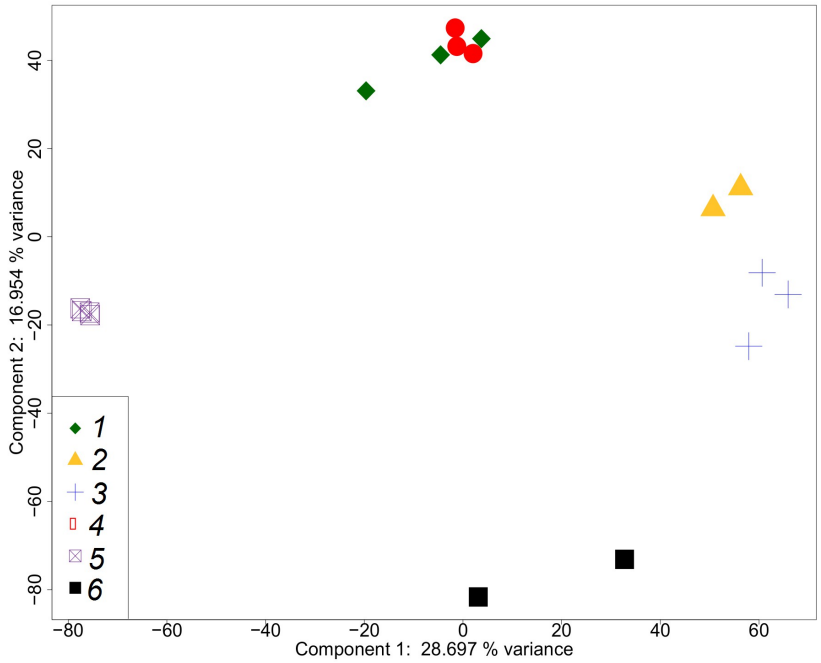
Estimation of variance – Recap

Variance?

PCA: Principle Component Analysis

- Convertion of set of (possibly correlated) values into set of values of linearly uncorrelated variables = principal components
- First principal component has the largest possible variance





Estimation of variance – Recap

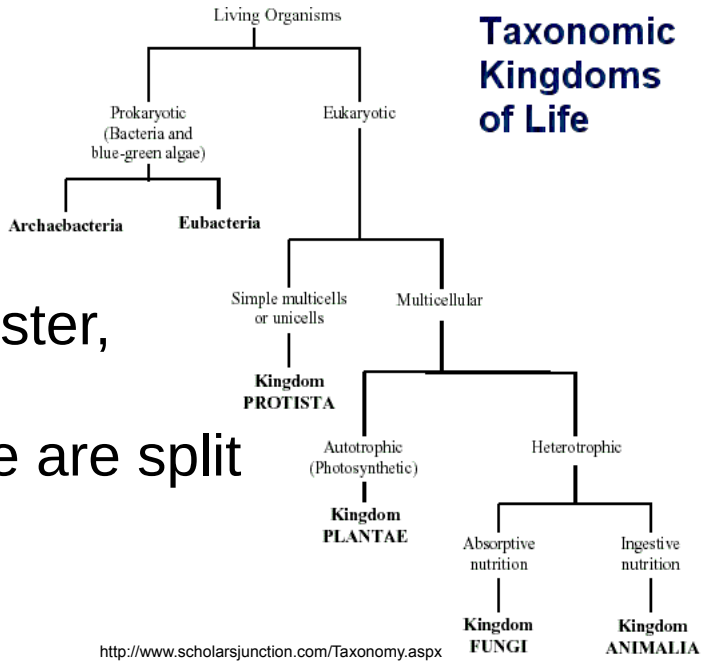
Variance?

Clustering: Finding similarities

- Taxonomy in biology based on clustering (cladistics)
- Different methods: hierarchical clustering, kmeans clustering...

a) Hierarchical clustering

- Known from taxonomy
- Build a hierarchy of clusters
- Agglomerative: each observation starts in own cluster, then those clusters are connected
- Divisive: all observations in one cluster, then those are split



http://www.scholarsjunction.com/Taxonomy.aspx

Estimation of variance – Recap

Clustering: Finding similarities

- Taxonomy in biology based on clustering (cladistics)
- Different methods: hierarchical clustering, kmeans clustering...

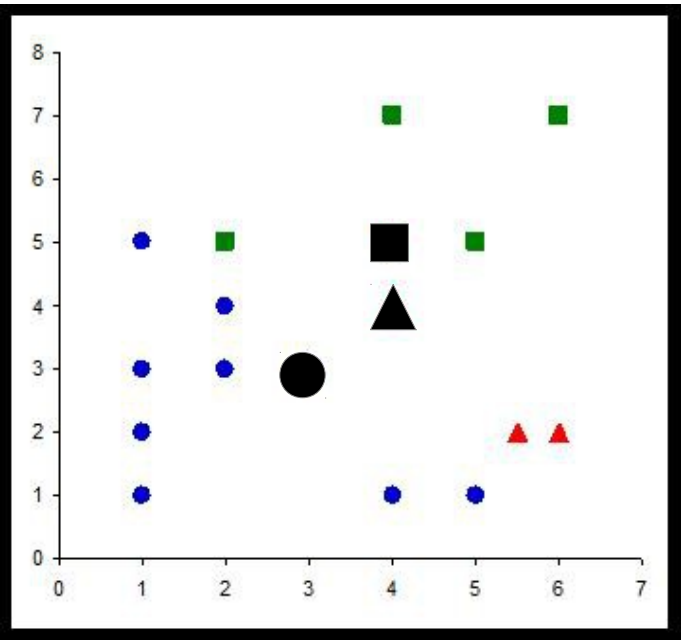
b) Kmeans clustering

- Number of clusters is known
- Each cluster has a center (=centroid)
- Iteratively: 1) choose centroids
  2) choose centroids, so that they are closer to your data points
  3) relate all data points to the closest centroids
  4) recalculate centroids
  ----> do so until the centroids do not change again

Estimation of variance – Recap

Variance?

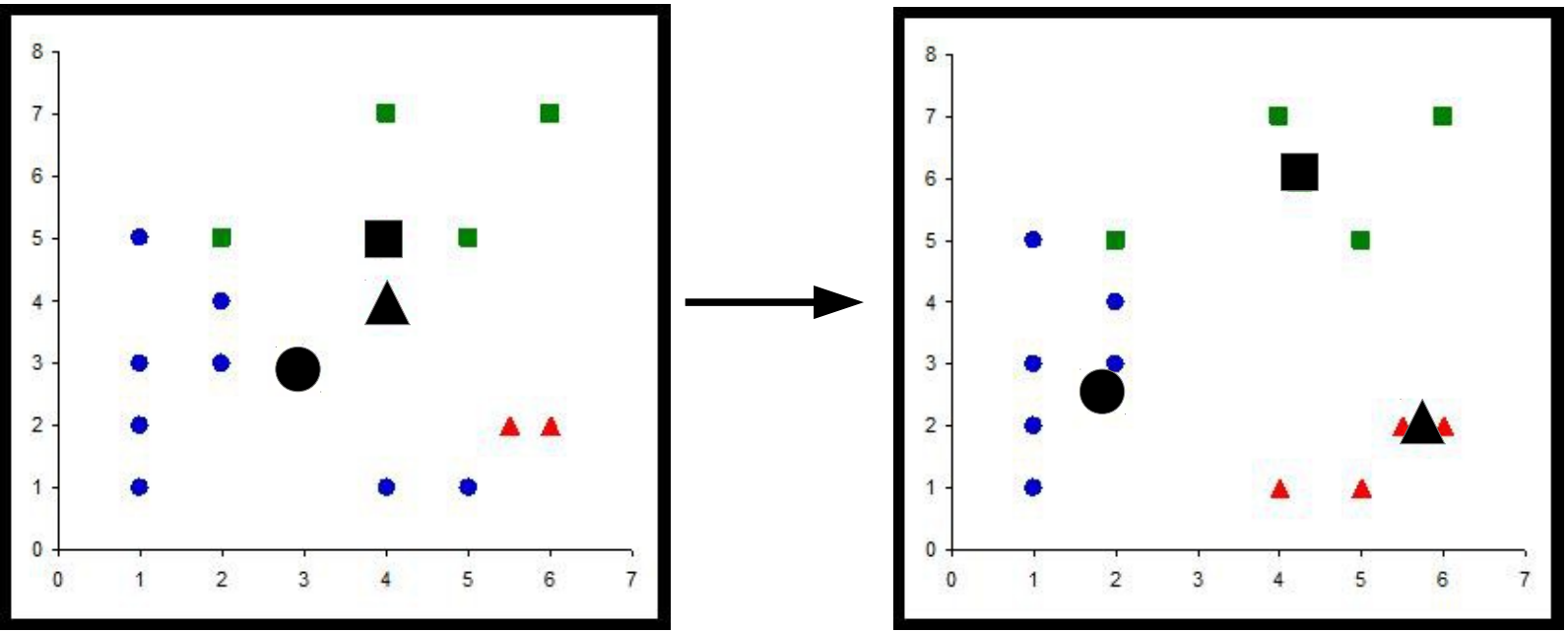
http://www-m9.ma.tum.de/material/felix-klein/clustering/Methoden/K-Means.php



Estimation of variance – Recap

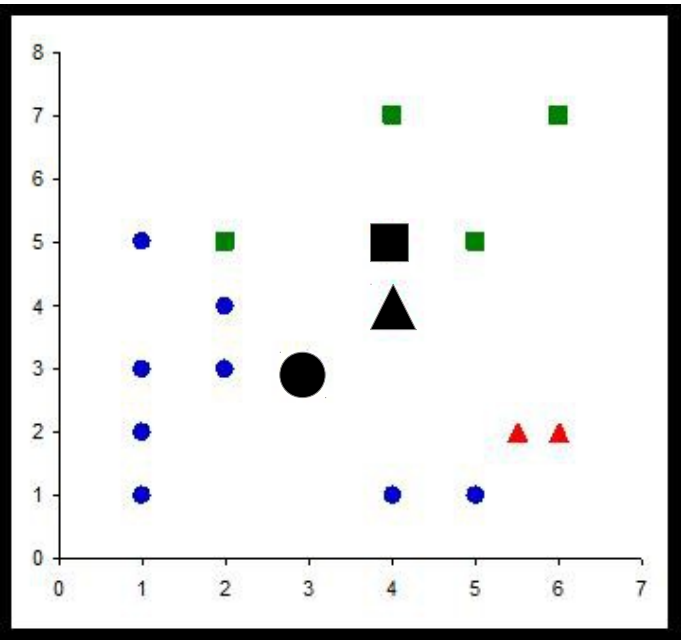
Variance?

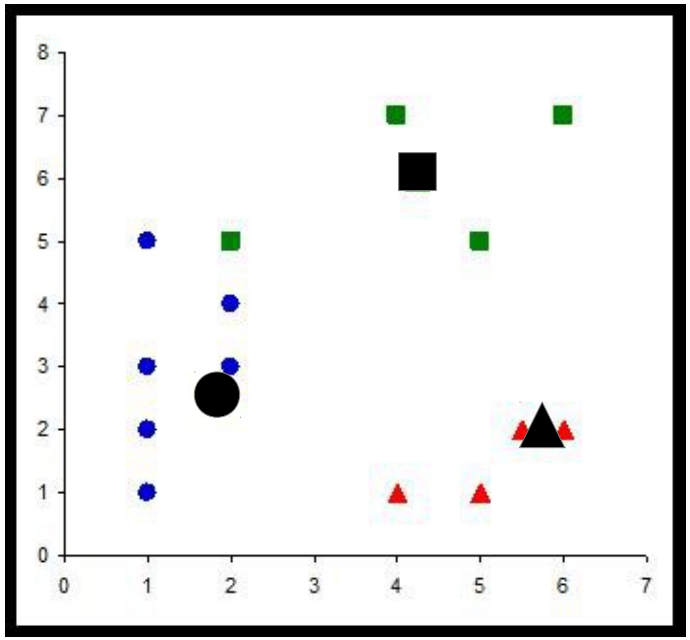
http://www-m9.ma.tum.de/material/felix-klein/clustering/Methoden/K-Means.php

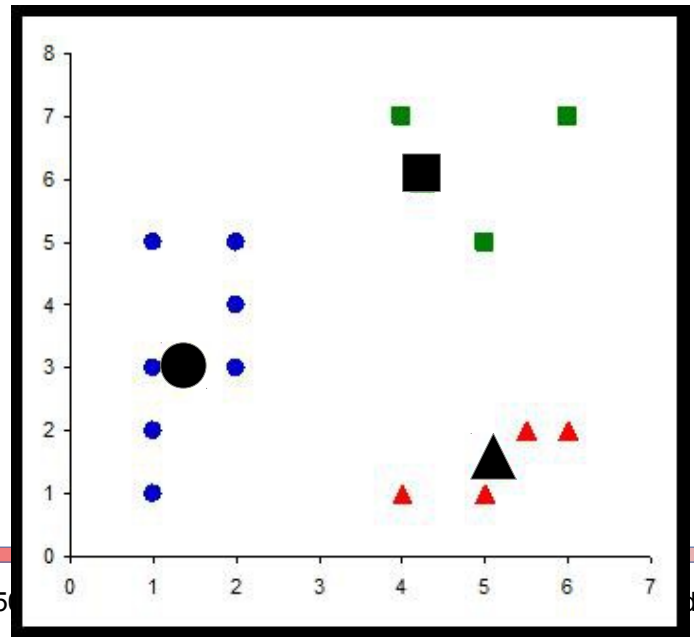


Estimation of variance – Recap

http://www-m9.ma.tum.de/material/felix-klein/clustering/Methoden/K-Means.php



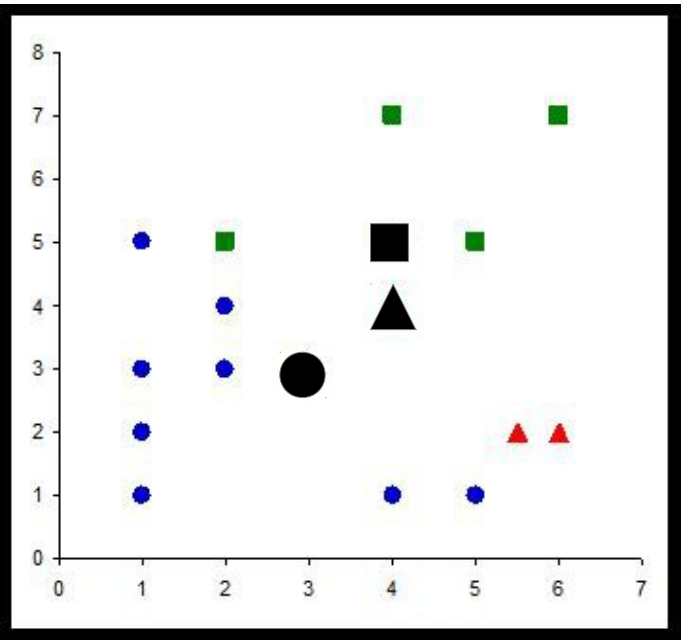


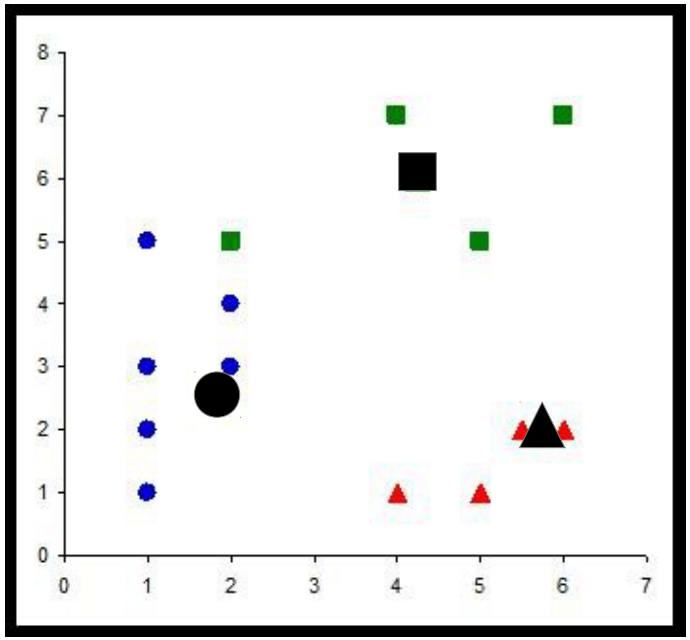


Variance?

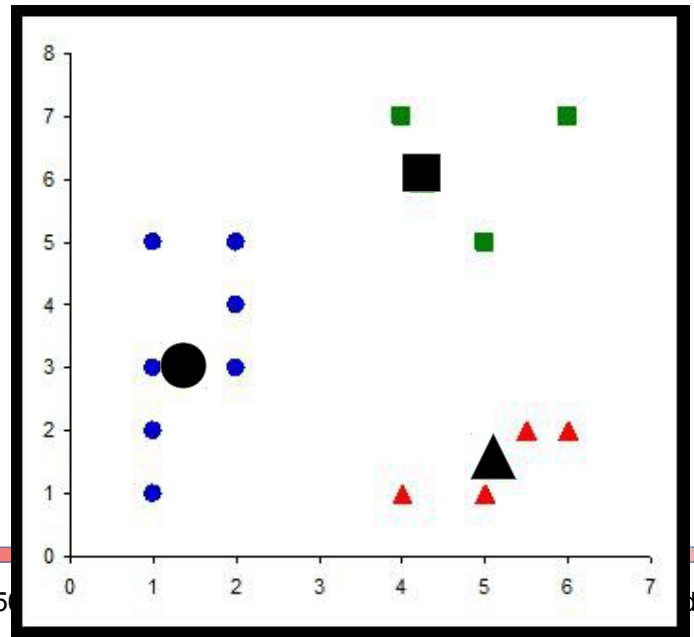
Estimation of variance – Recap

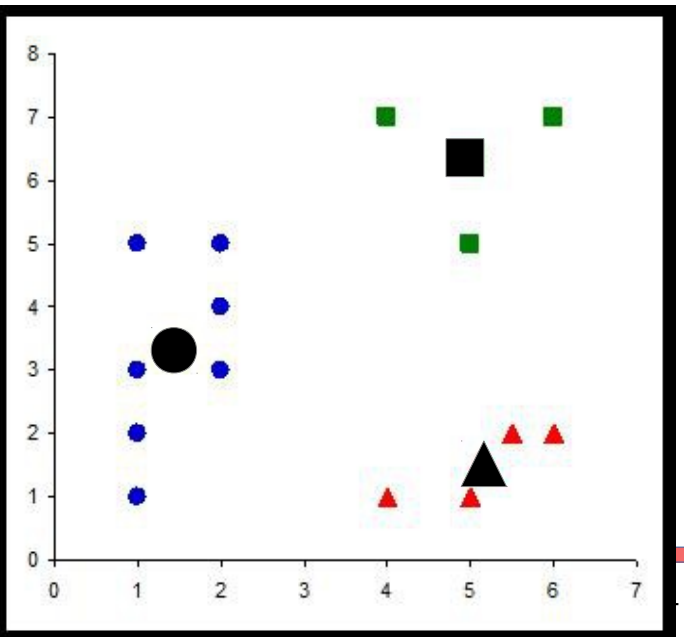
http://www-m9.ma.tum.de/material/felix-klein/clustering/Methoden/K-Means.php



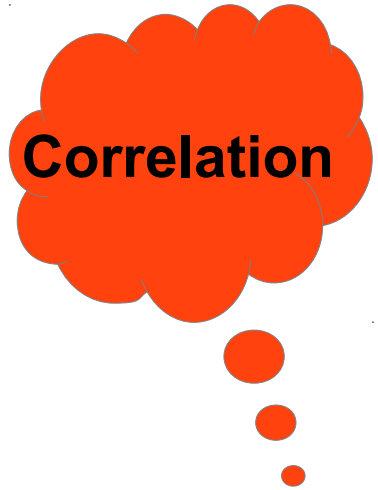


Variance?





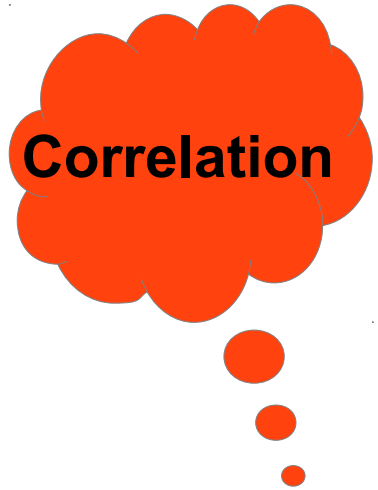
Estimation of variance – Correlation



Now: using default data set, provided by R
----> **mtcars**

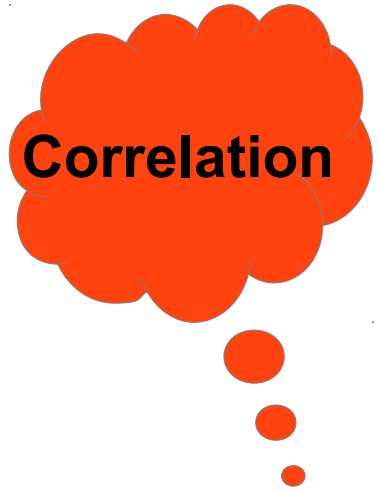
Estimation of variance – Correlation

Now: using d
----> **mtcars**




```
> mtcars
                   mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
Merc 280          19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
Merc 280C         17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
Merc 450SE        16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
Merc 450SL        17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
Merc 450SLC       15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
Fiat 128          32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
Honda Civic       30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
Toyota Corolla    33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
Toyota Corona     21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
Dodge Challenger  15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
AMC Javelin       15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
Camaro Z28        13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
Pontiac Firebird  19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2
Fiat X1-9         27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
Porsche 914-2     26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
Lotus Europa      30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
Ford Pantera L    15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
Ferrari Dino      19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
Maserati Bora     15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
Volvo 142E        21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```

Estimation of variance – Correlation

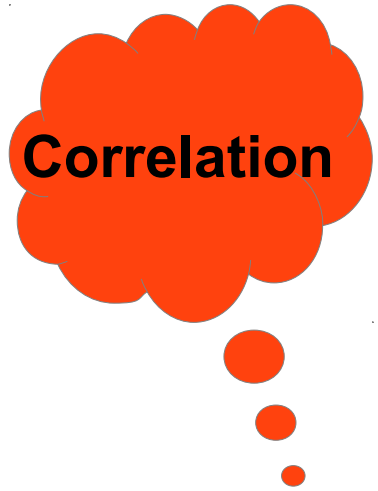


Now: using default data set, provided by R
----> **mtcars**

Command: cor()
---> check ?cor for settings
---> calculate correlation using Pearson

```
> cor(mtcars)
            mpg        cyl       disp         hp        drat         wt        qsec         vs          am       gear        carb
mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.68117191 -0.8676594  0.41868403  0.6640389  0.59983243  0.4802848 -0.55092507
cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.69993811  0.7824958 -0.59124207 -0.8108118 -0.52260705 -0.4926866  0.52698829
disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.71021393  0.8879799 -0.43369788 -0.7104159 -0.59122704 -0.5555692  0.39497686
hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.44875912  0.6587479 -0.70822339 -0.7230967 -0.24320426 -0.1257043  0.74981247
drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.00000000 -0.7124406  0.09120476  0.4402785  0.71271113  0.6996101 -0.09078980
wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.71244065  1.0000000 -0.17471588 -0.5549157 -0.69249526 -0.5832870  0.42760594
qsec  0.4186840 -0.5912421 -0.4336979 -0.7082234  0.09120476 -0.1747159  1.00000000  0.7445354 -0.22986086 -0.2126822 -0.65624923
vs    0.6640389 -0.8108118 -0.7104159 -0.7230967  0.44027846 -0.5549157  0.74453544  1.0000000  0.16834512  0.2060233 -0.56960714
am    0.5998324 -0.5226070 -0.5912270 -0.2432043  0.71271113 -0.6924953 -0.22986086  0.1683451  1.00000000  0.7940588  0.05753435
gear  0.4802848 -0.4926866 -0.5555692 -0.1257043  0.69961013 -0.5832870 -0.21268223  0.2060233  0.79405876  1.0000000  0.27407284
carb -0.5509251  0.5269883  0.3949769  0.7498125 -0.09078980  0.4276059 -0.65624923 -0.5696071  0.05753435  0.2740728  1.00000000
```

Estimation of variance – Correlation

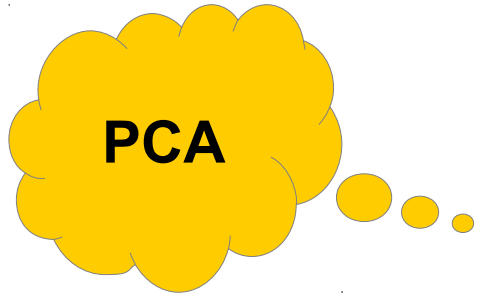


Now: using default data set, provided by R
----> **mtcars**

Command: cor()
---> check ?cor for settings
---> calculate correlation using Spearman

```
> cor(mtcars, method="spearman")
             mpg        cyl       disp         hp        drat         wt        qsec         vs          am       gear        carb
mpg    1.0000000 -0.9108013 -0.9088824 -0.8946646  0.65145546 -0.8864220  0.46693575  0.7065968  0.56200569  0.5427816 -0.65749764
cyl   -0.9108013  1.0000000  0.9276516  0.9017909 -0.67888119  0.8577282 -0.57235095 -0.8137890 -0.52207118 -0.5643105  0.58006798
disp  -0.9088824  0.9276516  1.0000000  0.8510426 -0.68359210  0.8977064 -0.45978176 -0.7236643 -0.62406767 -0.5944703  0.53977806
hp    -0.8946646  0.9017909  0.8510426  1.0000000 -0.52012499  0.7746767 -0.66660602 -0.7515934 -0.36232756 -0.3314016  0.73337937
drat   0.6514555 -0.6788812 -0.6835921 -0.5201250  1.00000000 -0.7503904  0.09186863  0.4474575  0.68657079  0.7448162 -0.12522294
wt    -0.8864220  0.8577282  0.8977064  0.7746767 -0.75039041  1.0000000 -0.22540120 -0.5870162 -0.73771259 -0.6761284  0.49981205
qsec   0.4669358 -0.5723509 -0.4597818 -0.6666060  0.09186863 -0.2254012  1.00000000  0.7915715 -0.20333211 -0.1481997 -0.65871814
vs     0.7065968 -0.8137890 -0.7236643 -0.7515934  0.44745745 -0.5870162  0.79157148  1.0000000  0.16834512  0.2826617 -0.63369482
am     0.5620057 -0.5220712 -0.6240677 -0.3623276  0.68657079 -0.7377126 -0.20333211  0.1683451  1.00000000  0.8076880 -0.06436525
gear   0.5427816 -0.5643105 -0.5944703 -0.3314016  0.74481617 -0.6761284 -0.14819967  0.2826617  0.80768800  1.0000000  0.11488698
carb  -0.6574976  0.5800680  0.5397781  0.7333794 -0.12522294  0.4998120 -0.65871814 -0.6336948 -0.06436525  0.1148870  1.00000000
```

Estimation of variance – PCA



Command: prcomp()

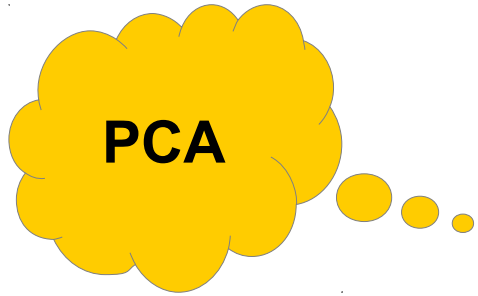
---> check ?prcomp for settings

---> calculate PCA for data set „USArrests“

---> prcomp advises to scale data before calculating PCA; data will have unit variance afterwards

---> thus, our command is: **prcomp(USArrests, scale=TRUE)**

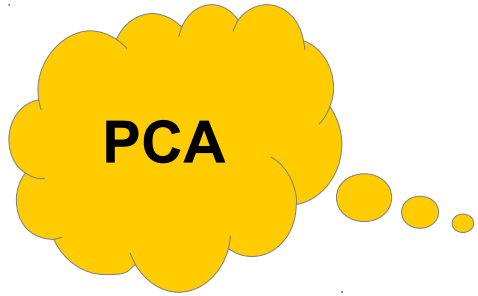
Estimation of variance – PCA



Command: prcomp()
---> check ?prcomp for settings
---> calculate PCA for data set „USArrests“
---> prcomp advises to scale data before calculating PCA; data will have unit variance afterwards
---> thus, our command is: **prcomp(USArrests, scale=TRUE)**

---> look at the summary to determine how strong your variance is in each component (e.g.)
Command: summary(prcomp(USArrests, scale=TRUE))
OR you could store the result of prcomp(USArrest, scale=TRUE) in a variable...

Estimation of variance – PCA

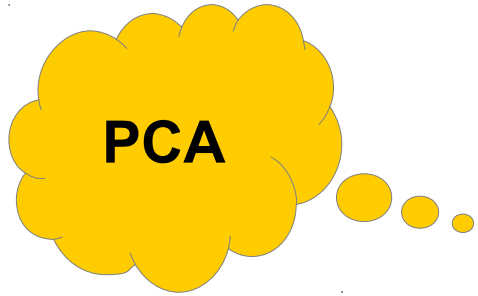


Command: prcomp()
---> check ?prcomp for settings
---> calculate PCA for data set „USArrests“
---> prcomp advises to scale data before calculating PCA; data will have unit variance afterwards
---> thus, our command is: **prcomp(USArrests, scale=TRUE)**

---> look at the summary to determine how strong your variance is in each component (e.g.)
Command: summary(prcomp(USArrests, scale=TRUE))

```
pr <- prcomp(USArrests, scale=TRUE)
summary(pr)
```

Estimation of variance – PCA



Command: prcomp()
---> a plot would be more informative!
---> **biplot(pr)**

Estimation of va

Command: prcon
---> a plot would b
---> **biplot(pr)**

Estimation of variance – Clustering

Hierarchical clustering

First, we need to calculate the distances in our data

Command: dist()
---> distance <-dist(USArrests)

Then, we can go on with clustering

Command: hclust()
---> hclust(distance)

Again, a plot would be nicer...

Estimation of variance – Clustering

Hierarchical clustering

First, we need to calculate the distances in our data

Command: dist()
---> distance <-dist(USArrests)

Then, we can go on with clustering

Command: hclust()
---> hclust(distance)

Again, a plot would be nicer...

Command: plot(hclust(distance))

Estimation of variance – Clustering

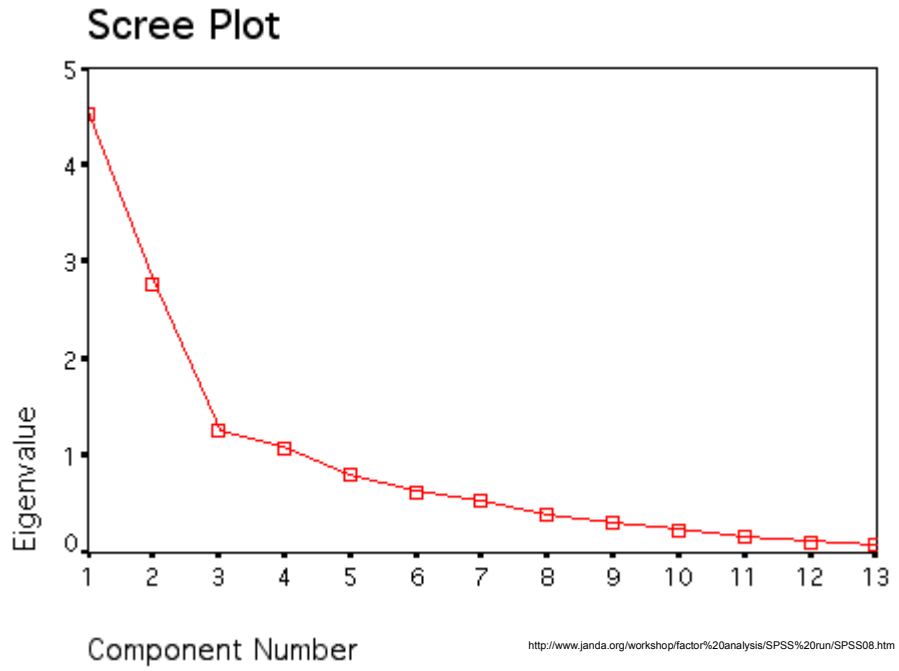
Kmeans clustering

How many centroids?

---> use e.g. cluster structure derived by hclust...
---> … do a scree plot …
---> …

Assuming 8 clusters

Command: kmeans()
---> read **?kmeans**



http://www.janda.org/workshop/factor%20analysis/SPSS%20run/SPSS08.htm

Estimation of variance – Clustering

Kmeans clustering

How many centroids?

---> use e.g. cluster structure derived by hclust...
---> … do a scree plot …
---> …

Assuming 8 clusters

Command: kmeans()
---> read **?kmeans**

---> kmeans(USArrests, 8)

Estimation of variance – Clustering

Kmeans clustering

How many centroids?

---> use e.g. cluster structure derived by hclust...
---> … do a scree plot …
---> …

Assuming 8 clusters

Command: kmeans()
---> where are the clusters?

Plot

Outlook: kmeans with random starts, hclust with different methods

Estimation of variance – Clustering

Kmeans clustering

How many centroids?

---> use e.g. cluster structure derived by hclust...
---> … do a scree plot …
---> …

Assuming 8 clusters

Command: kmeans()
---> where are the clusters?

---> kmeans(USArrests, 8)$cluster

Estimation of variance – Clustering

Kmeans clustering

How many centroids?

---> use e.g. cluster structure derived by hclust...
---> … do a scree plot …
---> …

Assuming 8 clusters

Command: kmeans()
---> where are the clusters?

---> kmeans(USArrests, 8)$cluster

A plot would be cool...

Estimation of variance – Clustering

Kmeans clustering

How many centroids?

---> use e.g. cluster structure derived by hclust...
---> … do a scree plot …
---> …

Assuming 8 clusters

Command: kmeans()
---> plot(USArrests, kmeans(USArrests, 8)$cluster)

With colors?

---> plot(USArrests, col=kmeans(USArrests, 8)$cluster)

Estimation of variance – Clustering

Kmeans clustering

How many centroids?

---> use e.g. cluster structure derived by hclust...
---> … do a scree plot …
---> …

Assuming 8 clusters

Command: kmeans()
---> plot(USArrests, kmeans(USArrests, 8)$cluster)

With names?

---> advanced, special libraries can simplify your task...

Estima

Kmeans

How ma

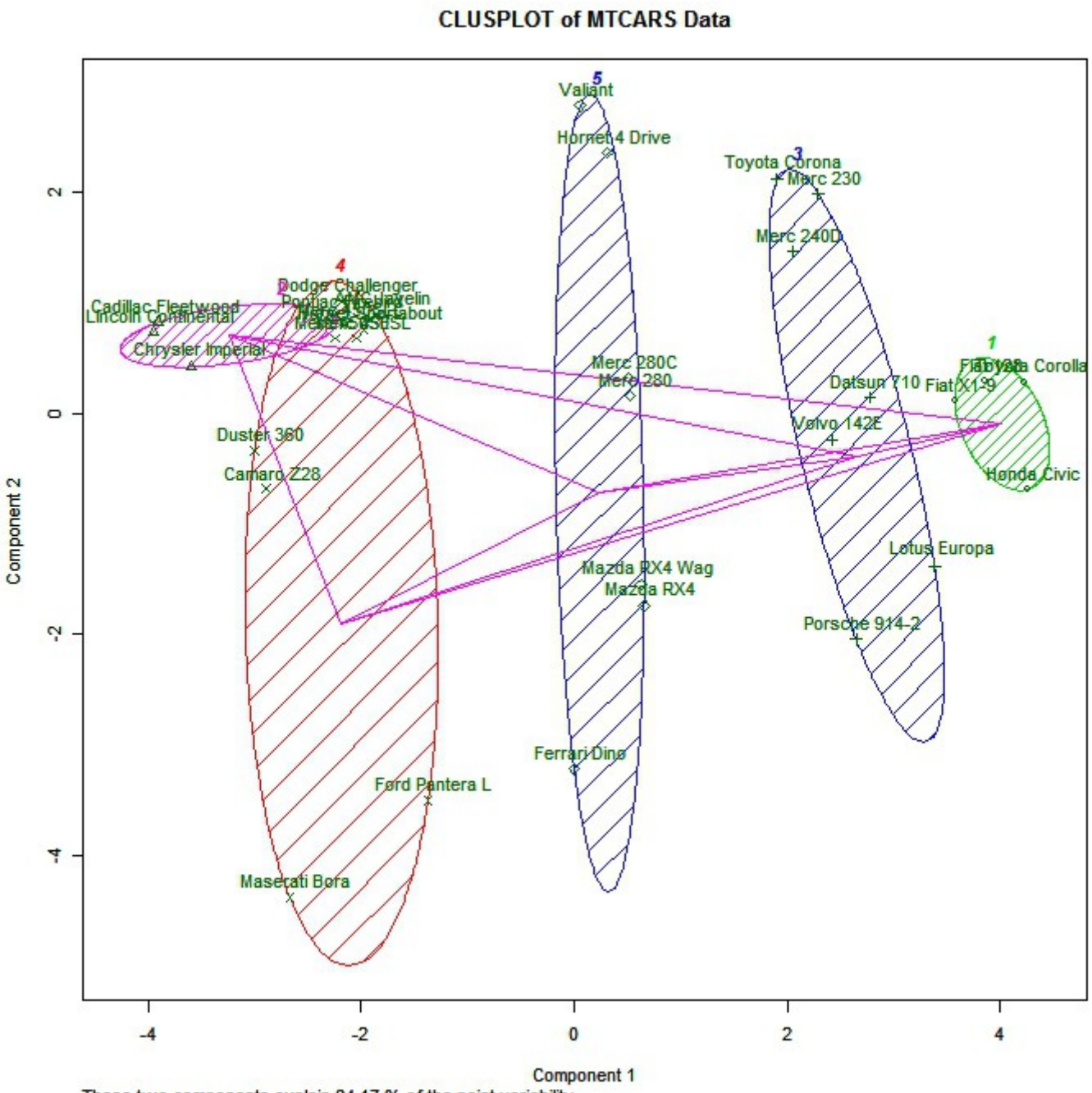
---> use
---> … (
---> …

Assumi

Comma
---> plo

With na

---> adv



Clustering

http://www.statmethods.net/advstats/images/cluster5.jpg