Adding Constraints to Situations When, In Addition to Intervals, We Also Have Partial Information about Probabilities

Martine Ceberio[1], Scott Ferson[2], Cliff Joslyn[3],
Vladik Kreinovich[1], and Gang Xiang[1]

[1]Department of Computer Science
University of Texas, El Paso, TX 79968, USA
[2]Applied Biomathematics
[3]Los Alamos National Laboratory

mceberio@utep.edu, scott@ramas.com, joslyn@lanl.gov
vladik@utep.edu, gxiang@utep.edu

1. Statistical Analysis Is Important

- *Fact:* statistical analysis of measurement and observation results is an important part of data processing and data analysis.
- *Specifics:*
  - when faced with new data,
  - engineers and scientists usually start with estimating standard statistical characteristics such as:
    * the mean E,
    * the variance V,
    * the probability distribution function (pdf) $F(x)$ of each variable, and
    * the covariance and correlation between different variables.

2. Limitations of Traditional Statistical Techniques and the Need to Consider Interval Uncertainty

- *Main assumption:* traditional statistical techniques assume that the measured values $\widetilde{x}_1, \ldots \widetilde{x}_n$ coincide with the actual values $x_1, \ldots, x_n$ of the measured quantities.
- *This assumption is often true:* if the variability of each variable is much higher than the measurement errors $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$.
- *Example:* the accuracy of measuring a person's height (≈ 1 cm) is $\ll$ variability in height.
- *Sometimes, this assumption is not true:* when the measurement errors Δx_i are of the same order of magnitude.
- *Conclusion:* Δx_i cannot be ignored in statistical analysis.
- *Frequent situation:* the only information about Δx_i is the upper bound Δ_i: $|\Delta x_i| \le \Delta_i$.
- *Interval uncertainty:* the only information about x_i is that $x_i \in \mathbf{x}_i \stackrel{\text{def}}{=} [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$.

3. Adding Interval Uncertainty to Statistical Techniques: What Is Known

- *We start with:* a statistic $C(x_1, \ldots, x_n)$, such as:
  - population mean $E = \dfrac{1}{n} \cdot \sum\limits_{i=1}^{n} x_i$;
  - population variance $V = \dfrac{1}{n} \cdot \sum\limits_{i=1}^{n} (x_i - E)^2$;
  - histogram pdf $F_n(x) = \dfrac{\# i : x_i \le x}{n}$;
  - population covariance $C_{x,y} = \dfrac{1}{n} \cdot \sum\limits_{i=1}^{n} (x_i - E_x) \cdot (y_i - E_y)$.
- *Interval extension:* find the range

$$\mathbf{C} = C(\mathbf{x}_1, \ldots, \mathbf{x}_n) \stackrel{\text{def}}{=} \{C(x_1, \ldots, x_n) : x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\}.$$

- *General case:* the general problem is NP-hard, even for V.
- *Conclusion:* in general, we can only compute an enclosure.
- *Specific cases:* efficient algorithms are possible: for $\mathbf{E}$, for $\underline{V}$, for $\overline{V}$ when $[\underline{x}_i, \overline{x}_i] \not\subseteq (\underline{x}_j, \overline{x}_j)$, etc.

4. Limitations of the Existing Approach

- *Currently used idea:*
  - we start with a statistic $C(x_1, \ldots, x_n)$ for estimating a given characteristic S;
  - we evaluate this statistic under interval uncertainty, resulting in $\mathbf{C} = C(\mathbf{x}_1, \ldots, \mathbf{x}_n)$.
- *First limitation of this idea:*
  - we know that $C(x_1, \ldots, x_n) \approx S$;
  - sometimes, the estimation error $C(x_1, \ldots, x_n) - S \neq 0$ is not always taken into account when estimating $\mathbf{C}$.
- *Solution:* instead of the original statistic C, we consider the bounds C^- and C^+ of the confidence interval.
- *Good news:* the interval $\left[\underline{C}^-, \overline{C}^+\right]$ is an enclosure for S (with appropriate certainty).
- *Remaining limitation:* excess width.
- *New idea:* find $\mathbf{S} = \{S(F) : F \text{ is possible}\}$.
- *Related problem:* how to describe class $\mathcal{F}$ of possible probability distributions F.

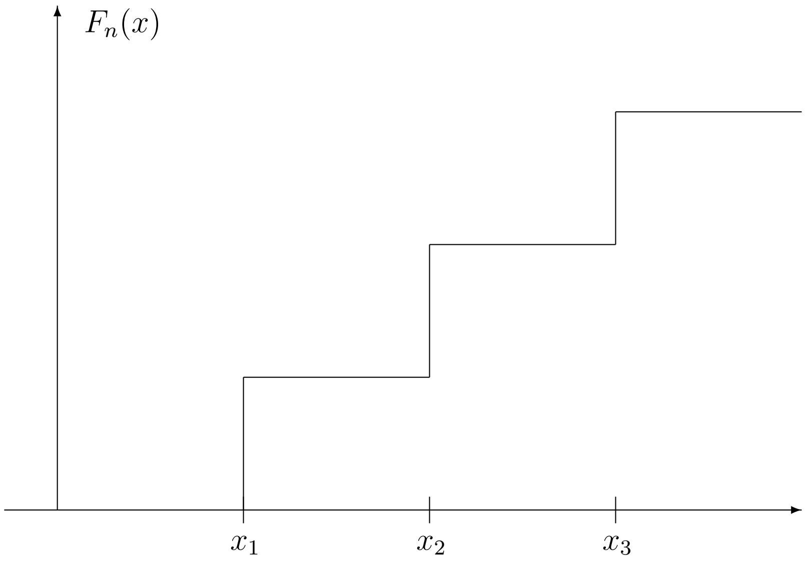
5. How to Describe Possible Probability Distributions: p-Boxes

- *General situation:*
  - we do not know the probability distribution of the actual values x_i;
  - we want to determine this distribution.
- *Question:* which characteristics of this distribution are practically useful?
- *Practical example:*
  - there is a critical threshold x_0 after which a chip delays too much, a panel cracks, etc.;
  - we want to make sure that the probability of exceeding x_0 is small.
- *Resulting characteristic:* $\text{Prob}(x_i \le x_0)$, i.e., cdf $F(x_0)$.
- *p-box:*
  - we cannot determine the *exact* values of $F(x)$;
  - thus, we should look for *bounds* $\mathbf{F}(x) = [\underline{F}(x), \overline{F}(x)]$;
  - the function $x \to \mathbf{F}(x)$ is called a *p-box*.

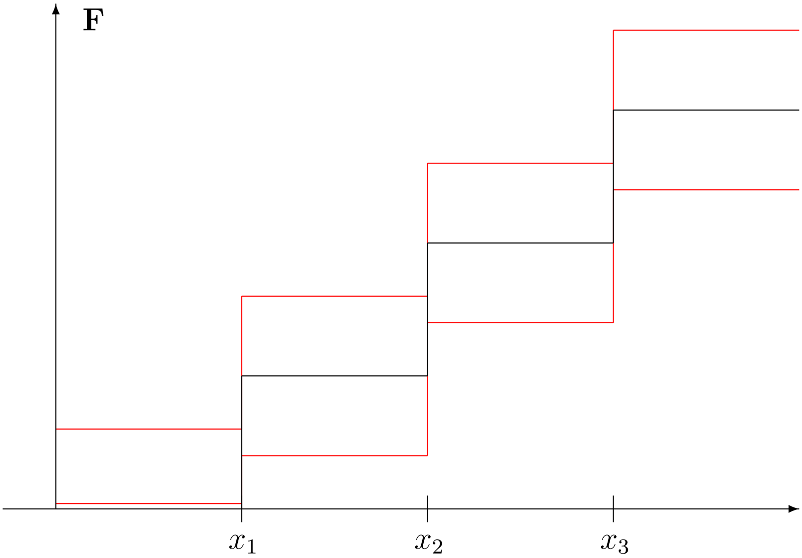
6. Kolmogorov-Smirnov (KS) p-Box

- *New idea (reminder):*
  - transform observations $x_1, \ldots, x_n$ into a p-box;
  - estimate a characteristic S based on the p-box.
- *How to transform:* use KS inequalities.
- *Main idea behind KS:* for each x_0, we have
  - actual (unknown) probability $p = F(x_0)$ that $x \le x_0$, and
  - frequency $F_n(x_0) = \dfrac{\#i : x_i \le x_0}{n}$.
- *Known:* for large n, $F_n(x_0) \approx$ normal, and with given certainty α, we have $p - k \cdot \sigma \le F_n(x_0) \le p + k \cdot \sigma$, where $\sigma = \sqrt{\dfrac{p \cdot (1-p)}{n}}$ and $k = k(\alpha)$.
- *Conclusion:* with certainty α, we get bounds on $p = F(x_0)$ in terms of $F_n(x_0)$.
- We use these bounds for $x_0 = x_i$ and use monotonicity to get bounds $[F_n(x) - \varepsilon, F_n(x) + \varepsilon]$ for all $x \in [x_i, x_{i+1}]$.

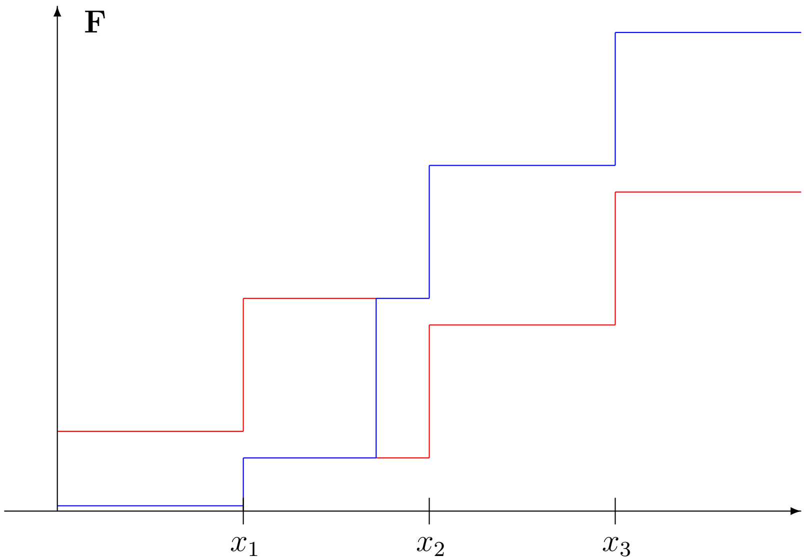
7. Illustration: Histogram Pdf



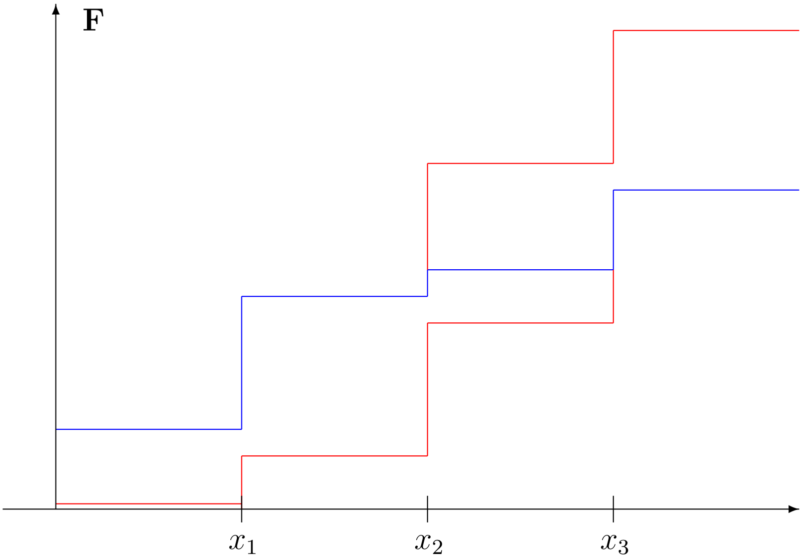
8. Illustration: Kolmogorov-Smirnov p-Box



9. Computing $\underline{V}$



10. Computing $\overline{V}$



11. Computational Complexity of Computing $\underline{V}$ and $\overline{V}$

- *Traditional method:*
  - we can compute $\underline{V}$ in linear time $O(n)$;
  - computing $\overline{V}$ is, in general, NP-hard;
  - when $[\underline{x}_i, \overline{x}_i] \not\subseteq (\underline{x}_j, \overline{x}_j)$, we can compute $\overline{V}$ is linear time.
- *Analysis:*
  - in effect, the variance of $F \in \mathbf{F}$ can be reduced to the variance over horizontal layers;
  - these layers satisfy the above "subset" property.
- *New method:*
  - we can compute $\underline{V}$ in linear time $O(n)$, and
  - we can compute $\overline{V}$ in linear time $O(n)$;

12. How to Handle Additional Constraints

- *Previously:* the only information we have is $F(x) \in \mathbf{F}(x)$.
- *Frequent situation:* we have additional information about $F(x)$.
- *Example:* we know the shape of $F(x)$, i.e., we know that $F(x) = F_0(x, a_1, \ldots, a_n)$ for known F_0 and $a_i \in [\underline{a}_i, \overline{a}_i]$.
- *Typical situation:* $F(x) = F_0\left(\sum\limits_{i=1}^n a_i \cdot e_i(x)\right)$.
- *Example:* Gaussian $F(x) = F_0\left(\dfrac{x-a}{\sigma}\right) = F_0(a_1 \cdot x + a_2)$.
- *p-box solution:* find a p-box containing all such $F(x)$, and estimate, e.g., $\mathbf{V}$, based on this p-box.
- *Drawback:* excess width.
- *Exact estimates:* $\underline{F}(x_i) \le F_0\left(\sum\limits_{i=1}^n a_i \cdot e_i(x_i)\right) \le \overline{F}(x_i)$, hence

$$F_0^{-1}(\underline{F}(x_i)) \le \sum_{i=1}^n a_i \cdot e_i(x_i) \le F_0^{-1}(\overline{F}(x_i)). \qquad (*)$$

- *Algorithm:* apply linear programming to (*) and $\underline{a}_i \le a_i \le \overline{a}_i$.

13. Gauging Amount of Uncertainty

- *Shannon's idea:* (average) number of "yes"-"no" (binary) questions that we need to ask to determine the object.
- *Fact:* after q binary questions, we have 2^q possible results.
- *Discrete case:* if we have n alternatives, we need q questions, where $2^q \geq n$, i.e., $q \sim \log_2(n)$.
- *Discrete probability distribution:* $q = -\sum p_i \cdot \log_2(p_i)$.
- *Continuous case – definition:* number of questions to find an object with a given accuracy ε.
- *Interval uncertainty:* if $x \in [a, b]$, then $q \sim S - \log_2(\varepsilon)$, with $S = \log_2(b - a)$.
- *Probabilistic uncertainty:* $S = -\int \rho(x) \cdot \log_2 \rho(x)\, dx$.

14. Case of a Continuous Probability Distribution

- Once an *approximate* value r is determined, possible *actual* values of x form an interval $[r-\varepsilon, r+\varepsilon]$ of width 2ε.
- So, we divide the real line into intervals $[x_i, x_{i+1}]$ of width 2ε and find the interval that contains x.
- The average number of questions is $S = -\sum p_i \cdot \log_2(p_i)$, where the probability p_i that $x \in [x_i, x_{i+1}]$ is $p_i \approx 2\varepsilon \cdot \rho(x_i)$.
- So, for small ε, we have

$$S = -\sum \rho(x_i) \cdot \log_2(\rho(x_i)) \cdot 2\varepsilon - \sum \rho(x_i) \cdot 2\varepsilon \cdot \log_2(2\varepsilon),$$

where the first sum in this expression is the integral sum for the integral $S(\rho) \stackrel{\text{def}}{=} -\int \rho(x) \cdot \log_2(\rho(x))\,dx$, so

$$S \approx -\int \rho(x) \cdot \log_2(\rho(x))\,dx - \log_2(2\varepsilon).$$

15. Case of a p-Box

- *Situation:* we know that

$$F(x) \in \mathbf{F}(x) = [F_0(x) - \Delta F(x), F_0(x) + \Delta F(x)],$$

where $F_0(x)$ is smooth, with $\rho_0(x) \stackrel{\text{def}}{=} F_0'(x)$.

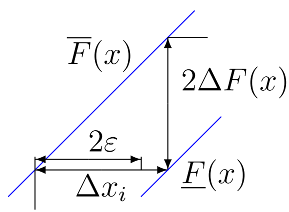
- *Problem:* find the range $[\underline{S}, \overline{S}] = \{S_\varepsilon(F) : F \in \mathbf{F}\}$.
- *Known result:* asymptotically,

$$\overline{S} \sim -\int \rho_0(x) \cdot \log_2(\rho_0(x))\, dx - \log_2(2\varepsilon).$$

- *New result:* $\underline{S} \sim -\int \rho_0(x) \cdot \log_2(\max(2\Delta F(x), 2\varepsilon \cdot \rho_0(x)))\, dx.$
- *Comment:* when $\varepsilon \to 0$, $\overline{S} \to \infty$ but $\underline{S}$ remains finite.
- *Idea of the proof:* $p_i \approx \rho_0(x_i) \cdot \Delta x_i$, hence

$$-\sum p_i \cdot \log_2(p_i) \approx -\int \rho_0(x) \cdot \log(\rho_0(x) \cdot \Delta x)\, dx.$$

Here, $\Delta x_i = \max\left(\dfrac{2\Delta F(x)}{\rho_0(x)}, 2\varepsilon\right)$:



16. Acknowledgments

This work was supported in part:

- by National Science Foundation grants EAR-0225670 and DMS-0532645 and
- by Texas Department of Transportation grant No. 0-5453

The authors are thankful to Bill Walster for fruitful discussions.

17. Shannon's Derivation: Reminder

- *Situation:* we know the probabilities $p_1, \ldots, p_n$ of different alternatives.
- We repeat the selection N times.
- Let N_i be number of times when we get A_i.
- For big N, the value N_i is $\approx$ normally distributed with average $a = p_i \cdot N$ and $\sigma = \sqrt{p_i \cdot (1 - p_i) \cdot N}$.
- With certainty depending on k_0, we conclude that

$$N_i \in [a - k_0 \cdot \sigma, a + k_0 \cdot \sigma].$$

- Let $N_{\text{con}}(N)$ be the number of situations for which N_i is within these intervals.
- Then, for N repetitions, we need $q(N) = \log_2(N_{\text{cons}})$ questions.
- Per repetition, we need $S = q(N)/N$ questions.

18. Shannon's Derivation (cont-d)

- *Shannon's theorem:* $S \to -\sum p_i \cdot \log_2(p_i)$.
- *Proof:*

$$N_{\text{cons}} \sim$$

$$\frac{N!}{N_1!(N-N_1)!} \cdot \frac{(N-N_1)!}{N_2!(N-N_1-N_2)!} \cdot \ldots =$$

$$\frac{N!}{N_1!N_2!\ldots N_n!}$$

where $k! \sim (k/e)^k$. So,

$$N_{\text{cons}} \sim \frac{\left(\frac{N}{e}\right)^N}{\left(\frac{N_1}{e}\right)^{N_1} \cdot \ldots \cdot \left(\frac{N_n}{e}\right)^{N_n}}.$$

Since $\sum N_i = N$, terms e^N and e^{N_i} cancel each other.

- Substituting $N_i = N \cdot f_i$ and taking logarithms, we get

$$\log_2(N_{\text{cons}}) \approx -N \cdot f_1 \cdot \log_2(f_1) - \ldots - N \cdot f_n \log_2(f_n).$$