COSU at Universite de Lausanne
Biostatistics Seminar 2019

Biostatistics in Molecular Plant Sciences

Ludwig A. Hothorn
(retired from) Leibniz University Hannover
hothorn@biostat.uni-hannover.de

April 14, 2019

Topics I

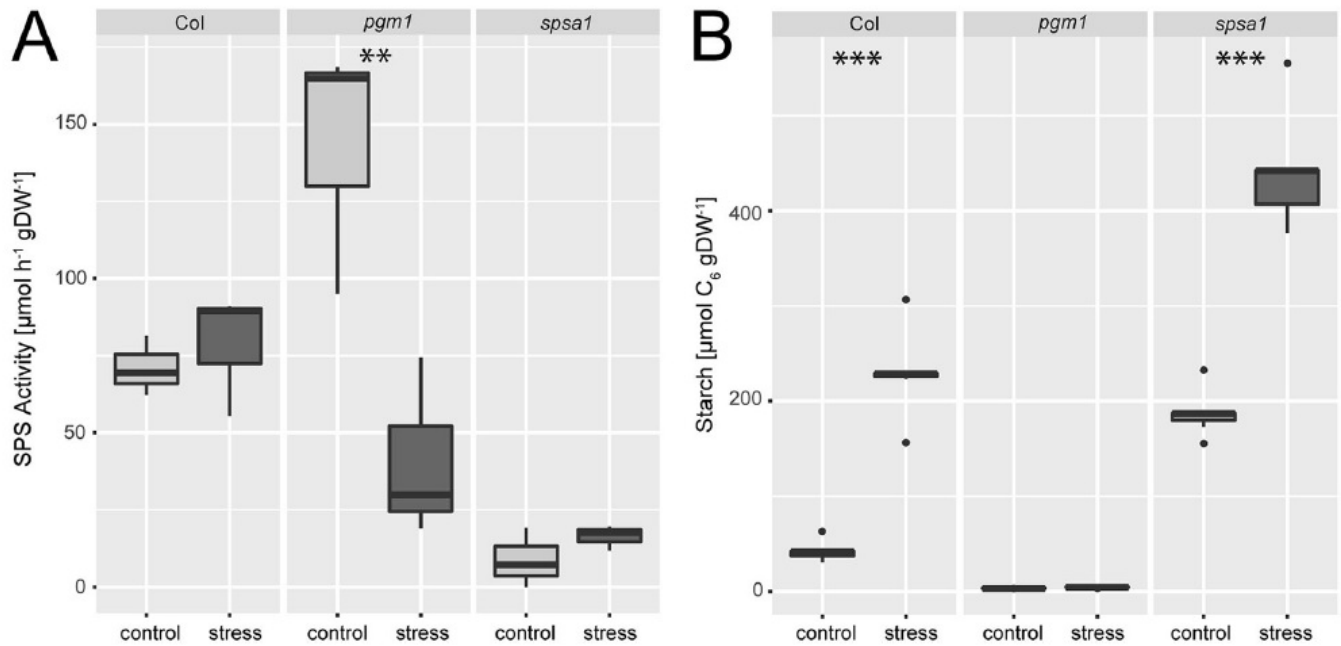
1. Introduction to biostatistics and the importance of statistical software (*Use R!: assuming R experience by CUSO2018 or next 19 20-23 May 2019; Advanced Statistics and Programming with R 4-7 June 2019 An Introduction to R*)
2. Statistical guidelines for publishing in high-ranked journals and their implications for data collection and analysis
3. First step of modeling: data structures in EXCEL (flat file format, unique sample ident, repeated measures, technical replicates, multiple factors) for an appropriate import into R
4. Graphical representation of grouped biological data using boxplots (R library(toxbox))

Topics II

5. Introduction to different statistical tests using real life data sets from CUSO member groups: Two-sample tests (t, Wilcoxon, Chi2, ratio-to-control tests): p-values vs. confidence intervals (effect sizes), tests for data containing technical replicates (mixed effect model), estimation of the required sample sizes: experimental design, the power concept
6. Multiple comparisons: i) comparing several treatments or doses vs. control (wild type,..), ii) BenjaminiHochberg procedure for high-dimensional data
7. Exercise I - case study: root growth assay as described in Hohmann et al., PNAS 115, 13, 3488-3493
8. Writing statistical method sections, reporting summaries and presenting graphics
9. Exercise II - to be determined (a selected case study)

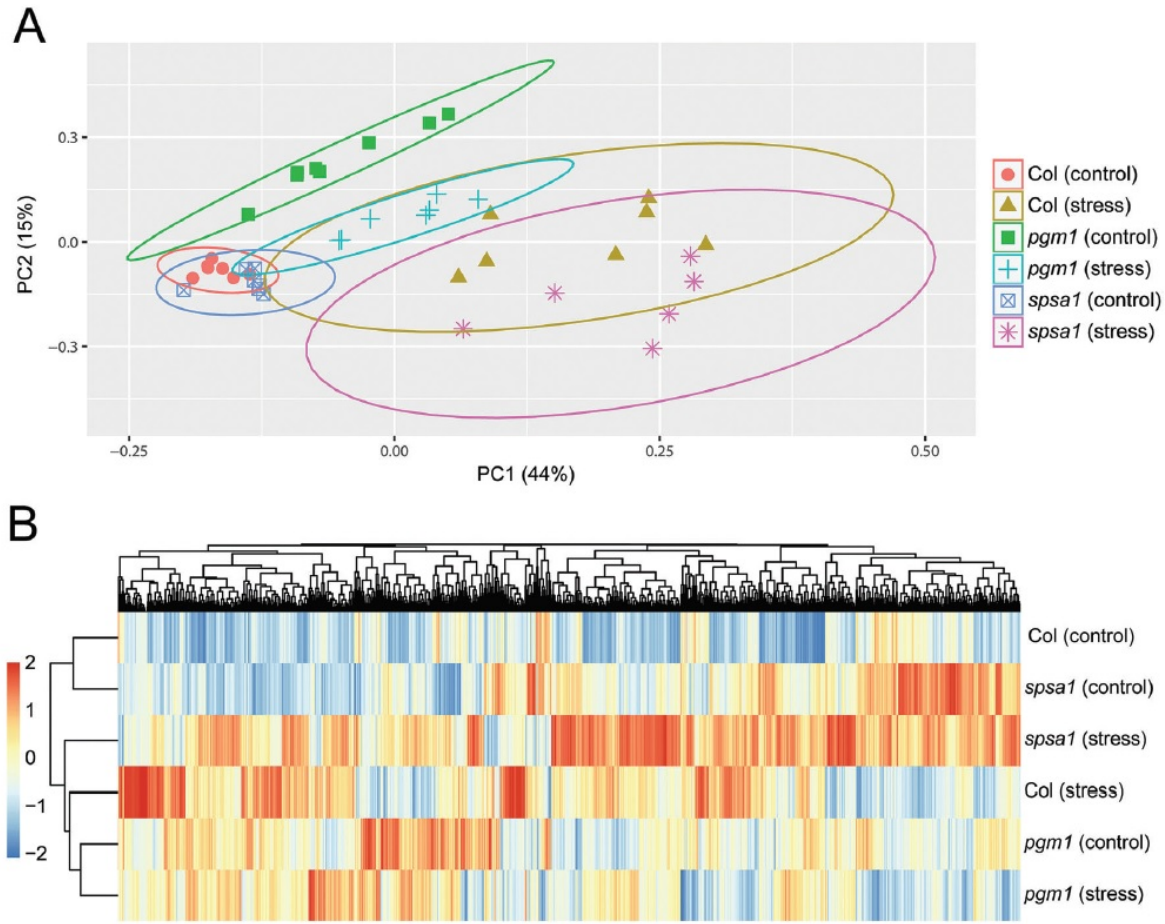
Scope and limitation I

- Recent ETH paper: Combined multivariate analysis and machine learning reveals a predictive module of metabolic stress response in Arabidopsis thaliana [FPS+18]
- **Covered:** Two factor exp (boxplots, ANOVA, multiple testing, interaction, confidence intervals)



Scope and limitation II

- Not covered: PCA, cluster analysis (may be 2020 course)



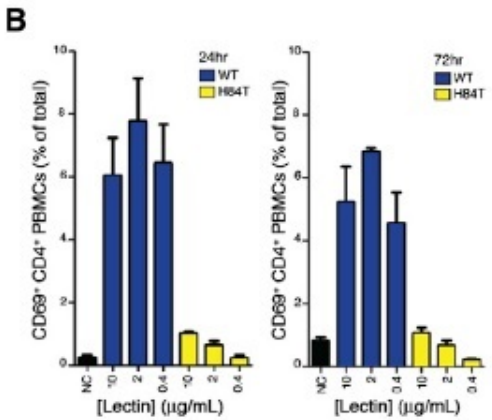
- Most Nature (etc) papers contain both levels

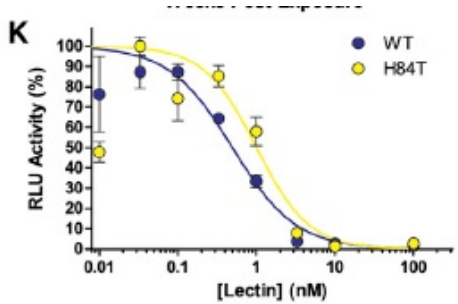
Exercise I: Using RStudio with a simple script and internal data I

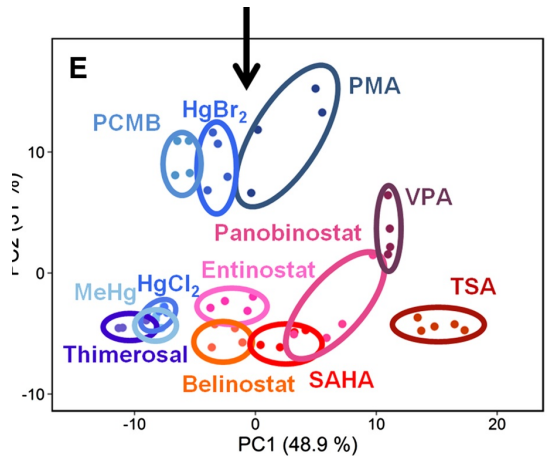
```
library(pairwiseCI) # a CRAN library or (bioconductor, github)
data(Oats) # two factor variety-by-nitro
Oats
apc <- pairwiseTest(yield ~ nitro, data=Oats,by="Variety", method="t.test")
apc
summary(apc) # generic function summary()
```

Motivating Examples I

- Crude structure of biostatistic methods in MolBio:
  i) testing, ii) modeling, iii) classification
- Examples

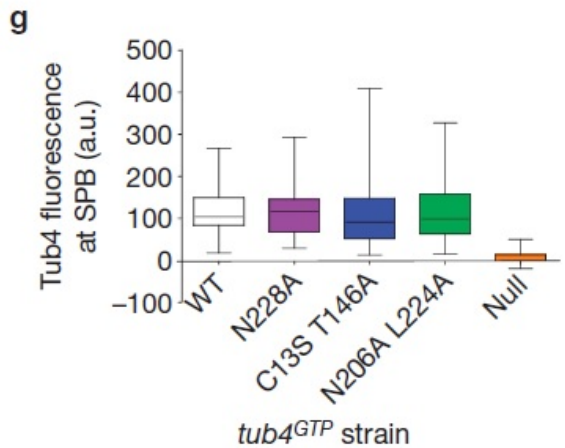






Motivating Examples II

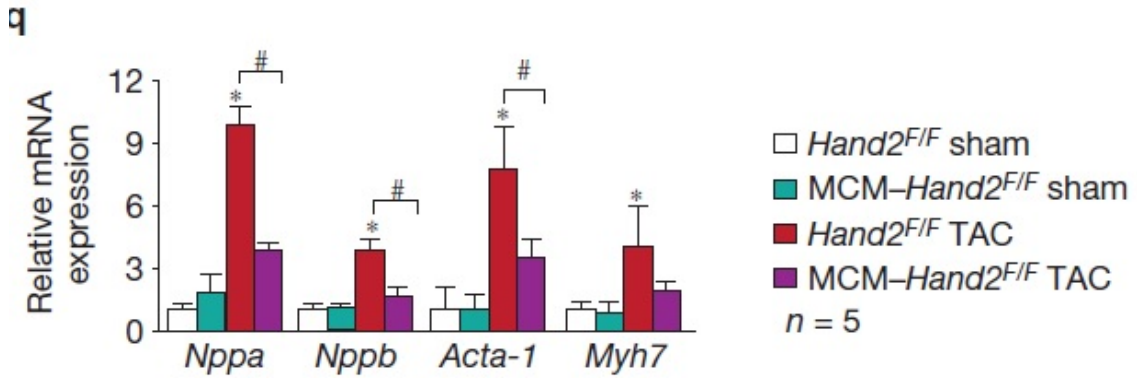
- Gombos et al. (GTP regulates the microtubule nucleation activity)
  Boxes represent upper and lower quartiles with a line at the median; whiskers extend from the min to max. (n=50 cells per strain per experiment) 5 independent experiments). Two-tailed, unpaired Student's t -test was used to obtain P values. tub4GTP strains were not significantly different from the wild type (P > 05) in contrast to the null mutant (P=0.0001).



- Q: i) what is the exp. unit: cell or exp? ii) no raw data in the boxplots

Motivating Examples III

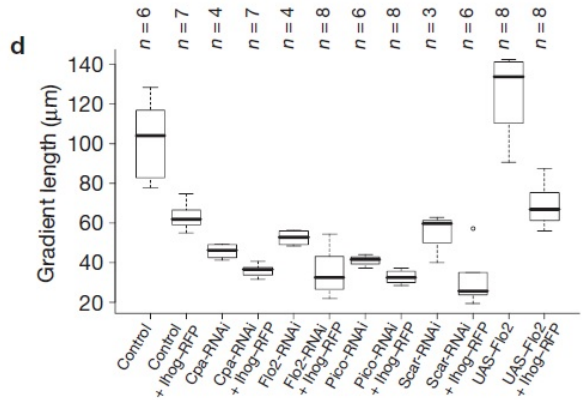
- Dirkx et al (Nfat and miR-25 cooperate to reactivate the transcription factor Hand2 in heart failure)
  Quantitative real-time PCR analysis of Nppa, Nppb, Acta1 and Myh7 in hearts from tamoxifen-treated Hand2FF and MCMHand2F mice after sham or TAC surgery; n, number of hearts. $*P < 0.05$ *versus corresponding control group; hashP* < 0.05 *versus experimental group (error bars are sem).*
  Q: error bars as sd or sem?



- Q: i) one- or two-way layout?, ii) adjustment against multiple testing? iii) raw data (both biol and techn.)?

Motivating Examples IV

- Bischoff et al. (Cytonemes are required for the establishment of a normal Hedgehog morphogen gradient)

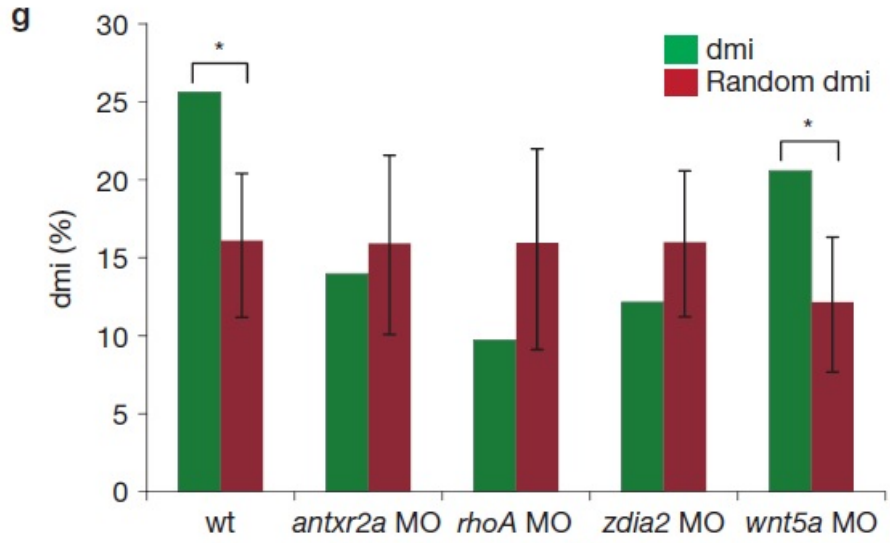


Box plot comparing gradient length between control discs and treatments. As for cytoneme length, there were significant differences between UAS.IhogRFP and the four RNAi treatments (Kruskal-Wallis test, $P < 0.001$ *and pairwise Wilcoxon rank sum test,* $P < 0.01$*)*

- Q: i) why nonparametric test, ii) why Kruskal-Wallis before Wilcoxon tests?, iii) multiplicity adjustment?

Motivating Examples V

- Castanon et al (Anthrax toxin receptor 2a controls mitotic spindle positioning)



- Q: i) is it an one-way layout?, ii) only comparison dmi/random, but against wt?

Recent challenges I

1. The standard measure for success, **the p-value**, is questioned or even banned. A simple alternative is not available yet
2. (Most) experiments in MolBio with **small sample sizes** n_i, e.g. 2,3,...,10.
   But I) (almost all) stat. approaches (and its software) base on large n_i (up to $n_i \to \infty$).
   II) The power trap: the smaller n_i, the more non-sign p-values.
   Today focusing on small n_i tests
3. The reproducibility crises
4. The power approach
5. Open-source software needed (useR!)
6. Open data sources (using standardized formats)

Editors recommendations I

Plant Cell or most recent Molecular Plant (2019 Jan 7) or the EMBO Journal

- Authors must ensure that appropriate experimental **design** and statistical analyses are carried out where necessary to support conclusions, such as large-scale analyses and experiments related to effects of various treatments, environmental conditions, or genotype on plant growth and development...
- In evaluating experiments, we will consider whether there is a clear and complete description of each experiment; whether technical, biological, and/or experimental **replicates** should have been used and if clearly defined; what statistical analysis has been performed and if clearly described; and where necessary, whether a **multiple comparison** correction has been used to control for Type I family-wise error.

Editors recommendations II

- A good understanding of the experimental design and any statistical analyses performed is critical both for proper interpretation of data and independent verification of claims
- Authors are encouraged to **involve statisticians** in both the design and analysis of experiments, to whatever extent is necessary, to properly interpret results.
- Figures and tables should include clearly defined **error bars** where appropriate

Editors recommendations III

Nature

- Every article that contains statistical testing should state the name of the statistical test, the **n value** for each statistical analysis, the comparisons of interest, a justification for the use of that test (including, for example, a discussion of the **normality** of the data when the test is appropriate only for normal data), the alpha level for all tests, whether the tests were **one-tailed or two-tailed**, and the **actual P value** for each test (not merely "significant" or "< 0.05
- It should be clear what statistical test was used to generate every P value. Use of the word "significant" should always be accompanied by a P value; otherwise, use "substantial," "considerable," etc.
- Data sets should be **summarized** with descriptive statistics, which should include the n value for each data set, a clearly labeled measure of center (such as the mean or the median), and a clearly labeled measure of variability (such as standard deviation or range). Ranges are more appropriate than standard deviations or standard errors for small data sets. **Graphs** should include clearly labeled error bars.

Editors recommendations IV

- Authors must state whether a number that follows the sign is a **standard error (s.e.m.) or a standard deviation (s.d.)**.
- Authors must justify the use of a particular test and explain whether their data conform to the assumptions of the tests. Three errors are particularly common: **Multiple comparisons:** When making multiple statistical comparisons on a single data set, authors should explain how they adjusted the alpha level to avoid an inflated Type I error rate, or they should select statistical tests appropriate for multiple groups (such as ANOVA rather than a series of t-tests). **Normal distribution**: Many statistical tests require that the data be approximately normally distributed; when using these tests, authors should explain how they tested their data for normality. If the data do not meet the assumptions of the test, then a non-parametric alternative should be used instead. **Small sample size:** When the sample size is small (less than about 10), authors should use tests appropriate to small samples or justify their use of large-sample tests. There is a checklist available to help authors minimize the chance of statistical errors.

Editors recommendations V

Key words:

- design
- technical, biological, and/or experimental replicates
- multiple comparison
- error bars, sem or sd
- n value, small sample size
- normality
- one-tailed or two-tailed test
- P value
- summarized data
- Graphs
- and ... involve statisticians

Nature's Statistics for biologists: Points of significance I

A collection of short papers: easy to understand and briefly written what to do. Unfortunately, not how to do (software solution missing) (in blue.. partly covered by this course):

- Importance of being uncertain - estimate uncertainty
- Significance, P values and t-tests: the concept of significance
- Design I: Power and sample size
- Visualizing samples with box plots
- Comparing samples I - two-sample t-test
- Comparing samples II - Adjustment for large numbers of tests
- Nonparametric tests - to robustly compare skewed or ranked data
- Design II: paired design
- Design III: Replication - Technical replication
- Design IV: Nested designs
- Design V: Two-factor designs
- Design VI: Sources of variation - randomization, blocking and replication
- Design VII: Split plot design

Nature's Statistics for biologists: Points of significance II

- Bayes theorem
- Bayesian statistics
- Sampling distributions and the bootstrap
- Bayesian networks - Model interactions between causes and effects
- Association, correlation and causation
- Simple linear regression
- Multiple linear regression
- Analyzing outliers: influential or nuisance
- Logistic regression
- Classification evaluation
- Model selection and overfitting
- Regularization
- P values and the search for significance
- Interpreting P values
- Tabular data

Nature's Statistics for biologists: Points of significance III

- Clustering
- Principal component analysis
- Classification and regression trees
- Ensemble methods: bagging and random forests
- Machine learning: a primer
- Machine learning: supervised methods
- Statistics versus machine learning
- The curse(s) of dimensionality
- Design VIII: Optimal experimental design

Using RStudio: simple data import I

- Focusing on common lab data, usually organized in EXECL
- Step I: within *.xls: organizing as complete flat file format

Using RStudio: simple data import II

Tabelle1

mutants	plate		concentration/µmol	Sepal lenght /mm
wt		1	0	12,5
		2	0	3,4
		3	10	missing
		4	10	0
		5	50	188
		6	50	
tr25&f1f14 abs		1	0	47,77
		2	0	23,8
		3	10	<3
		4	10	44,6
		5	50	123
		6	50	

Seite 1

Tabelle2

mutants	plate	conc	Sepalle	
wt	1	0	12,5	
wt	1	0	3,4	
wt	3	10	NA	
wt	4	10	0	
wt	5	50	188	
wt	6	50	NA	
tr2511	11	0	47,77	
tr2511	12	0	23,8	
tr2511	13	10	<3	
tr2511	14	10	44,6	
tr2511	15	50	123	
tr2511	16	50	NA	

Seite 1

Using RStudio: simple data import III

1. Each line must be filled in completely
2. The plate number must be unique
3. Variable names : short, without special characters
4. No text in numeric variables
5. Zero, no value, or detection limit uniquely coded

- Import as *.csv file format (simple, easy)
- A working example from Structural Plant Biology Laboratory, Univ Geneva

```
setwd("D:/externals/_CUSOLausanneApril2019") # use your folder here
f2c<-read.table("Genf2019Fig.csv", sep=";", dec="," ,header=TRUE)
f2c
```

Using RStudio: simple data import IV

- In xls file, already statistical design and model can be displayed:
  1. One-factor, two-factor, multi-factor layout
  2. Completely randomized vs. nested, vs. technical replicated
  3. Single endpoint, multiple endpoints, repeated measures
  4. Quantitative covariate, eg. dose $0, 1, 2, 10, 100$ or Dose $C-, 1, 2, 10, 100, C+$ or Combi $C, 1+1, 2+10, 10+100, C+$
- Data manipulation by non-stats users commonly in xls. But some relevant issues within R (out of inf)
  - subset
  - mean of replicates
  - repeated times as factor
  - xxxx

Visualization of grouped and clustered data I

- Visualization depends on data structure, statistical approach, story to tell (see the above ETH-example): boxplots AND cluster dendrogram
- Today common style of data presentation:
  Tables containing group-specific means, standard deviations, and sample sizes, commonly for multiple endpoints

Table 1 Effect of tBOOH on cellular integrity, redox status, and Ca^{2+} levels

	DMSO[a]	*t*BOOH[b]
Cellular viability (% of total cells)	96 ± 3	62 ± 3*
LDH release (% of total cell content)	32 ± 2	42 ± 3*
ATP cellular content (μmol/10^6 cells)	48 ± 6	43 ± 3
MDA cellular content/nmol/mg prot.)	1.3 ± 0.2	8.0 ± 0.3*
Total glutathione (mg/mg of protein)	27 ± 2	17 ± 1*
GSSG-to-total glutathione ratio (%)	3.1 ± 0.1	4.1 ± 0.2*
Cytosolic free Ca^{2+} (nM)	119 ± 11	3029 ± 474*

Cells were pretreated for 15 min either with *t*BOOH (500 μM) or with DMSO (controls)

LDH lactate dehydrogenase, *ATP* adenosine triphosphate, *MDA* malondialdehyde, *GSSG* oxidized glutathione

* $p < 0.05$ versus DMSO, for $n = 4$–10

Figure: Example for data summary table [TPB+14].

Visualization of grouped and clustered data II

- Bar charts, e.g., [DSP+12] used barcharts (including SEM) and letters of significance (Figure 2).

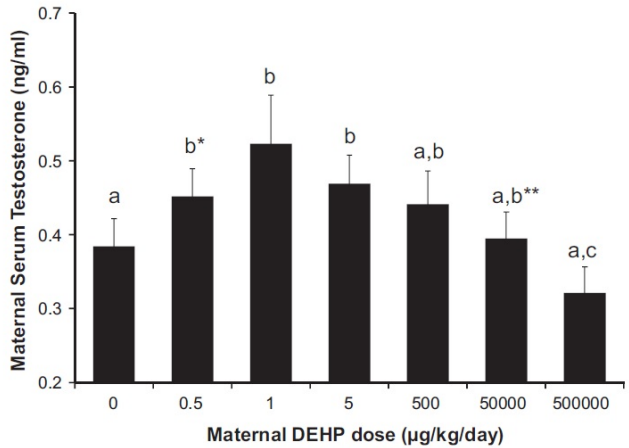


Fig. 2. Effect of different doses of DEHP on maternal serum testosterone concentrations on GD 18 (ANOVA on log-transformed data, $P < 0.05$). Values represent the mean ± SEM. Treatments with the same letter are not significantly different from each other but are statistically different from groups with other letters. b*, $P = 0.07$ relative to controls; b**, $P = 0.09$ relative to 500,000 group. Sample sizes were: oil $n = 20$; 0.5 µg, $n = 9$; 1 µg, $n = 11$; 5 µg, $n = 12$; 500 µg, $n = 13$; 50 mg, $n = 16$; 500 mg, $n = 17$.

Figure: Example of bar charts [DSP+12].

Visualization of grouped and clustered data III

- **Two major drawbacks:** i) they assume normally distributed data (and we know how often this is violated in real data) and ii) they do not allow access to the individual data.
- Individual datapoints have a special meaning in small n_i experiments, because sometimes the relevant information is contained just in a few extreme values —not necessarily in means.
- Rhodes et al. [RLG$^+$12] visualized just group-specific individual data for 20 rats together with the geom-mean

Visualization of grouped and clustered data IV

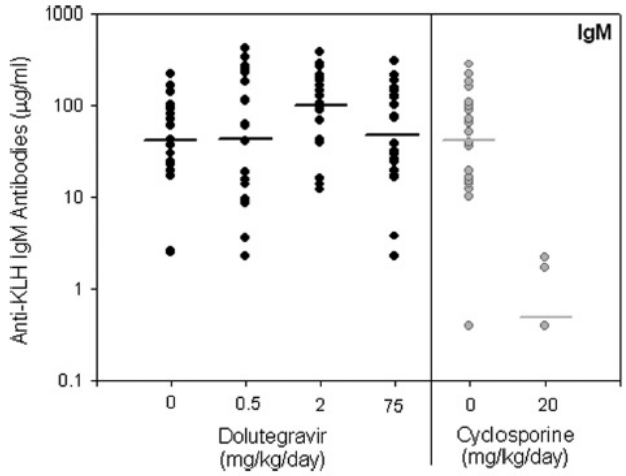


FIG. 2. IgM anti-KLH TDAR in juvenile rats treated with dolutegravir. Sera were harvested 5 days post-KLH immunization and assessed for anti-KLH IgM antibodies by quantitative electrochemiluminescent immunoassay. Circles represent individual animals (n = 20/group) with geometric mean indicated by the bars. There were no detectable dolutegravir-related effects on the anti-KLH IgM antibody response when juvenile animals received daily treatment. In rats given the positive control for immunosuppression (20 mg/kg/day of cyclosporine), there was a significant decrease ($p < 0.001$) in the level of anti-KLH IgM antibodies. Eighteen of 20 rats given 20 mg/kg/day of cyclosporine tested below the IgM assay lower limit of quantification and were assigned a value of 0.4 µg/ml for the purpose of calculating group geometric mean.

Figure: Example of individual data representation [RLG+12].

Visualization of grouped and clustered data V

- Measures of statistical significance
  - rejection/non-rejection of H_0 (whereby letters indicate non-distinguishable treatment groups)
  - rejection of H_0 for three α levels (0.05, 0.01, 0.001) visualized by stars $*, **, ***$,
  - p-value (p),
  - **confidence intervals**
- Nowadays, the use of p-values in medical journals was **seriously criticized** [Nuz14, Rot14, HCEVD15] up to banishing [ICM, TM15].
- Point I: *'... prior to publication, authors will have to remove all vestiges of the NHSTP (p-values, t-values, F-values, statements about significant differences'*

Visualization of grouped and clustered data VI

- Point II: *'Are any inferential statistical procedures required? No, because the state of the art remains uncertain. However, .. will require strong descriptive statistics, including effect sizes. We also encourage the presentation of frequency or distributional data when this is feasible'* - Extreme view in psychology
- Point III: Effect size $\mu_i - \mu_j$ ubiquitously in biomedical research. Is this appropriate? See Comet assay in the appendix $\Rightarrow$ extreme values may be important $\Rightarrow$ visualize it!

Visualization of grouped and clustered data VII

- Our example; using library(toxbox)

```
library(toxbox)
boxclust(data=f2c, outcome="value", treatment="type",xlabel="Genotypes
         option="color", hjitter=0.3, legpos="none", psize=1, printN=F
# PseudoN=8, but randomized n=4
boxclust(data=f2c, outcome="phos", cluster="replicate",treatment="type
         option="color", hjitter=0.3, legpos="top", psize=1, printN=FA
```

Modifying boxplots I

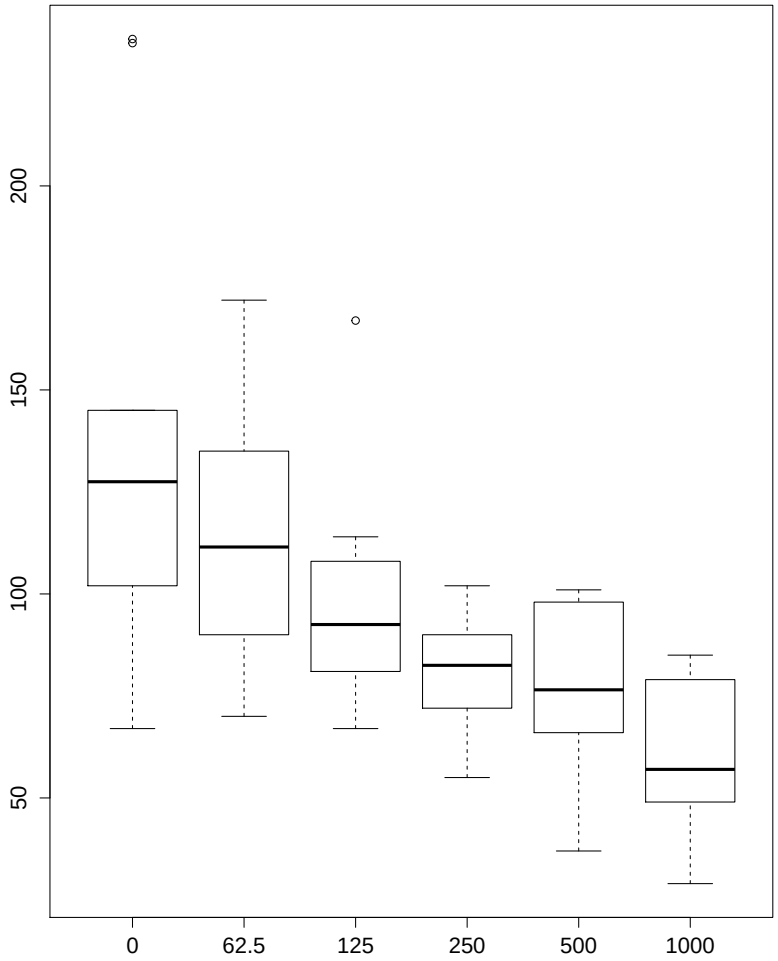
- A boxplot typically uses nonparametric measures of location and scale, namely **median** and **interquartile range** ($IQR = q_{0.75} - q_{0.25}$) as well as an **outlier rule** (represented by whiskers), e.g., $k * IQR; k = 1.5$ (notice, $q_{0.75}$ is the 75% percentile).
- Remember: what is the median? How it differ from $\bar{x}_i$?
- The boxplot provides simple information on group-specific location, variance, and asymmetry of distribution as well as existence of extreme values.
- For grouped data a specific jittered boxplot was developed in the `R library(toxbox)` and a Shiny-App (see details [PPR15])
- Grouped data: i) implies a **qualitative factor**, ii) *quantitative covariate* is commonly considered by regression models, iii) dose (with 0,10,50,100 mg/kg) can be factor or covariate. Discussion!

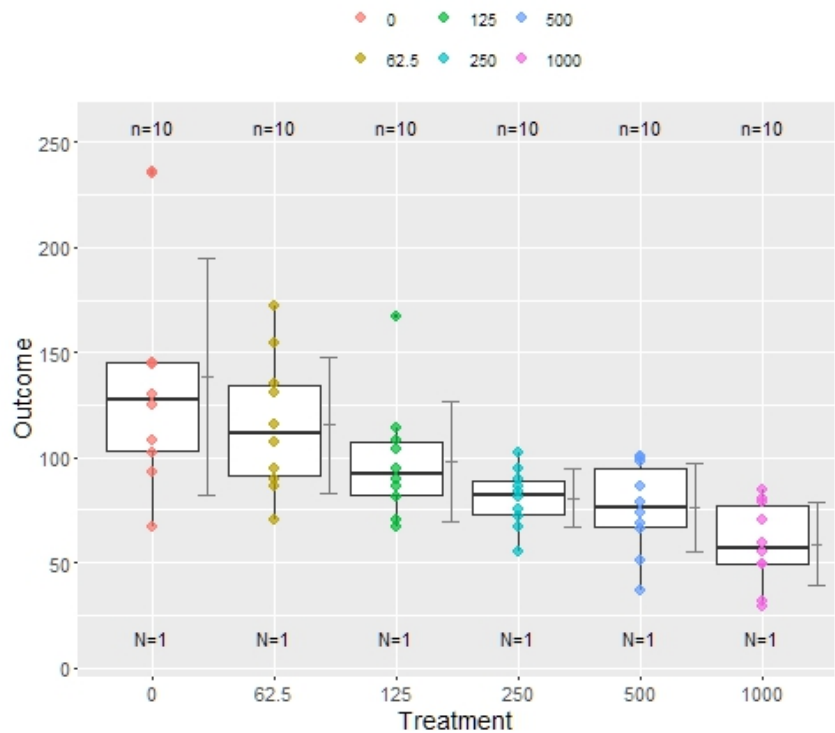
Modifying boxplots II

- Modification I: no whiskers and outlying values above/below. Outlier identification or even elimination may be problematic in small n_i experiments
- Modification II: adding $\hat{x} \pm SD$ as parametric measures for location and scale. Allows some insight into distribution, e.g. at least symmetric or not (but see n_i limitation below)
- Modification III: Include all raw data. Feasible for rather small sample sizes up to thousands (see case study genetics).

Modifying boxplots III

- Example: old vs. new



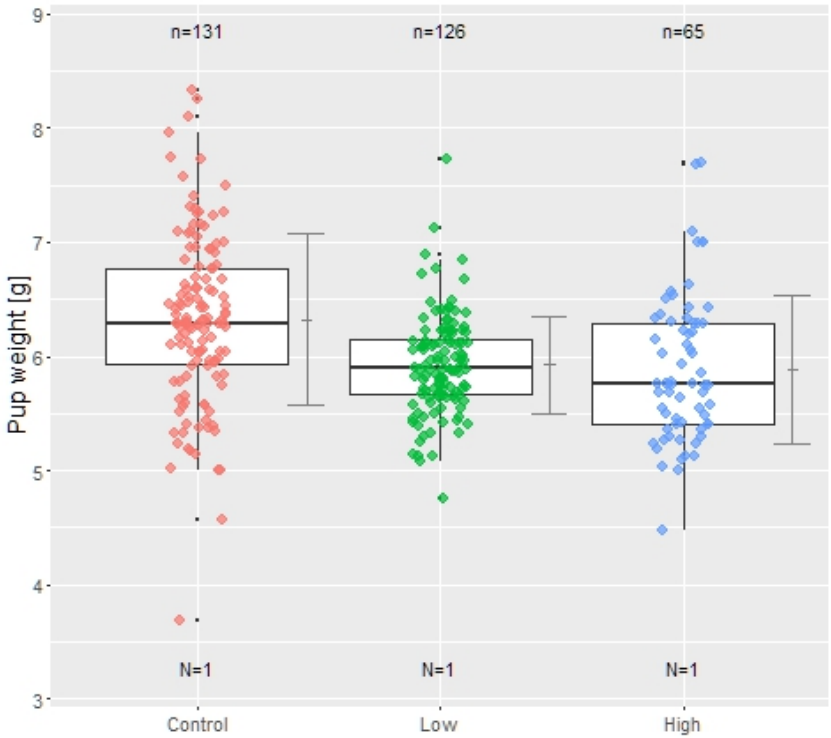


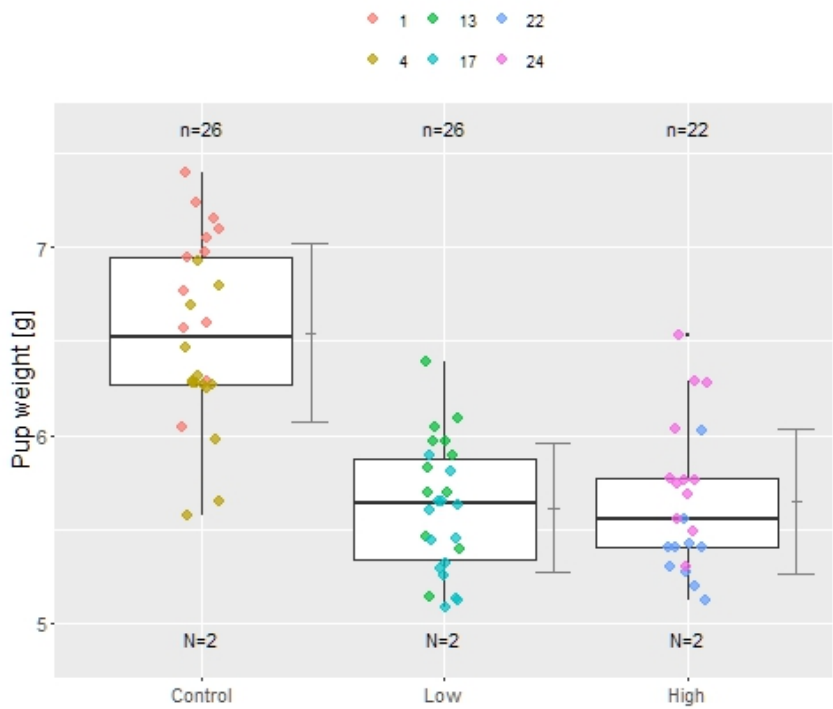
Modifying boxplots IV

- Modification IV: Highlighting of cluster and covariates
  - **Clustered** data: i) natural, such as pups within a female, ii) by design: technical replicates, such as tanks within a treatment groups in aquatic bioassay
  - Commonly: i) ignoring clustered data is a biased analysis (e.g. per-foetus analysis), ii) means within cluster as pseudo observation can be an appropriate approximation or NOT

Modifying boxplots V

- Example: Pup weight data: a) individual pup weights without females (=cluster) structure vs. b) with females structure (color)

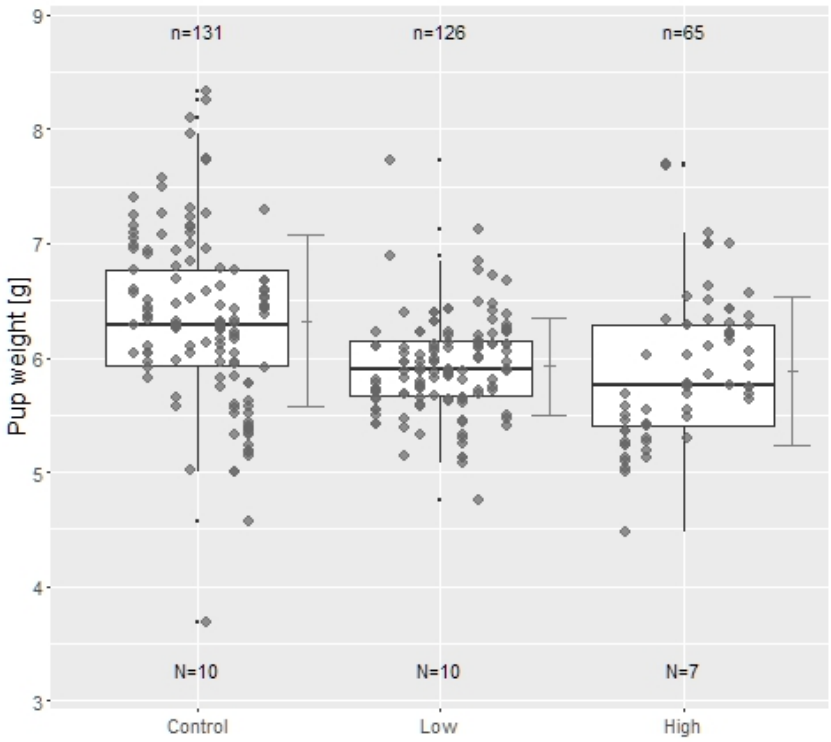


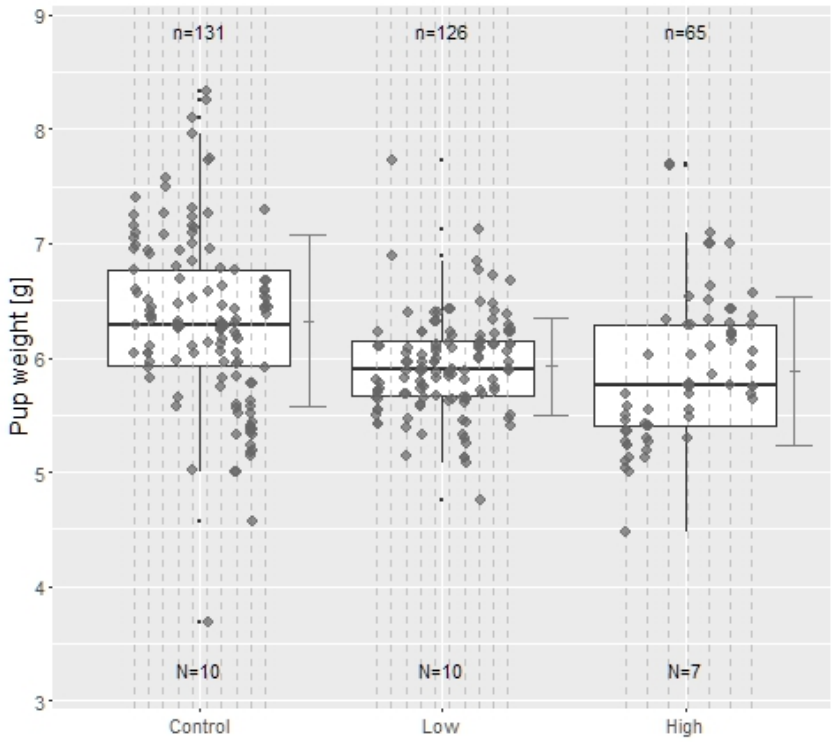


Notice: two sample sizes N_i, n_{ij}

Modifying boxplots VI

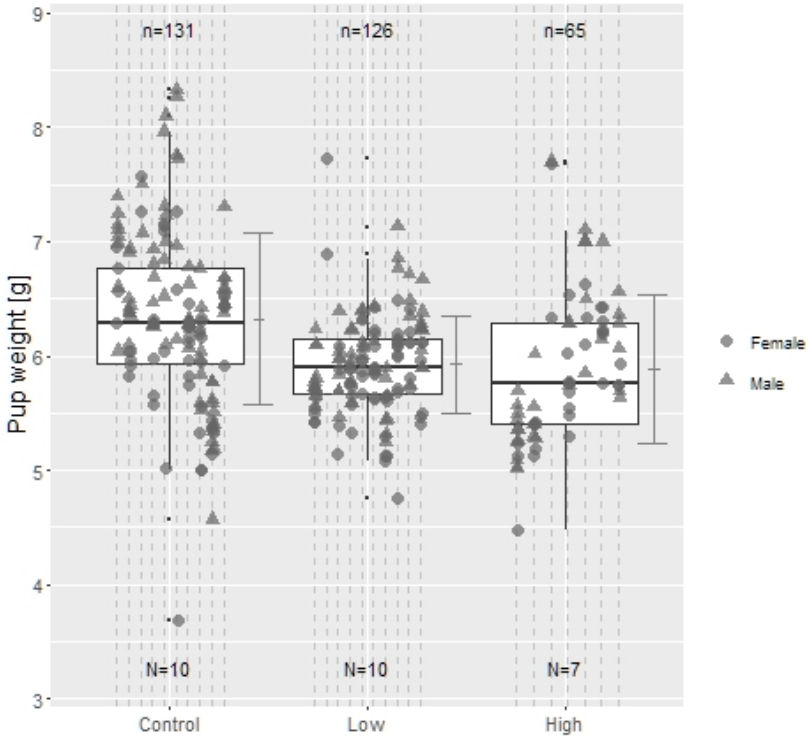
- Example: Pup weight data: c) with females structure instead color Cleveland plot, d) at lines (with additional jittering for equal value differentiation)

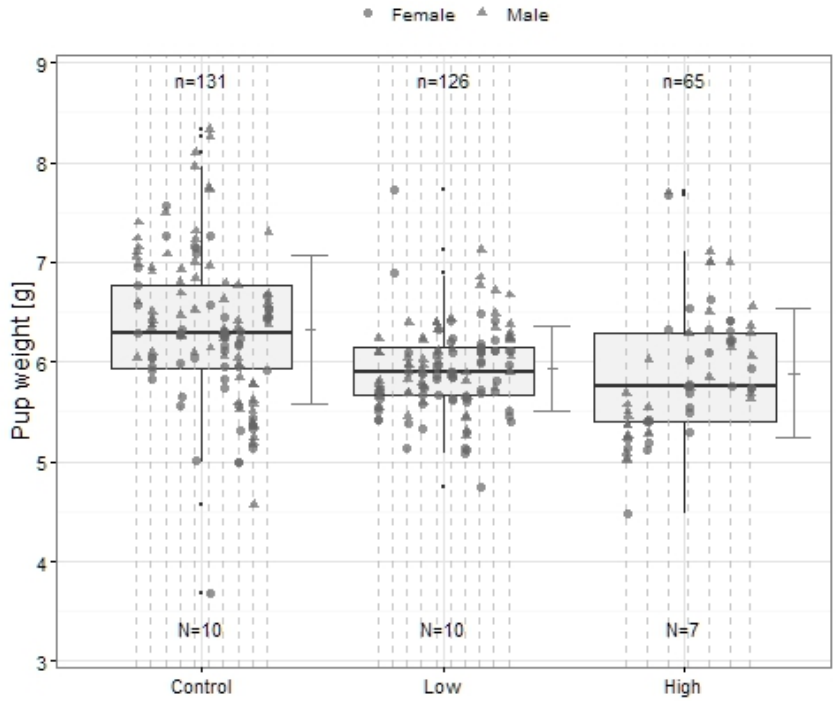




Modifying boxplots VII

- covariate: a secondary factor, e.g. sex (or time, location, organ, metabolic activation)

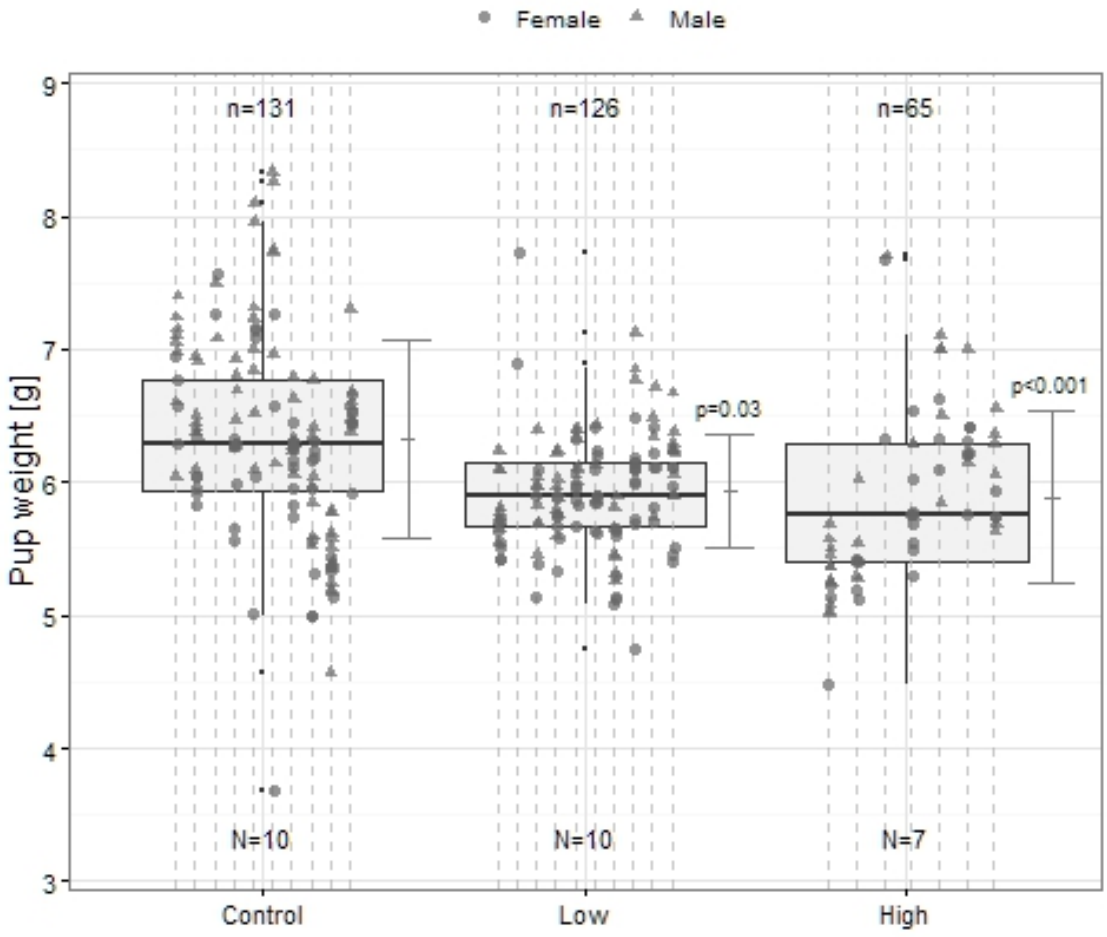




Modifying boxplots VIII

- Modification V: Including p-values for comparison against control(s)
  Which p-values?
  - **Primary:** Comparing against negative control
  - Three versions:
    1. Dunnett-type, control FWER: $f+$-rate controlled for exp, but increased $f-$-rate
    2. control CWER, i.e. Welch-t-tests (why Welch?) pairwise against control: $f+$-rate controlled for individual comparison only, therefore decreased $f-$-rate
    3. test on significant toxicity [DDZ11]: non-inferiority test with 80% threshold inhibition
  - Syntax flexible `pneg=c(0.03, 0.0004)` any p-value: k for k treatments against control
  - **Secondary:** Comparing against positive control (commonly non-inferiority test)
  - Syntax flexible `ppos=c(0.3, 0.4)`

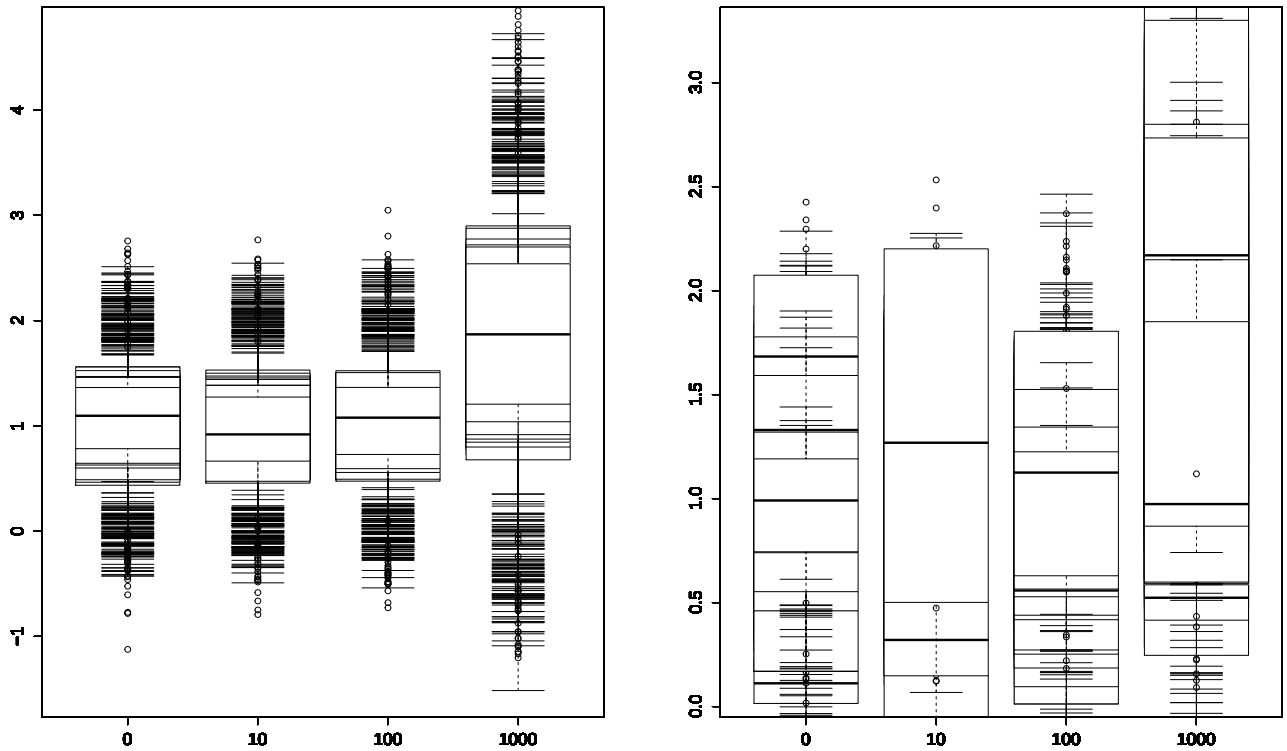
Modifying boxplots IX



- Modification VI: Including normal range see normal ranges below

Modifying boxplots X

- Boxplots for small sample sizes



Conclusion: Be carefully with conclusions drawn from 'small n_i boxplots' (rule: $n_i > 5$)

Modifying boxplots XI

- Two Shiny Apps: i) `https://lancs.shinyapps.io/ToxBox/`
  ii) BoxPlotR in Nature Methods

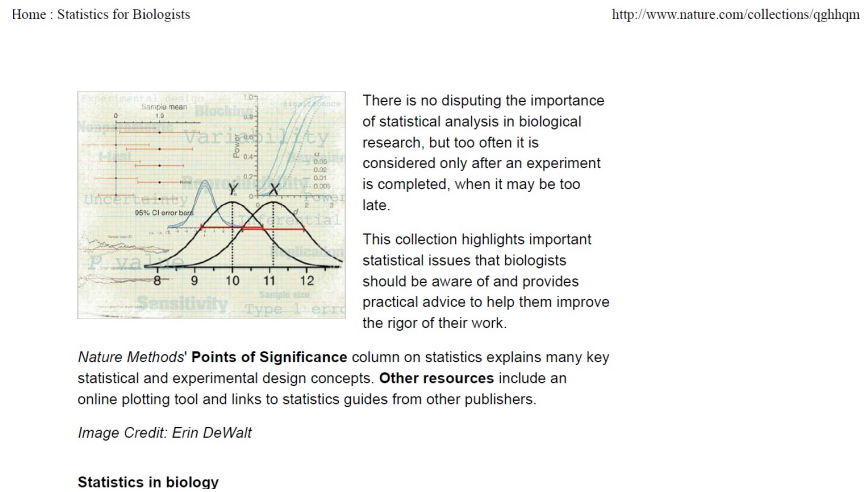
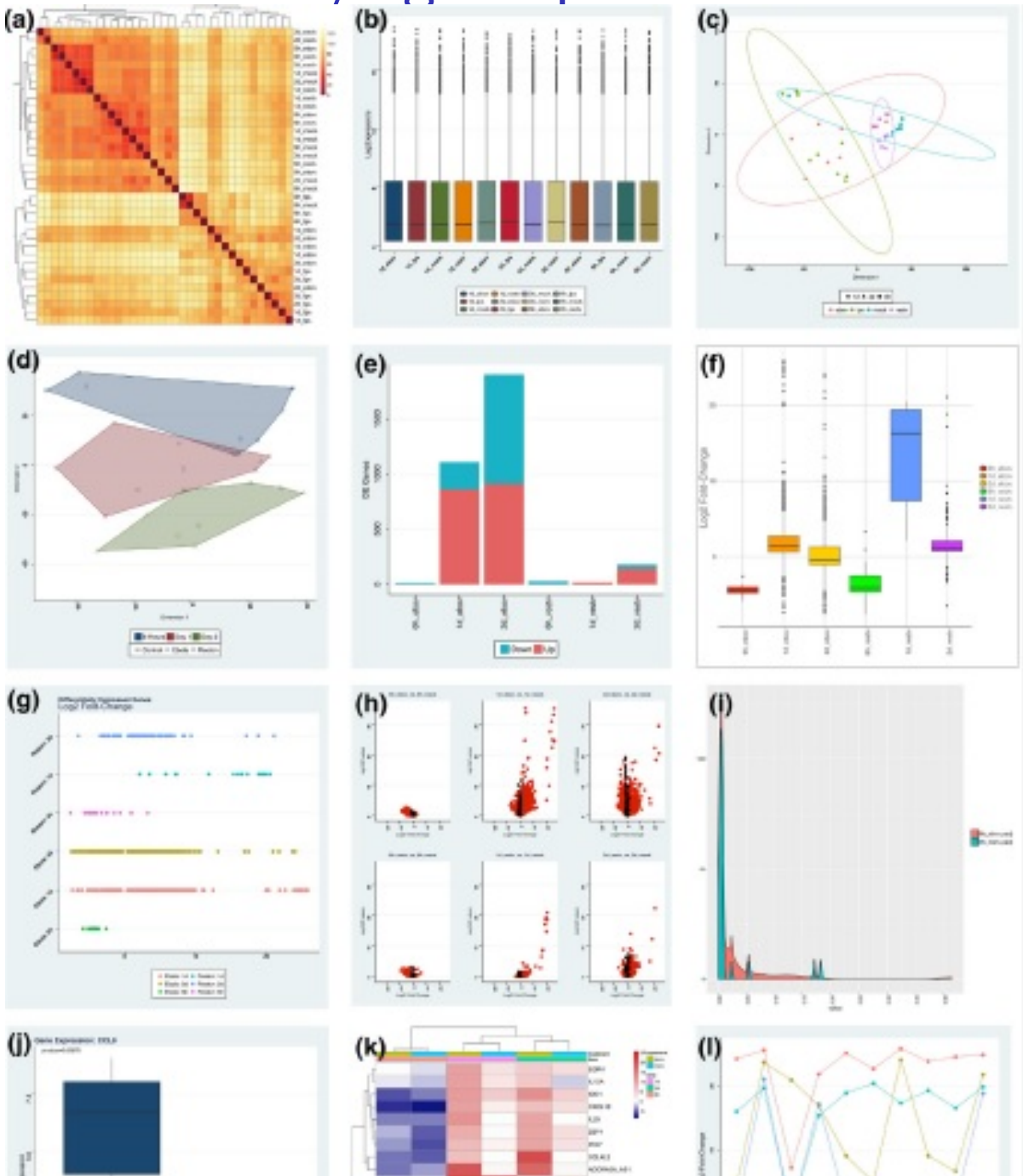


Figure: Stats Homepage Nature Methods.

See two further real data examples in the appendix

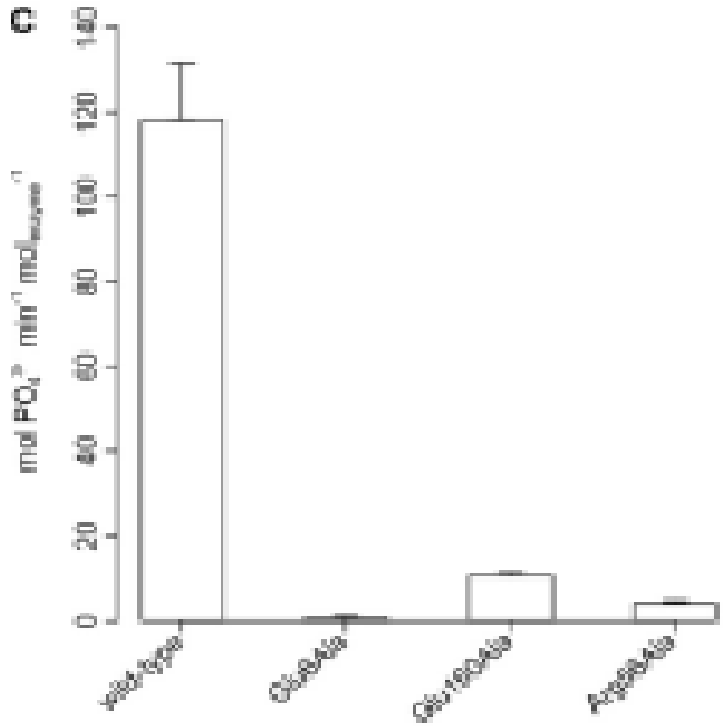
- Search, Ask, Inform (during conferences) for new visualization tools, e.g. recent package for expression data [PCG+19]

Modifying boxplots XII



Exercise II: Import xls data file, generate boxplots I

Use exa2.xls (Data from [MTH15] Fig 6)

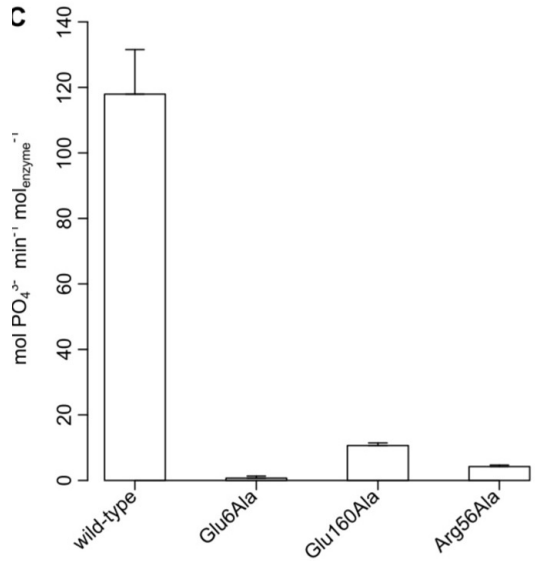


Take home visualization I

- Mandatory for publication: explain your story strongly summarized by raw data and appropriate statistics in a figure
- Depends on background (dendrogram, scatterplot, correlationplot,...)
  In mol. biology grouped data common: boxplot

Two-sample tests I

- A toy example, extracted from [MTH15]



- Common in paper: summary data only. For re-analysis generate random experiment
- Generate random normals data (notice: seed guarantee the same random variables)

Two-sample tests II

- Today rather elementary (a better version available): `simdat` contains pseudovalues, similar to raw values

```
set.seed(170549)
nsample1=10; nsample2=5;nsample3=5; nsample4=5;
mue1=104.2; mue2=3.5; mue3=31;mue4=7.9;  sigma1=19.5; sigma2=18.5; sigma3=28.5
u=rnorm(nsample1,mue1,sigma1); v=rnorm(nsample2,mue2,sigma2);x=rnorm(nsample3,
z=rnorm(nsample4,mue4,sigma4); # gaussian distr
ni<-c(nsample1,nsample2, nsample3, nsample4);
wt<-"wt"; glu6<-"glu6"; glu16<-"glu16"; arg<-"arg"
dose<-rep(c(wt,glu6,glu16,arg),ni)
PO4<-c(u,v,x,z); grp<-as.factor(rep(1:4, ni)); gr<-as.numeric(grp);
simdat<-data.frame(PO4=c(u,v,x,z), group=as.factor(rep(1:4, ni)),dose);
library(toxbox)
boxclust(data=simdat, outcome="PO4", treatment="group", printN="FALSE")
subdat<-droplevels(simdat[simdat$group %in% c("1", "3"),])
```

- Only **Wt vs. Glu160**: subdat R data manipulation

- Q: is the p-value of the t-test appropriate?

Two-sample tests III

```
Welch Two Sample t-test
data:  PO4 by dose
t = -4.6987, df = 8.4858, p-value = 0.001316
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -89.97262 -31.12726
```

- Q: Why Welch-t-test? Why confidence limits? Why one-sided tests? Can we check on normal distribution before using the test? Why not ratio-to-control for 2-fold rule? Why not confidence interval for ratio-to-control?

```
Ratio t-test for unequal variances
data:  wt and glu16
t = -2.6241, df = 4.1777, p-value = 0.05603
alternative hypothesis: true ratio of means is not equal to 5
95 percent confidence interval:
 1.378229 4.909020
sample estimates:
   mean wt mean glu16   wt/glu16
109.988851  49.438911   2.224743
```

Two-sample tests IV

```
> sci.ratioVH(PO4~dose, data=subdat, base=2)
Simultaneous 95% confidence intervals (heteroscedasticity)
Degree of freedom:4, quantile:2.776
          estimate  lower  upper
glu16/wt   0.4495 0.1859 0.7482
```

- Robustness of t-test. Alternative non-parametric test using ranks (explain!)
- Problem 1: for the common small sample sizes, no meaningful test on gaussian distribution, or not (and variance homogeneity, or heterogeneity) exists
- Recommendation 1:
  - i **trusting the robustness of t-test** (and related tests) or
  - ii using non-parametric tests or
  - iii modeling the distribution (e.g. quasi-Poisson)
- Using common t-test: p-value 0.0000012

Two-sample tests V

- Q.: Can we conclude a significant PO4 **reduction**, although the t-test was 2-sided (change only). **Yes, we can** (explain it)
- Look on boxplot: a serious variance heterogeneity occur (a "good" one ...)
- Problem 2: t-test is not robust against variance heterogeneity, particularly when $n_C >>, << n_D$. Explain why!
- **Use the Welch-t-test instead** Realize the bias of common t-test!
- No powerful test on variance heterogeneity for small n_i exists (as a pre-test)
- Therefore Recommendation 2: Use always the Welch-t-test (only minor power loss for homogeneous variances)
- Problem 3: Is the non-parametric Wilcoxon-test appropriate, e.g. when data are skewed or outlier occur? Counter-facts:
  1. WMW does not test mean differences (even not median differences). It tests stochastic order- hard to interpret (but see below relative effect size)

Two-sample tests VI

2. WMW is NOT robust against heterogeneous variances. Recently, a Behrens-Fisher modification is available using relative effect size `npar.t.test` (see below)
3. WMW is asymptotic only, i.e. requires large n_i, e.g. $n_i > 10$. Permutative modifications exist, but with disadvantages, e.g. conservativeness for rather small n_i (a serious problem in molecular biology)
4. WMW is defined for continuous data only, i.e. NOT for tied data (adjustments, permutative version)
5. (confidence intervals for WMW not common available)
6. Summary: for the common design with $n_i = 3...10$ and possible variance heterogeneity and tied data... **no standard carefree** non-parametric test available. Recommendation 3: Be careful when using and interpreting common WMW-test. Notice, an improved version, for relative effect size, exist [KH12b] - in a minute

Two-sample tests VII

- Coming back to Recommendation 2: Use always the Welch-t-test. Only two exceptions: i) serious outlier(s), ii) rather small n_i and approx variance homogeneity (in boxplots): use so-called t-test with common variance estimator.
- Problem 4: Outliers?
- Recommendation 4: Do not use formal outlier tests and be carefully when eliminate extreme values
- Problem 5: Using confidence intervals instead of of p-values?
  - Advantage of a p-value: measure of Popper's falsification principle
  - Disadvantage of a p-value:
    i) a probability $[0, 1]$ hard to interpret and rather skewed to zero
    ii) It is a monotonic function of n_i :$\Uparrow p \Leftrightarrow \Downarrow n_i$. Our example $p_{n_{500}=10} = 0.0027$; $p_{n_{500}=8} = 0.0044$; $p_{n_{500}=6} = 0.0064$,
    iii) commonly for a point-zero null-hypothesis $H_0 : \mu_T - \mu_C = 0$, but in biology we are never interested in tiny to zero true differences

Two-sample tests VIII

- A better alternative is the use of **effect sizes and their confidence intervals**
- Effect sizes for continuous data: $\mu_T - \mu_C$ or μ_T/μ_C
- Other: hazard ratio, odds ratio, risk ratio,....
- Confidence intervals (CI) for these measures by re-formulating the t-test: $\frac{\bar{x_T} - \bar{x_C}}{SD\sqrt{(2/n)}} = t_{df,1-p=min(\alpha)}$ into $(\mu_T - \mu_C) \pm SD\sqrt{(2/n)}t_{df,1-\alpha/2}$
- Sometimes, interpretation is easier as percentage change, e.g. k-fold rule in mutagenicity assays, and a confidence interval for μ_T/μ_C is recommended (switch from additive into multiplicative model). A bit more complicated (no formula here) according to Fieller [Fie54]
- Confidence interval approach for superiority and non-inferiority. Explain!

Two-sample tests IX

- Still the width of the confidence interval, i.e. $SD\sqrt{(2/n)}t_{df,1-\alpha}$ is a function of sample size, i.e. larger sample sizes, smaller (more significant) width (analogously the **smaller p-values**)- independent of effect size and variance.

- The sample size must be defined a-priori: i) guidelines, ii) power

- The common mis-understanding between: *statistical significance and biological relevance* results from inappropriate use of p-values, testing point-zero H_0, and un-designed experiments.

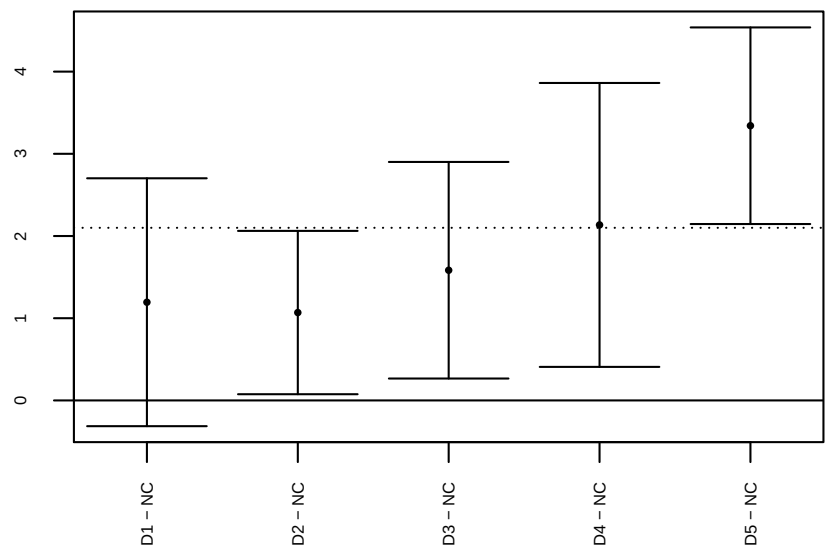
- **Therefore, biological experiments should be characterized by an appropriate effect measure and its confidence interval (two-sided)** or confidence limit (one-sided).

- The (possible) duality to significance test should be avoided by a significant/no significant decision whether or not 0 (more general the value of H_0) is contained in the interval.

Two-sample tests X

Hereby can five outcome types distinguished

- ★ statistically not significant `D1-NC`
- ★ significant without biological relevance `D2-NC`
- ★ not significantly less than threshold `D3-NC`
- ★ probably biologically significant effect `D4-NC`
- ★ large biologically significant effect `D5-NC`



Two-sample tests XI

- The scenario *probably biologically significant effect* (`D4-NC`) is of particular interest, i.e., the effect size (here mean difference) is above the relevance threshold (i.e., certain biological relevance) and the lower confidence limit larger than zero (i.e., formal statistical significant).

- For this scenario the sample size can be estimated [KFG13]. For example, for an assumed effect difference $\delta = 3.0$, a standard deviation of $\sqrt{\sigma^2} = 2.3$, a false positive rate $\alpha = 0.05$, and a false negative rate $\beta = 0.2$ a sample size for the point-zero hypothesis (i.e., scenario of just statistical significance `D2-NC`), a sample size of $n_i = 8$ is needed (two-sided t-test with balanced sample sizes).
- For a scenario of significance and relevance, i.e., probably clinically significant effect (`D4-NC`), a sample size of $n_i = 10$ is needed where the point estimator must be at least 2.1 (dotted line). These sample size estimations can be performed by the package `WinProb` using the concept of win probabilities [Hay13].
- Is a threshold available a-priori, such as 2-fold rule for the Ames assay such an approach can be used. In the most cases the threshold is unknown, and the lower limit should be interpreted from a biolog. perspective.

Two-sample tests XII

- Recommendation 5: Use confidence intervals to claim both stat significance and biological relevance
- Recommendation 6: reporting a p-value of a test based on an un-powered experiment is **misuse of statistics**. Even worse interpretation: *although a significant reduction was found* ($p < 0.001$)*, it is without any biological relevance*

Two-sample tests XIII

- Problem 5: (mostly hidden) almost used the effect size $\mu_i - \mu_0$. Really appropriate?
- Recommendation 7: Use different effect sizes, appropriate for scale, model and interpretation
  - common: $\mu_T - \mu_C$
  - rare, but relevant (k-fold change) μ_T/μ_C
  - for proportions: DR, RR, OR
  - relative effect size [BM00],[RA08]:

$$p_{01} = \int F_0 dF_1 = P(X_{01} < X_{11}) + 0.5P(X_{01} = X_{11}).$$

    - ★ Let $R_{sj}^{(0s)}$ denote the rank of X_{sj} among all $n_0 + n_s$ observations within the samples 0 and s.
    - ★ The rank means can be used to estimate p_{0s}

$$\widehat{p}_{0s} = \frac{1}{n_0}\left(\overline{R}_{s\cdot}^{(0s)} - \frac{n_s + 1}{2}\right).$$

Two-sample tests XIV

- Related approximate $(1-\alpha)100\%$ one-sided lower confidence limits are:

$$\left[\widehat{p}_i - t_{\nu,1-\alpha}\sqrt{S_i};\quad\right],$$

- Effect size p_i is *win probability* [Hay13] I.e. Under $H_0 : p = 0.5$ under $H_A : p = 0$ or $p = 1$
- Example: using library(nparcomp)

```
      Sample Size
glu16  glu16    5
wt        wt   10
```

```
       Effect Estimator Lower Upper      T p.Value
1 p(glu16,wt)      0.98 0.918 1.042 16.971       0
```

Two-sample tests XV

- Problem 6: Alternative to both p-value and CI: Bayes factor?
- Recommendation 8: Use Bayes factors for main findings as surprise for editors. Details in the section below

Nonparametric tests: the Wilcoxon test I

- WMW-test widely used, to be robust against non-normal distribution and variance heterogeneity
- But, it assumes continuous, homomorphic distributions (ie robust for counts and var hetero are myths), requires large n_i
- It base on ranks, which make it robust against extreme values (and skewed distributions)
- Permutative version for small n_i available (library(coin))
- Relative effect size version for var het and counts available (library(nparcomp), including confidence intervals, including small n_i)
- Trick: perform both t and WMW-test, choose min(p) test (max-T-stats of correlated tests behind). If $min(p)^{WMW}$ look on data an explain why, otherwise report t-test (trusting its robustness)

Nonparametric tests: the Wilcoxon test II

- Example: a) WMW and Welch-test have similar p-values... report Welch-t

```
library(coin)
myW<- wilcox_test(PO4~ dose, data = subdat,  distribution = "exact", c
myT<-t.test(PO4~dose, data=subdat)# similar p-values
```

- Manipulated example with an outlier and rather unbalanced: Subdat

```
library(coin)
set.seed(170549)
nsample1=20; nsample2=5;nsample3=5; nsample4=5;
mue1=104.2; mue2=3.5; mue3=31;mue4=7.9;  sigma1=19.5; sigma2=18.5; sig
u=rnorm(nsample1,mue1,sigma1); v=rnorm(nsample2,mue2,sigma2);x=c(rnorm
z=rnorm(nsample4,mue4,sigma4); # gaussverteilung, aber auch mischverte
ni<-c(nsample1,nsample2, nsample3, nsample4);
wt<-"wt"; glu6<-"glu6"; glu16<-"glu16"; arg<-"arg"
dose<-rep(c(wt,glu6,glu16,arg),ni)
PO4<-c(u,v,x,z); grp<-as.factor(rep(1:4, ni)); gr<-as.numeric(grp);
Simdat<-data.frame(PO4=c(u,v,x,z), group=as.factor(rep(1:4, ni)),dose)
Subdat<-droplevels(Simdat[Simdat$group %in% c("1", "3"),])
boxclust(data=Simdat, outcome="PO4", treatment="group", printN="FALSE"
```

Nonparametric tests: the Wilcoxon test III

```
myW<- wilcox_test(PO4~ dose, data = Subdat,  distribution = "exact", c
myT<-t.test(PO4~dose, data=Subdat)# similar p-values
myW
myT
library(nparcomp)
npar.t.test(PO4~dose, data=Subdat, method = "t.app",
            alternative = "two.sided", info=FALSE)
npar.t.test(PO4~dose, data=Subdat, method = "permu",
            alternative = "two.sided", info=FALSE)
myWa<- wilcox_test(PO4~ dose, data = Subdat,  distribution = "asymptot
myWaa<-wilcox.test(PO4~ dose, data = Subdat)
```

- Example: using relative effect size

```
library(nparcomp)
npar.t.test(PO4~dose, data=subdat, method = "permu",
            alternative = "two.sided", info=FALSE)
```

- Unbalanced example (20/5) with a single extreme value: confusing many versions of WMW test

Nonparametric tests: the Wilcoxon test IV

Welch	0.197
exact by coin	0.042
asym by coin	0.0415
exact by wilcox.test	0.0424
appr by nparcomp	0.208
permu by nparcomp	0.166

- Recommendation 9: Be careful with WMW test(s) (etc) (Enough confusion?)

Claiming for difference or equivalence by CIs I

- Problem 7: Almost all published test are for superiority, really all?
- Recommendation 10: Use CI for claiming equivalence
- Falsification teaches us: only one error rate can be controlled directly, i.e. for the more relevant question the alternative should be formulate
  Searching research for a difference
  but
  Safety assessment for the equivalence, e.g. no serious change of organ weight in a 90 days feeding study with GMO-corn
  $H_A : \theta < \mu_T/\mu_C AND \mu_T/\mu_C < 1/\theta$
- **We can use CIs for both aims**
  i) Superiority: **exclusion** of 1 (ratio) or 0(difference) of two-sided $1-\alpha$ CI.
  ii) Equivalence: **inclusion** within e.g. $0.7; 1/0.7$ of two-sided $1-2\alpha$ CI (idea: two-one-sided-tests, TOST).

Claiming for difference or equivalence by CIs II

- Even easier for one-sided decisions, e.g. for an increase
  i) superior when $1-\alpha$ lower limit ¿ δ.
  i) non-inferior (=non-hazardous) when $1-\alpha$ upper limit is ¡ θ.

Technical replicates I

- Coming back to our Geneva data: two technical replicates
- Fishers Analyse-as-randomize (see later)
- Assuming all individual values (of the techn. replicates) as randomized n_i increases the f^+ rate (because $> n_i$, and commonly $s_j^2 << s_i^2$
- Modeling replicates as random factor in a mixed effect model (can be complicated) or using means over replicates in the common test (approximate ok in unbalanced designs)
- Explain simple mixed effect model using `lme`

```
setwd("D:/externals/_CUSOLausanneApril2019") # use your folder here
f2c<-read.table("Genf2019.csv", sep=";", dec="," ,header=TRUE)
f2c
library("nlme")
library("multcomp")
mm1 <- lme(value ~ type, random=~1|replicate , data=f2c)
dfMM <-anova(mm1)$denDF[2]
p2c <-summary(glht(mm1,linfct = mcp(type = "Dunnett"), df=dfMM))
```

Technical replicates II

- Simple alternative: use mean values (but not to recommend)

```
f2cm <- aggregate(value ~ type+plant, data=f2c, mean)
mm2 <- lm(value ~ type, data=f2cm)
p2m <-summary(glht(mm2,linfct = mcp(type = "Dunnett")))
```

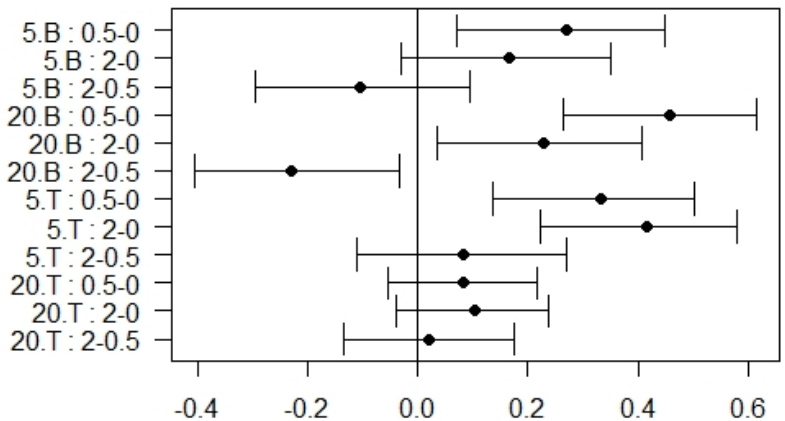
- More random factors: plates, runs, days,

Analysis of counts and proportions (briefly only) I

- Crude data structure: i) gaussian distributed , ii) no-gaussian, eg. skewed distr., iii) ordered categorical (count), iv) proportion
- Analysis of count data: nonparametric test up to generalized linear model (no today)
- Proportion: weak data quality, but high relevance (healing rate)
- Common data structure: 2 by k table; but originally : per unit 0 or 1
- Realize overdispersion (tumor rate in rat vs. skeletal variation in pub)
- Three effect sizes: risk difference, risk ratio, odds ratio
- Commonly used: Fishers exact test (rather conservative for small n_i) page
- Smaller n_i alternative: adjusted χ^2 tests or related CI (add2 intervals) (avoid really small n_i)

Analysis of counts and proportions (briefly only) II

```
library(pairwiseCI)
data(rooting)
rooting # specific data structure
aprootsRD<-pairwiseCI(cbind(root, noroot) ~ IBA,
            data=rooting, by=c("Age", "Position"),
            method="Prop.diff", CImethod="AC")
plot(aprootsRD)
```



- Recommendation 11: When analysing counts and proportion: ask

Bayesian view of two-sample tests I

- Bayesian paradigm as alternative the Neyman-Pearson hypothesis system
- Example: 2-sample test with unstandardized effect size $\mu_1 - \mu_0$ and noninformative Jeffreys prior
- **Likelihood ratio**: how many times more likely it is that a test result will occur in exp units in a new mutant than in wt (ie group 2 vs. group 1)
- **Bayes theorem**: $\frac{PPV}{(1-PPV)} = \frac{Sensitivity}{(1-Specificity)} x \frac{Prevalence}{(1-Prevalence)}$
  In words: **Posterior odds = Likelihood ratio x Prior odds**

Bayesian view of two-sample tests II

Definitions:

- **Prevalence**: $Prevalence = \frac{\text{No. of units with a certain effect}}{\text{No. units considered}}$ (e.g. mammary carcinoma in Germany $Prev_{MC} = \frac{117 cases}{100000 considered} = 0.00117$ per year, ie. standardized in medicine from 71000 new cases in 2010 (pop size 83 Mio))
- **Sensitivity** ... true positive rate TP (ie $1 - Sens = FP = f^+ = \alpha$)
- **Specificity**... true negative rate TN (ie. $1 - Spec = FN = f^- = \beta$ and Spec=Power
- **Positive predictive value** $PPV = \frac{TP}{(TP+FP)}$
- **Negative predictive value** $NPV = \frac{TN}{(TN+FN)}$
- Furthermore: **False discovery rate** $FDR = \frac{FP}{(FP+TP)} = 1 - PPV$

Bayesian view of two-sample tests III

- Variant I): Bayes theorem for $p < 0.05$ interpretation (Held 2018)
  $\frac{TRP}{(1-TRP)} = \frac{Power}{\alpha} x \frac{Pr(H_1)}{Pr(H_0)}$
  TRP ... true positive rate, FRP=1-TRP false positive rate
- Variant II): Bayes theorem for $p = 0.04321$ interpretation
  $\frac{FRP}{(1-FRP)} = \frac{f(p|H_0)}{f(p|H_1)} x \frac{Pr(H_0)}{Pr(H_1)}$
  = Bayes factor BF_{01} x prior odds
  False positive risk $FPR = Pr(H_0|p)$
- Furthermore minBF (Held2018)
- library(Bayesfactor)
- Classification

Bayes factor BF_{01}	Jeffreys(1961)	Held and Ott (2016)
1 to 3	Bare mention	weak
3 to 10	Substantial	Moderate
10 to 30	Strong	Substantial
30 to 100	Very strong	Strong
> 100	Decisive	Very strong

Bayesian view of two-sample tests IV

- Example data

```
library(BayesFactor)
myBF<-ttestBF(subdat$PO4,subdat$dose)
subdat23<-droplevels(simdat[simdat$group %in% c("2", "3"),])
myBF23<-ttestBF(subdat23$PO4,subdat23$dose)
t.test(PO4~dose, data=subdat23)
```

- Recommendation 12: Use BF additional (to p, CI). Surprise the Editor!

Take home: Summarized recommendations for using two-sample tests I

- Rec: Trust the robustness of t-test (as long as...)
- Rec: Use always the Welch-t-test
- Rec: Be careful with common WMW-test
- Rec: (Do not use formal outlier tests, be carefully when eliminate extreme values)
- Rec: Use confidence intervals to claim both stat significance and biological relevance, including for equivalence
- Rec: Report p-values (or CI) for powered exps only
- Rec: Use appropriate effect sizes (scale, model and interpretation)
- Rec: Use Bayes factor as an addition
- Rec: When analysing counts and proportion: ask (eg. ludwig@hothorn.de)
- Rec: Avoid series of two-sample tests: use multiple tests (see below)
- Rec: Use CRAN libraries (pairwiseCI, coin, nparcomp, BayesFactor,...)

Exercise III: two-sample tests I

- Questions so far?
- Analyse exa2 example by several tests and CI! Interpret results!

Multiplicity issues I

- **Counter example I:** Instead a single primary endpoint, use 10 (uncorrelated) endpoints, use independent t-tests (each at 0.05 level). With higher prob you will see at least a significant outcome (even when the true effect is tiny) because $f^(+)$ increases from 0.05 to 0.5
- **Counter example II:** Compare instead a single new mutant vs. wt, with say 20 mutants, use independent t-tests (each at 0.05 level). With higher prob you will see at least a significant outcome (even when the true effect is tiny) because $f^(+)$ increases from 0.05 to 0.99 (under some circumstances)
- In general: several endpoints, several treatment (or dose) comparisons, several time points, several subgroups (eg China, US, EU), several stats tests (eg t-test, Wilcoxon test), etc.... increase $f^(+)$ substantially.

Multiplicity issues II

- The concept of FWER should be used , ie keep $f^(+) = 0.05$ for the experiment, ie use smaller α_i, e.g. $\alpha_i = \alpha/$no. of comparisons (Bonferroni test)
- Bonferroni test is general and simple, but rather conservative when tests are correlated (when uncorrelated perfect!). E.g. genomewide association studies with 1000000 SNPs $\alpha_i = \alpha/100000 = 0.0000005$
- Therefore, i) take the correlations between the tests into account, ii) take less tests, iii) more...

The golden design: Multiple comparisons versus control (wt) I

- Control in molecular biology : wt
- Still better design: include a further positive control
- Two options:
  i) Proof assay sensitivity in advance (to limit false -),
  ii) to characterize a dose effect relative to C- **and** relative to C+.
- It is not enough to demonstrate to be better than control, even to be either slightly inferior to the competitor (C^+) (non-inferiority) or better superiority.
- Therefore, the most group comparisons in molecular biology is **simultaneous comparisons vs. control**

Simultaneous Confidence Intervals I

- Design: $[C, T_1, ..., T_k]$ resp. $[C, D_1, ..., D_k]$,i.e. comparing of treatments or doses versus C
- Alternative designs $[T_1, ..., T_k, R]$ resp. $[K, D_1, ..., D_k, C^+]$, i.e. comparisons vs. reference
- Claim of superiority or non-inferiority by means of simultaneous CI for difference or ratio vs. C
  i) Comparisons vs. C (many-to-one comparisons, simple tree alternative) for T_i or D_i,
  ii) Ordered alternative for designs with D_i
- Alternative: claim of non-inferiority vs. C^+
- Controversy on one/two-sided hypotheses formulation.
  But, alone for the perspective of power, hypotheses should be restricted: i) one-sided, ii) monotone; where two-sided hypotheses are only sometimes adequate,more a hint on uncertainty

Simultaneous Confidence Intervals II

- Typical point-zero-hypothesis:
  $H_0 : \mu_0 = ... = \mu_k$ vs. $H_1 : \mu_0 < \mu_i$ (at least one i, anyone)(0 ... index of control)
- Non-inferiority($\Uparrow$ toxic):
  $H_0 : \mu_i - \mu_0 \geq -\delta \forall i$ vs. $H_1 : \mu_i - \mu_0 < -\delta \; \exists i$
- Ordered alternative: $H_1 : \mu_0 \leq \mu_1 \leq ... \leq \mu_k$; at least $\mu_0 < \mu_k$
- Therefore only two methods , assuming $N(\mu_i, \sigma^2)$:
  i) **Dunnett (1955)** [Dun55] commonly one-sided,
  ii) **Williams (1971)** [Wil71], one-sided on monotone increase (or decrease)

Multiple Comparison Procedures for Differences - demonstrated as multiple contrast test I

- **Aim**: Simultaneous CI for $(\mu_i - \mu_{i'})$, using **linear** test statistics
- Special case: comparisons vs. C: $(\mu_i - \mu_0)$
- Simultaneous lower confidence limit acc. to Dunnett (1955) [Dun55]:
  $[\bar{x}_i - \bar{x}_0 - S\sqrt{n_i^{-1} + n_0^{-1}}\, t_{k,df,R,1-\alpha};]$
- A contrast is a suitable linear combination of means: $\sum_{i=0}^{k} c_i \bar{x}_i$.
  A contrast test is standardized $t_{Contrast} = \sum_{i=0}^{k} c_i \bar{x}_i / S\sqrt{\sum_i^k c_i^2/n_i}$
  where $\sum_{i=0}^{k} c_i = 0$ guaranteed a $t_{df,1-\alpha}$ distributed level-α-test.
- A multiple contrast test is defined as maximum test:
  $t_{MCT} = max(t_1, ..., t_q)$ which follows jointly $(t_1, \ldots, t_q)'$ a q-variate t- distribution with degree of freedom df and the correlation matrix R, with $\rho_{ab} = \frac{\sum_{i=1}^{k} a_i b_i / n_i}{\sqrt{\sum_{i=1}^{k} a_i^2/n_i \sum_{i=1}^{k} b_i^2/n_i}}$

Multiple Comparison Procedures for Differences - demonstrated as multiple contrast test II

- **Notice**: With increasing average correlation and lower number of contrasts q the q-variate t-distribution tends to the univariate t-distribution, i.e. the degree of adjustment reduces
- Question: which contrasts and how much? Aim: less, correlated contrasts, which a are relevant to molecular biology questions (see below)
- Simple examples (balanced design k=3)
- Dunnett one-sided

c_i	K	T_1	T_2
c_a	-1	0	1
c_b	-1	-1	0

Multiple Comparison Procedures for Differences - demonstrated as multiple contrast test III

- Tukey all pairs comparisons (two-sided)

c_i	K	T_1	T_2
c_a	-1	0	1
c_b	-1	1	0
c_c	1	0	-1
c_d	1	-1	0
c_e	0	1	-1
c_f	0	-1	1

- Williams Procedure as multiple contrast [Bre06]

c_i	K	T_1	T_2
c_a	-1	0	1
c_b	-1	1/2	1/2

- Two-sided CI: $[\sum_{i=0}^{k} c_i \bar{x}_i \pm St_{q,df,R,2-sided,1-\alpha} \sqrt{\sum_i^k c_i^2/n_i}]$

Multiple Comparison Procedures for Differences - demonstrated as multiple contrast test IV

- Our simulated example data (assuming control is treatment 1)

```
> library(multcomp)
> mod1<-lm(PO4~group, data=simdat)
> summary(glht(mod1, linfct = mcp(group = "Dunnett")))
 Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Dunnett Contrasts

Fit: lm(formula = PO4 ~ group, data = simdat)

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
2 - 1 == 0   -98.61      10.60  -9.300   <1e-04 ***
3 - 1 == 0   -60.55      10.60  -5.710   <1e-04 ***
4 - 1 == 0  -103.67      10.60  -9.777   <1e-04 ***
```

- Show simultaneous confidence intervals and their interpretation

Multiple Comparisons for ratios I

- **Aim**: simultaneous CI for (μ_i/μ_0)
- Trick: Re-formulation the ratios in a linear form $Z_{i0} = \bar{x}_i - \theta\bar{x}_0$ (Fieller, 1954) [Fie54] (Assumption $\theta = const.$)
- Therefore $Z_{i0} \sim N(0, \sigma^2_{Z_i0})$, where $\sigma^2_{Z_{i0}} = \left[\frac{1}{n_i} + \frac{\theta^2}{n_0}\right]\sigma^2$
- $t_{i0}(\theta) = \frac{\bar{x}_i - \theta\bar{x}_0}{S_{Z_{i0}}}$ is univariate t- distributed
- Simultaneous CI for the ratios $\gamma_{i0} = \mu_i/\mu_0$

$$\left\{(\widehat{\gamma}_i - G) \pm \left[(\widehat{\gamma}_i - G)^2 - (1-G)\left(\widehat{\gamma}_i^2 - \frac{N}{n_i}G\right)\right]^{\frac{1}{2}}\right\}/(1-G)$$

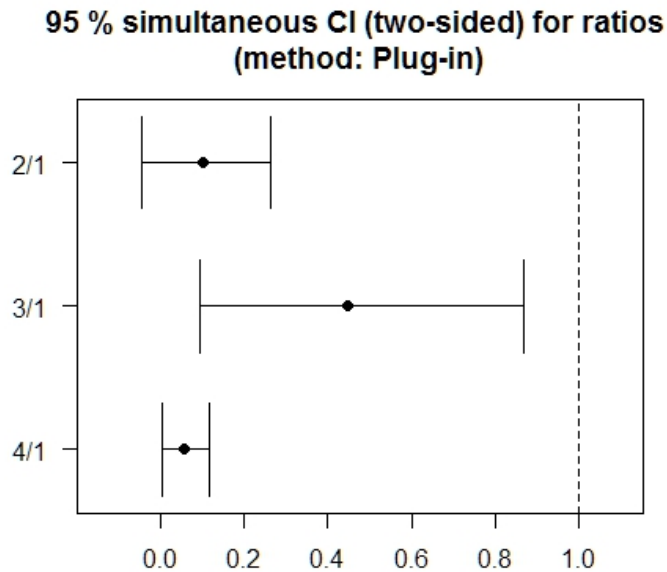
$i = 1, \ldots, q$, where $G = S^2 q^2_{\alpha,m,\nu,R}/(N\bar{x}_0^2)$

Multiple Comparisons for ratios II

- Notice, the equi-coordinate percentage point $t_{q,\nu,\mathbb{R},1-\alpha}$ depends on the unknown ratios γ_{i0} by the correlation matrix

- Our toy example using R package mratio [DSH07]

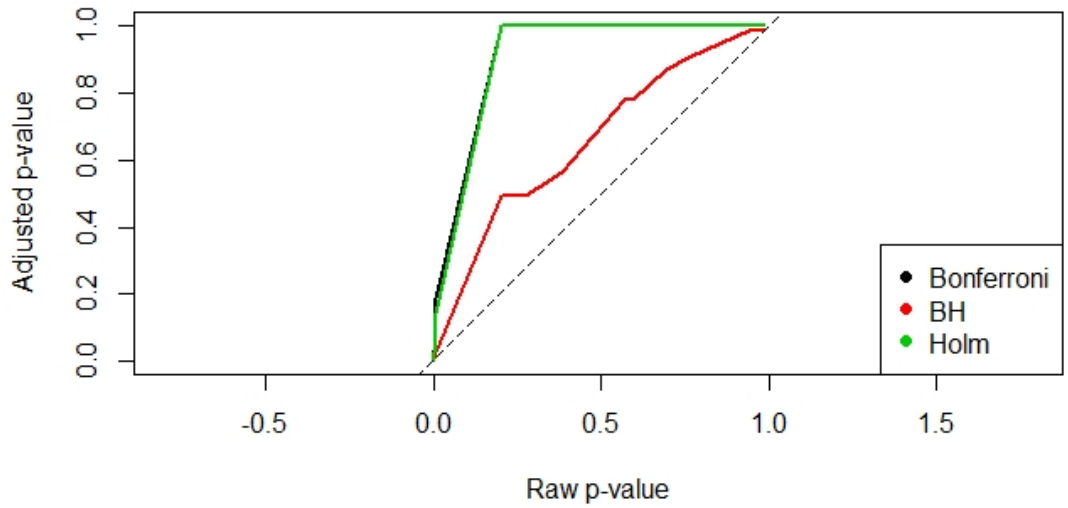
```
library(mratios)
> plot(sci.ratioVH(PO4~group, data=simdat, type="Dunnett")
```



Further multiplicity issues: BH procedure I

- Used for multiple endpoints, e.g. 100000 SNPs in GWAS
- The Benjamini Hochberg Procedure (BH) is a powerful tool that decreases the false discovery rate (see above)
- How to Run BH procedure
  - Put the individual p-values in ascending order
  - Assign ranks i to the p-values
  - Calculate each adjusted BH p-value $(i/m)Q$, where: m = total number of tests, Q = the false discovery rate (commonly 0.05)
- R example

Further multiplicity issues: BH procedure II



Take home: multiplicity issues I

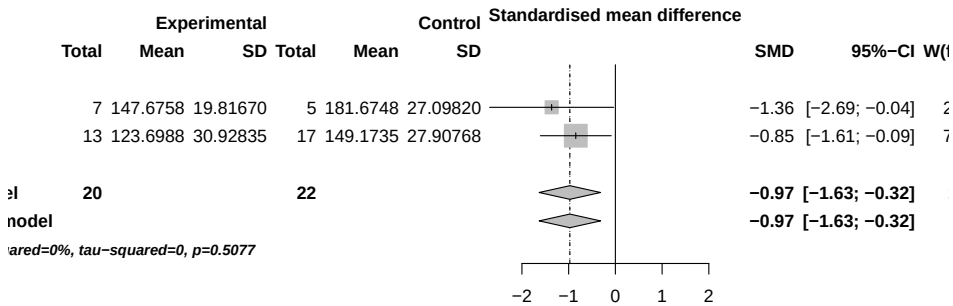
- When analysing multiple mutants (treatments), endpoints, times ... adjust against multiplicity to avoid too high f^+ rate
- Take the correlation between the tests into account- which reduce conservativeness, still controlling 0.05
- Use library(multicomp,nparcomp)

Exercise IV: Comparison with wt I

- Questions so far?
- Analyse data(f2m) example: adjusted p-value and simulatneous CI. Interpret results!

One big study or two smaller studies? Repeatability I

- Two (or more) smaller studies allows to test repeatability by means of Q-test and between-study variability (e.g. DerSimonian esti.)
- Approach is call meta-analysis (be careful: manly for binomial data) using library(meta) Q-test (Thompson1999). Homogeneous $p \geq 0.10$
- A teaching example



One big study or two smaller studies? Repeatability II

$p_1 = 0.049, p_2 = 0.028, p_{joint} = 0.012$, and homogeneous, i.e. repeatable

- Interim analysis. Idea: i) perform a first study with arbitrarily chosen, but small n_i^1. Estimate sd and n_i^2 to achieve a certain effect size for global study (possible taken from the first data); ii) stop, if hopeless ($p > 0.5$) or effective ($p < 0.023$), otherwise iii) plan n_i^2 using adjusted level α^*, iv) report the p-value for both studies, v) for the second study we modify design: skip treatment arms, skip endpoints.

  Recommendation from a Bayesian perspective in hypothesis testing: Held2018 Use test with $p < 0.005$ as measure for a **reproducible result**

Principles of experimental design I

- Molecular biology has (compared with epidemiology) the privilege of customized randomized design
- Before: design depends on stats methods to analyze data: t-test will require a different design as linear regression (and for common complex analyses no design approaches exist; simulation models can be used instead)
- **First issue** (focusing on significance tests): choose appropriate n_i: the higher n_i, the monotone lower p (or $f-$) for constant effect size (commonly $\delta = \mu_i - \mu_0$). **Ie.** n_i **is the major success factor for your research**. But: The lowdown on very low p-values Paul Eilers (Goettingen talk, 2016)
- **Second issue** reduce variance (because $n_i = 2\sigma^2(t^2_{1-alpha} + t^2_{1-\beta})/\delta^2$:
  1. σ^2 is a variable-specific property,
  2. reduce by inclusion/exclusion criteria (counter-example the German corn field trial)

Principles of experimental design II

3. use technical replicates (possibly less than randomized!)
4. use repeated measures
5. use secondary factors, e.g. sex in in vivo studies
6. use blocks
7. compare $s_j^2 with s_i^2$ to explain a part of variability; but df-effect when rather small sample sizes
8. adjust against covariates

- **Third issue:** use one-sided tests or confidence intervals
- **Fourth issue:** adjust against unit-specific covariates
- **Fifth issue:** use as less as factor levels are needed
- **Six issue:** use as less as secondary factors are needed (interaction issue)

Principles of experimental design III

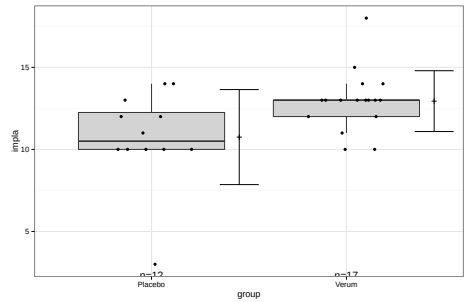
- **Seventh issue:** select a primary endpoint (out of q): most predictive, normal distributed scaled (before non-normal, before censored, before many missings, before graded findings, before proportion), before multiple correlated, before multiple uncorrelated- with the smallest CV ($CV = \sigma_0/\mu_0$)
- **Eight issue:** use appropriate statistical methods
- **Ninth issue:** Fisher's principle *Analyse as randomize* ... definition of exp unit (which is randomized) not always obvious (compared with patients in RCT). Randomization should reduce structural bias by exp units (chronological bias (undesired time effects) selection bias (blinding))
- **Further issues:** I) One-factor, two-factor, multi-factor layout, II) Completely randomized vs. nested, vs. technical replicated, III) Single endpoint, multiple endpoints, repeated measures, IV) Quantitative covariate, eg. dose $0, 1, 2, 10, 100$ or Dose $C-, 1, 2, 10, 100, C+$ or Combi $C, 1+1, 2+10, 10+100, C+$, V) etc.

The power approach: A-priori sample size estimation I

- Popper's falsification principle: *we can never prove an effect directly, only by the unlikeliness of its opposite*
- Neyman-Pearson test theory: $H_A : \mu_1 - \mu_2 > 0$ vs. $H_0 : \mu_1 - \mu_2 \leq 0$
- Two error rates exists: $f^+ ... \alpha$, $f^- ... \beta$. Only one can be controlled directly, commonly $f^+ = 0.05$
- Power $\pi = 1 - \beta$, NOT defined. Commonly 0.80, but at least > 0.5
- t-test $\frac{\bar{x_1} - \bar{x_2}}{SD\sqrt{(1/n_1 + 1/n_2)}} \propto t_{df, 1-\alpha)}$
- Assumptions
  1. randomized two-group design
  2. independent (opposite: paired, matched)
  3. Gaussian distribution of the error
  4. Homogeneous variances
  5. $min(N) = 3$
  6. difference to ZERO? But commonly δ-relevance $\mu_1 - \mu_2 > \delta$
  7. effect size $eff = \mu_1 - \mu_2$

The power approach: A-priori sample size estimation II

- Using R. Example data: no. implantation in mice [KH12a]



- not really small n_i, but $n_1 \neq n_2$
- normal distribution?
- tied data (counts)
- variance homogeneity?
- One-sided Welch-t-test p=0.017
- Controversy one- vs. two-sided testing
- Confidence intervals for 8 tests

```
pairwiseCI(impla~group, data = impla, method = "Param.diff", var.equal=FALSE, alternative="greate
pairwiseCI(impla~group, data = impla, method = "Param.ratio", var.equal=TRUE, alternative="greate
pairwiseCI(impla~group, data = impla, method = "Param.ratio", var.equal=FALSE, alternative="great
pairwiseCI(impla~group, data = impla, method = "Lognorm.diff", var.equal=FALSE, alternative="grea
pairwiseCI(impla~group, data = impla, method = "Lognorm.ratio", var.equal=FALSE, alternative="gre
pairwiseCI(impla~group, data = impla, method = "HL.diff", alternative="greater")
pairwiseCI(impla~group, data = impla, method = "HL.ratio", alternative="greater")
```

The power approach: A-priori sample size estimation III

Definition	Method	Effect size	lower confidence limit
Diff	t-test	2.19	0.69
Diff	Welch-t-test	2.19	0.54
Diff	Lognormal diff	1.91	-1.57
Diff	HL diff	2.0	0
Ratio	Ratio2C	1.20	1.06
Ratio	Ratio2C var het	1.20	1.04
Ratio	Lognormal ratio	1.17	0.89
Ratio	HL ratio	1.18	1.0

- Power formula $n_i = 2\frac{\sigma^2}{\Delta^2}(t_{df=n_1+n_2-2,1-\alpha} + t_{df,1-\beta})^2$
  1. a-priori fixing of f^+, f^-, commonly 0.05, 0.80
  2. a-priori knowledge of variance σ^2: i) pre-exp, ii) publication, iii) interim analysis
  3. Notice: this formula can be solved for n, or Δ, or f^- (even for σ^2 or f^+). The corn field study
  4. Several extension available (techn. replicates?) e.g. complicated for Wilcoxon test
- Using R. SD= 2.90 from own control or 1.9 treatment? (Published SD 3.2), $\Delta = 2, 3, 4$ (see number of implants)

The power approach: A-priori sample size estimation IV

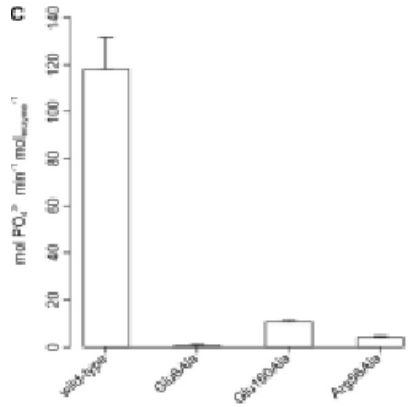
```
power.t.test(power = .70, delta = 2, sd=3.2, alternative = "one.sided")
```

SD	Δ	f^-	n_i
3.2	2	0.20	33
	3	0.20	15
	4	0.20	9
2.9	2	0.20	27
	3	0.20	13
	4	0.20	8
1.9	2	0.20	12
	3	0.20	6
	4	0.20	4
3.2	2	0.30	25
	3	0.30	12
	4	0.30	7
2.9	2	0.30	21
	3	0.30	10
	4	0.30	7
1.9	2	0.30	9
	3	0.30	5
	4	0.30	4

- Conclusion: i) $\Delta \geq 3 \Rightarrow n_i \leq 10$ possible, ii) select one scenario and report it in suppl. material
- With relevance shift $\mu_1 - \mu_2 > \delta$ follows $\Delta_{relevance} = \Delta - \delta$, e.g. $4 = 5 - 1$

Exercise V: Power I

- Coming back to [MTH15]



- In a future experiment in at least one (anyone) of 10 mutants a significant decrease against wt should be demonstrated from normal distributed $\delta = x_{wt} = 104.2 - x_i = 31 \approx 75$ with a $SD = 29, ..., 19$

```
mue1=104.2; mue2=3.5; mue3=31;mue4=7.9;
sigma1=19.5; sigma2=18.5; sigma3=28.5;sigma4=2.5
```

- Approach I: Bonferroni-type many-to-one comparisons for difference
- Approach II: Ratio-to-control approach

Final analysis I

- Geneva 2018 exp: exa6.xls
- R-script, Figures, Summary, Methods

Take home I

- Biostatistics in molecular biology today: using R
- Investment in some data manipluation (xls and within R) gives sense
- Search for appropriate libraries
- Use it in daily routine

Appendices I

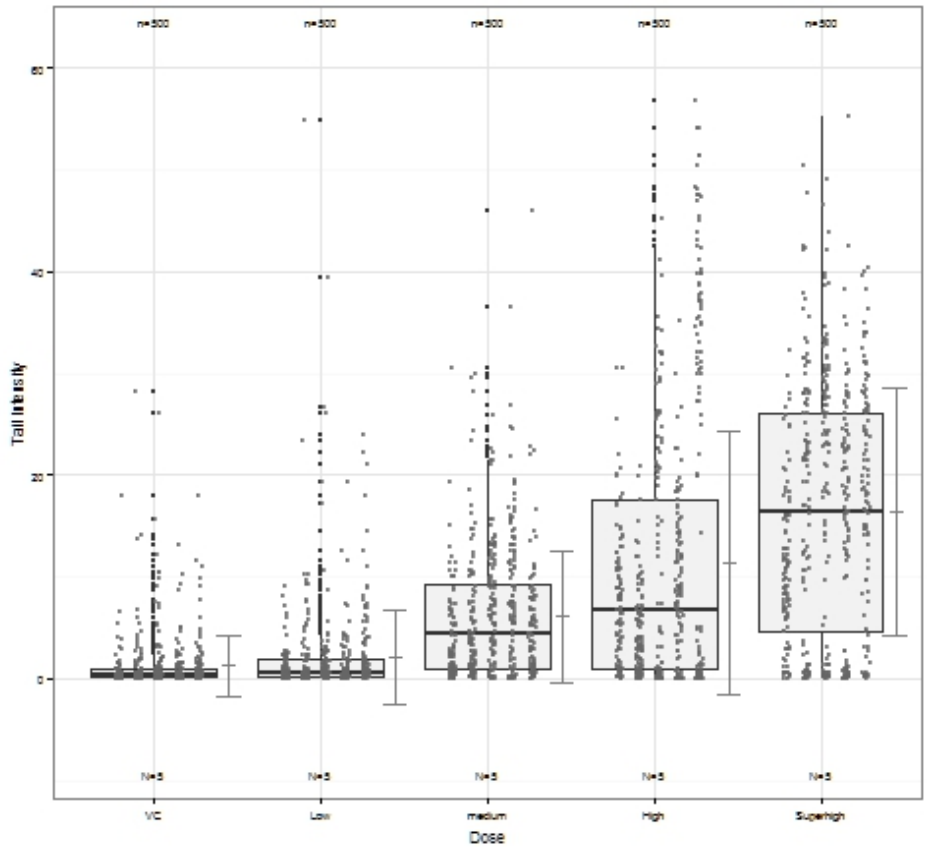
Appendix I: Analysis of interactions I

- By means of exa5.xls and library(statint)

Appendix II: Case study Comet Assay I

- DNA damage: free DNA-loops form comet-shaped structures in gel electrophoresis
- Several parameters are derived, such as tail length, moment or **intensity** [WA03].
- Specific: i) Hierarchical design:
  treatment $\supset$ *animal* $\supset$ *organ* $\supset$ *sample* $\supset$ *slide* $\supset$ *cell* exists ii) the distribution of the endpoints is neither symmetric nor uni-modal, e.g., the % tail DNA in liver is extreme skewed [LO08]. [WA03] 90^{th} percentile, capturing the upper tail of the distribution
  Example: tail intensities for liver in each 5 animals, 2 samples and each 50 cells [NG14].

Appendix II: Case study Comet Assay II

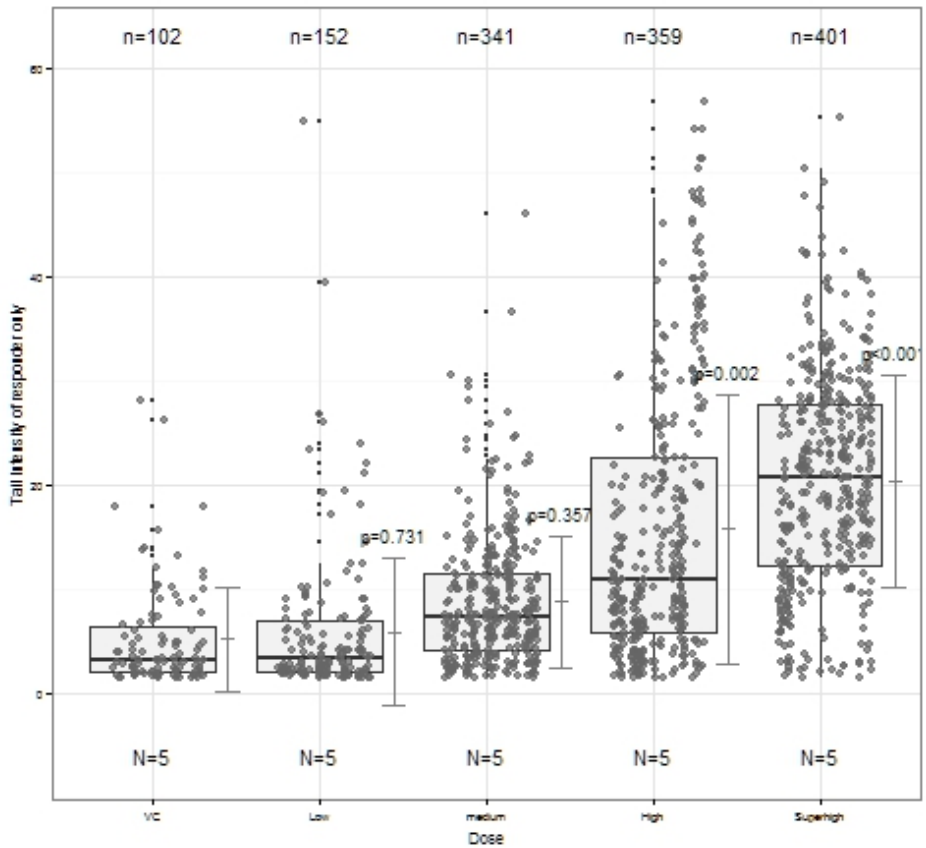


- The boxplots show dose-dependent skewness and bimodality and the between-animal variability.
- High sophisticated analysis:

Appendix II: Case study Comet Assay III

- assuming a bimodal mixing distribution of two normally distributed variables: responder and non-responder
- Responder category is estimated data-dependent by model based-clustering using the R package `flexmix` [GL07],
- not just by a naive percentile rule
- selecting the responder values only (see the seriously unbalanced design in the boxplots) and estimated Dunnett-type p-values using a mixed model
- p-values plugged-in into boxplot

Appendix II: Case study Comet Assay IV



Appendix III: Case study genetic association study I

- Three major designs of genetic association studies:
  i) Case-control, ii) Family-specific, iii) cohort design for multiple quantitative phenotypes (traits): **one-way layout for the factor genotype score**
- Single phenotype with a diallelic marker: *a* is the high-risk candidate allele and *A* is any of the other alleles. A one-way layout with three levels follows: one heterozygotic and two homozygotic groups.
- Two models: i) genotype as factor or ii) as covariate with the scores 0,0.5,1

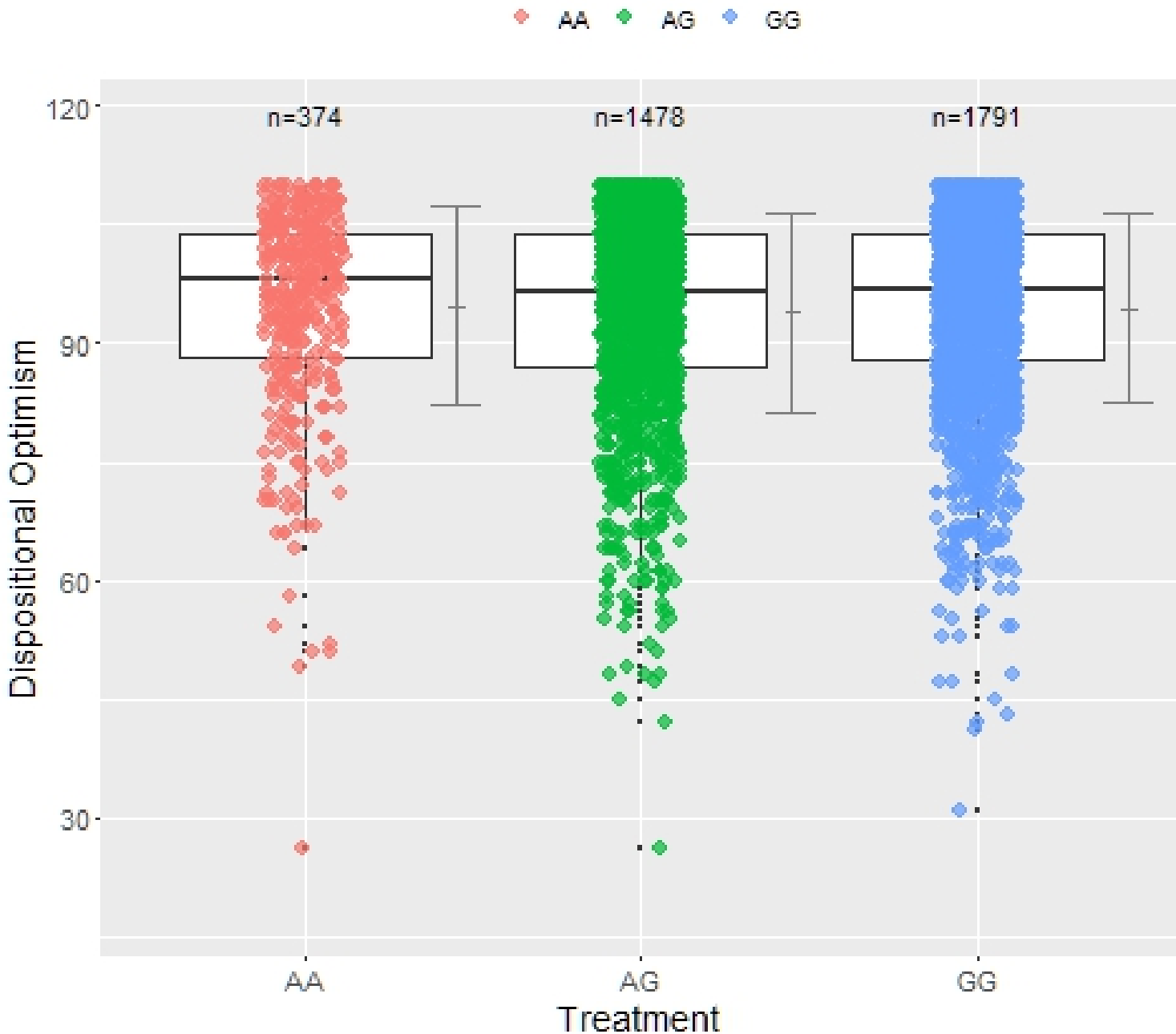
Table: Numeric scores for different modes of inheritance.

genetic model	variable (j)	AA	Aa	aa
dominant	x(1)	0	1	1
additive	x(2)	0	0.5	1
recessive	x(3)	0	0	1

Appendix III: Case study genetic association study II

- Three basic modes of inheritance (add, dom, rec)
- Variance heterogeneity for Mendelian inheritance likely
- Visualization of a study in psychiarty (Strohmaier et al. The psychiatric vulnerability gene CACNA1C and its sex-specific relationship with personality traits, resilience factors and depressive symptoms in the general population. Mol Psychiatry. 2012)

Appendix III: Case study genetic association study III



Appendix IV: Recommendations in Nat.Cell Biol I

Reporting Life Sciences Research

This non-exhaustive list summarizes several elements of methodology that are frequently poorly reported. Inconsistent reporting may lead to incorrect interpretation of results and a lack of reproducibility. To improve the transparency and the reproducibility of published results, we ask that authors include in their manuscripts relevant details about these elements of their experimental design. During peer review, authors confirm via the Reporting Checklist For Life Sciences Articles that this information is reported.

Reporting Experimental Design

Sample size: When confirming an effect of known size, it is considered best practice to estimate before conducting the experiments what sample size is needed to ensure statistical power of detection. If no sample size calculation was performed, the authors should report why they think

the absence of a statement is taken to mean that there was no randomization.

Blinding: Whenever possible, the investigator should be unaware of the sample group allocation during the experiment and when assessing its outcome. Although we realize

Appendix IV: Recommendations in Nat.Cell Biol II

- **I) Sample size**:
  1. *When confirming an effect of known size, it is considered best practice to estimate before conducting the experiments what sample size is needed to ensure statistical power* of detection
  2. *If no sample size calculation was performed, the authors should report why they think the sample size is adequate to measure their effect size*
  3. For animal studies, authors must report whether statistical methods have been used to predetermine sample size
  4. For all experiments, the sample size (n) must be reported as an exact number(not a range)
  5. Investigators should define the criteria for identifying and dealing with outliers before running the experiments
  6. When reporting the results, they must explain any discrepancy between sample size at the beginning and end of each analysis due to attrition or exclusion
- II) Randomization (not today)
- III) Blinding (not today)

Appendix IV: Recommendations in Nat.Cell Biol III

- **IV) Replication**
  1. It is often unclear whether replicates represent biological or technical replicates
  2. In reporting their results, authors should provide enough details about the sample collection to distinguish between independent data points and technical replicates
  3. Depending on the experimental design, technical replicates will reflect the variation of the assay and/or sample preparation by assaying a sample from the same source multiple times
  4. Biological replicates are intended to reflect the biological variability and require processing samples from different sources
  5. Experimental design should be taken into account to define biological replicates, for example, they may require animals from different litters
  6. Therefore, careful reporting of the experimental conditions and nature of replicates is essential
  7. When showing a representative experiment, authors must specify the number of times this experiment was successfully repeated and discuss any limitations in repeatability

  - We learn: $no_{exp} \succ no_{animals} \succ no_{techn.replicates}$. The definition of randomized unit is essential

Appendix IV: Recommendations in Nat.Cell Biol IV

- We learn: repeatability should be demonstrated, at least by two similar exp's

- **V) Statistical tests**

1. Authors must describe the statistical tests used during the analysis and justify their choices
2. Many statistical tests require that the data be approximately normally distributed; when using these tests, authors should explain how they tested their data for normality, which may be difficult if sample sizes are small. robust test for small n_i is a challenge
3. If the data do not meet the assumptions of the tests, then a nonparametric alternative should be used instead really?
4. If the distribution is not normal, mean and standard deviation calculations are not appropriate. yeh: use median and mean!
5. Authors should specify whether the tests are one-sided or two-sided explain!

Appendix IV: Recommendations in Nat.Cell Biol V

6. They should also estimate the variation within each experimental group and ensure that the variance is similar for groups that are being statistically compared. commonly heterogeneous: use Welch-t-test
7. When making multiple statistical comparisons on a single data set, authors should explain how they adjusted the alpha level to avoid an inflated Type I error rate, or they should select statistical tests appropriate for multiple groups (such as ANOVA rather than a series of t-tests) Topic IX
8. Statistical measures, such as center (mean, median) and error bars (standard duration, standard error of the mean), used to describe a dataset must be stated
9. The P value for each test must be reported regardless of overall significance really?

Appendix IV: Recommendations in Nat.Cell Biol VI

10. When the sample size is small, authors should use tests appropriate to small samples or justify their use of largesample tests a challenge!
11. Mean and standard deviation are not appropriate with small sample sizes, and bar graphs are often misleading yes, but what else?
12. Plotting independent data points is usually more informative
13. When technical replicates are reported, error and significance measures reflect the experimental variability, not the variability of the biological process; it is misleading not to state this clearly absolutely; see Topic V

References I

[BM00] Brunner, E. ; Munzel, U.: The nonparametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation. In: *Biometrical Journal* 42 (2000), Nr. 1, S. 17–25

[Bre06] Bretz, Frank: An Extension of the Williams Trend Test to General Unbalanced Linear Models. In: *Computional Statistics and Data Analysis* 50 (2006), Nr. 7, S. 1735–1748

[DDZ11] Denton, D. L. ; Diamond, J. ; Zheng, L.: Test of significance in toxicity: A statistical application for assessing whether an effluent or site water is truly toxic. In: *Environ Toxicol Chem* 30 (2011), MAY, Nr. 5, S. 1117–1126. http://dx.doi.org/{10.1002/etc.493}. – DOI 10.1002/etc.493. – ISSN 0730–7268

[DSH07] Dilba, G. ; Schaarschmidt, F. ; Hothorn, L.A.: Inferences for ratios of normal means. In: *R News* 7 (2007), S. 20–23

[DSP+12] Do, R. P. ; Stahlhut, R. W. ; Ponzi, D. ; Saal, F. S. ; Taylor, J. A.: Non-monotonic dose effects of in utero exposure to di(2-ethylhexyl) phthalate (DEHP) on testicular and serum testosterone and anogenital distance in male mouse fetuses. In: *Reproductive Toxicology* 34 (2012), Dezember, Nr. 4, S. 614–621. http://dx.doi.org/10.1016/j.reprotox.2012.09.006. – DOI 10.1016/j.reprotox.2012.09.006

[Dun55] Dunnett, C. W.: A Multiple Comparison Procedure For Comparing Several Treatments With A Control. In: *J Am Stat Assoc* 50 (1955), Nr. 272, S. 1096–1121

[Fie54] Fieller, E. C.: Some problems in interval estimation. In: *Journal of the Royal Statistical Society Series B-Statistical Methodology* 16 (1954), Nr. 2, S. 175–185

[FPS+18] Furtauer, L. ; Pschenitschnigg, A. ; Scharkosi, H. ; Weckwerth, W. ; Nagele, T.: Combined multivariate analysis and machine learning reveals a predictive module of metabolic stress response in Arabidopsis thaliana. In: *Molecular Omics* 14 (2018), Dezember, Nr. 6, S. 437–449. http://dx.doi.org/10.1039/c8mo00095f. – DOI 10.1039/c8mo00095f

[GL07] Grün, B. ; Leisch, F.: FlexMix: An R Package for Finite Mixture Modelling. In: *R News* 7 (2007), April, Nr. 1, 8–13. #http#

[Hay13] Hayter, A. J.: Inferences on the difference between future observations for comparing two treatments. In: *Journal of Applied Statistics* 40 (2013), APR 1, Nr. 4, S. 887–900. http://dx.doi.org/{10.1080/02664763.2012.758245}. – DOI 10.1080/02664763.2012.758245. – ISSN 0266–4763

References II

[HCEVD15] Halsey, Lewis G. ; Curran-Everett, Douglas ; Vowler, Sarah L. ; Drummond, Gordon B.: The fickle P value generates irreproducible results. In: *Nature Methods* 12 (2015), März, Nr. 3, S. 179–185. http://dx.doi.org/10.1038/nmeth.3288. – DOI 10.1038/nmeth.3288

[ICM] *Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals. ICMJE (2015).* : *Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals. ICMJE (2015)*

[KFG13] Kieser, M. ; Friede, T. ; Gondan, M.: Assessment of statistical significance and clinical relevance. In: *Statistics in Medicine* 32 (2013), Mai, Nr. 10, S. 1707–1719. http://dx.doi.org/10.1002/sim.5634. – DOI 10.1002/sim.5634

[KH12a] Konietschke, F. ; Hothorn, L.A.: Evaluation of Toxicological Studies Using a Non-Parametric Shirley-type Trend Test for Comparing Several Dose Levels With a Control Group. In: *Stat Biopharm Res* 4 (2012), S. 14–27

[KH12b] Konietschke, F. ; Hothorn, L.A.: Rank-based multiple test procedures and simultaneous confidence intervals. In: *Electron J Stat* 6 (2012), S. 738–759. http://dx.doi.org/{10.1214/12-EJS691}. – DOI 10.1214/12–EJS691. – ISSN 1935–7524

[LO08] Lovell, D. P. ; Omori, T.: Statistical issues in the use of the Comet assay. In: *Mutagenesis* 23 (2008), S. 1–12

[MTH15] Martinez, J. ; Truffault, V. ; Hothorn, M.: Structural Determinants for Substrate Binding and Catalysis in Triphosphate Tunnel Metalloenzymes. In: *Journal of Biological Chemistry* 290 (2015), September, Nr. 38, S. 23348–23360. http://dx.doi.org/10.1074/jbc.M115.674473. – DOI 10.1074/jbc.M115.674473

[NG14] Nazarov, M. ; Geys, H.: New R Routines for Facilitating Comet Assay Studies in Toxicology M. In: *Nonclinical Statistics Conference Brugge*, 2014

[Nuz14] Nuzzo, Regina: Statistical Errors. In: *Nature* 506 (2014), Februar, Nr. 7487, S. 150–152

[PCG+19] Price, A. ; Caciula, A. ; Guo, C. ; Lee, B. ; Morrison, J. ; Rasmussen, A. ; Lipkin, W. I. ; Jain, K.: DEvis: an R package for aggregation and visualization of differential expression data. In: *Bmc Bioinformatics* 20 (2019), März, S. 110. http://dx.doi.org/10.1186/s12859-019-2702-z. – DOI 10.1186/s12859–019–2702–z

[PPR15] Pallmann, P. ; Pretorius, M. ; Ritz., C.: Simultaneous comparisons of treatments at multiple time points: combined marginal models versus joint modeling. In: *Statistical Methods in Medical Research, accepted for publication.* (2015)

References III

[RA08] Ryu, E. J. ; Agresti, A.: Modeling and inference for an ordinal effect size measure. In: *Statistics in Medicine* 27 (2008), Mai, Nr. 10, S. 1703–1717

[RLG+12] Rhodes, M. ; Laffan, S. ; Genell, C. ; Gower, J. ; Maier, C. ; Fukushima, T. ; Nichols, G. ; Bassiri, A. E.: Assessing a Theoretical Risk of Dolutegravir-Induced Developmental Immunotoxicity in Juvenile Rats. In: *Toxicological Sciences* 130 (2012), November, Nr. 1, S. 70–81. http://dx.doi.org/10.1093/toxsci/kfs220. – DOI 10.1093/toxsci/kfs220

[Rot14] Rothman, Kenneth J.: Six Persistent Research Misconceptions. In: *Journal of General Internal Medicine* 29 (2014), Juli, Nr. 7, S. 1060–1064. http://dx.doi.org/10.1007/s11606-013-2755-z. – DOI 10.1007/s11606–013–2755–z

[TM15] Trafirmow, D. ; Marks, M.: Editorial. In: *Basic Appl. Psych* (2015)

[TPB+14] Toledo, F. D. ; Perez, L. M. ; Basiglio, C. L. ; Ochoa, J. E. ; Pozzi, E. J. S. ; Roma, M. G.: The Ca2+-calmodulin-Ca2+/calmodulin-dependent protein kinase II signaling pathway is involved in oxidative stress-induced mitochondrial permeability transition and apoptosis in isolated rat hepatocytes. In: *Archives of Toxicology* 88 (2014), September, Nr. 9, S. 1695–1709. http://dx.doi.org/10.1007/s00204-014-1219-5. – DOI 10.1007/s00204–014–1219–5

[WA03] Wiklund, S. J. ; Agurell, E.: Aspects of design and statistical analysis in the Comet assay. In: *Mutagenesis* 18 (2003), März, Nr. 2, S. 167–175

[Wil71] Williams, D.A.: A test for differences between treatment means when several dose levels are compared with a zero dose control. In: *Biometrics* 27 (1971), Nr. 1, S. 103–117