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

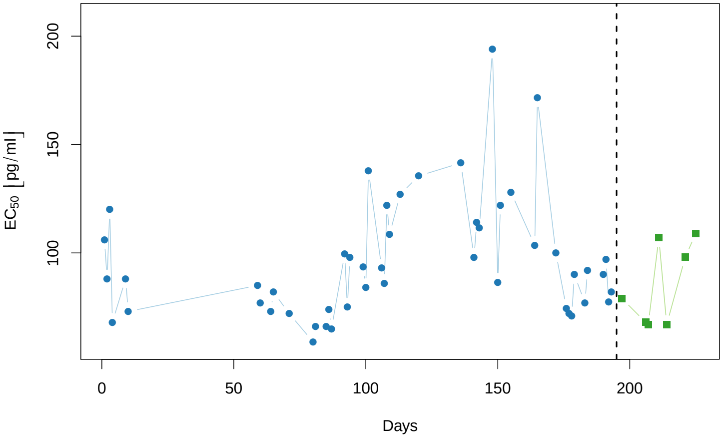
Tolerance Intervals

- With probability α, the probability that a future observation y_i falls in the interval $[\delta_{lower}; \delta_{upper}]$ is at least β

Prediction Intervals

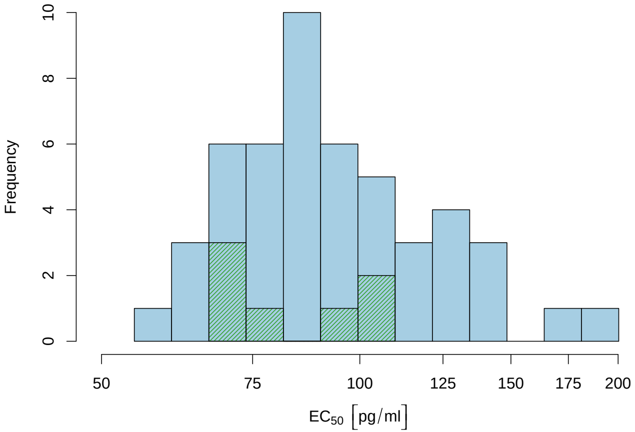
- With probability α, a proportion β (or equally k out of n) future observations $y_1, \ldots, y_n$ will fall in the interval $[\delta_{lower}; \delta_{upper}]$
- With probability α, the mean/median of n future observations $y_1, \ldots, y_n$ will fall in the interval $[\delta_{lower}; \delta_{upper}]$

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 $i = 1, \ldots, m$

$$\left[\hat{\delta}_{lower}; \hat{\delta}_{upper}\right] = \bar{x} \pm q_{1-\alpha,m,n,k}\, s \sqrt{1 + \frac{1}{m}}$$

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

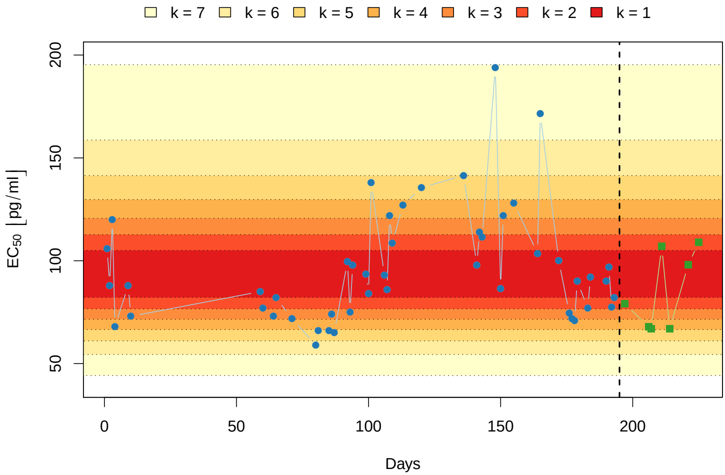
$q_{1-\alpha,m,n,k}$ is a two-sided $1-\alpha$ quantile of a multivariate t-distribution considering the restricted number of future observations contained in the interval

▸ Quantile calculation

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 $\boldsymbol{R}$, which is a $n \times n$ matrix with off-diagonal elements $\rho = \frac{1}{1+m}$ (Chew 1968)

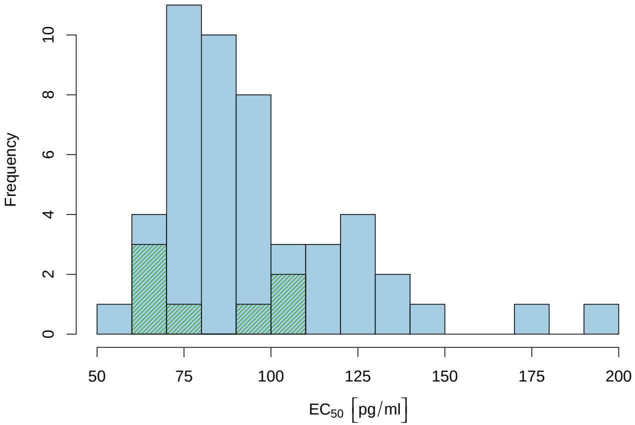
Prediction interval

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



Data Example

original scale



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:

$$P\left(x_r < y_p, \ldots, y_n\right) = \binom{p+m-r}{p}\binom{n-p+r-1}{n-p} \Big/ \binom{n+m}{n}$$

Searching the r that satisfies

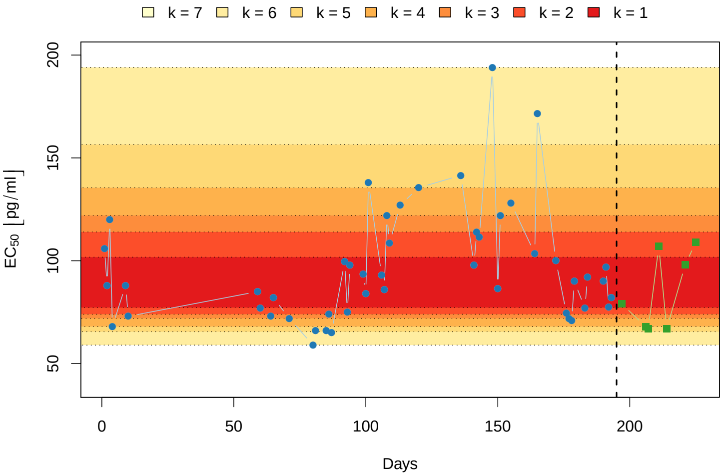
$$\sum_{k=p}^{n} P\left(x_r < y_p, \ldots, y_n\right) \leq 1-\alpha$$

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

At $r = 0$ the limits are set to $-\infty$ and ∞.
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)

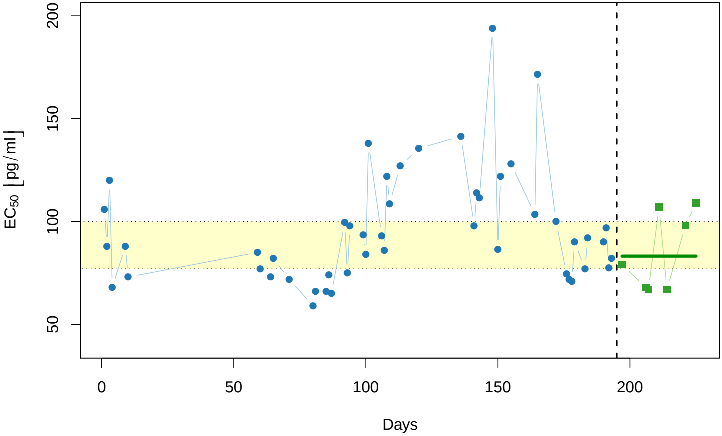
$$\left[\hat{\delta}_{lower}; \hat{\delta}_{upper}\right] = \bar{x} \pm t\, s \sqrt{\frac{1}{m} + \frac{1}{n}}$$

$\bar{x}$, 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:

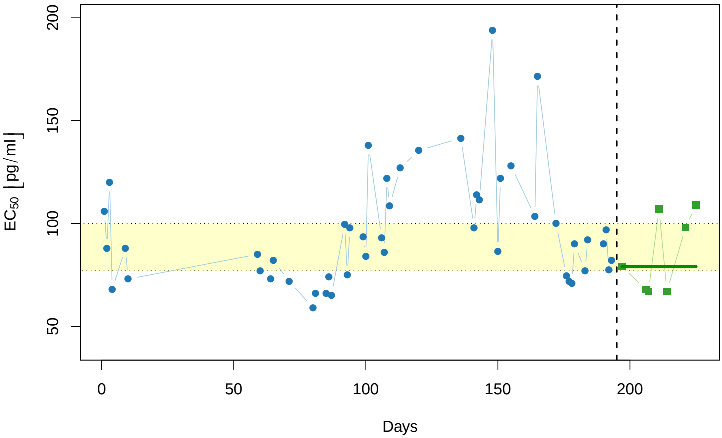
$$P\left(x_1 \leq \cdots \leq x_r \leq y_p\right) = \binom{p+r-1}{r}\binom{m+n-p-r}{m-r} \Big/ \binom{n+m}{m}$$

Prediction interval limits $[x_l; x_u]$ are found by

$$\sum_{r=l}^{u+1} P\left(x_r < y_p, \ldots, y_n\right) \geq 1 - \alpha$$

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

```
> 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

$x_i \sim$ Normal

- Parametric PI
- Non-Parametric PI

$x_i \sim$ 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

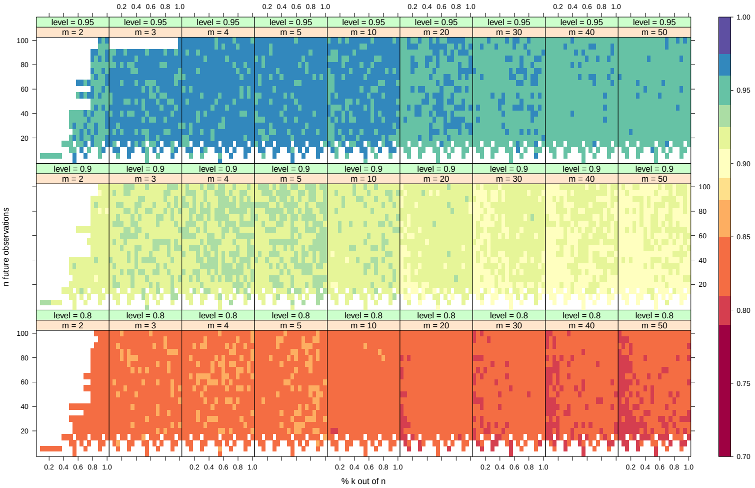
- Dependent on the actual sample ($m > 20$ needed)
- Distribution-free

References

- HAHN, GJ AND MEEKER, WQ (1991): *Statistical Intervals*. Wiley, New York.
- CHAKRABORTI, S, VAN DER LAAN, P, VAN DE WIEL, MA (2004): A class of distribution-free control charts. *Applied Statistics* 53(3):443-462.
- DANZIGER, L AND DAVIS, SA (1964): Tables of distribution-free tolerance limits. *Annals of Mathematical Statistics* 35(3):1361-1365.
- HOTHORN, LA, GERHARD, D, HOFMANN, M (2009): Parametric and non-parametric prediction intervals based phase II control charts for repeated bioassay data. *Biologicals* (5):323-330.
- 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.
- R DEVELOPMENT CORE TEAM (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

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

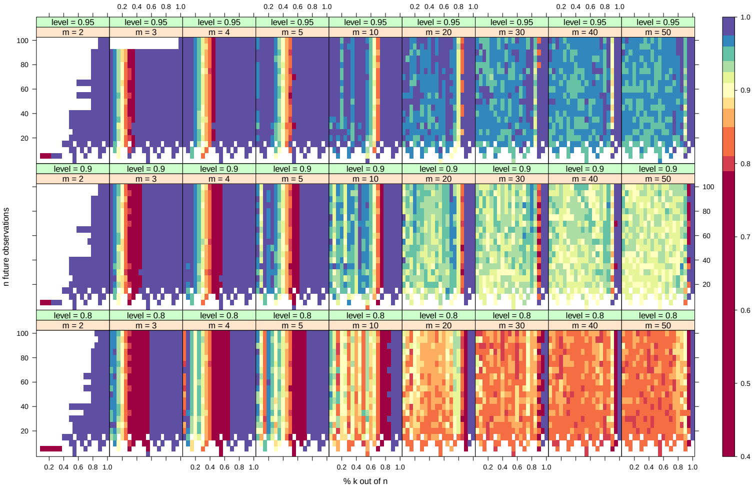
Parametric prediction interval



Return

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

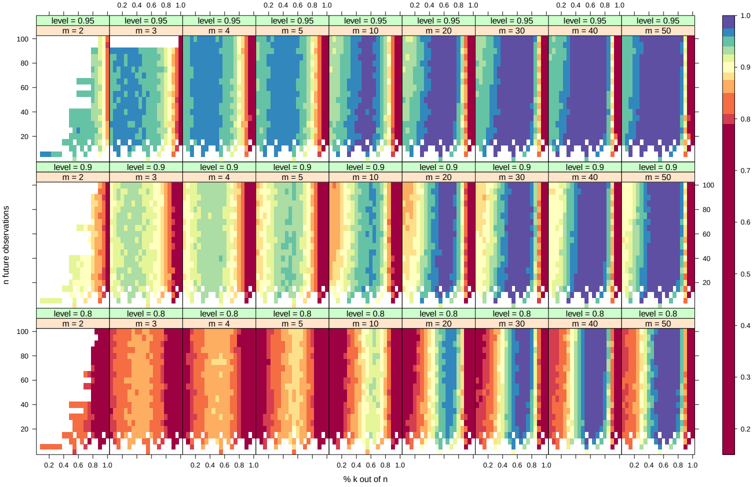
Non-Parametric prediction interval



Return

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

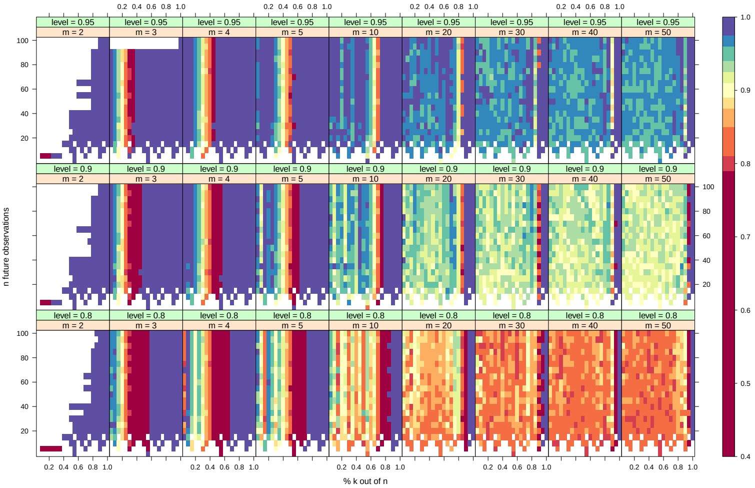
Parametric prediction interval



Return

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

Non-Parametric prediction interval



Return

Quantile calculation for a prediction interval

to include *k* out of *n* future observations (Odeh 1990)

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

$$P\left(f_j\left(u^\star\right)\right) = \int_0^\infty \left\{ \int_{-\infty}^{\infty} \binom{n}{j} \left[\Phi(b) - \Phi(a)\right]^j \times \left[\Phi(b) - \Phi(a)\right]^{n-j} \phi(y) dy \right\} f_\nu(s) ds$$

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

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

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

Return