# Nonparametric Predictive Inference for Future Order Statistics

Frank Coolen

joint with Tahani Maturi

Royal Holloway
22 June 2010

www.npi-statistics.com

# Nonparametric predictive inference (NPI)

- NPI is based on Hill's assumption $A_{(n)}$
- $X_1, \ldots, X_n, X_{n+1}$ are real-valued and exchangeable random quantities
- $x_1 < x_2 < \ldots < x_n$ are the ordered observed values ($x_0 = -\infty$ and $x_{n+1} = \infty$)
- For $X_{n+1}$, $A_{(n)}$ is given by

$$P(X_{n+1} \in I_j = (x_{j-1}, x_j)) = \frac{1}{n+1} \quad , \quad j = 1, \ldots, n+1$$

NPI is frequentist and exactly calibrated (with a 'Bayesian' justification), has strong consistency properties in theory of interval probability, and has been developed for many inference problems, using lower and upper probabilities if needed, e.g.:

- Bernoulli data
- Multinomial data
- Right-censored data
- Applications in Statistics, Reliability, Risk, Operations Research

# NPI for order statistics

- we have real-valued ordered data $x_1 < x_2 < \ldots < x_n$.
- The $n$ observations partition the real line into $n+1$ intervals $I_j = (x_{j-1}, x_j)$ for $j = 1, \ldots, n+1$. We are interested in $m \geq 1$ future observations, $X_{n+i}$ for $i = 1, \ldots, m$.
- We link the data and future observations via Hill's assumption $A_{(n)}$, via $A_{(n+m-1)}$ (which implies $A_{(n+k)}$ for all $k = 0, 1, \ldots, m-2$)
- Let $S_j = \#\{X_{n+i} \in I_j,\ i = 1, \ldots, m\}$, then inferences about these $m$ future observations, assuming $A_{(n+m-1)}$, can be based on the probabilities

$$P(\bigcap_{j=1}^{n+1} \{S_j = s_j\}) = \binom{n+m}{n}^{-1}$$

Let $X_{(r)}$, for $r = 1, \ldots, m$, be the $r$-th ordered future observation, so $X_{(r)} = X_{n+i}$ for one $i = 1, \ldots, m$ and $X_{(1)} < X_{(2)} < \ldots < X_{(m)}$. Then,

$$P(X_{(r)} \in I_j) = \binom{j+r-2}{j-1} \binom{n-j+1+m-r}{n-j+1} \binom{n+m}{n}^{-1}$$

The limiting result

$$\lim_{m \to \infty} P(X_{(\theta m)} \in I_j) = \binom{n}{j-1} \theta^{j-1} (1-\theta)^{n-j+1}$$

# Comparing two groups

- $X$ and $Y$ independent groups of real-valued observations, their ordered observed values are $x_1 < x_2 < \ldots < x_{n_x}$ and $y_1 < y_2 < \ldots < y_{n_y}$.
- And let $I_{j_x}^x = (x_{j_x-1}, x_{j_x})$ and $I_{j_y}^y = (y_{j_y-1}, y_{j_y})$.
- We are interested in $m \geq 1$ future observations from each group (i.e. $m_x = m_y = m$)

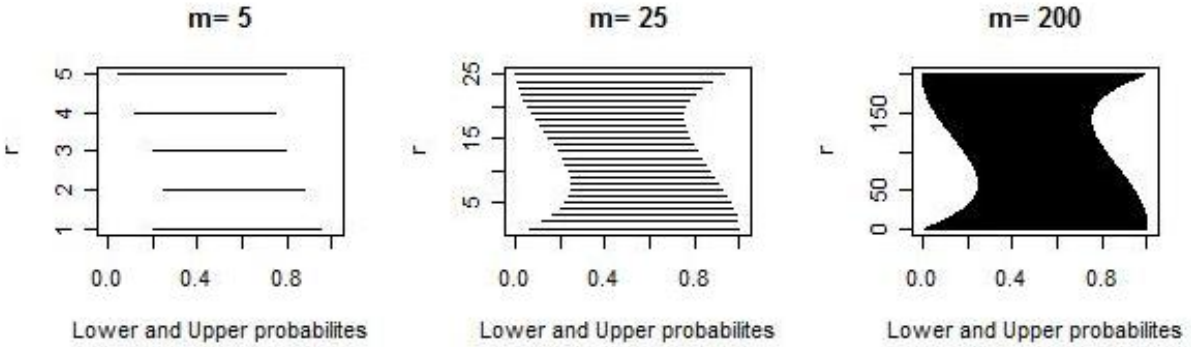
$$\underline{P}(X_{(r)} < Y_{(r)}) = \sum_{j_x=1}^{n_x+1} \sum_{j_y=1}^{n_y+1} \mathbf{1}\{x_{j_x} < y_{j_y-1}\} P(X_{(r)} \in I_{j_x}^x) P(Y_{(r)} \in I_{j_y}^y)$$

$$\overline{P}(X_{(r)} < Y_{(r)}) = \sum_{j_x=1}^{n_x+1} \sum_{j_y=1}^{n_y+1} \mathbf{1}\{x_{j_x-1} < y_{j_y}\} P(X_{(r)} \in I_{j_x}^x) P(Y_{(r)} \in I_{j_y}^y)$$

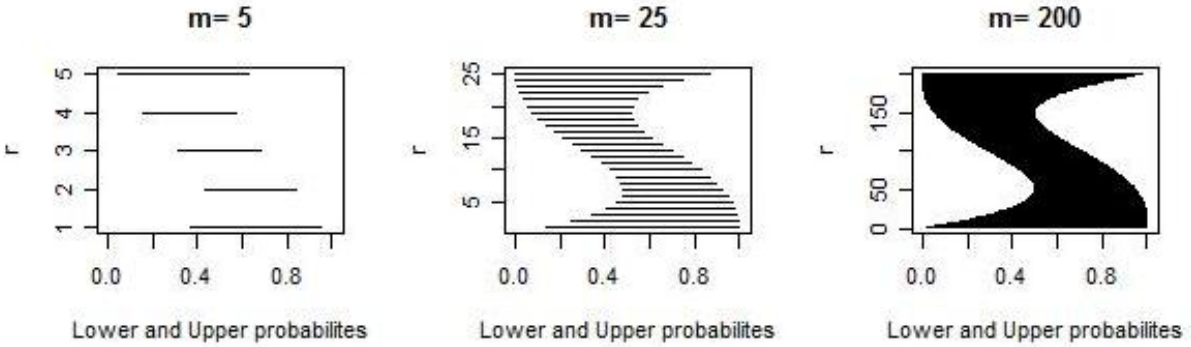
Example 1

# Example 1

X: 1 4
Y: 2 3

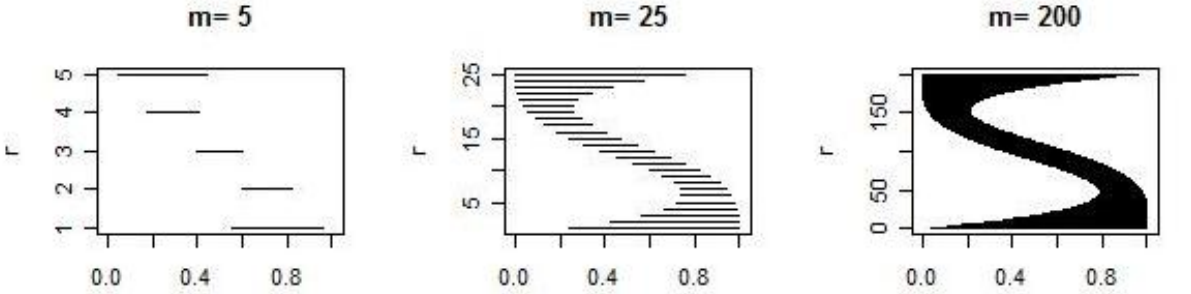



X: 1 2 7 8
Y: 3 4 5 6




X: 1 2 3 4
13 14 15 16

Y: 5 6 7 8
9 10 11 12

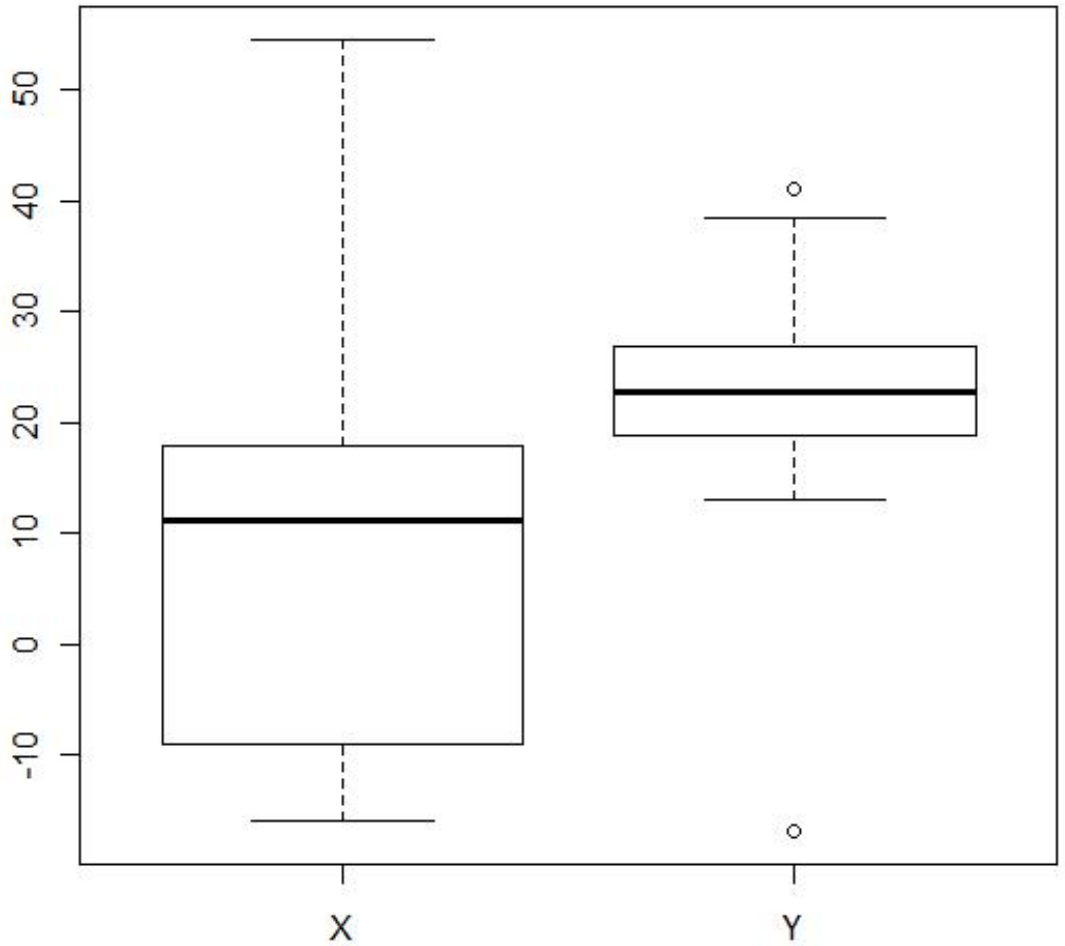


# Example 2

We consider the data set of a study of the effect of ozone environment on rats growth. One group of 22 rats were kept in an ozone environment and the second group of 23 similar rats were kept in an ozone-free environment.

| Ozone group ($X$) | | | | | Ozone-free group ($Y$) | | | | |
|---|---|---|---|---|---|---|---|---|---|
| -15.9 | -14.7 | -12.9 | -9.9 | -9.0 | -16.9 | 13.1 | 15.4 | 17.4 | 17.7 |
| -9.0 | 6.1 | 6.6 | 6.8 | 7.3 | 18.3 | 19.2 | 21.4 | 21.8 | 21.9 |
| 10.1 | 12.1 | 14.0 | 14.3 | 15.5 | 22.4 | 22.7 | 24.4 | 25.9 | 26.0 |
| 15.7 | 17.9 | 20.4 | 28.2 | 39.9 | 26.0 | 26.6 | 27.3 | 27.4 | 28.5 |
| 44.1 | 54.6 | | | | 29.4 | 38.4 | 41.0 | | |

## Example 2

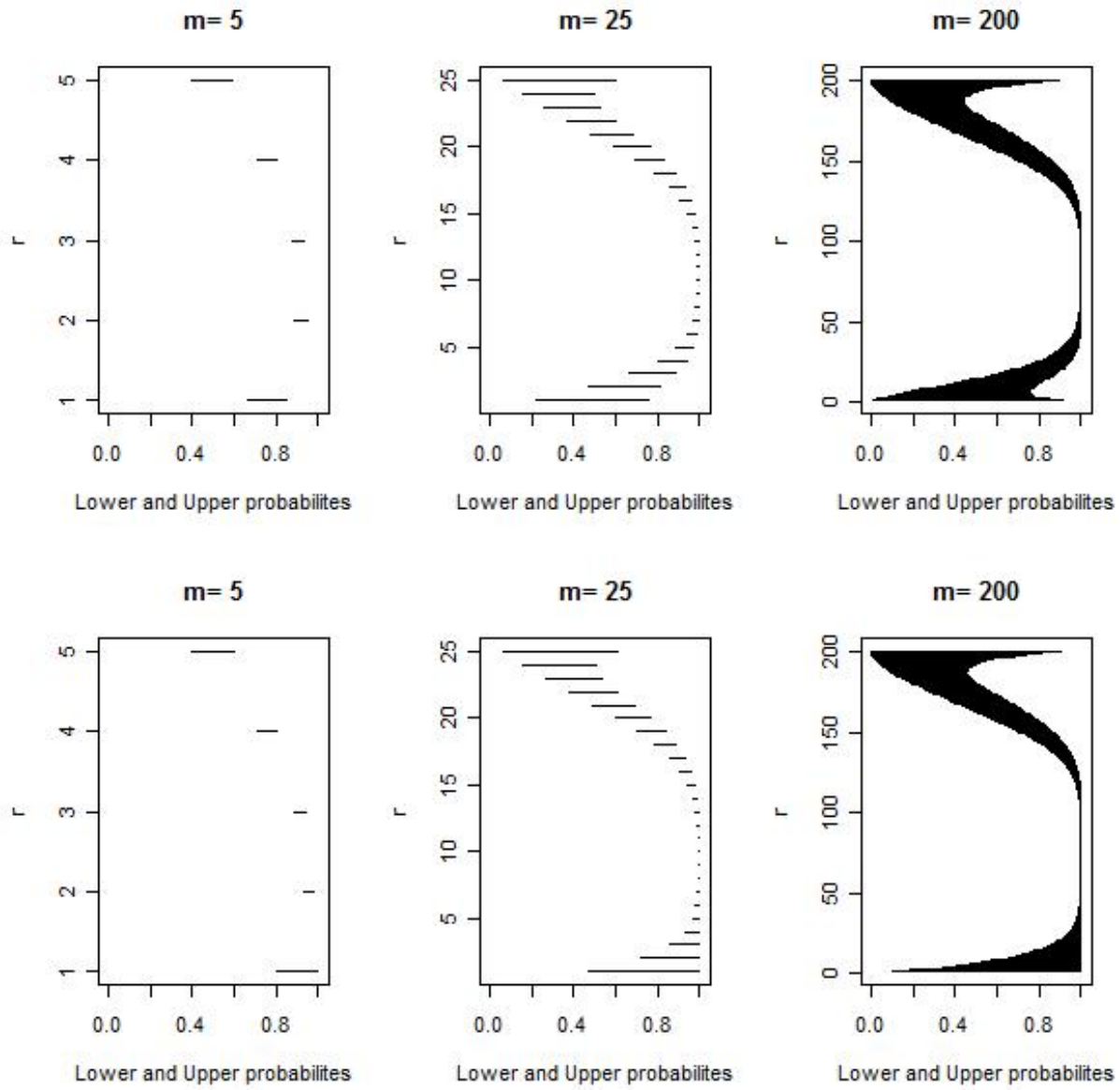


Example 2

| | $r$ | $\underline{P}$ | $\overline{P}$ | $\underline{P}^*$ | $\overline{P}^*$ |
|---|---|---|---|---|---|
| $m = 5$ | | | | | |
| | 1 | 0.661 | 0.849 | 0.805 | 0.993 |
| | 2 | 0.888 | 0.951 | 0.934 | 0.980 |
| | 3 | 0.877 | 0.930 | 0.890 | 0.937 |
| | 4 | 0.707 | 0.803 | 0.711 | 0.806 |
| | 5 | 0.396 | 0.593 | 0.399 | 0.599 |
| $m = 25$ | | | | | |
| | 1 | 0.224 | 0.756 | 0.468 | 1.000 |
| | 6 | 0.936 | 0.985 | 0.982 | 1.000 |
| | 11 | 0.990 | 0.997 | 0.993 | 0.998 |
| | 16 | 0.906 | 0.958 | 0.907 | 0.958 |
| | 21 | 0.483 | 0.683 | 0.488 | 0.690 |
| $m = 200$ | | | | | |
| | 1 | 0.010 | 0.911 | 0.099 | 1.000 |
| | 41 | 0.956 | 0.993 | 0.991 | 1.000 |
| | 81 | 0.999 | 1.000 | 1.000 | 1.000 |
| | 121 | 0.966 | 0.989 | 0.966 | 0.989 |
| | 161 | 0.515 | 0.748 | 0.522 | 0.756 |

$^*$ after removing the outlier.

## Example 2


m= 5
m= 25
m= 200
r
Lower and Upper probabilites
0.0
0.4
0.8
m= 5
m= 25
m= 200
r
Lower and Upper probabilites
0.0
0.4
0.8

# References


- Hill, B.M.: Posterior distribution of percentiles: Bayes' theorem for sampling from a population. JASA 63, 677–691 (1968)
- Augustin, T., Coolen, F.P.A.: Nonparametric predictive inference and interval probability. JSPI 124, 251–272 (2004)
- Coolen, F.P.A: On nonparametric predictive inference and objective Bayesianism. JLLI 15, 21–47 (2006)
- Coolen, F.P.A., Maturi, T.A.: Nonparametric predictive inference for order statistics of future observations. Proceedings SMPS, Sept 2010, to appear.
- **www.npi-statistics.com**
- *Nonparametric Predictive Inference* - Research Monograph, to appear (Wiley - 2011??)