Statistical Evaluation of Biotests

Glyceria Ringtest Workshop 2021

Part IV: Statistical Design and Importance of Uniformity

Dr. Monika Ratte
ToxRat Solutions GmbH & Co. KG

May 10 2021

Outline

• Statistical Design and validity criteria of the Glyceria Test
• Why are Coefficients of Variation (CoV) relevant at all and why is the critical value set to 35%?
• What can we learn from previous ring tests results so far?
• Sources of variability in Glyceria testing
• A common error about sample size and variability
• How to reduce CoV – which parameters affect variability?
• The issue of calculating yields and growth rates – how to measure initial weights?
• The issue of „representative plants“ – what are they important for?
• Conclusion: The Importance of Uniformity

Always the talk of „uniformity“. It all has to be practicable, too!
Experimental Design Glyceria Test

Test protocol:

“The test should incorporate six replicate test vessels for the untreated control and four replicate test vessels for each of a minimum of six concentration levels. Each test vessel represents a replicate containing one plant pot. A single pot will contain one shoot, attached to a short section of rhizome.”

Validity criteria

- Doubling of FW until d14
- CoV% Y FW ≤ 35%

If these designated validity criteria were applied,

- 72% of the tests (8 out of 11) were found to be valid with respect of doubling time for FW
- but only 18% of the tests (2 out of 11) would meet the maximum coefficient of variation for YFW of 35%.

→ To adopt a maximum variability of 35% for YFW as validity criterion, the intra-laboratory variability of YFW must be reduced.
Statistical Evaluation of Biotests

Why are coefficients of variation (CoV) relevant at all and why is the critical value set to 35%?

Example: Yield FW data from ring test with Imazapyr

<table>
<thead>
<tr>
<th>Treatment</th>
<th>500</th>
<th>1200</th>
<th>2300</th>
<th>3200</th>
<th>4100</th>
<th>5200</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.51</td>
<td>0.588</td>
<td>0.387</td>
<td>0.581</td>
<td>0.158</td>
<td>0.258</td>
</tr>
<tr>
<td>2</td>
<td>0.900</td>
<td>0.235</td>
<td>0.015</td>
<td>0.168</td>
<td>0.274</td>
<td>0.372</td>
</tr>
<tr>
<td>3</td>
<td>0.000</td>
<td>0.277</td>
<td>0.740</td>
<td>0.523</td>
<td>0.657</td>
<td>0.109</td>
</tr>
<tr>
<td>4</td>
<td>1.188</td>
<td>0.546</td>
<td>0.411</td>
<td>0.591</td>
<td>0.417</td>
<td>0.303</td>
</tr>
<tr>
<td>5</td>
<td>0.003</td>
<td>0.021</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.025</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Effect size: 26% 50% 53% 65% 74% 83%

Mean: 1.00 g  CoV: 30.4%

Dunnett’s Multiple t-test Procedure

Dunnett’s multiple t-test procedure with yield fix at 14.0 d. Comparison of treatments with “Control”. Significance was Alpha = 0.050, one-sided (multiple level). Mean; arithmetic mean, n; sample size, s; standard deviation, MDD; minimum decisionable difference to Control (in percent of Control), t; sample t; t*; critical t for H0; μ1 = ... = μk, the differences are significant in case |t| > |t*| (The residual variance of an ANOVA was applied; df = N - k; N: sum of treatment replicates n(k), k: number of treatments).

A NOEC of 60.000 μg/L is suggested by the program.
**Statistical Evaluation of Biotests**

Same means, same effect sizes, but....

<table>
<thead>
<tr>
<th>Time</th>
<th>0.0</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>1.56</td>
<td>0.58</td>
<td>0.58</td>
<td>0.58</td>
<td>0.58</td>
<td>0.58</td>
</tr>
<tr>
<td>1</td>
<td>1.85</td>
<td>0.58</td>
<td>0.58</td>
<td>0.58</td>
<td>0.58</td>
<td>0.58</td>
</tr>
<tr>
<td>2</td>
<td>0.6</td>
<td>1.27</td>
<td>0.46</td>
<td>0.62</td>
<td>0.58</td>
<td>0.58</td>
</tr>
<tr>
<td>3</td>
<td>1.85</td>
<td>0.58</td>
<td>0.58</td>
<td>0.58</td>
<td>0.58</td>
<td>0.58</td>
</tr>
<tr>
<td>4</td>
<td>0.6</td>
<td>1.27</td>
<td>0.46</td>
<td>0.62</td>
<td>0.58</td>
<td>0.58</td>
</tr>
<tr>
<td>5</td>
<td>1.85</td>
<td>0.58</td>
<td>0.58</td>
<td>0.58</td>
<td>0.58</td>
<td>0.58</td>
</tr>
<tr>
<td>6</td>
<td>0.6</td>
<td>1.27</td>
<td>0.46</td>
<td>0.62</td>
<td>0.58</td>
<td>0.58</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Replicates</th>
<th>Mean</th>
<th>Std Dev</th>
<th>TV (%)</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.85</td>
<td>0.74</td>
<td>68.8</td>
<td>26%</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.58</td>
<td>16.0</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>1.85</td>
<td>0.58</td>
<td>58.1</td>
<td>53%</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.58</td>
<td>16.0</td>
<td>65%</td>
</tr>
<tr>
<td></td>
<td>1.85</td>
<td>0.58</td>
<td>58.1</td>
<td>74%</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.58</td>
<td>16.0</td>
<td>83%</td>
</tr>
</tbody>
</table>

... higher variability in control: CoV = 59.1%

**Dunnett's Multiple t-test Procedure**

Dunnett’s multiple t-test procedure with yield 11 at 10.6 d. Comparison of treatments with
“Control”. Significance was Alpha = 0.05, one-sided smaller (multiple level). Mean: arithmetic mean, n: sample size, s: standard deviation, MDD: minimum detectable difference to Control (in percent of Control), t*: sample t, t*: critical t for H0: μ1 = μ2 = ... = μk; the differences are significant in case if |t*| > |t| (The residual variance of an ANOVA was applied; df = N - k; k: number of treatment replicates n; k). k = number of treatments).

<table>
<thead>
<tr>
<th>Treatment [µg/L]</th>
<th>Mean</th>
<th>s</th>
<th>df</th>
<th>MDD</th>
<th>t*</th>
<th>Sign.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>0.00</td>
<td>0.334</td>
<td>23</td>
<td>-2.32</td>
<td>-2.43</td>
<td>-</td>
</tr>
<tr>
<td>50.000</td>
<td>0.74</td>
<td>0.334</td>
<td>23</td>
<td>-2.32</td>
<td>-2.43</td>
<td>-</td>
</tr>
<tr>
<td>120.000</td>
<td>0.50</td>
<td>0.334</td>
<td>23</td>
<td>-2.32</td>
<td>-2.43</td>
<td>-</td>
</tr>
<tr>
<td>223.000</td>
<td>0.47</td>
<td>0.334</td>
<td>23</td>
<td>-2.32</td>
<td>-2.43</td>
<td>+</td>
</tr>
<tr>
<td>781.000</td>
<td>0.35</td>
<td>0.334</td>
<td>23</td>
<td>-2.32</td>
<td>-2.43</td>
<td>+</td>
</tr>
<tr>
<td>1692.000</td>
<td>0.26</td>
<td>0.334</td>
<td>23</td>
<td>-2.32</td>
<td>-2.43</td>
<td>+</td>
</tr>
<tr>
<td>2727.000</td>
<td>0.17</td>
<td>0.334</td>
<td>23</td>
<td>-2.32</td>
<td>-2.43</td>
<td>-</td>
</tr>
</tbody>
</table>

* significant, - not significant

A NOEC of 120.000 µg/L is suggested by the program.
Statistical Evaluation of Biotests

<table>
<thead>
<tr>
<th>Yield FW control</th>
<th>CoV%</th>
<th>NOEC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00 g</td>
<td>30.4</td>
<td>50 µg/L</td>
</tr>
<tr>
<td>1.00 g</td>
<td>59.1</td>
<td>120 µg/L</td>
</tr>
</tbody>
</table>

Why are Coefficients of Variation relevant and why is the critical value set to 35%?

- The higher the CoVs, the higher the MDDs and the lower the statistical power.
- Variability directly affects statistical power of multiple testing, i.e. the informative value of NOECs.
- High variability also hampers ECx determination by regression analysis and causes wide confidence limits.
- CoVs of 35% proved to be a reasonable compromise between statistical requirements and practicability and therefore were prescribed as validity criterion in various biotests.

MDD = minimum detectable difference ("detection limit of a statistical test")

MDD = 36.9%

MDD = 52.3%
What can we learn from existing ring test data so far?
Results from ring test 2, Imazapyr
→ CoV range of control plants for each assessment parameter in each lab

<table>
<thead>
<tr>
<th>Lab Rank</th>
<th>Growth rate</th>
<th>Yield</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FW DW LL</td>
<td>FW DW LL</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Key:**
- CoV < 35%
- CoV 35-40%
- CoV 40-100%
- CoV >100%

• Yield FW is the most variable parameter at all → demanding, to use YFW as validity criterion
• Four labs exceeding 100% for CoV in most cases → points to laboratory specific issues with test performance and/or plant growth
• Four labs with CoVs < 35% in nearly all cases → CoVs < 35% can be achieved, even for Yield FW!

Sources of variability in *Glyceria* growth inhibition test

- inherent variability
  even plants with exactly the same initial size will grow differently
- „technical” variability:
  • usage of representative plants
  • heterogeneity of plants at test start
Common error about variability:

„The higher the sample size, the lower the variability!“

Simulation using Excel

fictitious sample population: 28 plants, height 5-31 cm
sample (= number of replicates)
n= 3
n=6
n=12

→ Variability?
Higher sample size does not necessarily mean lower CoV. But: the higher the sample size, the smaller is the range of obtained CoV (span between min and max) → the obtained CoV is more precise.

Measure for scattering of obtained standard deviations: standard error. → Higher sample size reduces the standard error of the obtained standard deviation and CoV (i.e. its accuracy is increased)

→ Higher sample size does not reduce the absolute (mean) value of CoV. On average, it cannot become lower than that of the sample population.

How to reduce variability?

"Let's increase sample size" → does not work*

"Let's use pseudoreplicates" → works, but: increases number of plants, i.e. costs, efforts and work load

Glyceria: one shoot per pot and three pots per vessel = 3 pseudoreplicates → each replicate value = mean of the 3 pseudoreplicates

"Let's unify sample population"

* if basic population is still heterogeneous
Simulation using Excel

**Fictitious sample population:** 28 plants, height 15-25 cm

**Sample (= number of replicates)**
- n = 3
- n = 6
- n = 12

→ **Variability?**

**Conclusions about sample size and variability (2)**

Increasing sampling size increases the precision of the CoV%, but does not necessarily decreases the CoV%

On average, the CoV% of a sample cannot become lower than the CoV% of the sample population.

The preferable way to decrease CoVs is decreasing initial variability of tests plants, i.e. unifying plant size and quality.
Test protocol:
„Initial plants might have heights between 15-25 cm, but in one test, height should be x±10%, e.g. 20 cm ± 2 cm“

→ sounds very strict – why? → unavoidable variability:

→ CoV% of initial plants should be clearly lower than 35%, because total CoV% must not exceed 35% including additional, unavoidable variability.

The issue of calculating yields and growth rates.

Yield = (weight d14) – (weight d0)

GR [d-1] = (ln (weight d14) – ln (weight d0)) / 14

→ measurement of initial weight destroys the plant

→ the measurements must be made on other plants as a substituted measure

→ assessment of initial FW and DW using „representative“ plants

→ unavoidable source of additional variability „technical variability“ in addition to „inherent variability“
The mean FW and DW of representative plants is used as initial weight for all test plants.

This number is wrong in any case! It can at best be an approximation of the individual initial weights of the test plants.

The higher the initial uniformity ...
- within test plants
- within representative plants
- between test plants and representative plants

... the better fits the mean initial weight of representative plants to individual start weights of the test plants

... the more precise are the calculated individual yields and growth rates

... the lower is the unavoidable „technical variability“ of yields and growth rates

How to ensure that representative plants are really representative?

- Direct comparison of initial weights impossible
- Indirect comparison using height
- Measuring initial shoot height of both test plants and representative plants enables to check, whether at least initial heights are uniform.
- If uniform number of leaves, plants of same height are assumed to have similar weight
- The only way to achieve at least fairly uniform initial weights is to select plants of same height and leaf number both for test plants and representative plants.
Conclusions: The importance of uniformity

- Different sources of variability: inherent variability + „technical” variability (use of representative plants; variability of plants at test start)
- Higher sample size reduces the standard error of the obtained standard deviation and CoV (i.e. its accuracy is increased)
- Higher sample size does not necessarily reduce the absolute value of CoV. On average, it cannot become lower than that of the sample population.
- The preferable way to decrease CoVs is decreasing initial variability of plants, i.e. unifying plant size and quality.
- The higher the uniformity within test plants, within representative plants and between test plants and representative plants, the lower is the „technical variability“ compared to the inherent variability.
- To achieve overall CoVs < 35%, it is essential to place the highest value on uniformity of both representative and test plants plants at test start (height, number of leaves, condition).

Importance of uniformity

Thank you very much for your attention!
Statistical Evaluation of Biotests

Questions?