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

Matthias C. M. Troffaes    Tathagata Basu

Department of Mathematical Sciences
Durham University, UK

July 2019
Outline

1  Cantelli’s inequality and p-boxes

2  Contributions

3  Conclusions
Cantelli’s inequality & induced p-box

- random variable $X$, known mean $\mu$ and variance $\sigma^2$
- Cantelli’s inequality [1]:

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

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

- induces a p-box (lower & upper cdf, bounding a set of probability measures)

**Issue**

What if only sample mean and sample standard deviation are known?
Contributions: Problem Statement

Assumptions

\( X_1, \ldots, X_n, X_{n+1} \) is a finite sequence of discrete exchangeable random variables.

Notation

- \( \bar{X} := \frac{1}{n} \sum_{j=1}^{n} X_j \)
- \( S^2 := \frac{\sum_{j=1}^{n} (X_j - \bar{X})^2}{(n-1)} \)

Objective

Find functions \( f \) and \( \bar{f} \) such that

\[
 f(\lambda, n) \leq P \left( \frac{X_{n+1} - \bar{X}}{S} \leq \lambda \right) \leq \bar{f}(\lambda, n) \quad (2)
\]
**Contributions: Cantelli-type inequality**

**Theorem**

For every $\lambda \geq 0$,

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

(3)

where $\lambda_n := \frac{n}{\sqrt{n^2-1}} \lambda$ and $\Delta_n := \sqrt{\frac{n+1}{n-1}} (\max X_j - \min X_j)$.

Similarly, for every $\lambda \leq 0$,

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

(4)

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

Plotting left and right hand sides from inequalities in theorem:
Contributions: P-box and Prediction Interval

Inequalities from theorem induce the following:

Non-parametric p-box

...on the random variable \( Z_{n+1} = \frac{X_{n+1} - \bar{X}}{S + \frac{\Delta_n}{\sqrt{n}}} \).

Not a p-box on \( X_{n+1} \) directly!

Prediction interval

...on the random variable \( X_{n+1} \).

For any \( \ell_1 \) and \( \ell_2 \), we can calculate \( \alpha_1 \) and \( \alpha_2 \) such that

\[
\alpha_1 \leq P(\bar{X} - \ell_1 S_n < X_{n+1} \leq \bar{X} + \ell_2 S_n) \leq \alpha_2
\]

(5)

where, \( S_n := S + \frac{\Delta_n}{\sqrt{n}} \).
Conclusions I

- novel Cantelli-type inequality
- induces non-parametric p-box and prediction interval
- only assumes exchangeability (rather than conditional independence)
- only uses sample mean and sample standard deviation
- similar to Saw [2] (but Saw does not induce a p-box)
- useful for modelling when only sample mean and sample standard deviation are known (e.g. measurement problems)
Conclusions II

- p-box only on $Z_{n+1}$ and not on $X_{n+1}$
- use of prediction interval not entirely clear
- p-box on $X_{n+1}$, conditional on $\bar{X}$ and $S$, using exchangeability, remains an open problem (might be impossible as pointed out by a kind reviewer)
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 the set of all possible re-orderings of \( X_1, X_2, \ldots, X_n \) is a group with identity element \( (X_1, X_2, \ldots, X_n) \).

\[
\text{Independent} \implies \text{Exchangeable}
\]

Cantelli’s inequality [1] and p-boxes [2]

Let \( X \sim \mathcal{N}(\mu, \sigma^2) \) be a random variable with known mean and known variance \( \sigma^2 \). Then the Cantelli’s inequality is given by:

\[
\Pr(X < \mu + \sigma) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\mu + \sigma} e^{-\frac{1}{2}x^2} dx \leq 1 - \frac{1}{2} \Phi \left( \frac{\mu + \sigma - \mu}{\sigma} \right) = 1 - \frac{1}{2} \Phi \left( \frac{\sigma}{\sqrt{\sigma^2} \sqrt{2\pi}} \right) = 1 - \frac{1}{2} \Phi \left( 1 \right).
\]

A p-box is specified by two cumulative distribution functions \( F \) and \( F' \) and represents the set of all cumulative distribution functions bounded by \( F \) and \( F' \):

\[ F(x) \leq F(x) \leq F'(x) \leq F'(x) \leq F'(x) \text{ for } x \in \mathbb{R} \]

Cantelli’s inequality gives distribution-free p-box for known \( \mu \) and \( \sigma^2 \):

\[
\Pr(X < \mu + \sigma) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\mu + \sigma} e^{-\frac{1}{2}x^2} dx \leq 1 - \frac{1}{2} \Phi \left( \frac{\mu + \sigma - \mu}{\sigma} \right) = 1 - \frac{1}{2} \Phi \left( \frac{\sigma}{\sqrt{\sigma^2} \sqrt{2\pi}} \right) = 1 - \frac{1}{2} \Phi \left( 1 \right)
\]

Cantelli’s inequality can also be written as a p-box on \( Y \):

\[
\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\mu + \sigma} e^{-\frac{1}{2}x^2} dx \leq 1 - \frac{1}{2} \Phi \left( \frac{\sigma}{\sqrt{2\pi}} \right) = 1 - \frac{1}{2} \Phi \left( 1 \right)
\]

Saw’s inequality (Chebyshev) [4]

Let \( X_1, X_2, \ldots, X_n \) be a sequence of independent random variables. Define \( Y = \sum_{i=1}^n X_i - \bar{X} \) and \( \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \). Then for every \( h \geq 0 \),

\[
\Pr(|Y| > h) = \Pr(|\bar{X} - \sum_{i=1}^n \frac{X_i}{n}| > h) = \frac{1}{n} \sum_{i=1}^n \Pr(|X_i - \bar{X}| > \frac{h}{n}) \leq \frac{1}{n} \sum_{i=1}^n \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\mu + \sigma} e^{-\frac{1}{2}x^2} dx \leq 1 - \frac{1}{2} \Phi \left( \frac{\sigma}{\sqrt{2\pi}} \right) = 1 - \frac{1}{2} \Phi \left( 1 \right)
\]

where \( \lambda_i = \frac{\sigma}{\sqrt{2\pi}} \) and \( \lambda = \max_i \lambda_i = \max_i \frac{\sigma}{\sqrt{2\pi}} \).

Main Result

Let \( X_1, X_2, \ldots, X_n \) be a finite sequence of independent random variables. Let \( \Delta \geq 2 \) denote the range of the \( Y_i \), i.e., \( \Delta \) is the maximum value that can be attained by \( Y_i \) and \( \bar{X} \), the minimum value. Let \( \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \). Then for every \( h \geq 0 \),

\[
\Pr(|Y| > h) = \Pr(|\bar{X} - \sum_{i=1}^n \frac{X_i}{n}| > h) = \frac{1}{n} \sum_{i=1}^n \Pr(|X_i - \bar{X}| > \frac{h}{n}) \leq \frac{1}{n} \sum_{i=1}^n \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\mu + \sigma} e^{-\frac{1}{2}x^2} dx \leq 1 - \frac{1}{2} \Phi \left( \frac{\sigma}{\sqrt{2\pi}} \right) = 1 - \frac{1}{2} \Phi \left( 1 \right)
\]

where \( \lambda_i = \frac{\sigma}{\sqrt{2\pi}} \) and \( \lambda = \max_i \lambda_i = \max_i \frac{\sigma}{\sqrt{2\pi}} \). Similarly, for \( \Delta \leq 0 \),

\[
\Pr(|Y| > h) = \Pr(|\bar{X} - \sum_{i=1}^n \frac{X_i}{n}| > h) = \frac{1}{n} \sum_{i=1}^n \Pr(|X_i - \bar{X}| > \frac{h}{n}) \leq \frac{1}{n} \sum_{i=1}^n \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\mu + \sigma} e^{-\frac{1}{2}x^2} dx \leq 1 - \frac{1}{2} \Phi \left( \frac{\sigma}{\sqrt{2\pi}} \right) = 1 - \frac{1}{2} \Phi \left( 1 \right)
\]

Here, \( \max \{ -\Delta \leq X \leq \Delta \} = \Delta \) and \( \Delta = -\Delta \).

P-box and Imprecise Prediction Interval

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

\[
X_{\text{imp}} = \frac{\bar{Y}}{\max_i \lambda_i}
\]

However, this cannot be used as a p-box for \( X_{\text{imp}} \). We cannot substitute the observed values for \( \bar{Y} \) and \( \lambda_i \) in the equation.

Our result gives, for every \( \lambda_i \) and \( \lambda \), the following bounds:

\[
\lambda \leq P(X_{\text{imp}} < \lambda) \leq P(X_{\text{imp}} < \lambda) \leq P(X_{\text{imp}} < \lambda) \leq P(X_{\text{imp}} < \lambda)
\]

This gives us the following bounds on the coverage probability:

\[
\Pr(Y < \Delta) = \int_{-\Delta}^{\Delta} f_Y(y) dy \leq \int_{-\Delta}^{\Delta} f_Y(y) dy \leq \int_{-\Delta}^{\Delta} f_Y(y) dy \leq \int_{-\Delta}^{\Delta} f_Y(y) dy
\]

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 \( n = 1 \), this value is approximately \( 0.605 \Delta \), whereas for \( n = 250 \), it is only \( 0.605 \Delta \).

Illustration

We look forward to seeing you at our poster!