A Cantelli-type inequality for constructing non-parametric p-boxes based on exchangeability

Matthias C. M. Troffaes and Tathagata Basu

Department of Mathematical Sciences, Durham University, UK

Introduction

- We derive a Cantelli-type inequality to produce a non-parametric p-box and prediction interval.
- Based on sample mean and sample standard deviation only (i.e. no parametric assumptions).
- We assume exchangeability (rather than conditional independence).
- Useful for modelling when only sample mean and sample standard deviation are known (e.g. measurement problems).

Exchangeability [3]

We say a finite sequence $X_1, X_2, \ldots, X_n$ of discrete random variables is exchangeable if, for all $\sigma \in \Sigma_n$ and all $x_1, x_2, \ldots, x_n \in \mathbb{R}$

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) = P(X_1 = x_{\sigma(1)}, \cdots, X_n = x_{\sigma(n)})$$

Independent $\implies$ Exchangeable

Cantelli's inequality [1] *and p-boxes* [2]

Let X be a random variable with known mean μ and known variance σ^2. Then the Cantelli's inequality is given by:

$$0 \leq P\left(\frac{X-\mu}{\sigma} \leq \lambda\right) \leq \frac{1}{1+\lambda^2} \qquad \text{if } \lambda \leq 0,$$

$$\frac{\lambda^2}{1+\lambda^2} \leq P\left(\frac{X-\mu}{\sigma} \leq \lambda\right) \leq 1 \qquad \text{if } \lambda \leq 0.$$

A p-box is specified by two cumulative distribution functions $\underline{F}$ and $\overline{F}$, and represents the set of all cumulative distribution functions bounded by $\underline{F}$ and $\overline{F}$:

$$\{F \in \mathcal{F} \colon \underline{F}(x) \leq F(x) \leq \overline{F}(x), \forall x \in \mathbb{R}\}.$$

Cantelli's inequality gives distribution free p-box for known μ and σ^2:

Cantelli's inequality can also be written as a p-box on X:

$$0 \leq P(X \leq x) \leq \frac{\sigma^2}{\sigma^2 + (x-\mu)^2} \qquad \text{if } x \leq \mu$$

$$\frac{(x-\mu)^2}{\sigma^2 + (x-\mu)^2} \leq P(X \leq x) \leq 1 \qquad \text{if } x \geq \mu.$$

Saw's inequality (Chebyshev) [4]

Let $X_1, \ldots, X_n, X_{n+1}$ be a sequence of discrete exchangeable random variables. Define $\overline{X} := \frac{1}{n}\sum_{j=1}^{n} X_j$, $S^2 := \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2$ and $Q^2 := \frac{n+1}{n}S^2$. Then for every $\lambda \geq 1$

$$P\left(\left|X_{n+1} - \overline{X}\right| > \lambda Q\right) \leq \frac{1}{n+1}\left\lfloor \frac{n+1}{\lambda_n^2} \right\rfloor_*$$

where $\lambda_n := \sqrt{\frac{n\lambda^2}{n-1+\lambda^2}}$ and $\lfloor x \rfloor_* := \max\{n \in \mathbb{Z} \colon n < x\}$.

Main Result

Let, $X_1, \ldots, X_n, X_{n+1}$ be a finite sequence of discrete exchangeable random variables. Let $\Delta \in \mathbb{R}$ denote the range of the X_j i.e. $\Delta := \max X_j - \min X_j$ where $\max X_j$ is the maximum value that can be attained by X_j, and $\min X_j$ is the minimum value. Let $\overline{X} := \sum_{j=1}^{n} X_j/n$, and $S^2 := \sum_{j=1}^{n}(X_j - \overline{X})^2/(n-1)$. Then for every $\lambda \geq 0$,

$$\frac{1}{n+1}\left\lceil \frac{(n+1)\lambda_n^2}{\lambda_n^2+1} \right\rceil \leq P\left(\frac{X_{n+1} - \overline{X}}{S + \frac{\Delta_n}{\sqrt{n}}} < \lambda\right) \leq 1 \tag{3}$$

where $\lambda_n := \frac{n}{\sqrt{n^2-1}}\lambda$ and $\Delta_n := \sqrt{\frac{n+1}{n-1}}\Delta$. Similarly, for $\lambda \leq 0$,

$$0 \leq P\left(\frac{X_{n+1} - \overline{X}}{S + \frac{\Delta_n}{\sqrt{n}}} \leq \lambda\right) \leq \frac{1}{n+1}\left\lfloor \frac{n+1}{\lambda_n^2+1} \right\rfloor. \tag{4}$$

Here, $\lfloor x \rfloor := \max\{n \in \mathbb{Z} \colon n \leq x\}$ and $\lceil x \rceil := -\lfloor -x \rfloor$.

P-box and Imprecise Prediction Interval

We use our main result to construct a **p-box** for the random variable:

$$Z_{n+1} = \frac{X_{n+1} - \overline{X}}{S + \frac{\Delta_n}{\sqrt{n}}}$$

However, this cannot be used as a p-box for X_{n+1}. We cannot substitute the observed values for $\overline{X}$ and S in the equation.
Our result allows to construct an asymmetric **prediction interval**. Our result gives, for every ℓ_1 and ℓ_2, the following bounds:

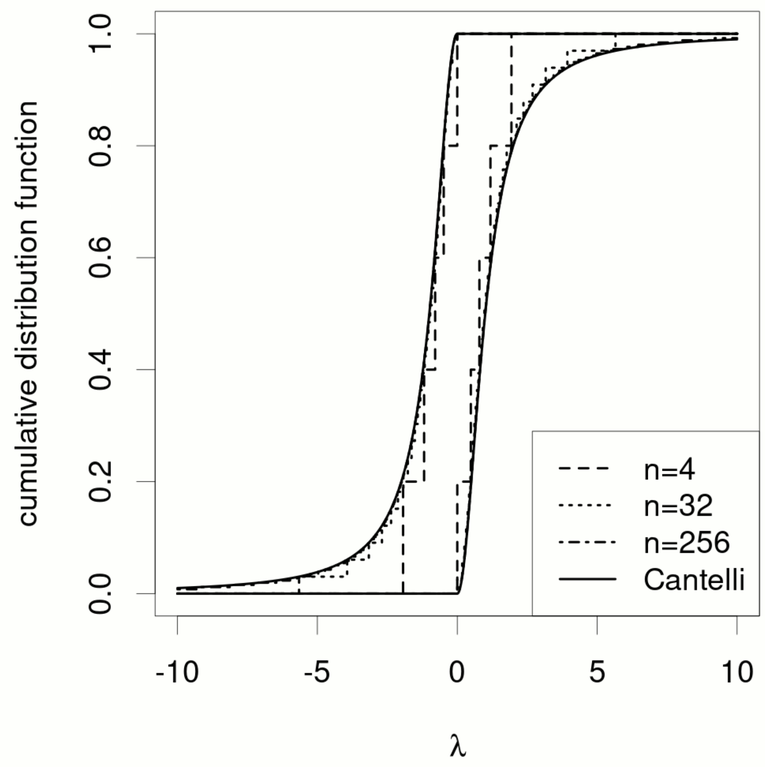
$$\underline{p}_1 \leq P(Z_{n+1} \leq \ell_1) \leq \overline{p}_1$$
$$\underline{p}_2 \leq P(Z_{n+1} \leq \ell_2) \leq \overline{p}_2$$

This gives us the following bounds on the coverage probability:

$$\underline{p}_2 - \overline{p}_1 \leq P(\overline{X} + \ell_1 S_n < X_{n+1} \leq \overline{X} + \ell_2 S_n) \leq \overline{p}_2 - \underline{p}_1$$

where, $S_n := S + \frac{\Delta_n}{\sqrt{n}}$.

Illustration



We compare our bounds with Cantelli's bounds. The correction term is larger for smaller sample sizes, which results in tighter bounds for smaller samples. For instance, for $n = 4$ this value is approximately $0.645 \times \Delta$, whereas for $n = 256$ it is only $0.063 \times \Delta$.

Discussion

- We provide a Cantelli type inequality based on exchangeability and construct a non-parametric p-box.
- We propose a prediction interval based on exchangeability.
- We can construct p-box only for the random variable Z_{n+1}.
- Finding bounds on the cumulative distribution of X_{n+1}, conditional on $\overline{X}$ and S using exchangeability remains and open problem.

References

[1] F. P. Cantelli. 'Sui confini della probabilità'. In: Atti del Congresso Internazional del Matematici 6 (1928). Bologna, pp. 47–59.

[2] Scott Ferson et al. Constructing Probability Boxes and Dempster-Shafer Structures. Tech. rep. SAND2002–4015. Unabridged version. Sandia National Laboratories, Jan. 2003.

[3] J. F. C. Kingman. 'Uses of Exchangeability'. In: The Annals of Probability 6.2 (Apr. 1978), pp. 183–197. DOI: 10.1214/aop/1176995566.

[4] John G. Saw, Mark C. K. Yang and Tse Chin Mo. 'Chebyshev Inequality with Estimated Mean and Variance'. In: The American Statistician 38.2 (1984), pp. 130–132. URL: http://www.jstor.org/stable/2683249.

This work is funded by the European Commissions H2020 programme, through the UTOPIAE Marie Curie Innovative Training Network, H2020-MSCA-ITN-2016, Grant Agreement number 72273.