Assessing quality control for repeated bioassay data by parametric and non-parametric prediction intervals

Daniel Gerhard

NCSC, Lyon 2010
Outline

- Evaluating quality control of repeated bioassay data
  - Multiple *historical* observations to characterize bioassay variability
  - Test sample to judge about process control
- Using prediction intervals to define a tolerable region
- Retrieving the test sample in the tolerable region

- R package `predIntervals`
- GUI available [Rohmeyer, Gerhard 2008]
Tolerance- vs. Prediction-Intervals

Tolerance Intervals
- With probability $\alpha$, the probability that a future observation $y_i$ falls in the interval $[\delta_{\text{lower}}; \delta_{\text{upper}}]$ is at least $\beta$

Prediction Intervals
- With probability $\alpha$, a proportion $\beta$ (or equally $k$ out of $n$) future observations $y_1, \ldots, y_n$ will fall in the interval $[\delta_{\text{lower}}; \delta_{\text{upper}}]$
- With probability $\alpha$, the mean/median of $n$ future observations $y_1, \ldots, y_n$ will fall in the interval $[\delta_{\text{lower}}; \delta_{\text{upper}}]$
Data Example
Data Example

log-transformation
Prediction interval

to include at least \( k \) out of \( n \) future observations (Odeh 1990)

historical sample \( x_i, \) with \( i = 1, \ldots, m \)

\[
\begin{bmatrix}
\hat{\delta}_{\text{lower}}; \hat{\delta}_{\text{upper}} \\
\end{bmatrix} = \bar{x} \pm q_{1-\alpha,m,n,k} s \sqrt{1 + \frac{1}{m}}
\]

\( \bar{x}, s \) is the arithmetic mean and the estimated standard error of the historical observations

\( q_{1-\alpha,m,n,k} \) is a two-sided \( 1 - \alpha \) quantile of a multivariate \( t \)-distribution considering the restricted number of future observations contained in the interval

If \( k = n \), \( q_{1-\alpha,m,n} \) is a two-sided \( 1 - \alpha \) quantile of a multivariate \( t \)-distribution with \( df = m - 1 \) and correlation \( R \), which is a \( n \times n \) matrix with off-diagonal elements \( \rho = \frac{1}{1+m} \) (Chew 1968)
Prediction interval
to include at least \( k \) out of \( n \) future observations (Odeh 1990)
Data Example

original scale

EC50 [pg/ml]

Frequency

50 75 100 125 150 175 200
Nonparametric prediction interval

to include at least \( k \) out of \( n \) future observations (Danziger, Davis 1964)

ordered historical sample \( x_1 \leq \cdots \leq x_m \)

ordered test sample \( y_1 \leq \cdots \leq y_n \)

Probability that \( p \) of \( n \) future observations are larger than the historical observation \( x_r \):

\[
P(x_r < y_p, \ldots, y_n) = \binom{p + m - r}{p} \binom{n - p + r - 1}{n - p} / \binom{n + m}{n}
\]

Searching the \( r \) that satisfies

\[
\sum_{k=p}^{n} P(x_r < y_p, \ldots, y_n) \leq 1 - \alpha
\]

Prediction interval limits are found as \( [x_{r/2}; x_{m-r/2+1}] \).

At \( r = 0 \) the limits are set to \(-\infty\) and \( \infty \).
If \( r/2 \) is not an integral number, the mean of the observations with the neighboring ranks are chosen.
Nonparametric prediction interval
to include at least $k$ out of $n$ future observations (Danziger, Davis 1964)
Prediction interval
to include the mean of \( n \) future observations (Hahn, Meeker 1991)

\[
\left[ \hat{\delta}_{\text{lower}}, \hat{\delta}_{\text{upper}} \right] = \bar{x} \pm t \, s \sqrt{\frac{1}{m} + \frac{1}{n}}
\]

\( \bar{x}, s \) is the arithmetic mean and the estimated standard error of the historical observations

\( t \) is a two-sided \( 1 - \alpha \) quantile of a univariate \( t \)-distribution with \( df = m - 1 \)
Prediction interval to include the mean of $n$ future observations (Hahn, Meeker 1991)
Nonparametric prediction interval
to include the median of \( n \) future observations (Chakraborti et al. 2004)

ordered historical sample \( x_1 \leq \cdots \leq x_m \)
ordered test sample \( y_1 \leq \cdots \leq y_n \)

Probability that \( p \) of \( n \) future observations are larger than the historical observation \( x_r \):

\[
P (x_1 \leq \cdots \leq x_r \leq y_p) = \binom{p + r - 1}{r} \binom{m + n - p - r}{m - r} / \binom{n + m}{m}
\]

Prediction interval limits \([x_l; x_u]\) are found by

\[
\sum_{r=l}^{u+1} P (x_r < y_p, \ldots, y_n) \geq 1 - \alpha
\]
Nonparametric prediction interval
to include the median of $n$ future observations (Chakraborti et al. 2004)
Package *predIntervals*

on R-Forge:

http://predintervals.r-forge.r-project.org

**R Functions**

```r
> predint(x, k, m, level=0.95,
   alternative="two.sided", quantile=NULL)
> nparpredint(x, k, m, level=0.95,
   alternative="two.sided")

> precint(x, m, level=0.95,
   alternative="two.sided")
> nparprecint(x, m, level=0.95,
   alternative="two.sided")
```
Coverage Simulations

\[ x_i \sim \text{Normal} \]
- Parametric PI
- Non-Parametric PI

\[ x_i \sim \text{logNormal} \]
- Parametric PI
- Non-Parametric PI
Discussion

- At least \( m \approx 20 \) observations needed to obtain accurate intervals
- Better performance at small \( k \)

Parametric intervals

- Calculation inaccuracy at small \( n \)
- Dependent on parametric assumptions

Nonparametric intervals

- Dependent on the actual sample \( (m > 20 \text{ needed}) \)
- Distribution-free
References


ODEH, RE (1990): 2-Sided prediction intervals to contain at least k out of m future observations from a normal distribution. *Technometrics* 32(2): 203-216.

Simulated coverage: $x_i \sim \text{Normal}(0,1)$

Parametric prediction interval

Return
Simulated coverage: $x_i \sim \text{Normal}(0,1)$

Non-Parametric prediction interval

Return
Simulated coverage: $x_i \sim \logNormal(0,1)$

Parametric prediction interval

<table>
<thead>
<tr>
<th>m</th>
<th>Level</th>
<th>n future observations</th>
<th>% k out of n</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.8</td>
<td>20</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>0.8</td>
<td>40</td>
<td>0.4</td>
</tr>
<tr>
<td>4</td>
<td>0.8</td>
<td>60</td>
<td>0.6</td>
</tr>
<tr>
<td>5</td>
<td>0.8</td>
<td>80</td>
<td>0.8</td>
</tr>
<tr>
<td>10</td>
<td>0.8</td>
<td>100</td>
<td>1.0</td>
</tr>
<tr>
<td>20</td>
<td>0.8</td>
<td>20</td>
<td>0.2</td>
</tr>
<tr>
<td>30</td>
<td>0.8</td>
<td>40</td>
<td>0.4</td>
</tr>
<tr>
<td>40</td>
<td>0.8</td>
<td>60</td>
<td>0.6</td>
</tr>
<tr>
<td>50</td>
<td>0.8</td>
<td>80</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>100</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>0.9</td>
<td>20</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>0.9</td>
<td>40</td>
<td>0.4</td>
</tr>
<tr>
<td>4</td>
<td>0.9</td>
<td>60</td>
<td>0.6</td>
</tr>
<tr>
<td>5</td>
<td>0.9</td>
<td>80</td>
<td>0.8</td>
</tr>
<tr>
<td>10</td>
<td>0.9</td>
<td>100</td>
<td>1.0</td>
</tr>
<tr>
<td>20</td>
<td>0.9</td>
<td>20</td>
<td>0.2</td>
</tr>
<tr>
<td>30</td>
<td>0.9</td>
<td>40</td>
<td>0.4</td>
</tr>
<tr>
<td>40</td>
<td>0.9</td>
<td>60</td>
<td>0.6</td>
</tr>
<tr>
<td>50</td>
<td>0.9</td>
<td>80</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>0.95</td>
<td>100</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>0.95</td>
<td>20</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>0.95</td>
<td>40</td>
<td>0.4</td>
</tr>
<tr>
<td>4</td>
<td>0.95</td>
<td>60</td>
<td>0.6</td>
</tr>
<tr>
<td>5</td>
<td>0.95</td>
<td>80</td>
<td>0.8</td>
</tr>
<tr>
<td>10</td>
<td>0.95</td>
<td>100</td>
<td>1.0</td>
</tr>
<tr>
<td>20</td>
<td>0.95</td>
<td>20</td>
<td>0.2</td>
</tr>
<tr>
<td>30</td>
<td>0.95</td>
<td>40</td>
<td>0.4</td>
</tr>
<tr>
<td>40</td>
<td>0.95</td>
<td>60</td>
<td>0.6</td>
</tr>
<tr>
<td>50</td>
<td>0.95</td>
<td>80</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>0.95</td>
<td>100</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Simulated coverage: $x_i \sim \log\text{Normal}(0,1)$

Non-Parametric prediction interval

% $k$ out of $n$

$n$ future observations

0.2 0.4 0.6 0.8 1.0

$m = 2$
level = 0.8
$m = 3$
level = 0.8
$m = 4$
level = 0.8
$m = 5$
level = 0.8
$m = 10$
level = 0.8
$m = 20$
level = 0.8
$m = 30$
level = 0.8
$m = 40$
level = 0.8
$m = 50$
level = 0.8

$m = 2$
level = 0.9
$m = 3$
level = 0.9
$m = 4$
level = 0.9
$m = 5$
level = 0.9
$m = 10$
level = 0.9
$m = 20$
level = 0.9
$m = 30$
level = 0.9
$m = 40$
level = 0.9
$m = 50$
level = 0.9

$m = 2$
level = 0.95
$m = 3$
level = 0.95
$m = 4$
level = 0.95
$m = 5$
level = 0.95
$m = 10$
level = 0.95
$m = 20$
level = 0.95
$m = 30$
level = 0.95
$m = 40$
level = 0.95
$m = 50$
level = 0.95
Quantile calculation for a prediction interval
to include $k$ out of $n$ future observations (Odeh 1990)

$$r = \sqrt{\frac{m + 1}{m}} u^* \quad \text{satisfying} \quad \sum_{j=k}^{m} P(f_j(u^*)) = 1 - \alpha$$

$$P(f_j(u^*)) = \int_{0}^{\infty} \left\{ \int_{-\infty}^{\infty} \binom{n}{j} [\Phi(b) - \Phi(a)]^j \times [\Phi(b) - \Phi(a)]^{n-j} \phi(y)dy \right\} f_\nu(s)ds$$

$$a = -us + \frac{\sqrt{\rho y}}{\sqrt{1 - \rho}} \quad b = us + \frac{\sqrt{\rho y}}{\sqrt{1 - \rho}} \quad \rho = \frac{1}{m + 1}$$

$\Phi(\cdot)$, $\phi(\cdot)$ are the standard normal density and distribution functions

$f_\nu(s)$ is the density function of $S$, where $\nu S^2$ is $\chi^2$ distributed with $df = m - 1$ and $\nu = m - 1$