Nonparametric Predictive Inference: Overview and Recent Results

Frank Coolen

Department of Mathematical Sciences
Durham University
frank.coolen@durham.ac.uk

DStatG, Leipzig, September 2011

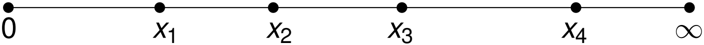
joint work with Tahani Coolen-Maturi

Nonparametric Predictive Inference (NPI):

- A nonparametric approach: few modelling assumptions
- Predictive: inference on the next future observation(s)
- Depends on Hill's assumption $A_{(n)}$ (Hill 1968)
- Uses lower and upper probabilities to quantify uncertainty
- Frequentist approach but no counter-factuals used and exactly calibrated (for any sample size)

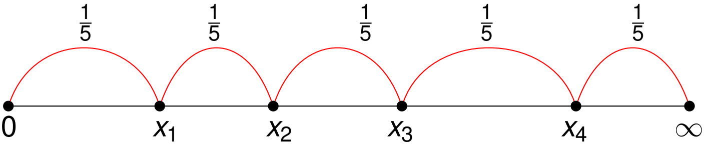
Hill's assumption $A_{(n)}$ (Hill 1968)

$n = 4$



Hill's assumption $A_{(n)}$ (Hill 1968)

$n = 4$

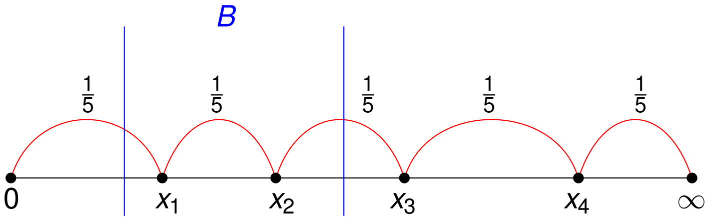


$P(X_5 \in (0, x_1)) = \frac{1}{5}$ $\qquad$ $P(X_5 \in (x_4, \infty)) = \frac{1}{5}$

$P(X_5 \in (x_i, x_{i+1})) = \frac{1}{5} \quad , \quad i = 1, 2, 3$

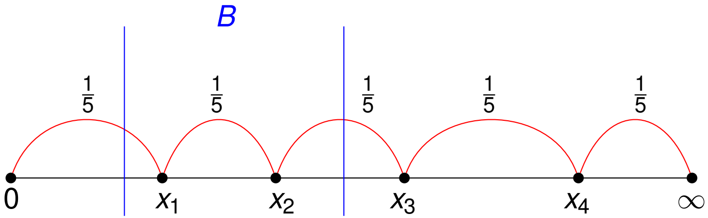
Hill's assumption $A_{(n)}$ (Hill 1968)

$n = 4$



Hill's assumption $A_{(n)}$ (Hill 1968)

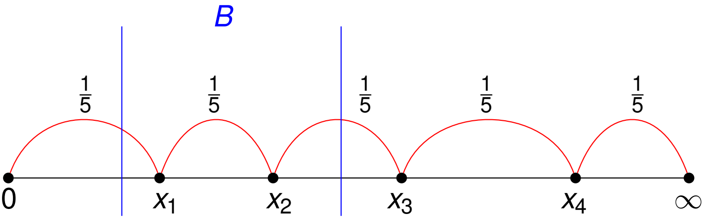
$n = 4$



$\underline{P}[X_5 \in B] = \frac{1}{5}$ $\overline{P}[X_5 \in B] = \frac{3}{5}$

Hill's assumption $A_{(n)}$ (Hill 1968)

$n = 4$



$\underline{P}[X_5 \in B] = \frac{1}{5}$ $\overline{P}[X_5 \in B] = \frac{3}{5}$

$Imprecision = \overline{P} - \underline{P} = \frac{2}{5} = 0.4$

Right-censored data

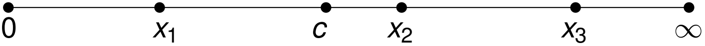
Lifetime data often contain right-censored observations, for example

- Units may not have failed by the end of the experiment
- Units may fail due to other reasons than those of interest
- Units may be withdrawn from the experiment

We assume throughout that the censoring process is independent of the failure process.

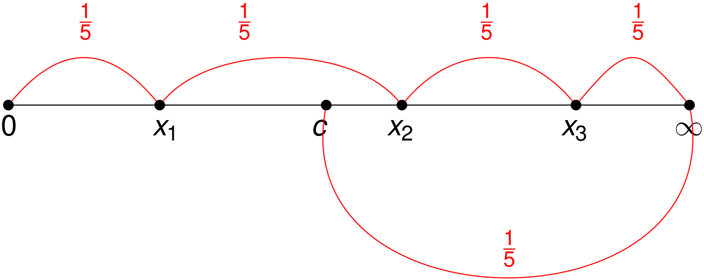
rc-$A_{(n)}$ assumption (Coolen & Yan, 2004)

$n = 4$



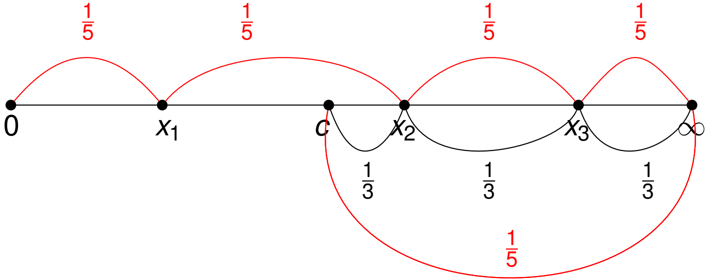
rc-$A_{(n)}$ assumption (Coolen & Yan, 2004)

$n = 4$



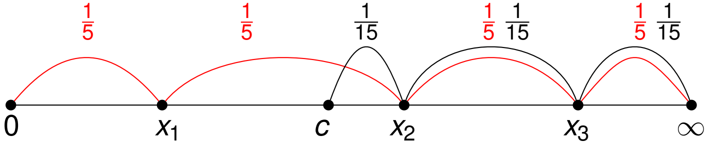
rc-$A_{(n)}$ assumption (Coolen & Yan, 2004)

$n = 4$



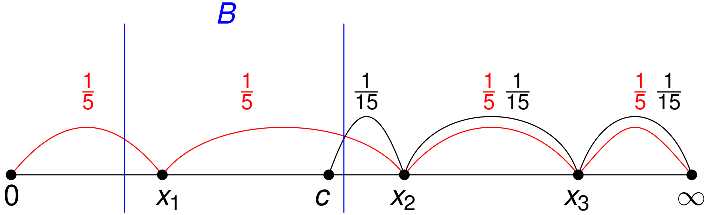
rc-$A_{(n)}$ assumption (Coolen & Yan, 2004)

$n = 4$



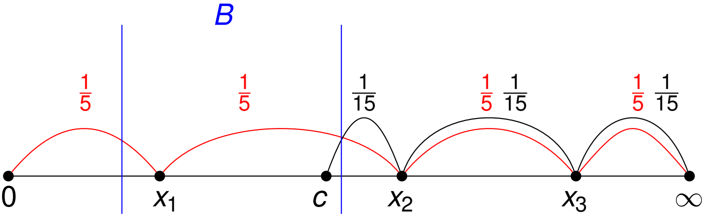
rc-$A_{(n)}$ assumption (Coolen & Yan, 2004)

$n = 4$



rc-$A_{(n)}$ assumption (Coolen & Yan, 2004)

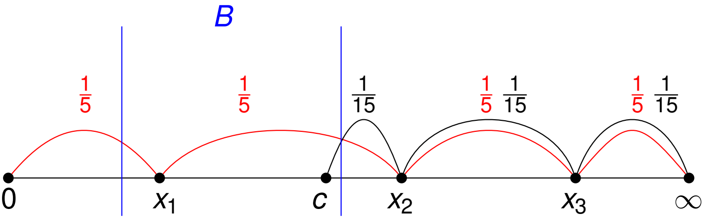
$n = 4$



$\underline{P}[X_5 \in B] = 0$ $\overline{P}[X_5 \in B] = \frac{7}{15}$

rc-$A_{(n)}$ assumption (Coolen & Yan, 2004)

$n = 4$

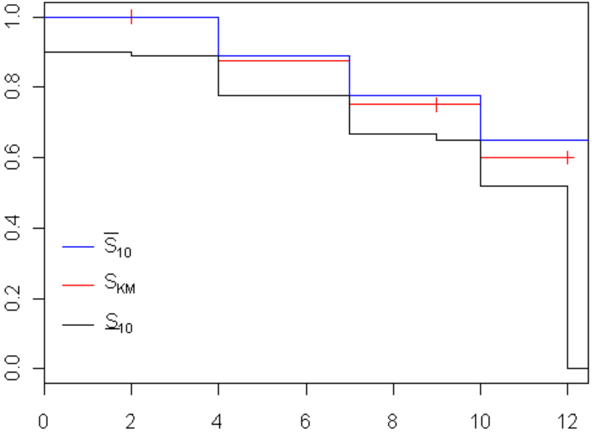


$\underline{P}[X_5 \in B] = 0$ $\overline{P}[X_5 \in B] = \frac{7}{15}$

$Imprecision = \overline{P} - \underline{P} = \frac{7}{15} = 0.4667$

Example

NPI lower and upper survival functions and the Kaplan-Meier estimator, for data $2^c, 4, 7, 9^c, 10, 12^c, 12^c, 12^c, 12^c$



Comparing two independent groups, X and Y

We have two independent groups X and Y:

$$x_1 < x_2 < \ldots < x_{n_x} \qquad \text{and} \qquad y_1 < y_2 < \ldots < y_{n_y}$$

The classical methods test $H_0 : F_X = F_Y$.

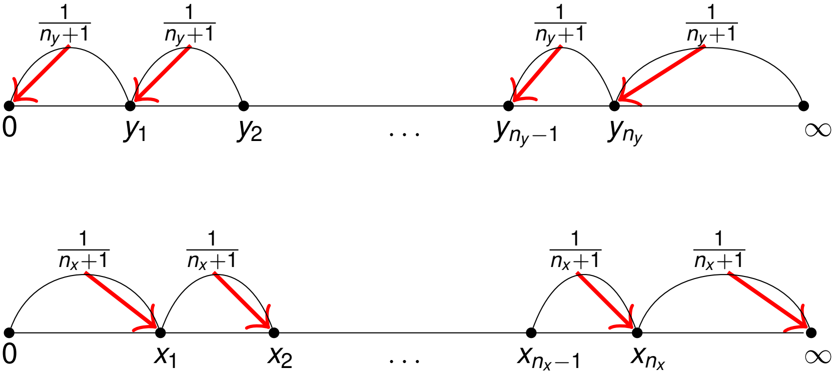
For complete data, Coolen (1996) introduced NPI to compare two independent groups depending on $A_{(n)}$. This is given via the lower and upper probabilities

$$\underline{P}(Y_{n_y+1} > X_{n_x+1}) \qquad \overline{P}(Y_{n_y+1} > X_{n_x+1})$$

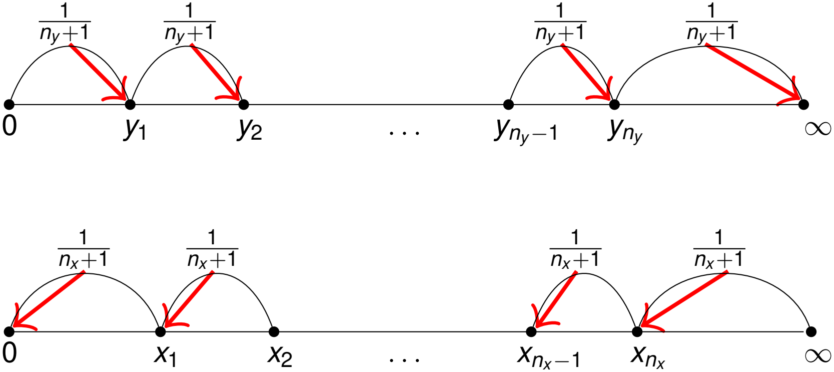

$\frac{1}{n_y+1}$
$\frac{1}{n_y+1}$
$\frac{1}{n_y+1}$
$\frac{1}{n_y+1}$
0
y_1
y_2
$\dots$
y_{n_y-1}
y_{n_y}
∞


$\frac{1}{n_x+1}$
$\frac{1}{n_x+1}$
$\frac{1}{n_x+1}$
$\frac{1}{n_x+1}$
0
x_1
x_2
$\dots$
x_{n_x-1}
x_{n_x}
∞

Lower Probability, $\underline{P}(Y_{n_y+1} > X_{n_x+1})$



Upper Probability, $\overline{P}(Y_{n_y+1} > X_{n_x+1})$



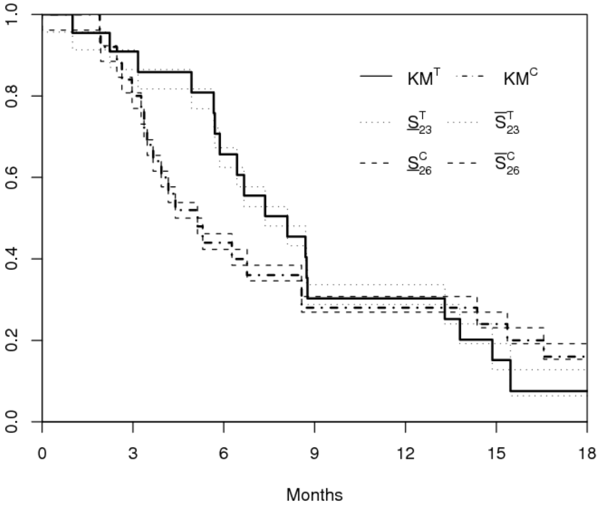
Example (Fleming, et al 1980)

In a Mayo Clinic study, patients with bile duct cancer were followed to determine whether those treated with a combination of radiation treatment (RoRx) and 5-Fluorouracil (5-FU) would survive significantly longer than a control population.

The treatment sample (days):
$30, 67, 79^+, 82^+, 95, 148, 170, 171, 176, 193, 200, 221, 243, 261, 262, 263, 399, 414, 446, 446^+, 464, 777.$

The control (days):
57, 58, 74, 79, 89, 98, 101, 104, 110, 118, 125, 132, 154, 159, 188, 203, 257, 257, 431, 461, 497, 723, 747, 1313, 2636.

Figure: Kaplan Meier, the NPI lower and upper survival functions for both groups (Control and Treatment)



T_0 months	$\underline{P}(X_{T,23} > X_{C,26})$	$\overline{P}(X_{T,23} > X_{C,26})$
1	0	0.9582
3	0.1704	0.9197
5	0.3903	0.8587
6	0.4495	0.7977
8	0.4938	0.7441
9	0.5271	0.6887
15	0.5345	0.6517
18	0.5419	0.6394
25	0.5468	0.6394
88	0.5468	0.6345

Table: Lower and upper probabilities that the lifetime of the next patient from the treatment group is greater than the lifetime of the next patient from the control group; experiment assumed ended after T_0 months.

Durham
University

k independent groups

For $k \geq 2$ groups, classical methods test the null hypothesis

$$H_0 : F_1 = F_2 = \ldots = F_k$$

against the alternative that population *i* is the best, that is,

$$H_{Ai} : F_i < F_j \text{ for all } j \neq i \text{ and } j = 1, \ldots, k$$

For each group j $(j = 1, 2, \ldots, k)$, the data consist of

$$x_{j,1} < x_{j,2} < \ldots < x_{j,n_j}$$

For NPI: future observations from each group

$$X_{1,n_1+1} \quad X_{2,n_2+1} \quad \cdots \quad X_{j,n_j+1} \quad \cdots \quad X_{k,n_k+1}$$

The event of interest is that l is the best group (longest lifetime)

$$X_{l,n_l+1} = \max_{1 \le j \le k} X_{j,n_j+1} \quad (l = 1, 2, \ldots, k)$$

The lower and upper probabilities for this event are

$$\underline{P}^{(l)} = \underline{P}\left(X_{l,n_l+1} = \max_{1 \le j \le k} X_{j,n_j+1}\right)$$

$$\overline{P}^{(l)} = \overline{P}\left(X_{l,n_l+1} = \max_{1 \le j \le k} X_{j,n_j+1}\right)$$

Example (Desu & Raghavarao, 2004)

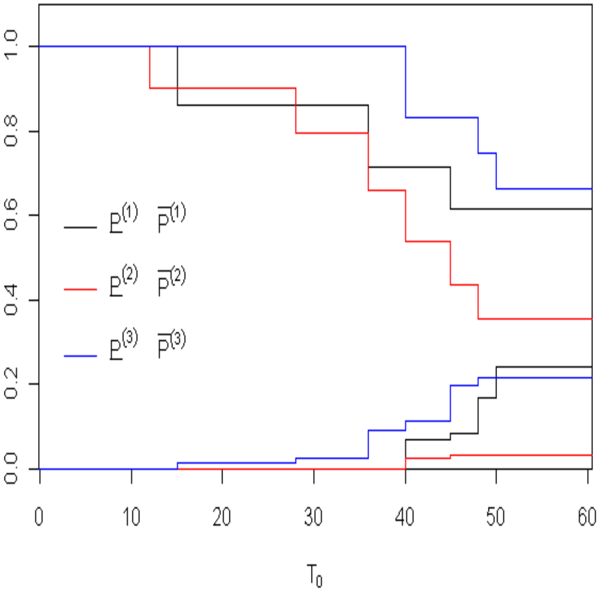
Times (months) until promotion at a large company, for 19 employees in $k = 3$ departments, ($n_1 = 6$, $n_2 = 9$ and $n_3 = 4$):

Dept 1: 15, 20^+, 36, 45, 58, 60

Dept 2: 12, 25^+, 28, 30^+, 30^+, 36, 40, 45, 48

Dept 3: 30^+, 40, 48, 50

Suppose that observation actually ends at time T_0.


1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
T_0
$\underline{P}^{(1)}$ $\overline{P}^{(1)}$
$\underline{P}^{(2)}$ $\overline{P}^{(2)}$
$\underline{P}^{(3)}$ $\overline{P}^{(3)}$

Competing Risks

- For n independent units, a unit subject to failure from one of k different (independent) failure causes (Failure Modes).
- For each unit, X_j represents the unit's time to failure under the assumption that failure could only occur due to failure mode j.
- The failure time of the unit is $X = \min(X_1, \ldots, X_j, \ldots, X_k)$.
- In NPI, let the failure time of the future unit be

$$X_{n+1} = \min(X_{1,n+1}, \ldots, X_{j,n+1}, \ldots, X_{k,n+1})$$

with $X_{j,n+1}$ similar as X_j above. Then

$$\underline{P}(X_{n+1} = \min_{1 \le j \le k} \{X_{j,n+1}\}) \; , \; \overline{P}(X_{n+1} = \min_{1 \le j \le k} \{X_{j,n+1}\})$$

Example

36 units of a small electrical appliance are tested. There are 18 different ways in which a unit could fail (18 FM), Lawless (2003).

# cycles	FM	# cycles	FM	# cycles	FM
11	1	1990	9	3034	9
35	15	2223	9	3034	9
49	15	2327	6	3059	6
170	6	2400	9	3112	9
329	6	2451	5	3214	9
381	6	2471	9	3478	9
708	6	2551	9	3504	9
958	10	2565	-	4329	9
1062	5	2568	9	6367	-
1167	9	2702	10	6976	9
1594	2	2761	6	7846	9
1925	9	2831	2	13403	-

FM9	1167	1925	1990	2223	2400	2471
	2551	2568	3034	3034	3112	3214
	3478	3504	4329	6976	7846	
OFM	11	35	49	170	329	381
	708	958	1062	1594	2327	2451
	2702	2761	2831	3059		
RC	2565	6367	13403			

$$\underline{P}(X_{37}^{FM9} < X_{37}^{OFM}) = 0.4358 \quad \text{and} \quad \overline{P}(X_{37}^{FM9} < X_{37}^{OFM}) = 0.5804$$

$$\underline{P}(X_{37}^{OFM} < X_{37}^{FM9}) = 0.4196 \quad \text{and} \quad \overline{P}(X_{37}^{OFM} < X_{37}^{FM9}) = 0.5642$$

‘Weak evidence’ that it is more likely that unit 37 will fail due to *FM*9 than due to *OFM*.

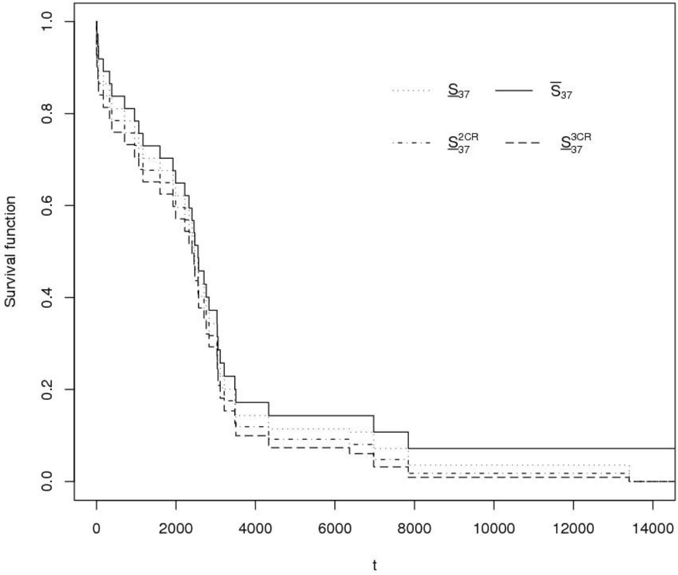
FM9	1167	1925	1990	2223	2400	2471	2551
	2568	3034	3034	3112	3214	3478	3504
	4329	6976	7846				
FM6	170	329	381	708	2327	2761	3059
OFM	11	35	49	958	1062	1594	2451
	2702	2831					
RC	2565	6367	13403				

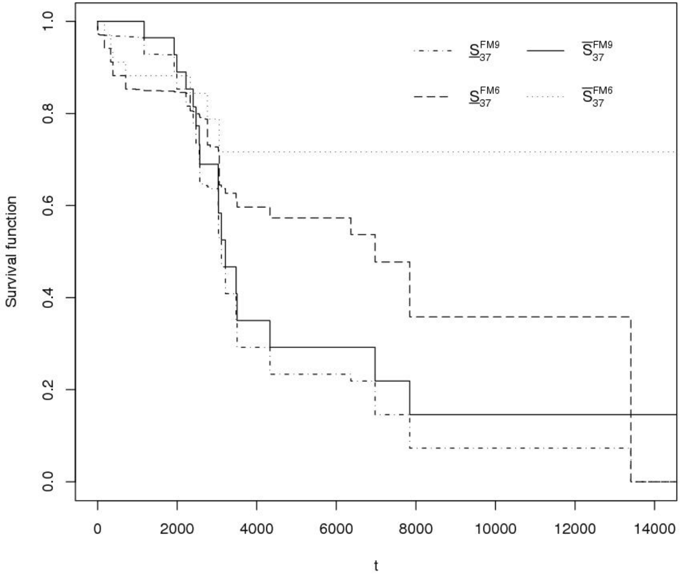
$$\underline{P}\left(X_{37}^{FM9} < \min\left\{X_{37}^{FM6}, X_{37}^{OFM}\right\}\right) = 0.3915\ ,\ \overline{P}\left(X_{37}^{FM9} < \min\left\{X_{37}^{FM6}, X_{37}^{OFM}\right\}\right) = 0.5804$$

$$\underline{P}\left(X_{37}^{FM6} < \min\left\{X_{37}^{FM9}, X_{37}^{OFM}\right\}\right) = 0.1749\ ,\ \overline{P}\left(X_{37}^{FM6} < \min\left\{X_{37}^{FM9}, X_{37}^{OFM}\right\}\right) = 0.3279$$

$$\underline{P}\left(X_{37}^{OFM} < \min\left\{X_{37}^{FM6}, X_{37}^{FM9}\right\}\right) = 0.2265\ ,\ \overline{P}\left(X_{37}^{OFM} < \min\left\{X_{37}^{FM6}, X_{37}^{FM9}\right\}\right) = 0.3808$$

‘Strong evidence’ that unit 37 is more likely to fail due to *FM*9 than due to *FM*6 or due to *OFM*.


Survival function
t
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
10000
12000
14000
$\underline{S}_{37}$
$\overline{S}_{37}$
$\underline{S}_{37}^{2CR}$
$\underline{S}_{37}^{3CR}$


$\underline{S}_{37}^{FM9}$
$\overline{S}_{37}^{FM9}$
$\underline{S}_{37}^{FM6}$
$\overline{S}_{37}^{FM6}$
Survival function
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
10000
12000
14000
t

Unobserved failure modes

If n units all failed, then for any unobserved failure mode $U \in \{1, \ldots, k\}$, the NPI lower and upper probabilities for the event that the next unit will fail due to U are

$$\underline{P}^{(U)} = 0$$

$$\overline{P}^{(U)} = \frac{1}{n+1} \sum_{i=1}^{n+1} \frac{1}{i}$$

If one failure mode caused all n observed units to fail, the NPI lower and upper probabilities for the event that the next unit will fail also due to this failure mode are

$$\underline{P}^{(O)} = \sum_{i=1}^{n} \frac{i}{n+1} \left\{ \left[\frac{n+1-i}{n+2-i} \right]^{k-1} - \left[\frac{n-i}{n+1-i} \right]^{k-1} \right\}$$

$$\overline{P}^{(O)} = 1$$

Unknown failure modes

There may be unknown risks, that is failure modes that are possible but that have not been defined or even recognized.

For such a failure mode to cause the next failure, the NPI lower and upper probabilities are the same as the $\underline{P}^{(U)}$ and $\overline{P}^{(U)}$ for unobserved failure modes - note that the number of failure modes, k, did not affect these.

For other failure modes (with observed failures), the NPI lower probability decreases as function of k, but the upper probability remains unchanged.

Multinomial data

Joint work with Thomas Augustin (Munich)

Probability wheel representation assumed for data

Prediction of unobserved category: lower probability 0, upper probability positive and depending on number of observations and number of categories observed thus far (and on total number of categories if known)

As for competing risks: a more detailed data representation leads to more (or equal) imprecision in prediction

Concluding remarks

Many research challenges!
In particular: generalization for multi-variate data

Regression .... *just solved!*

Monograph introducing NPI theory, methods and applications is in development - expected to appear (Wiley) in 2013.

www.npi-statistics.com

Introduction to Imprecise Probabilities (eds Augustin, Coolen, de Cooman, Troffaes) to appear (Wiley) in 2012.

www.sipta.org