Statistics and experimental design in silage research: Some comments on design and analysis of comparative silage experiments

Bärbel Kroschewski¹, Kirsten Weiß¹ & Horst Auerbach²

¹Humboldt Universität zu Berlin ²ISC Wettin-Löbejün
1. Introduction

“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

P1+P2+P3 = MIX

CON

without silage additive

1 ... 10

LAB

biological additive

1 ... 10

Lactobacillus buchneri
CNCM-I 4323 and
Pediococcus acidilactici
DSM 11673

CHEM

chemical additive

1 ... 10

sodium nitrite,
hexamethylenetetramine
and potassium sorbate
2. Description of the grassland trial

Lab-scale ensiling trial

Traits → fresh forage

per sampling point (P1, P2, P3)
n=5

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

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

<table>
<thead>
<tr>
<th>Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DM</strong></td>
</tr>
<tr>
<td><strong>WSC</strong></td>
</tr>
<tr>
<td><strong>NO₃⁻</strong></td>
</tr>
<tr>
<td><strong>BC</strong></td>
</tr>
<tr>
<td><strong>Yeast</strong></td>
</tr>
<tr>
<td><strong>Lactobac</strong></td>
</tr>
</tbody>
</table>

**Known to have an influence on the fermentation process**

**traits → silage**

<table>
<thead>
<tr>
<th>Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>pH</strong></td>
</tr>
<tr>
<td><strong>LA</strong></td>
</tr>
<tr>
<td><strong>AA</strong></td>
</tr>
<tr>
<td><strong>1,2-PD</strong></td>
</tr>
<tr>
<td><strong>WSC</strong></td>
</tr>
<tr>
<td><strong>ETOH</strong></td>
</tr>
<tr>
<td><strong>PROP</strong></td>
</tr>
<tr>
<td><strong>ASTA</strong></td>
</tr>
<tr>
<td><strong>DML</strong></td>
</tr>
</tbody>
</table>

**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.

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

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Variance component</th>
<th>percentage %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Field sampling point</td>
<td>12.42</td>
<td>93</td>
</tr>
<tr>
<td>Residual (Fermentation process)</td>
<td>0.93</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td>13.35</td>
<td>100</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Ethanol</th>
<th>ASTA</th>
</tr>
</thead>
<tbody>
<tr>
<td>CON</td>
<td>11.4 b</td>
<td>5.7 a</td>
</tr>
<tr>
<td>LAB</td>
<td>17.1 c</td>
<td>8.7 a</td>
</tr>
<tr>
<td>CHEM</td>
<td>2.0 a</td>
<td>8.4 a</td>
</tr>
<tr>
<td>HSD (α=5%)</td>
<td>1.4</td>
<td>3.1</td>
</tr>
<tr>
<td>s% Residual</td>
<td>12.2</td>
<td>37.1</td>
</tr>
</tbody>
</table>
3.2 Results – grass mixture (CON, LAB, CHEM)

Observations come from populations with normally distributed data?

Ethanol

Shapiro-Wilk-Test
W=0.975
P=0.689

ASTA

Shapiro-Wilk-Test
W=0.921
P=0.030*

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?

<table>
<thead>
<tr>
<th>Trait</th>
<th>Residual variance</th>
<th>AIC (fit criteria)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CON</td>
<td>LAB</td>
</tr>
<tr>
<td>Acetic acid</td>
<td>4.0</td>
<td>31.1</td>
</tr>
<tr>
<td>Ethanol</td>
<td>2.61</td>
<td>1.98</td>
</tr>
<tr>
<td>ASTA</td>
<td>4.9</td>
<td>11.3</td>
</tr>
<tr>
<td>DM losses</td>
<td>0.002</td>
<td>0.090</td>
</tr>
</tbody>
</table>

Anova approach, assuming variance homogeneity, average residual variance used.

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

With \( n=3/6 \) also reliable estimations?

ANOVA approach, assuming variance heterogeneity, individual residual variances used.
### 3.2 Results – grass mixture (CON, LAB, CHEM)

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

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Replication</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>CON</td>
<td>9.9 10.7 10.4 13.9 9.0 11.8 12.5</td>
</tr>
<tr>
<td>LAB</td>
<td>16.4 17.6 19.1 15.4 14.9 17.8 17.0</td>
</tr>
<tr>
<td>CHEM</td>
<td>2.0 2.0 1.7 1.8 2.2 2.0 2.1</td>
</tr>
</tbody>
</table>

**subset 1 ...**

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Replication</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>CON</td>
<td>9.9 10.7 10.4 13.9</td>
</tr>
<tr>
<td>LAB</td>
<td>16.4 17.6 19.1 15.4</td>
</tr>
<tr>
<td>CHEM</td>
<td>2.0 2.0 1.7 1.8</td>
</tr>
</tbody>
</table>

**subset 1 ...**

**... subset 120**

**Subset n=3**

**Subset n=6**

**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!

average variances far too large for CHEM; too small for CON, LAB
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**

<table>
<thead>
<tr>
<th></th>
<th>Comparison</th>
<th>Difference</th>
<th>P-value</th>
<th>HSD ((\alpha=5%))</th>
<th>Confidence limits</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ASTA (days, n=10)</strong></td>
<td>LAB – CON</td>
<td>3.0</td>
<td>0.060</td>
<td>3.1</td>
<td>[-0.1 ; +6.1]</td>
</tr>
<tr>
<td></td>
<td>CHEM – CON</td>
<td>2.7</td>
<td>0.094</td>
<td>3.1</td>
<td>[-0.4 ; +5.9]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>No significance !!!</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>No relevance ???</strong></td>
<td></td>
</tr>
</tbody>
</table>

**EFSA (2008):** “... additive shall be stable two days longer 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 | + | +

ASTA (days) as difference to CONTROL ± HSD

EFSA (2008): “... additive shall be stable two days longer than untreated control ...”
3.2 Results – grass mixture (CON, LAB, CHEM)

Comparison of treatments – Significance versus Relevance

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Difference</th>
<th>P-value</th>
<th>HSD (α=5%)</th>
<th>Confidence limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAB – CON</td>
<td>3.0</td>
<td>0.060</td>
<td>3.1</td>
<td>[-0.1 ; +6.1]</td>
</tr>
<tr>
<td>CHEM – CON</td>
<td>2.7</td>
<td>0.094</td>
<td>3.1</td>
<td>[-0.4 ; +5.9]</td>
</tr>
</tbody>
</table>

EFSA (2008): “... additive shall be stable two days longer than untreated control ...”. 

---

**ASTA**

*(days, n=10)*
3.2 Results – grass mixture (CON, LAB, CHEM)

Comparison of treatments – Significance versus Relevance

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

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Difference</th>
<th>P-value</th>
<th>HSD (α=5%)</th>
<th>Confidence limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAB – CON</td>
<td>5.7</td>
<td>&lt;0.001</td>
<td>1.8</td>
<td>[3.9 ; 7.5]</td>
</tr>
<tr>
<td>CHEM – CON</td>
<td>-9.4</td>
<td>&lt;0.001</td>
<td>1.4</td>
<td>[-10.7 ; -8.0]</td>
</tr>
</tbody>
</table>

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 + number of response variables)*

<table>
<thead>
<tr>
<th>All pairwise comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 LAB – CON</td>
</tr>
<tr>
<td>2 CHEM – CON</td>
</tr>
<tr>
<td>3 CHEM – LAB</td>
</tr>
</tbody>
</table>

\[ \alpha^* = 1 - (1 - \alpha)^c \]

\( \alpha^* \) - Experiment-wise Type I error rate
\( \alpha \) - Comparison-wise Type I error rate

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

<table>
<thead>
<tr>
<th>number of treatments</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>c (pairwise comparisons)</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>10</td>
<td>15</td>
<td>28</td>
<td>45</td>
<td>66</td>
<td>91</td>
</tr>
<tr>
<td>( \alpha^* ) (( \alpha = 0.05 ))</td>
<td>0.05</td>
<td>0.14</td>
<td>0.26</td>
<td>0.40</td>
<td>0.54</td>
<td>0.76</td>
<td>0.90</td>
<td>0.97</td>
<td>0.99</td>
</tr>
</tbody>
</table>

only here: \textit{t-test} = \textit{Tukey’s test}

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

The more comparisons, the more findings of something in the data!
### 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 **two days longer** than untreated control …”.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Difference</th>
<th>Tukey (HSD_{5%} = 3.1)</th>
<th>Dunnett (GD_{5%} = 2.9)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n=10</td>
<td>P-value</td>
<td>Conf. limits</td>
</tr>
<tr>
<td>LAB – CON</td>
<td>3.0</td>
<td>0.060</td>
<td>[-0.1 ; +6.1]</td>
</tr>
<tr>
<td>CHEM – CON</td>
<td>2.7</td>
<td>0.094</td>
<td>[-0.4 ; +5.9]</td>
</tr>
</tbody>
</table>

*all pairwise comparisons, two-sided*  

*Comparisons versus control, one-sided*

---

**Relevant + significant**

**Relevant + not significant**
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

<table>
<thead>
<tr>
<th>Treatment</th>
<th>LS Mean</th>
<th>Rank mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>CON</td>
<td>5.7</td>
<td>9.6</td>
</tr>
<tr>
<td>LAB</td>
<td>8.7</td>
<td>18.6</td>
</tr>
<tr>
<td>CHEM</td>
<td>8.4</td>
<td>18.4</td>
</tr>
</tbody>
</table>

(n=10)

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Contrasts (Bonferroni correction)</th>
<th>P-value (1)</th>
<th>P-value (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAB – CON</td>
<td></td>
<td>0.071</td>
<td>0.024</td>
</tr>
<tr>
<td>CHEM – CON</td>
<td></td>
<td>0.031</td>
<td>0.010</td>
</tr>
</tbody>
</table>

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**

<table>
<thead>
<tr>
<th>Treatment</th>
<th>observations (days), n=10</th>
</tr>
</thead>
<tbody>
<tr>
<td>CON</td>
<td>1.6 ... 8.2</td>
</tr>
<tr>
<td>LAB</td>
<td>3.0 ... 8.4 <strong>12.1 12.1 12.1 12.1</strong></td>
</tr>
<tr>
<td>CHEM</td>
<td>3.0 ... 11.3</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Treatment</th>
<th>observations (days), n=3</th>
</tr>
</thead>
<tbody>
<tr>
<td>CON</td>
<td>4.0 4.8 2.8</td>
</tr>
<tr>
<td>Additive</td>
<td>7.0 7.0 7.0</td>
</tr>
</tbody>
</table>
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!

Meta - analysis as magic recipe

Are aggregated results correct ??????
References

Bello, N. M. & Renter, D. G. (2017)

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


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