Statistics and experimental design in silage research:

Some comments on design and analysis of comparative silage experiments

Bärbel Kroschewski[1], **Kirsten Weiß**[1] & **Horst Auerbach**[2]

[1]Humboldt Universität zu Berlin [2]ISC Wettin-Löbejün

1. Introduction

J. Dairy Sci. 101:1–23
https://doi.org/10.3168/jds.2017-13978

Invited review: **Reproducible research from noisy data: Revisiting key statistical principles for the animal sciences**

Nora M. Bello*†‡[1] and David G. Renter‡§
*Department of Animal Science, University of Wisconsin, Madison, WI 53706
†Department of Statistics,
‡Center for Outcomes Research and Epidemiology, and
§Department of Diagnostic Medicine and Pathobiology, Kansas State University, Manhattan 66506

"Reproducible results define the very core of scientific integrity in modern research"

... **but across all scientific disciplines only too a little number of trials generate reproducible results.**

- **improper scope of inference**
- **failure in understanding data generation process**
- **unclear distinction between experimental and observational units**
- **insufficient number of experimental replication**
- **inferential misinterpretations associated with use of P-values ...**
- **error inflation due to multiple comparisons**
- **selective reporting**
- **...**

1. Introduction

Aim: Critical findings on design, analysis, and interpretation of results will be addressed based on comparative silage experiments.

Scientific papers on silage trials during the last 8 years published in

- *Journal of Dairy Science*
- *Grass and Forage Science*
- *Agricultural and Food Science*

- Experiments with **1 up to 3 factors** (sometimes even more),
- **3 to 6 replicates** per treatment,
- Statistical analyses:

  frequently performed by **parametric analysis of variance**,
  followed by pairwise comparisons (**LSD, Tukey**, Bonferroni, Sidak, Duncan, ...),

  sometimes by non-parametric procedures (Wilcoxon and others).

1. Introduction

Aim: Critical findings on design, analysis, and interpretation of results will be addressed based on comparative silage experiments.

Lab-scale ensiling trial on biostatistical questions in 2017

(1) What can the scope of inference for one ensiling experiment with mini-silos be?

- material taken from **different field locations** versus a **composite sample,**
- impact of location and fermentation process on silage traits.

(2) Is the frequently used (low) number of replications sufficient regarding significance and relevance of results?

Do the traits meet the assumptions of normal distribution and variance homogeneity?

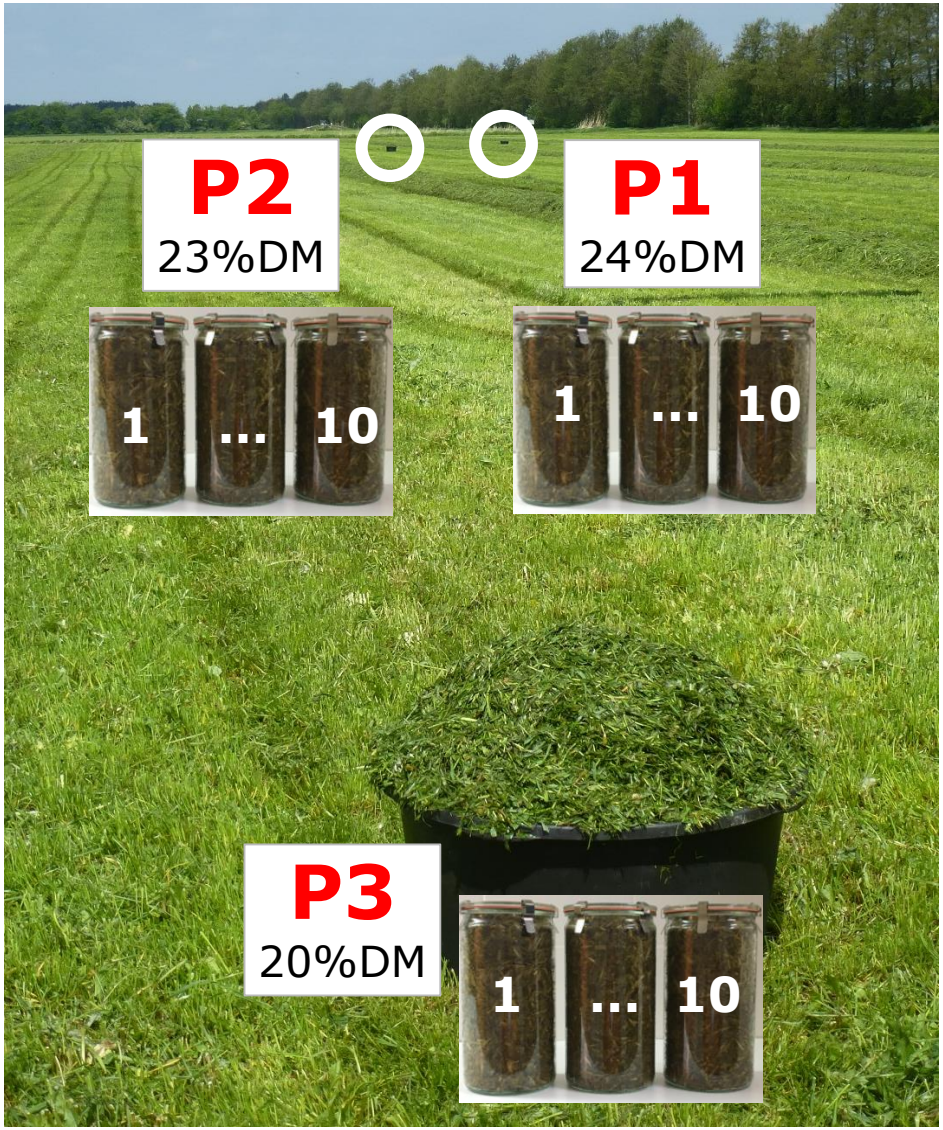
- **three treatments**: untreated control, biological, and chemical silage additive,
- **ten mini-silos as replicates** *(composite sample = restricting experimental input),*
- samples of smaller size (n=3, n=6) extracted:
  to **contrast the results of statistical analyses for different sample sizes.**

2. Description of the grassland trial

Lab-scale ensiling trial

(1) with respect to
field sampling locations

- material taken from three randomly selected sampling points (P1, P2, P3) of natural grassland
- per sampling point: ten 1.5-L jars (=mini-silos) filled with grass material
- without silage additives (= **CON**)

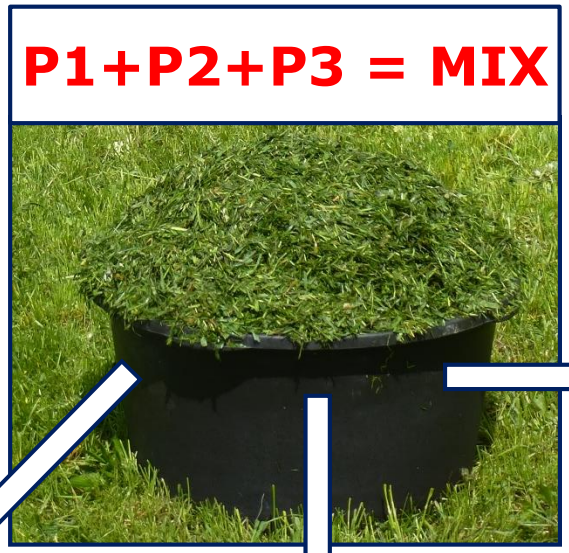


2. Description of the grassland trial

Lab-scale ensiling trial

(2) with grass mixture

- composite grassland sample, mixing identical quantities from three sampling points
- three treatments (silage additive)
- ten 1.5-L jars per treatment



LAB

biological additive

Lactobacillus buchneri CNCM-I 4323 and *Pediococcus acidilactici* DSM 11673

CON

without silage additive

CHEM

chemical additive

sodium nitrite, hexamethylene tetramine and potassium sorbate

2. Description of the grassland trial

Lab-scale ensiling trial

Traits → fresh forage

per sampling point
(P1, P2, P3)

n=5

DM	dry matter
WSC	water-soluble carbohydates
NO_3^-	nitrate
BC	buffering capacity
Yeasts	yeast count
Lactobac	lactic acid bacteria

known to have an influence on the fermentation process

Traits → silage (after 121 days of storage at 22ºC)

- per sampling point (P1, P2, P3)
- mixture (MIX): per treatment (CON, CHEM, LAB)
- n=10 -> total sample size N=60

pH	
LA	lactic acid
AA	acetic acid
1,2-PD	1,2-propanediol
WSC	water-soluble carbohydates
ETOH	ethanol
PROP	n-propanol
ASTA	aerobic stability
DML	anaerobic DM losses

butyric acid, counts of yeasts and moulds: small values

3.1 Results – field sampling locations (CON)

Scope of inference

⇩

- Population, to which the results from a research study are applicable.
- Ideally, this population is sampled at random.

⇩

- „Where can I reasonably expect results to reproduce?“

  = Degree of generalization.

3.1 Results – field sampling locations (CON)

Scope of inference

(1) One field sampling point

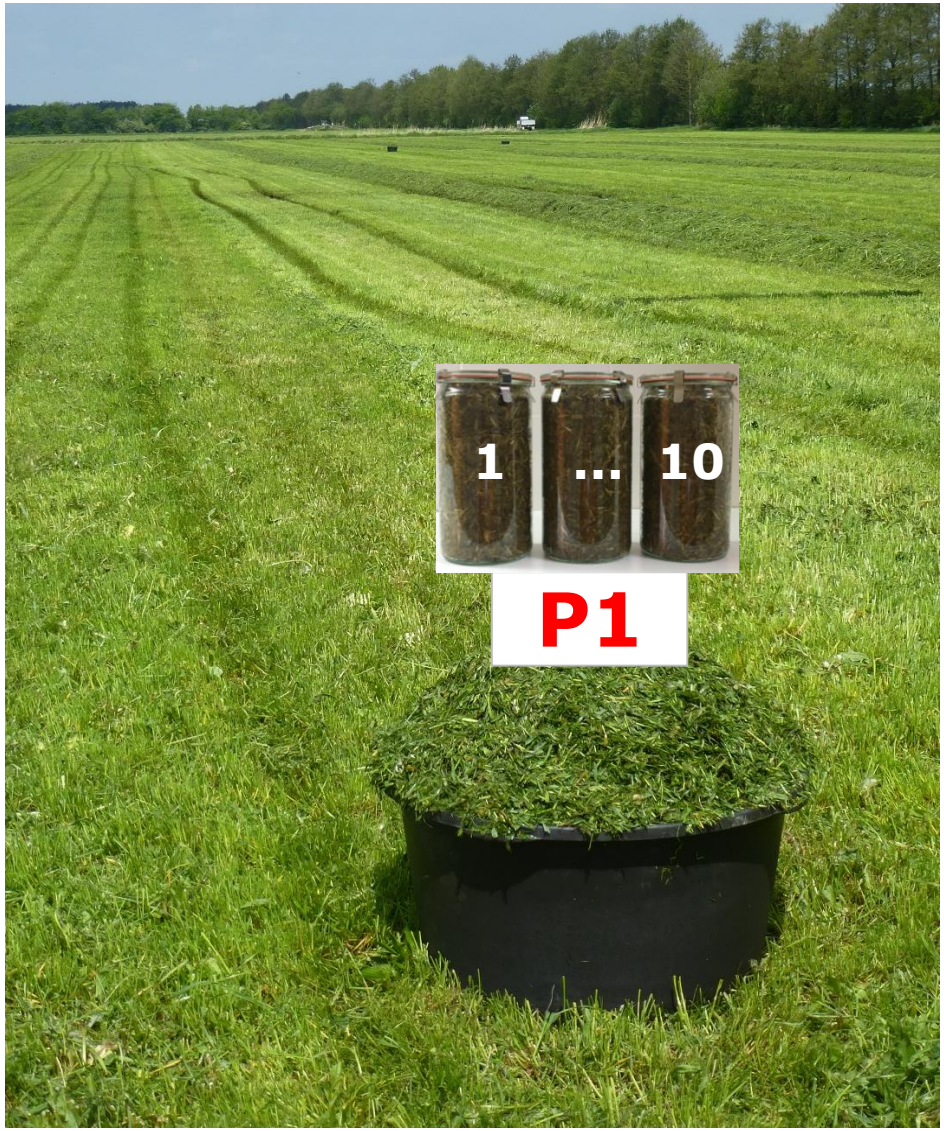
Forage material represents exactly this field location.

⇩

Variability of measurements reflects only the different fermentation processes in the replicated mini-silos per treatment for one material.

different location

farm silo



3.1 Results – field sampling locations (CON)

Scope of inference

(2) Mixture of field sampling points

Composite sample of forage material represents an average field situation.

⇩

Variability of measurements reflects

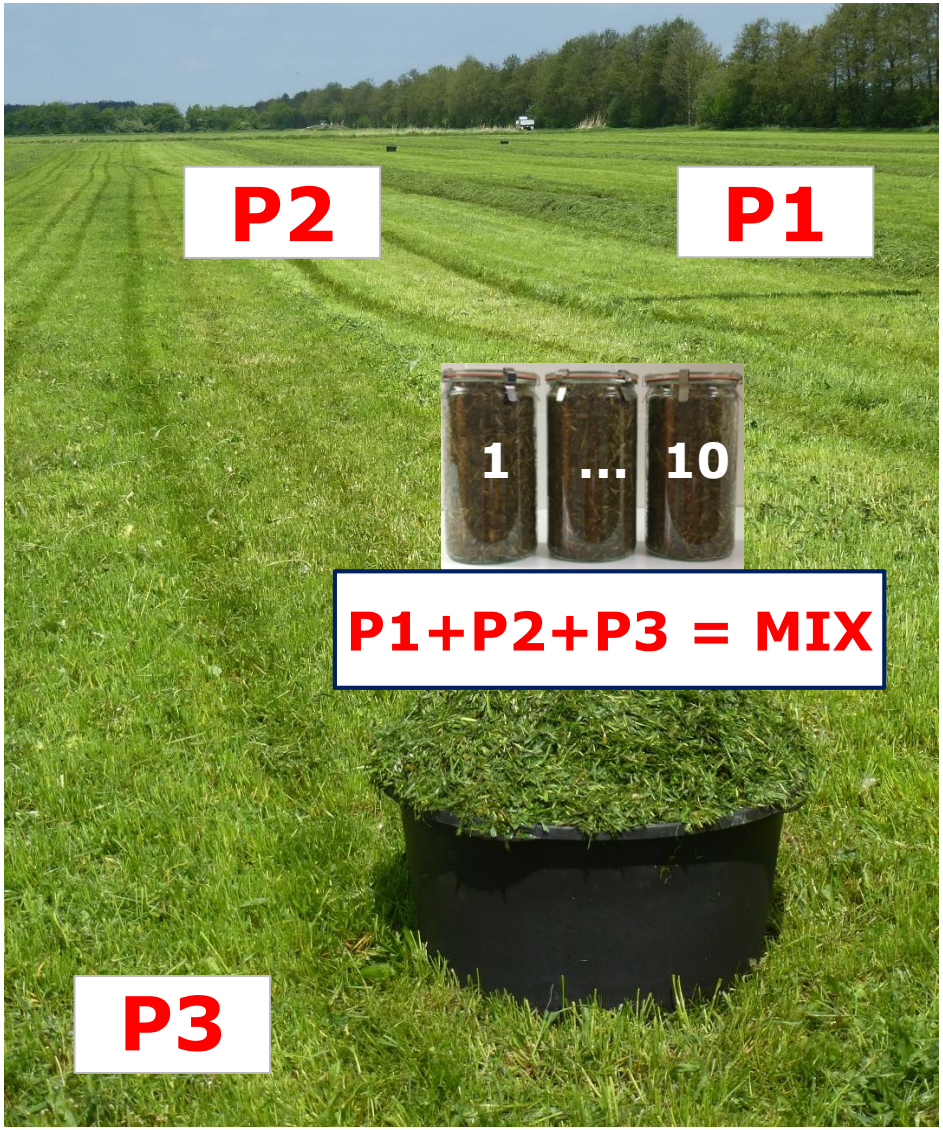
only the different fermentation processes

in the replicated mini-silos per treatment

for one composed material.

different location

farm silo



3.1 Results – field sampling locations (CON)

Scope of inference

(3) Several field sampling points, n=1

Forage material represents the field.

⇩

Variability of measurements reflects

arbitrary field locations and the
different fermentation processes

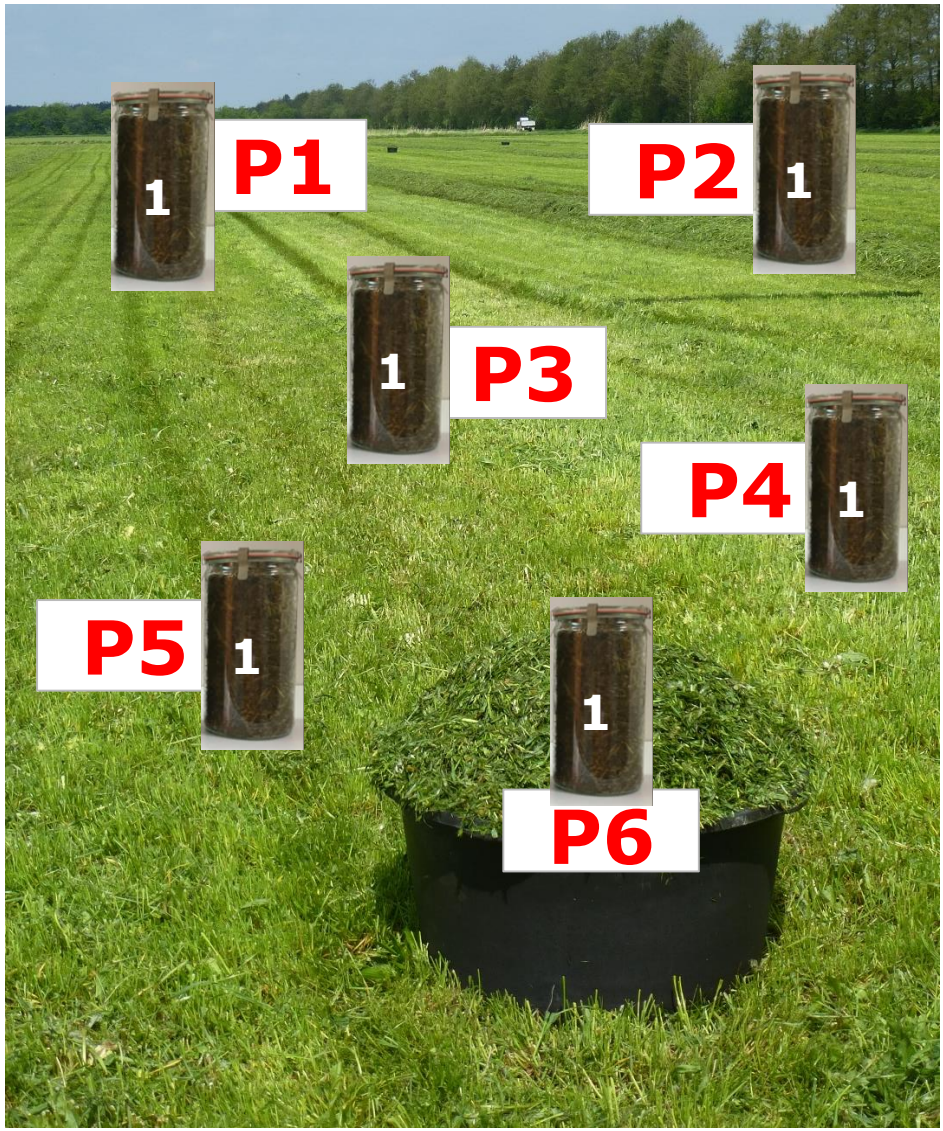
in the replicated mini-silos per treatment.

⇩

Both effects are confounded
(no separation possible).

arbitrary location

farm silo



3.1 Results – field sampling locations (CON)

Scope of inference

(4) Several field sampling points, n>1

Forage material represents the field.

⇩

Variability of measurements reflects

arbitrary field locations and the
different fermentation processes

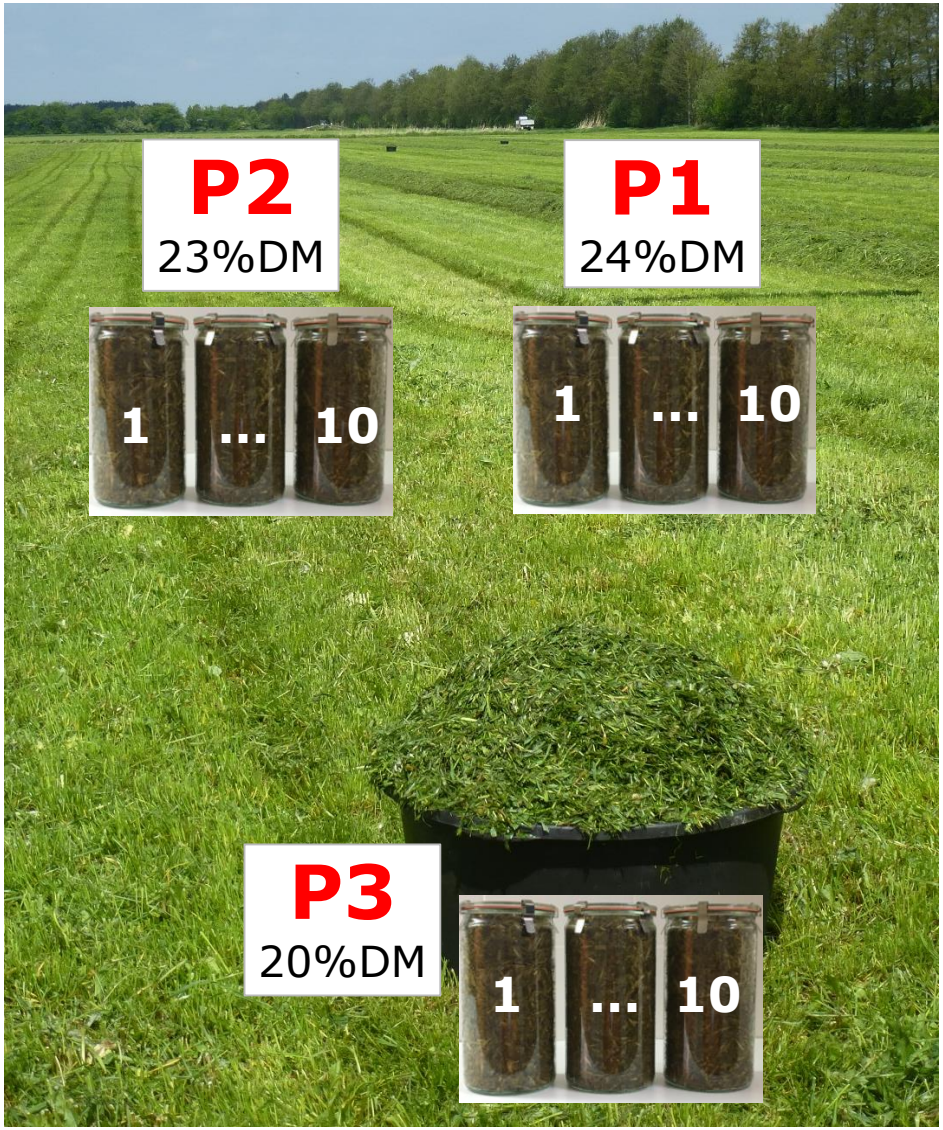
in the replicated mini-silos per treatment.

⇩

Both effects are not confounded
(separation is possible).

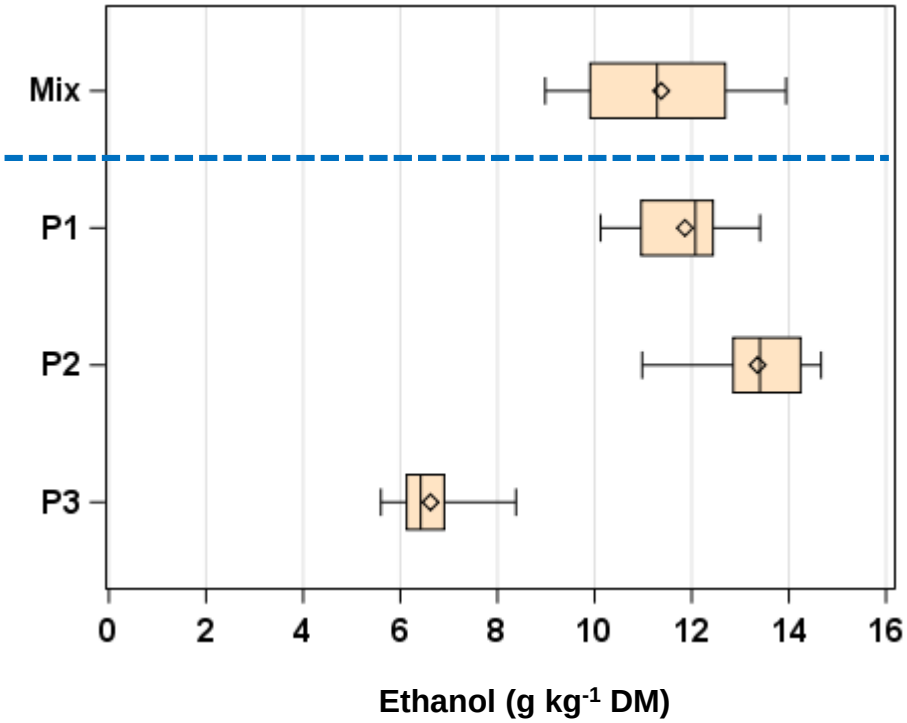
arbitrary location

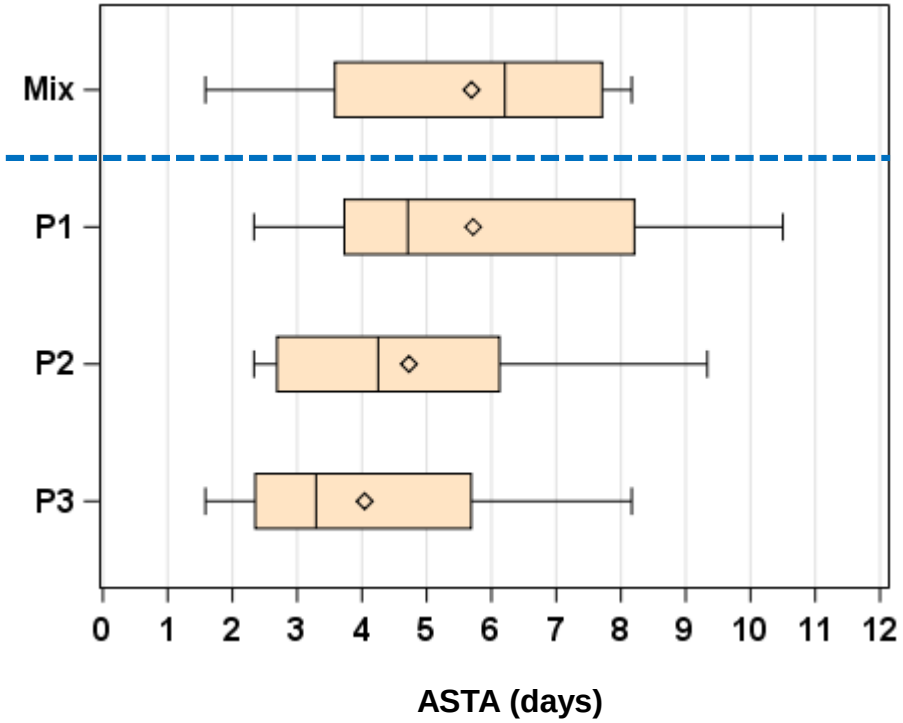
farm silo



3.1 Results – field sampling locations (CON)

Box-Whisker-Plot for three sampling points and grass mixture (MIX), n=10

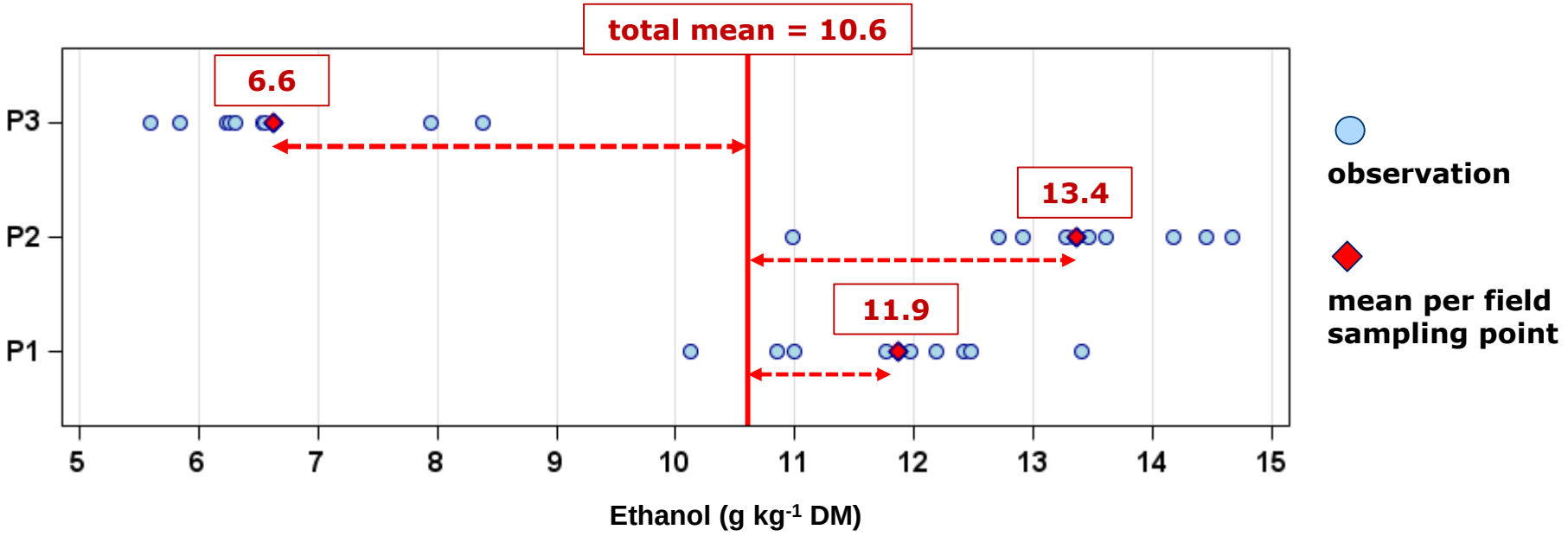




- except ASTA: observations more or less different between sampling points (e.g. ethanol)
- all traits: values from grassland mixture reflect average situation

3.1 Results – field sampling locations (CON)

Variation between field sampling points and within sampling points for Ethanol

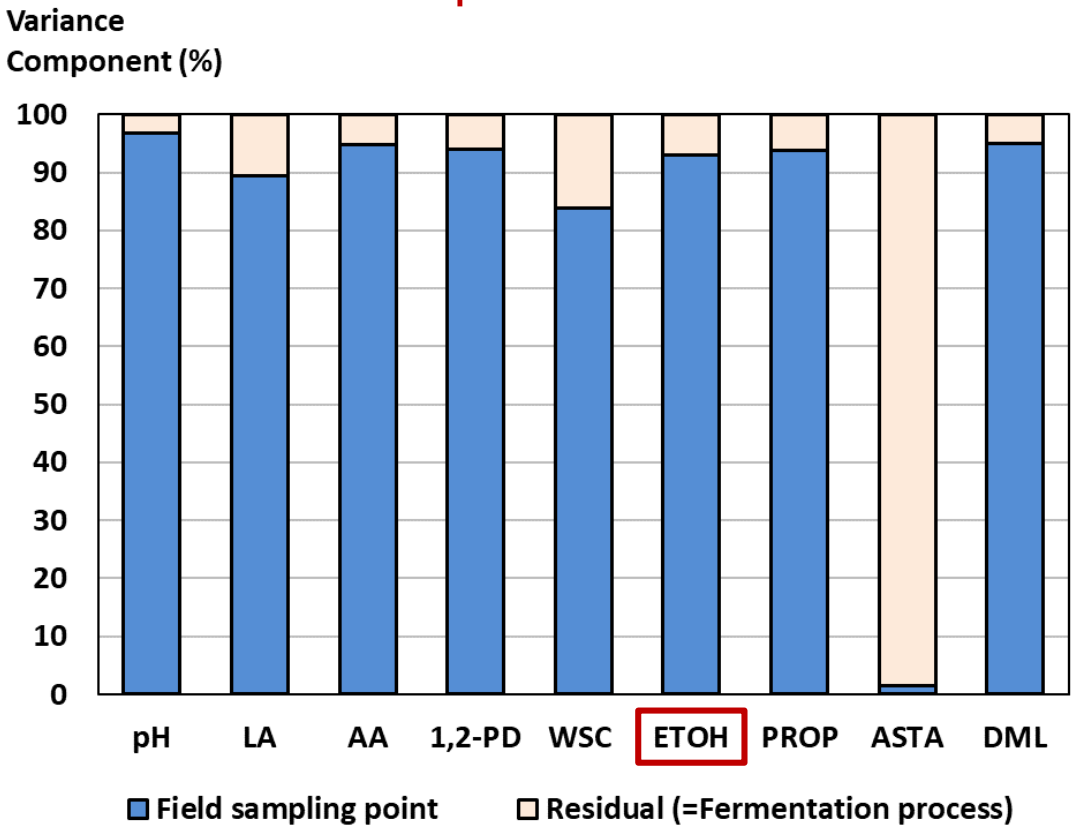


Decomposition of total variability *(random effects model)*

Source of variation	Variance component	percentage %
Field sampling point	12.42	93
Residual (Fermentation process)	0.93	7
Total	13.35	100

3.1 Results – field sampling locations (CON)

Decomposition of total variability of observed values



All traits *(except ASTA)*:

- largest fraction of variation was caused by field sampling point,
- remaining residual variation was related to effects of fermentation process of the ten replicates per sampling point.

ASTA:

- was affected almost completely by the fermentation process.

Note: How far the efficacy of silage additives will be affected by sampling points cannot be shown in our study.

But: Final evaluation of silage additive effects should request more than one trial (EFSA).

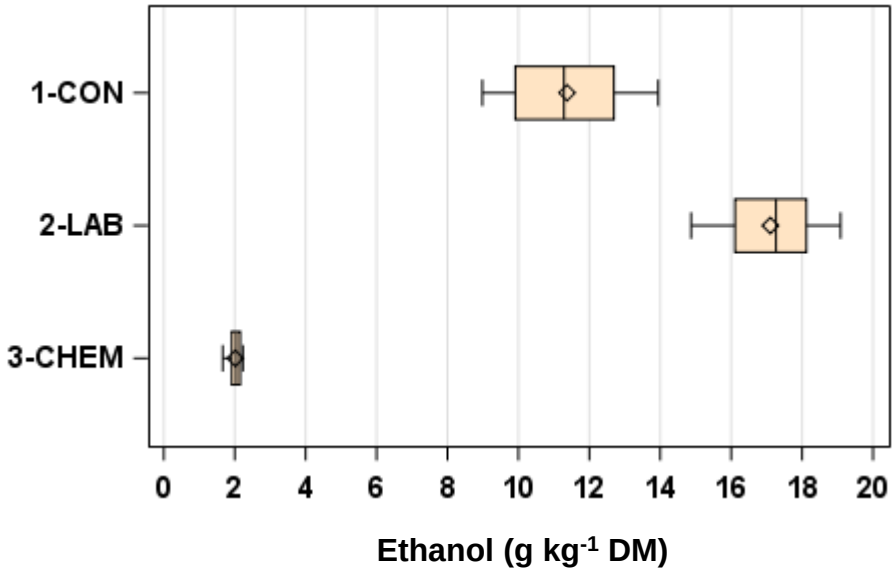
3.2 Results – grass mixture (CON, LAB, CHEM)

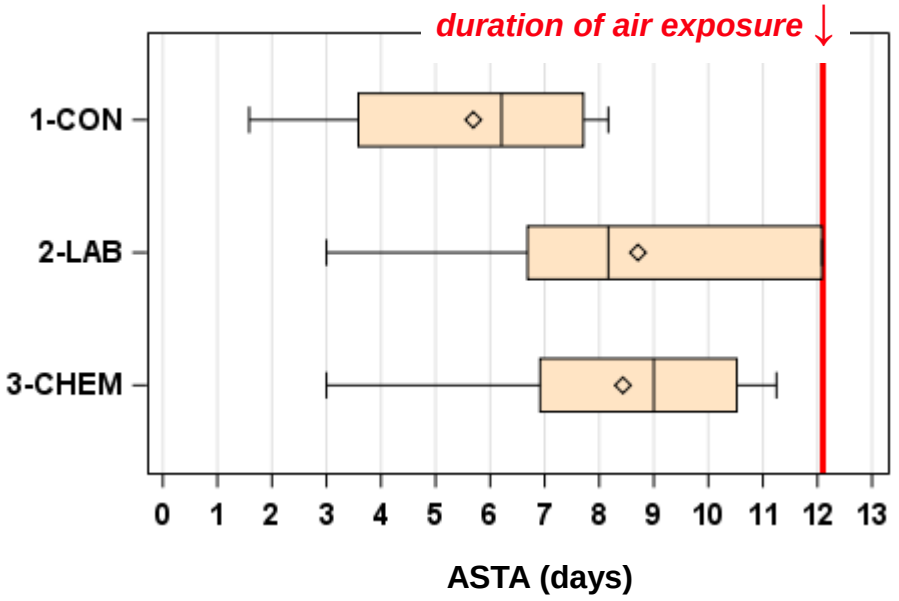
Comparison of treatments

- *fixed effects model,*
- *Anova + Tukey's test procedure,*
- *assuming normally distributed data, variance homogeneity,*
- *scope of inference: one composed material, mini-silos.*

Table of LSMeans, n=10

Treatment	Ethanol	ASTA
CON	11.4 b	5.7 a
LAB	17.1 c	8.7 a
CHEM	2.0 a	8.4 a
HSD (α=5%)	1.4	3.1
s% Residual	12.2	37.1

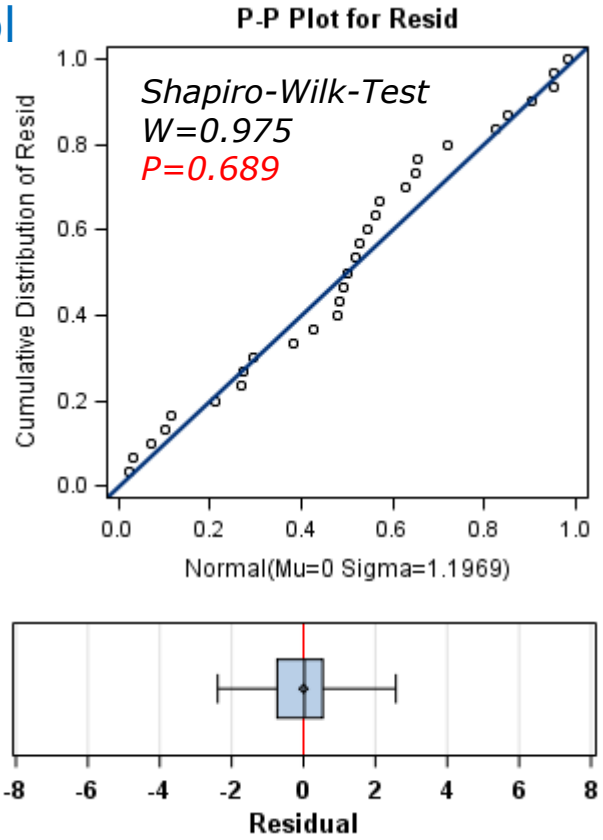




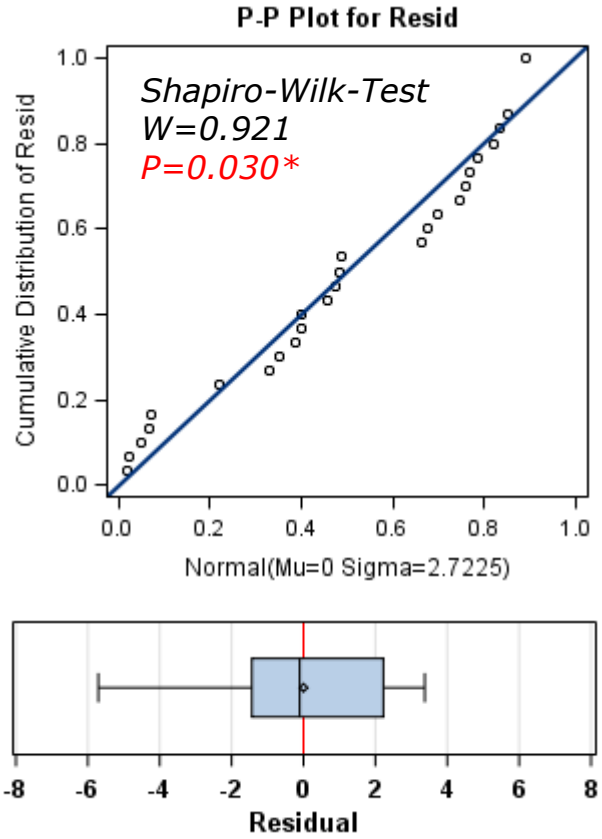
3.2 Results – grass mixture (CON, LAB, CHEM)

Observations come from populations with normally distributed data?

Ethanol



ASTA



Here: 3 treatments x (n=10) → 30 residuals

But if: 3 treatments x (n= 3) → 9 residuals ⇒ **How reliable are test and graphs?**

3.2 Results – grass mixture (CON, LAB, CHEM)

Observations come from populations with homogeneous variances?

Trait	Residual variance				AIC (fit criteria) *"smaller is better"*	
	CON	LAB	CHEM	**Average**	**Var.hom.**	**Var.het.**
Acetic acid	4.0	31.1	5.1	13.4	155.6	147.6
Ethanol	**2.61**	**1.98**	**0.03**	**1.54**	**97.2**	**73.1**
ASTA	**4.9**	**11.3**	**7.7**	**8.0**	**141.5**	**144.0**
DM losses	0.002	0.090	0.007	0.033	-6.3	-30.3

Anova approach, assuming variance homogeneity, average residual variance used

ANOVA approach, assuming variance heterogeneity, individual residual variances used

!

Here: treatment variances estimated from n=10 (as basis for inferences).

With n=3/6 also reliable estimations?

3.2 Results – grass mixture (CON, LAB, CHEM)

Extraction of subsets from the whole sample (n=10) – e.g. Ethanol

n=3

Treatment	Replication									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
CON	9.9	10.7	10.4	13.9	9.0	11.8	12.5	9.8	12.6	12.9
LAB	16.4	17.6	19.1	15.4	14.9	17.8	17.0	16.4	19.1	17.5
CHEM	2.0	2.0	1.7	1.8	2.2	2.0	2.1	2.1	2.1	2.2
	subset 1 ...							**... subset 120**		

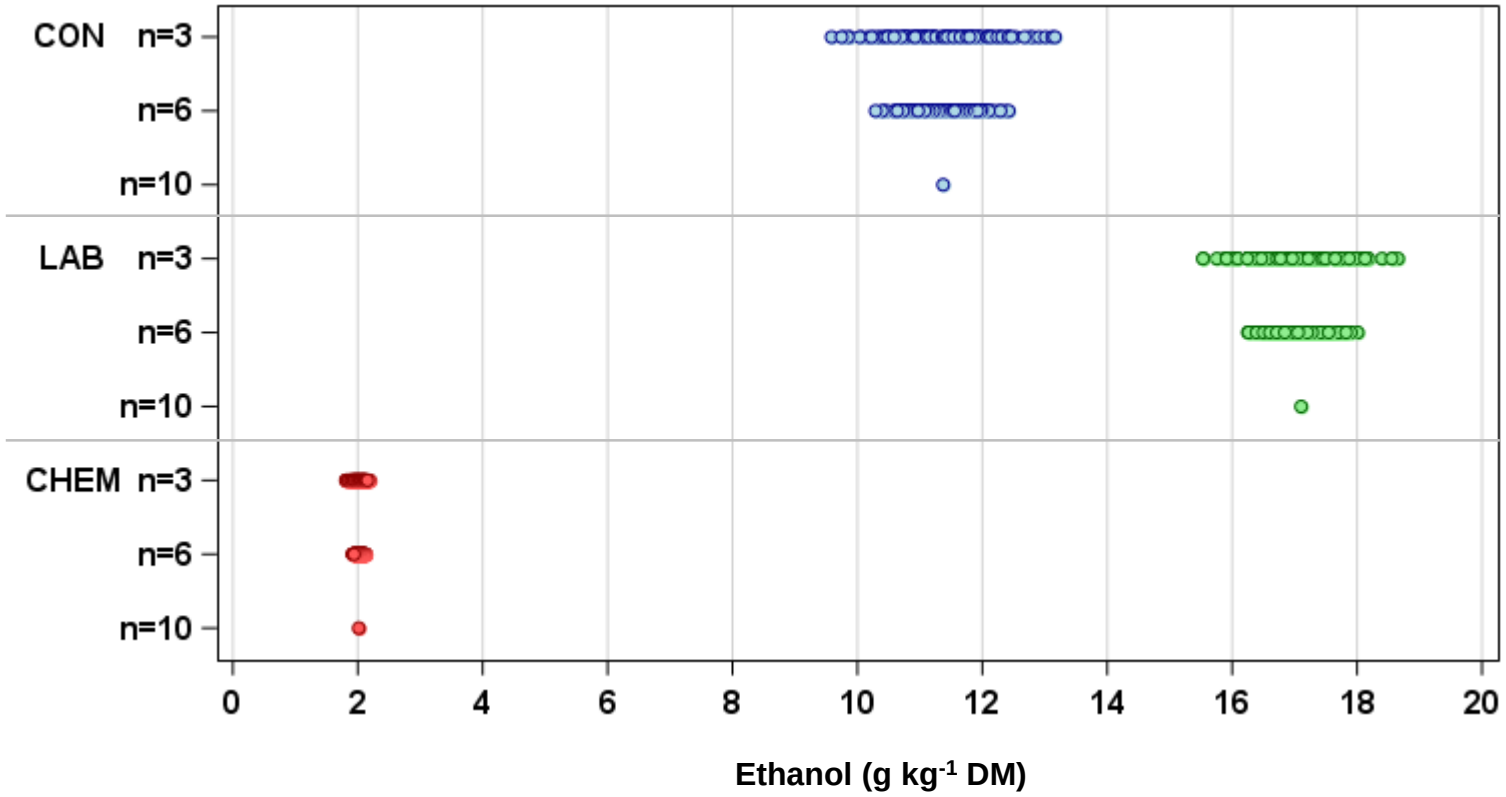
n=6

Treatment	Replication									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
CON	9.9	10.7	10.4	13.9	9.0	11.8	12.5	9.8	12.6	12.9
LAB	16.4	17.6	19.1	15.4	14.9	17.8	17.0	16.4	19.1	17.5
CHEM	2.0	2.0	1.7	1.8	2.2	2.0	2.1	2.1	2.1	2.2
	subset 1 ...				**... subset 210**					

⇨ separate data analyses for all subsamples

3.2 Results – grass mixture (CON, LAB, CHEM)

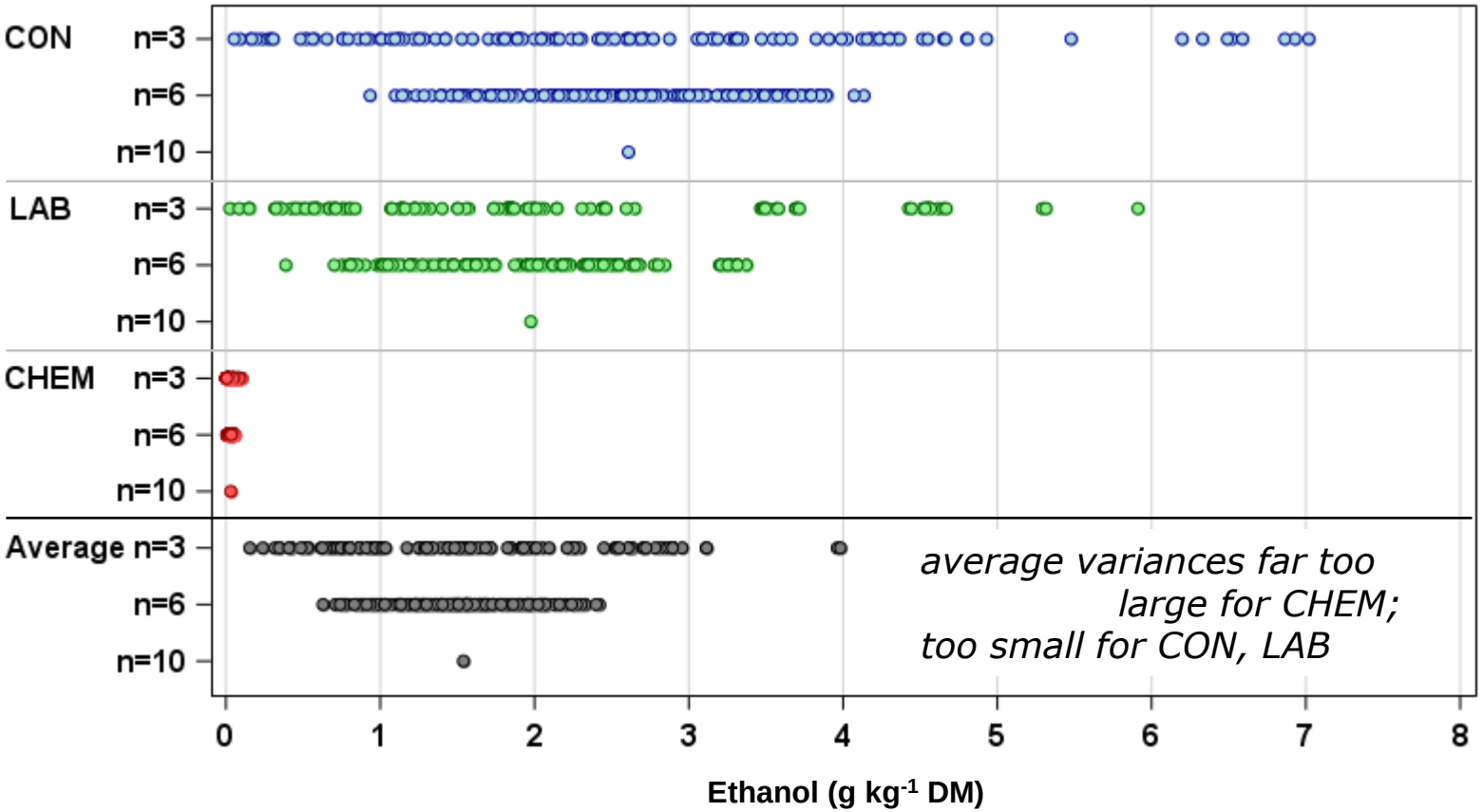
LSMeans for total sample and subsets – **Ethanol**



Compared to CON, LAB increases ethanol content, whereas CHEM decreases ...

3.2 Results – grass mixture (CON, LAB, CHEM)

Residual variances for total sample and subsets – **Ethanol**

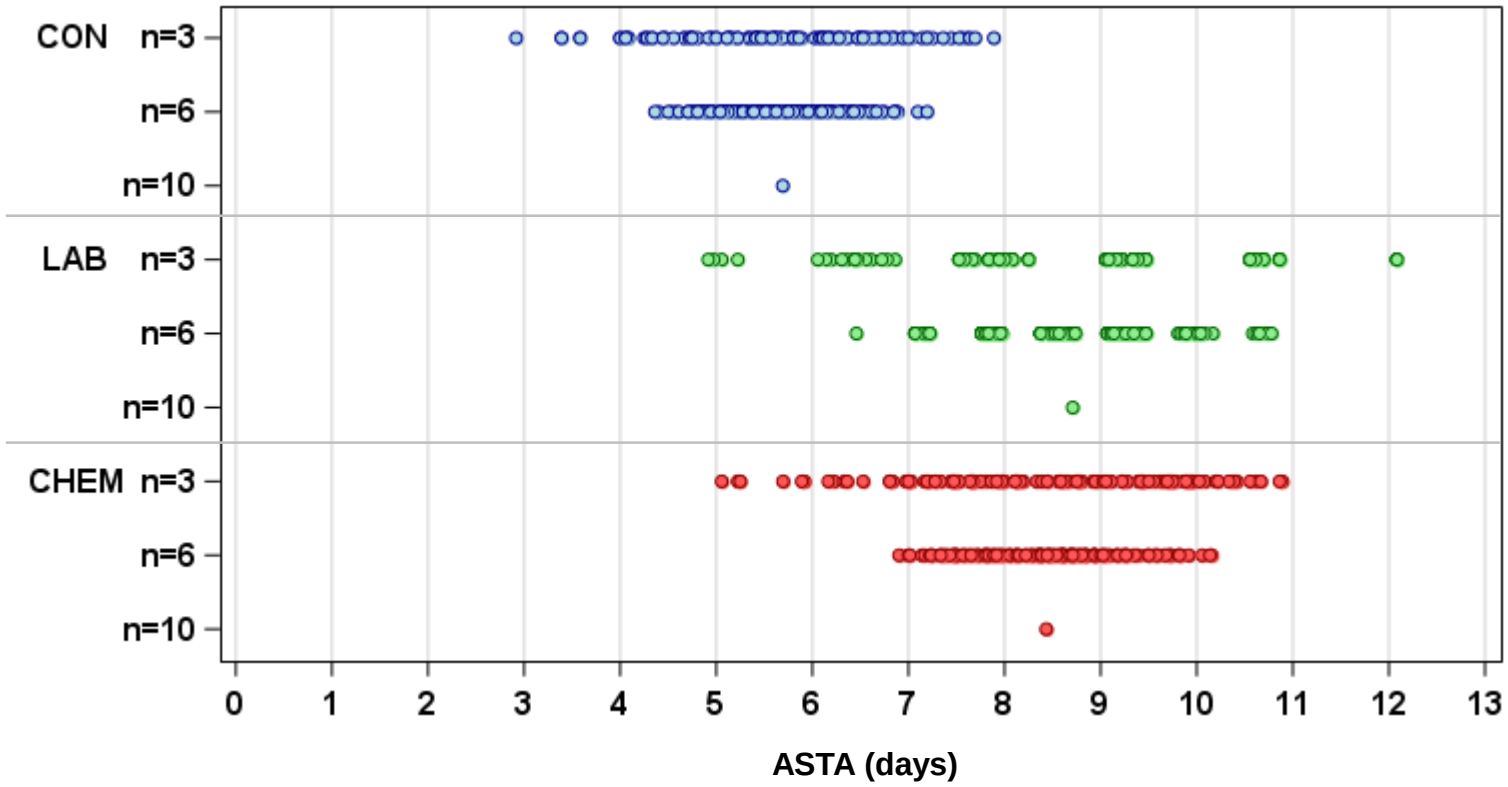


Compared to CON, LAB increases ethanol content, whereas CHEM decreases ...

CHEM reduces the variability dramatically!

3.2 Results – grass mixture (CON, LAB, CHEM)

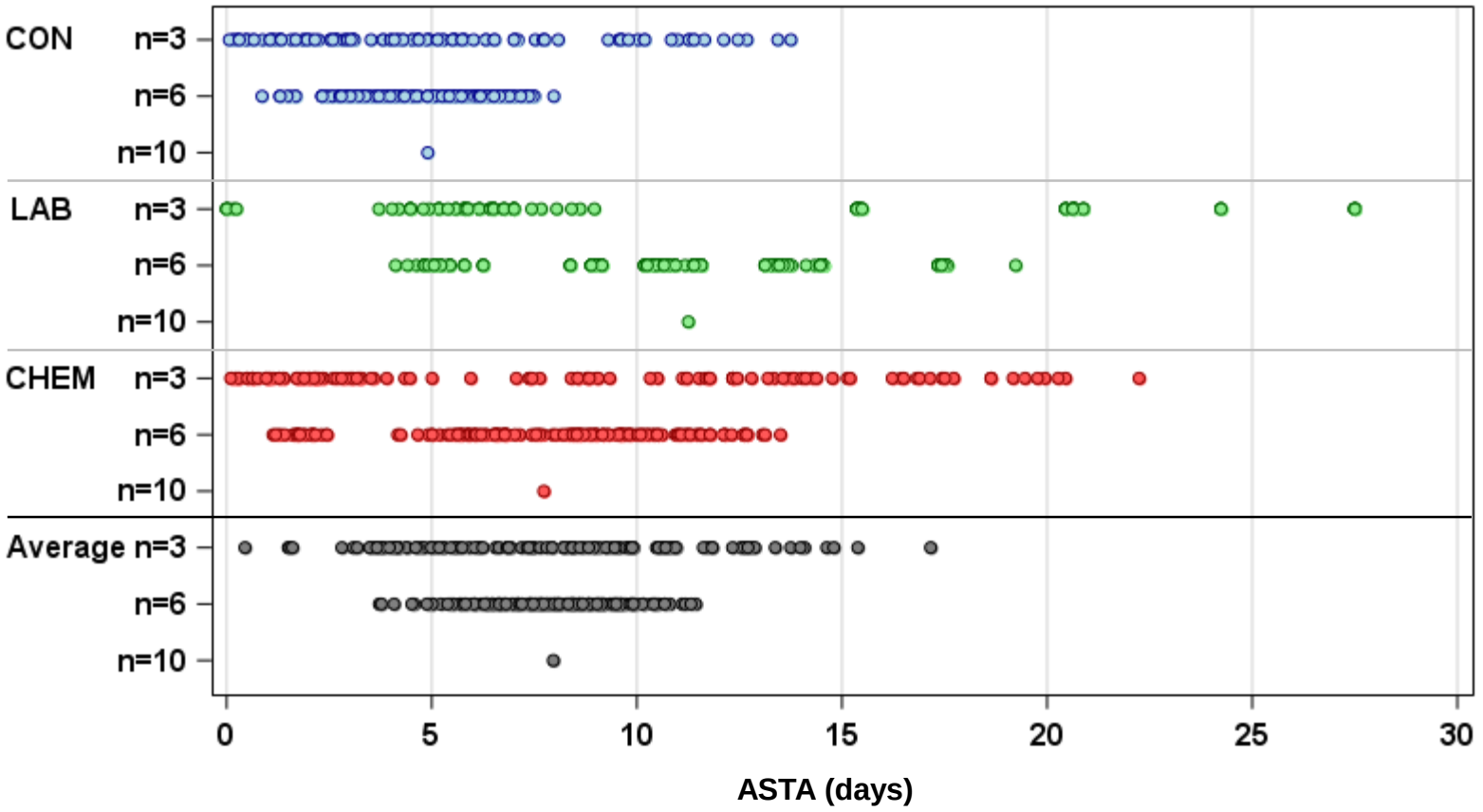
LSMeans for total sample and subsets – **ASTA**



Compared to CON, LAB and CHEM show slightly higher aerobic stability ...

3.2 Results – grass mixture (CON, LAB, CHEM)

Residual variances for total sample and subsets – **ASTA**



Compared to CON, LAB and CHEM show slightly higher stability ...

variability similar.

3.2 Results – grass mixture (CON, LAB, CHEM)

Comparison of treatments – Significance versus Relevance

ASTA

(days, n=10)

Comparison	Difference	assuming variance homogeneity P-value	HSD (α=5%)	Confidence limits
LAB – CON	3.0	0.060	3.1	[-0.1 ; +6.1]
CHEM – CON	2.7	0.094	3.1	[-0.4 ; +5.9]

No relevance ???

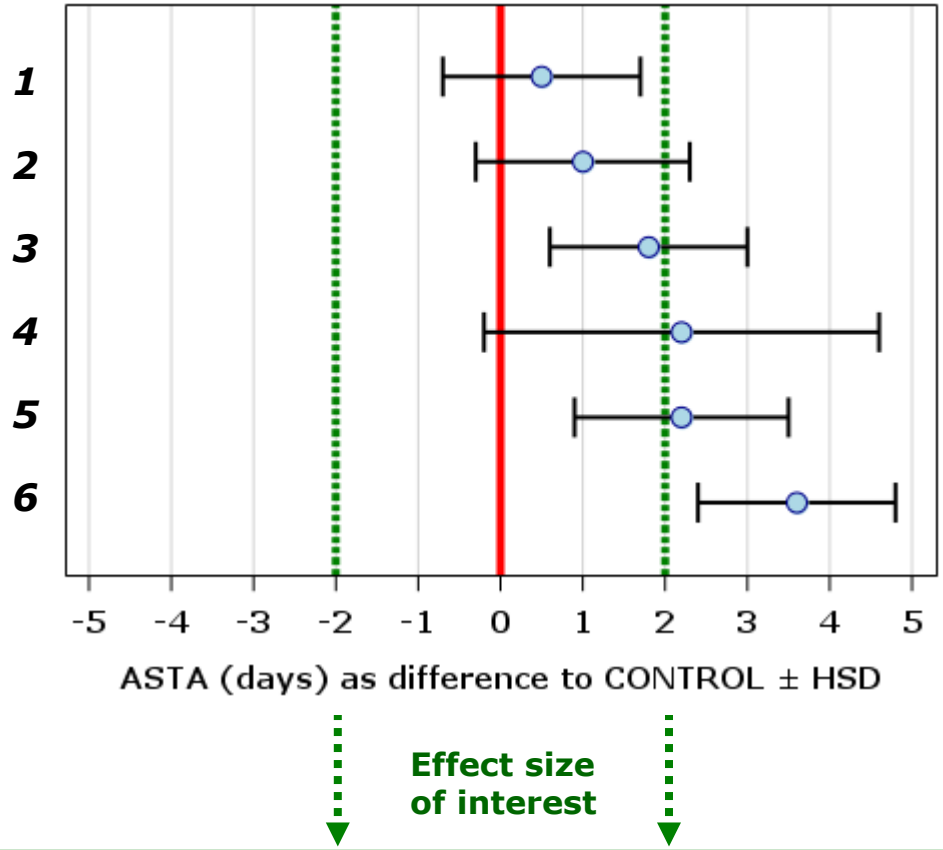
No significance !!!

EFSA (2008): **"... additive shall be stable <u>two days longer</u> than untreated control ..."**.

Effect size of interest

3.2 Results – grass mixture (CON, LAB, CHEM)

Comparison of treatments – Significance versus Relevance



Scenario	Significance	Relevance
1	-	-
2	-	(+)
3	+	(+)
4	-	+
5	+	+
6	+	+

EFSA (2008): **"... additive shall be stable <u>two days longer</u> than untreated control ..."**.

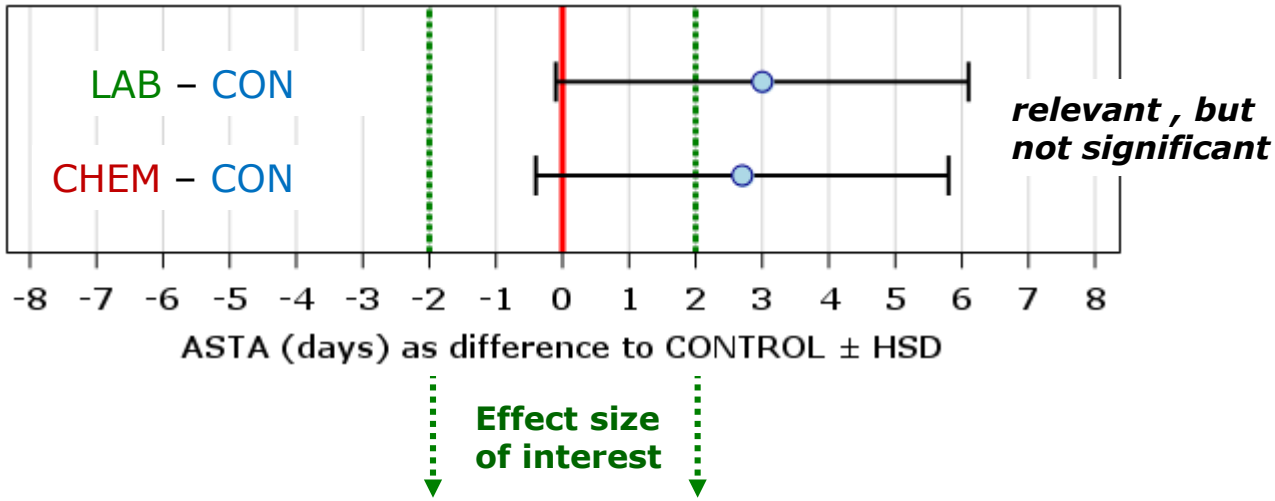
3.2 Results – grass mixture (CON, LAB, CHEM)

Comparison of treatments – Significance versus Relevance

ASTA

(days, n=10)

Comparison	Difference	assuming variance homogeneity P-value	HSD (α=5%)	Confidence limits
LAB – CON	3.0	0.060	3.1	[-0.1 ; +6.1]
CHEM – CON	2.7	0.094	3.1	[-0.4 ; +5.9]



EFSA (2008): **"... additive shall be stable <u>two days longer</u> than untreated control ..."**.

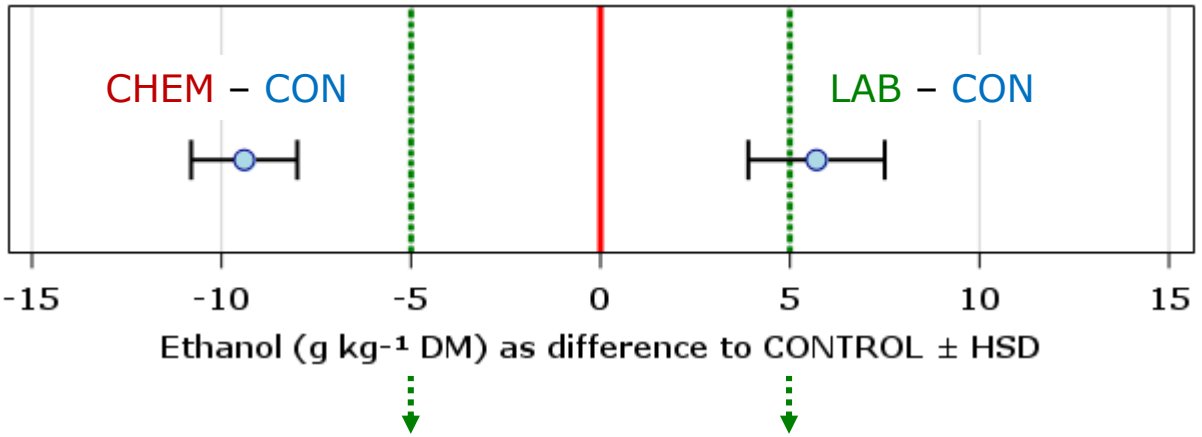
3.2 Results – grass mixture (CON, LAB, CHEM)

Comparison of treatments – Significance versus Relevance

Ethanol

(g kg^{-1} DM, n=10)

Comparison	Difference	assuming variance heterogeneity P-value	HSD (α=5%)	Confidence limits
LAB – CON	5.7	<0.001	1.8	[3.9 ; 7.5]
CHEM – CON	-9.4	<0.001	1.4	[-10.7 ; -8.0]



Which effect size of interest is relevant ???
Necessary for interpretation!

3.2 Results – grass mixture (CON, LAB, CHEM)

Comparison of treatments – problem of multiplicity

(caused by number of comparisons + numer of response variables)

All pairwise comparisons

1	LAB	– CON
2	CHEM	– CON
3	CHEM	– LAB

$$\alpha^* = 1-(1-\alpha)^c$$

α* - Experiment-wise Type I error rate

α - Comparison-wise Type I error rate

Same sample used for several tests – results are not independent!

number of treatments	2	**3**	4	5	6	8	10	12	14
c (pairwise comparisons)	1	**3**	6	10	15	28	45	66	91
α* (α=0.05)	0.05	**0.14**	0.26	0.40	0.54	0.76	0.90	0.97	0.99

only here: **t-test = Tukey's test**

The more comparisons, the more findings of something in the data!

Error inflation particularly problematic,
when large variability + small sample size come together.

3.2 Results – grass mixture (CON, LAB, CHEM)

ASTA as one of the most important responses

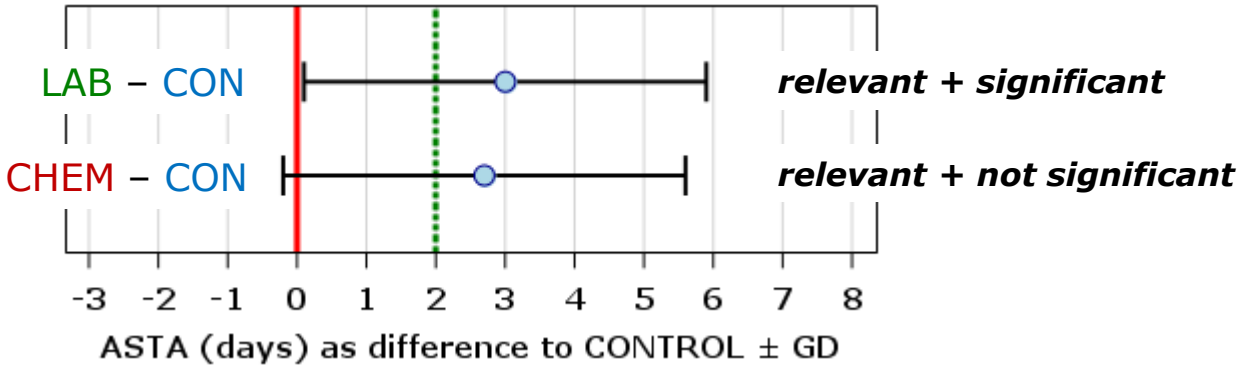
(1) assuming normally distributed values → parametric data analysis

EFSA (2008): **"... additive shall be stable <u>two days longer</u> than untreated control ..."**.

~~*all pairwise comparisons, two-sided*~~

Comparisons versus control, one-sided

Comparison	Difference (n=10)	Tukey ($HSD_{5\%}$ = 3.1) P-value	Tukey Conf. limits	Dunnett ($GD_{5\%}$ = 2.9) P-value	Dunnett Conf. limits
LAB – CON	3.0	0.060	[-0.1 ; +6.1]	0.044	[+0.1 ; +6.0]
CHEM – CON	2.7	0.094	[-0.4 ; +5.9]	0.070	[-0.2 ; +5.7]



3.2 Results – grass mixture (CON, LAB, CHEM)

ASTA as one of the most important responses

(2) assuming non-normally distributed values → nonparametric data analysis

→ **rank procedure with ANOVA-Typ-Statistics** (SAS, Proc Mixed)

- for ≥ 1 treatment factor, variance heterogeneity of ranks considered, identical observations no problem, ...
- but: minimal sample size for reliable results about n=10

Treatment	LS Mean	Rank mean
CON	5.7	9.6
LAB	8.7	18.6
CHEM	8.4	18.4

(n=10)

Comparison	Contrasts *(Bonferroni correction)* P-value (1)	P-value (2)
LAB – CON	0.071	0.024
CHEM – CON	0.031	0.010

all pairwise comparisons, two-sided *„Tukey“*

comparisons versus Control, one-sided *„Dunnett“*

3.2 Results – grass mixture (CON, LAB, CHEM)

ASTA as one of the most important responses

(3) How to consider identical observations for treatments?

- grass mixture: duration of air exposure 12.1 days

Treatment	observations (days), n=10
CON	1.6 ... 8.2
LAB	3.0 ... 8.4 12.1 12.1 12.1 12.1
CHEM	3.0 ... 11.3

- often situation more extreme (Weiss et. al 2016): duration of air exposure 7 days

Treatment	observations (days), n=3
CON	4.0 4.8 2.8
Additive	7.0 7.0 7.0

Normal distribution ?

Homogeneous variances ?

Rank procedures with n=3 ?

But: Practical conclusions are totally clear !

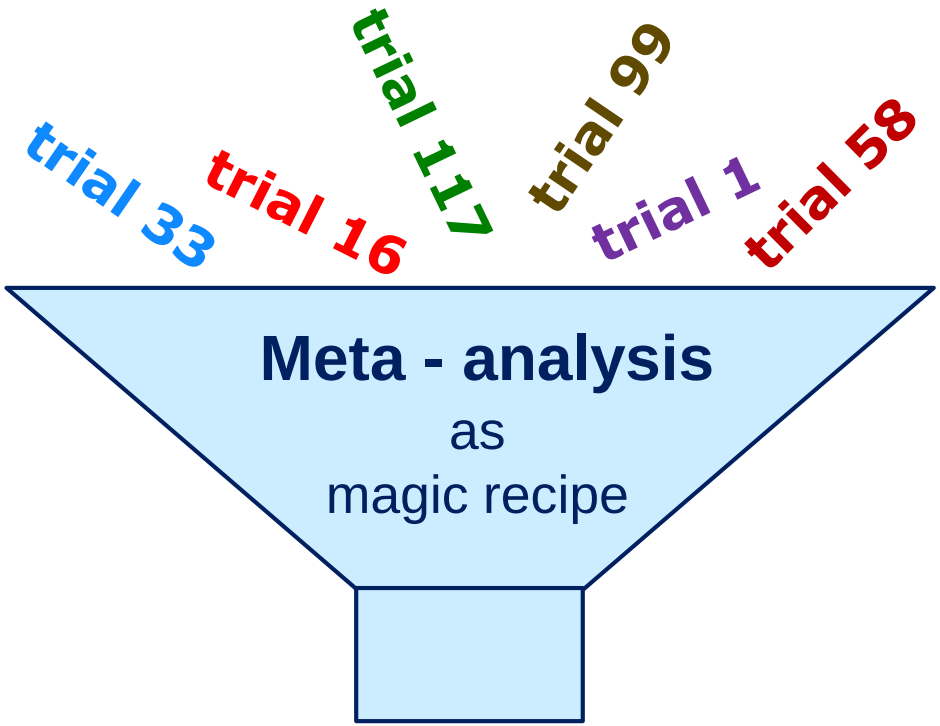
4. Summary

- Comparative silage experiments are most frequently performed with **few replications**.
- The traits of interest for the evaluation of treatments do **not meet** in each case the **assumptions** for the chosen statistical analysis procedures.
- Moreover, **checking the assumptions** by statistical preliminary tests and **performing the analysis** on the **same data** are **problematic**.
- Often, the experimenter has **information beforehand** about treatments which have an impact not only on the magnitude but also on variation.
- Both, **rank procedures** in case of non-normality as well as reliable estimations of individual treatment variances in case of **variance heterogeneity** demand **higher sample sizes**.
- Together with well justified effect sizes of interest, the trials should be designed in such a way that **significance** and **relevance** of results **come together**.

4. Summary

Problematic single trials?

Don‘t worry! Put together!



Are aggregated results correct ?????

References

Bello, N. M. & Renter, D. G. (2017)
Invited review: Reproducible research from noisy data: Revisiting key statistical principals for the animal sciences. Journal of Dairy Science, 101, 1-23.

EFSA (2008)
Guidance for the preparation of dossiers for technological additives. The EFSA Journal, 774, 121.

Kaiser, E. & Weiss, K. (1997)
Zum Gärungsverlauf bei der Silierung von nitratarmem Grünfutter.
2.Mitt.: Gärungsverlauf bei Zusatz von Nitrat, Nitrit, Milchsäurebakterien und Ameisensäure.
Archives of Animal Nutrition, 50, 187-200.

Nuzzo, R. (2015)
Fooling ourselves. Nature, 526, 182-185.

Udén, P. & Robinson, P. H. (2015)
Design and statistical issues in silage experiments.
In: Nussio, L. G., de Sousa, D. O., Gritti, V. C., Salvati, G. G., Santos, W. P. & Salvo, P. A. R. (eds.). V International Symposium on Forage Quality and Conservation, Piracicaba, SP, Brazil, 2015, 166-179.

Weiss, K., Kroschewski, B. & Auerbach, H. (2016)
Effects of air exposure, temperature and additives on fermentation characteristics, yeast count, aerobic stability and volatile organic compounds in corn silage. Journal of Dairy Science, 99, 1-17