9:00 – 17:00

10:30 – 10:45 Coffee break
12:30 – 13:00 Lunch break
14:30 – 14:45 Coffee break

# INTRODUCTION TO STATISTICAL DATA ANALYSIS USING R

ANJA EGGERT

https://github.com/AnjaEggert/StatisticsInR_Intro

Universität Rostock Traditio et Innovatio

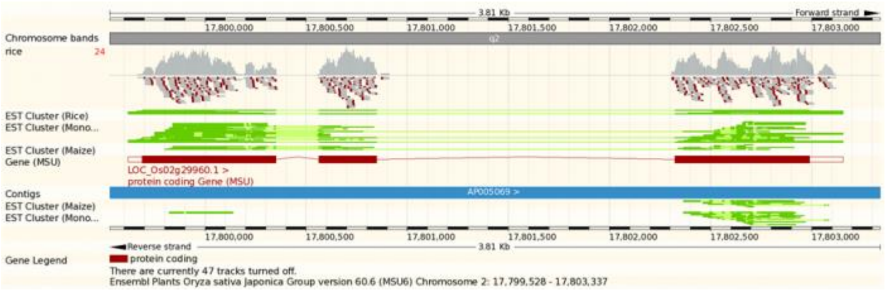
## Outline

- **Genomic data**
  - HowTo use Bioconductor
- **Cluster analysis**
- **Multivariate analysis**
  - PCA dimension reduction
- **Non-linear regression**
- **Design types**
  - Blocking, fixed and random effects
  - Completely Randomized Design, Randomized Complete Block Design

  - Demographic data,

# GENOMIC DATA

## Genomic data



There are currently 47 tracks turned off.
Ensembl Plants Oryza sativa Japonica Group version 60.6 (MSU6) Chromosome 2: 17,799,528 - 17,803,337

Screenshot from Ensembl genome browser, showing gene annotation of a genomic region as well as a read pile-up visualization of an RNA-Seq experiment.

## ggbio: visualization toolkits for genomic data

- How to make plots of genomic data…

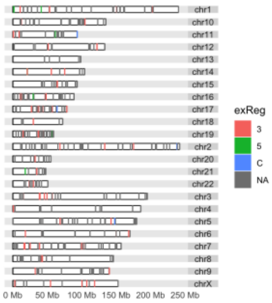
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

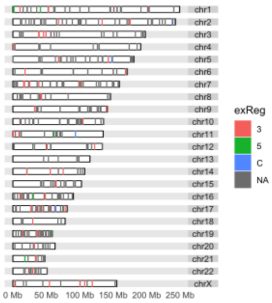
## Question for you

- What type of object is darned_hg19_subset500?

## Question for you

- How do you re-order the chromosomes?

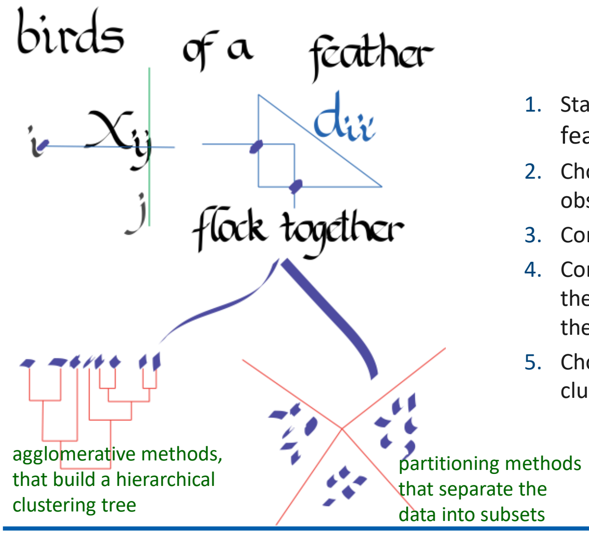




Th darned_hg19_subset500 lists a selection of 500 RNA editing sites in the human genome. It was obtained from the *Database of RNA editing in Flies, Mice and Humans* (DARNED, http://darned.ucc.ie).

# CLUSTER

## Steps of a cluster analysis



agglomerative methods, that build a hierarchical clustering tree

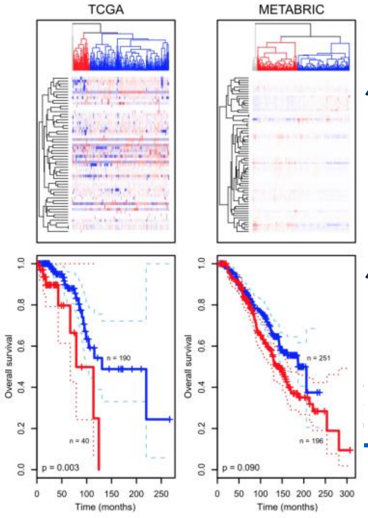
partitioning methods that separate the data into subsets

1. Starting from an observations-by-features rectangular table X
2. Choosing an observations-to-observations distance measure
3. Computing the distance matrix
4. Constructing the clusters based on the distances are used to construct the clusters.
5. Choice to be made: the number k of clusters

## Medical cluster example: breast cancer (Aure et al. 2017)

The Cancer Genome Atlas

Molecular Taxonomy of Breast Cancer International Consortium

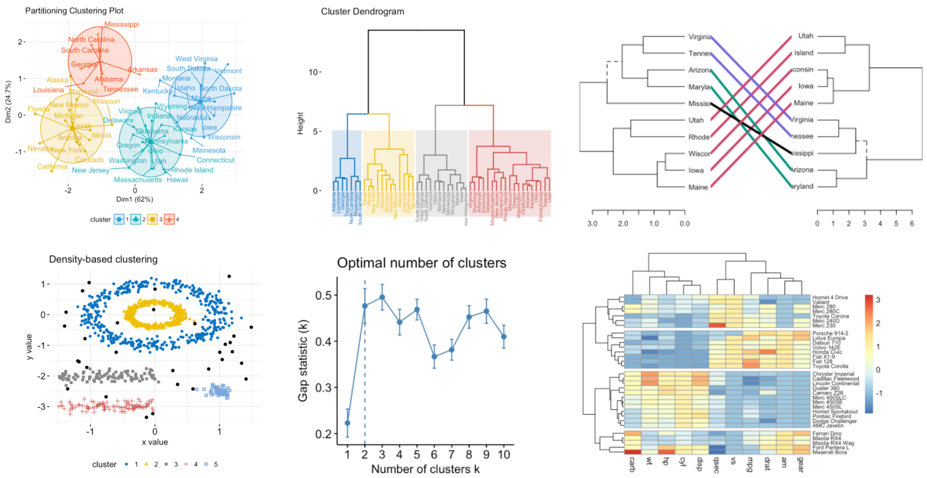


The breast cancer samples can be split into groups using their miRNA expression.

Survival times in the two groups were different.

Thus clusters are biologically and clinically relevant !!

## Clustering in R – many options



## Task for you

- Look up the **BiocViews Clustering** or the **Cluster view on CRAN** and count the number of packages providing clustering tools.

# MULTIVARIATE ANALYSIS

## Several variables measured on the same set of subjects

- E.g. biometric characteristics for thousand patients:
  - Height
  - weight
  - Age
  - blood pressure
  - blood sugar
  - heart rate
  - genetic data
- Multivariate analysis is the investigation of connections or associations between the different variables measured.
- Data reported in a tabular data structure with one row for each subject and one column for each variable

## Dimension reduction

- **If the columns of the matrix are all independent**
  - study each column separately and do standard “univariate” statistics

- **If there are patterns and dependencies**
  - need to use “multivariate” statistics

- **Example:**
  - Gene expression of 25,000 genes (columns) on 1000 patients (rows)
  - each row representing the many measurements made on the same observational unit
  - genes often are either positively correlated or they are anti-correlated
  - Reduce 25,000 dimensions to a smaller number of important dimensions, without losing too much information

## Load data sets and do the tasks

- **Turtles**
  - Show the first 4 rows of the data set

- **Athletes**
  - Plot the performance for the 100 m with ggplot

- **Gene expression**
  - Round the numbers to 2 digits

- **Bacterial Species Abundances**
  - Extract the otu_table as a matrix
  - Use the function phyloseq::otu_table ()

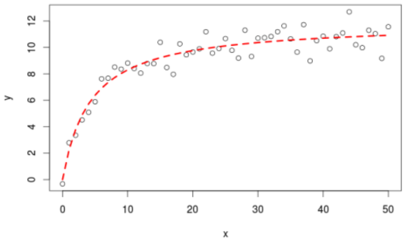
## Tasks for you

- **Compute the matrix of all correlations between the measurements from the turtles data. What do you notice ?**
- **Produce all pairwise scatterplots, as well as the one-dimensional histograms on the diagonal, for the turtles data. Use the package GGally.**
- **Make a heatmap of the athletes data. What do you notice? Use the package pheatmap.**

# NON-LINEAR REGRESSION

## Fit a non-linear function

- **Non-linear regression: specify a function with a set of parameters to fit to the data**
- **Estimate such parameters is to use a non-linear least squares approach**
  - nls() in R
- **The estimated parameters have a clear interpretation (Vmax in a Michaelis-Menten model is the maximum rate) which would be harder to get using linear models on transformed data for example.**



## Finding the right starting values

- **Finding good starting values is very important in non-linear regression to allow the model algorithm to converge. If you set starting parameters values completely outside of the range of potential parameter values the algorithm will either fail or it will return non-sensical parameter.**

- **The best way to find correct starting value is to “eyeball” the data, plotting them and based on the understanding that you have from the equation find approximate starting values for the parameters.**

**Completely Randomized Design, Randomized Complete Block Design,**
**Split-Plot Design, Latin Square, alpha-Lattice**

# DESIGN TYPES

## Completely randomized design (CRD)

- **The simplest type**
- **Treatments are assigned to the experimental units completely at random**
- **Assumption: any other external conditions affect treatment conditions equally**
- **Thus: any significant effect will be attributed to the tested treatment/factor**
- **Most useful if:**
  - Experimental units are homogeneous
  - Experiments are relatively small
  - Number of treatments is relatively small

Maybe not valid in many field experiments?

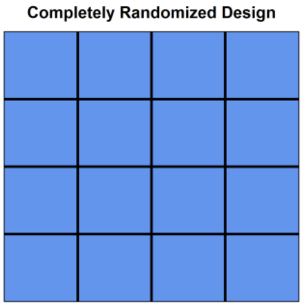
## CRD cont.

- **Advantages:**
  - Flexible design
  - Number of treatments and replicates only limited by the available number of experimental units
  - Simple statistical analysis

- **Disadvantages:**
  - If experimental units are not homogeneous this will cause loss of statistical power

## CRD example

- You want to study the effect of 4 different treatments A, B, C & D on growth of bacteria.
- You can examine 16 cultures in petri dishes.
- The petri dishes will be kept in a cultivation room on 4 shelves above each other and each shelf has room for 4 petri dishes.

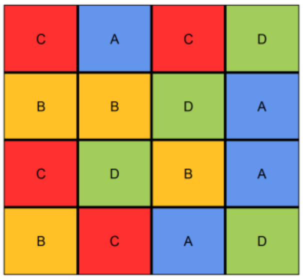


Remember! You have to make complete randomization, not only partial.

How should we use **randomization** to assign the cultures to the 4 treamtents and to the 16 petri dishes ?

## CRD example

- You want to study the effect of 4 different treatments A, B, C & D on growth of bacteria.
- You can examine 16 cultures in petri dishes.
- The petri dishes will be kept in a cultivation room on 4 shelves above each other and each shelf has room for 4 petri dishes.



Completely Randomized Design

| | | | |
|---|---|---|---|
| C | A | C | D |
| B | B | D | A |
| C | D | B | A |
| B | C | A | D |

```
library(agricolae)
design.crd(c("A","B","C","D"),c(4,4,4,4),
           serie=2,seed=13,"Super-Duper")

library(desplot)
desplot(data=dat,form=diet~col+row,text = diet)
```

## CRD Data analysis

| | Diet A | Diet B | Diet C | Diet D |
|---|---|---|---|---|
| | 61 | 63 | 62 | 62 |
| | 56 | 66 | 62 | 60 |
| | 54 | 63 | 61 | 67 |
| | 58 | 59 | 61 | 64 |
| **Group mean** | **57.25** | **62.75** | **61.50** | **63.25** |

```
# Linear model
mod <- lm(mass ~ diet)
```

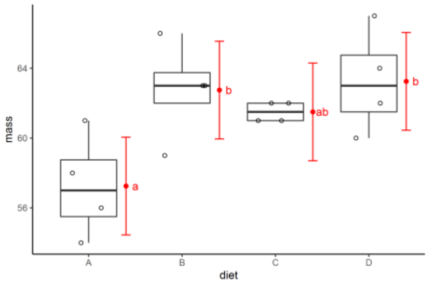
- **Linear model equation with one factor:** $\boldsymbol{Y_{ij}} = \mu + \boldsymbol{T_j} + \boldsymbol{e_{ij}}$

$Y_{ij}$ the i<sup>th</sup> observation under the j<sup>th</sup> treamtent
μ global mean
$T_j$ the effect of the j<sup>th</sup> treatment
$e_{ij}$ random error associated with the i<sup>th</sup> observation under the j<sup>th</sup> treatment

## CRD Data analysis (in *R*)

```
# Analysis of Variance Table
Response: mass
          Df Sum Sq Mean Sq F value  Pr(>F)
diet       3 89.188 29.7292  4.5016 0.02455 *
Residuals 12 79.250  6.6042

# Adjusted means and contrasts
emmeans(mod, pairwise ~ diet, adjust = "tukey“)
$`emmeans`
# diet emmean       SE df lower.CL upper.CL
# A     57.25 1.284929 12 54.45038 60.04962
# B     62.75 1.284929 12 59.95038 65.54962
# C     61.50 1.284929 12 58.70038 64.29962
# D     63.25 1.284929 12 60.45038 66.04962
#
$contrasts
# contrast estimate       SE df t.ratio p.value
# A - B       -5.50 1.817164 12  -3.027  0.0452
# A - C       -4.25 1.817164 12  -2.339  0.1434
# A - D       -6.00 1.817164 12  -3.302  0.0280
# B - C        1.25 1.817164 12   0.688  0.8997
# B - D       -0.50 1.817164 12  -0.275  0.9923
# C - D       -1.75 1.817164 12  -0.963  0.7723
```

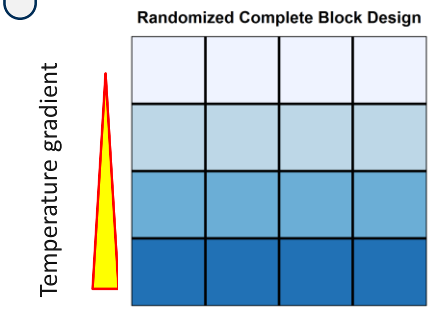


**Completely Randomized Design, Randomized Complete Block Design,** **Split-Plot Design, Latin Square, alpha-Lattice**

# DESIGN TYPES

## RCBD example

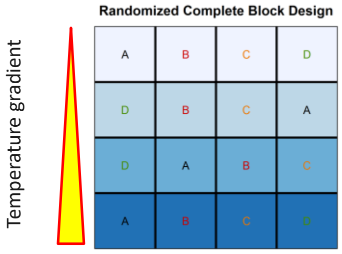
- You want to study the effect of 4 different treatments A, B, C & D on growth of bacteria.
- You can examine 16 cultures in petri dishes.
- The petri dishes will be kept in a cultivation room on 4 shelves above each other and each shelf has room for 4 petri dishes.



CRD **does not** account for possible (likely?) microclimate differences on the shelves in the room.
If we want to acknowledge the potential effect of a vertical temperature gradient, we should organize the experiment using a RCBD !

## RCBD example

- You want to study the effect of 4 different treatments A, B, C & D on growth of bacteria.
- You can examine 16 cultures in petri dishes.
- The petri dishes will be kept in a cultivation room on 4 shelves above each other and each shelf has room for 4 petri dishes.



| | | | |
|---|---|---|---|
| A | B | C | D |
| D | B | C | A |
| D | A | B | C |
| A | B | C | D |

Wothin each block (here shelf), treatments are randomly assigned.

One treatment of each type will be used on each of the 4 shelfs = RCBD!

## Decrease residual errors by blocking

"Block what you can; randomize what you cannot."

- **Disadvantages of CBD:**
  - If experimental units are not homogeneous **and you fail to minimize this variation using blocking**, this will cause loss of statistical power

- Blocking can be used to reduce the contribution to the residual error contributed by nuisance factors by creating more homogeneous groups (blocks) in which the nuisance factors are held constant and the factor of interest is allowed to vary.

- Blocks usually represent levels of naturally-occurring differences or sources of variation that are unrelated to the treatments, and the characterization of these differences is not (so much) of interest to the researcher.

## Blocking

- **Blocking = grouping similar experimental units together and assigning the treatments of interest (randomly) to the experimental units within such groups.**

- **Variation among blocks can be partitioned out of the residual error.**

- **Blocking reduces the experimental error and increases the power of the test:**

CRD: $SS_{residuals} = SS_{total} - SS_{treatments}$

RCBD: $SS_{residuals} = SS_{total} - SS_{treatments} - \mathbf{SS_{blocks}}$

- **But blocking also reduces degrees of freedom of** $SS_{residuals}$ → only include blocking factor if ther are differences among blocks

## Experimental Blocking – why do we do this?

- **Example from environmental science:** Often we have environmental gradients in our sites that cause variation or effect our results, so to account for this we can place our blocks in such a way so that we minimize environmental heterogeneity (difference). Therefore our results are more likely due to your treatment effect.
- **Example from other experiments?**

  – Heterogeneous animals (age/weight/sex)
  – Different shelves/rooms
  – Natural structure (litters)
  – Split experiment in time

## CRD *versus* RCBD

- Experimental units are distributed at random
- Treatment levels are typically equally represented

| | | |
|---|---|---|
| A | B | B |
| C | A | C |
| B | B | A |
| C | A | C |

**Example**
Varieties are planted randomly

- Experimental units are assigned to blocks, then randomly to treatment levels
- The representation of treatment levels in each block are equal

| Block 1 | Block 2 | Block 3 |
|---|---|---|
| A | B | B |
| C | A | C |
| B | B | A |
| C | A | C |

**Example**
Animals split by family, then given a treatment

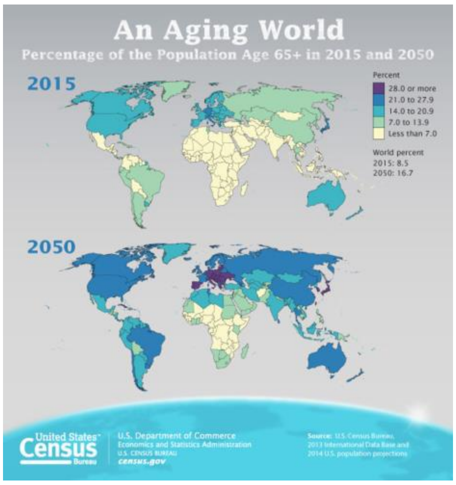
- Experimental units assigned to blocks, then randomly to treatment levels
- The representation of treatment levels in each block are NOT equal

| Block 1 | Block 2 | Block 3 |
|---|---|---|
| A | B | B |
| C | A | C |

**Example**
When the number of available animals per family is not big enough to accommodate all treatments

# DEMOGRAPHIC DATA

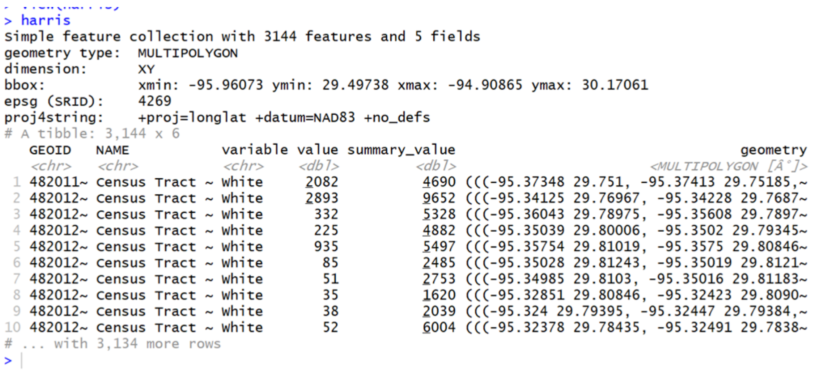
## United States Census Bureau



Source: U.S. Census Bureau, 2013 International Data Base and 2014 U.S. population projections

**Download & map Census data with the 'tidycensus' package.**

## Tidyverse and tibble



```
> harris
Simple feature collection with 3144 features and 5 fields
geometry type:  MULTIPOLYGON
dimension:      XY
bbox:           xmin: -95.96073 ymin: 29.49738 xmax: -94.90865 ymax: 30.17061
epsg (SRID):    4269
proj4string:    +proj=longlat +datum=NAD83 +no_defs
# A tibble: 3,144 x 6
   GEOID   NAME              variable value summary_value                                  geometry
   <chr>   <chr>             <chr>    <dbl>         <dbl>                       <MULTIPOLYGON [Â°]>
 1 482011~ Census Tract ~ White    2082          4690 (((-95.37348 29.751, -95.37413 29.75185,~
 2 482012~ Census Tract ~ White    2893          9652 (((-95.34125 29.76967, -95.34228 29.7687~
 3 482012~ Census Tract ~ White     332          5328 (((-95.36043 29.78975, -95.35608 29.7897~
 4 482012~ Census Tract ~ White     225          4882 (((-95.35039 29.80006, -95.3502 29.79345~
 5 482012~ Census Tract ~ White     935          5497 (((-95.35754 29.81019, -95.3575 29.80846~
 6 482012~ Census Tract ~ White      85          2485 (((-95.35028 29.81243, -95.35019 29.8121~
 7 482012~ Census Tract ~ White      51          2753 (((-95.34985 29.8103, -95.35016 29.81183~
 8 482012~ Census Tract ~ White      35          1620 (((-95.32851 29.80846, -95.32423 29.8090~
 9 482012~ Census Tract ~ White      38          2039 (((-95.324 29.79395, -95.32447 29.79384,~
10 482012~ Census Tract ~ White      52          6004 (((-95.32378 29.78435, -95.32491 29.7838~
# ... with 3,134 more rows
> 
```

- **Tibbles are a modern take on data frames, but crucially they are still data frames. Well, what’s the difference then?**

- **“keeping what time has proven to be effective, and throwing out what is not”.**