
Vladik Kreinovich\textsuperscript{1}, Nitaya Buntao\textsuperscript{2}, and Olga Kosheleva\textsuperscript{1}

\textsuperscript{1}University of Texas at El Paso
El Paso, TX 79968, USA
vladik@utep.edu, olgak@utep.edu

\textsuperscript{2}Department of Applied Statistics
King Mongkut’s University of Technology
North Bangkok
Bangkok 10800 Thailand
taltanot@hotmail.com
1. Formulation of the Problem

• Traditionally, most statistical techniques assume that the random variables are normally distributed.

• For such distributions:
  – a natural characteristic of the “average” value is the mean, and
  – a natural characteristic of the deviation from the average is the variance.

• In practice, we encounter heavy-tailed distributions, with infinite variance; what are analogs of:
  – “average” and deviation from average?
  – correlation?
  – how to take into account interval uncertainty?
2. Normal Distributions Are Most Widely Used

• Most statistical techniques assume that the random variables are normally distributed:

\[ \rho(x) = \frac{1}{\sqrt{2\pi \cdot V}} \cdot \exp\left(-\frac{(x - m)^2}{2V}\right). \]

• For such distributions:
  – a natural characteristic of the “average” value is the mean \( m \overset{\text{def}}{=} E[x] \), and
  – a natural characteristic of the deviation from the average is the variance \( V \overset{\text{def}}{=} E[(x - m)^2] \).

• It is known that a normal distribution is uniquely determined by \( m \) and \( V \).

• Thus, each characteristic (mode, median, etc.) is uniquely determined by \( m \) and \( V \).
3. Estimating the Values of the Characteristics: Case of Normal Distributions

- We have a sample consisting of the values \( x_1, \ldots, x_n \).
- We can use the Maximum Likelihood Method: \( m \) and \( V \) maximizing

\[
L = \rho(x_1) \cdots \rho(x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi \cdot V}} \cdot \exp \left( -\frac{(x_i - m)^2}{2V} \right). 
\]

- Maximizing \( L \) is equivalent to minimizing

\[
\psi \overset{\text{def}}{=} -\ln(L) = \sum_{i=1}^{n} \left[ 1 \cdot \ln(2\pi \cdot V) + \frac{(x_i - m)^2}{2V} \right].
\]

- Equating derivatives to 0, we get:

\[
\hat{m} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i; \quad \hat{V} = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \hat{m})^2.
\]
4. In Many Practical Situations, We Encounter Heavy-Tailed Distributions

- In the 1960s, Benoit Mandelbrot empirically studied fluctuations.
- He showed that larger-scale fluctuations follow the power-law distribution \( \rho(x) = A \cdot x^{-\alpha} \), with \( \alpha \approx 2.7 \).
- For this distribution, variance is infinite.
- Such distributions are called heavy-tailed.
- Similar heavy-tailed laws were empirically discovered in other application areas.
- These result led to the formulation of fractal theory.
- Since then, similar heavy-tailed distributions have been empirically found:
  - in other financial situations and
  - in many other application areas.
5. First Problem: How to Characterize Such Distributions?

- Usually, \textit{variance} is used to describe deviation from the average.
- For heavy-tailed distributions, variance is \textit{infinite}.
- So, we \textit{cannot} use variance to describe the deviation from the “average”.
- Thus, we need to come up with \textit{other} characteristics for describing this deviation.
- We will \textit{describe} such characteristics in the first part of this talk.
- We will also describe \textit{how we can estimate} these characteristics.
6. How to Describe Deviation from the “Average” for Heavy-Tailed Distributions: Analysis

- A standard way to describe preferences of a decision maker is to use the notion of utility $u$.
- According to decision theory, a user prefers an alternative for which the expected value
  \[ \int \rho(x) \cdot u(x) \, dx \to \max. \]
- Alternative, the expected value $\int \rho(x) \cdot U(x) \, dx$ of the disutility $U \overset{\text{def}}{=} -u$ is the smallest possible.
- If we replace $x \to m \approx x$, there is disutility $U(x - m)$.
- So, we choose $m$ s.t. $\int \rho(x) \cdot U(x - m) \, dx \to \min$.
- The resulting minimum describes the deviation of the values from this “average”.

\[ \int \rho(x) \cdot u(x) \, dx \to \max. \]
7. Resulting Definitions

- Let $U : \mathbb{R} \to \mathbb{R}_0$ be a function for which:
  - $U(0) = 0$,
  - $U(d)$ is (non-strictly) increasing for $d \geq 0$, and
  - $U(d)$ is (non-strictly) decreasing for $d \leq 0$.
- For a distribution $\rho(x)$, by a $U$-mean, we mean the value $m_U$ that minimizes $\int \rho(x) \cdot U(x - m) \, dx$.
- By a $U$-deviation, we mean
  
  $V_U \overset{\text{def}}{=} \int \rho(x) \cdot U(x - m_U) \, dx$.

- When $U(x) = x^2$, $m_U$ is mean, and $V_U$ is variance.
- When $U(x) = |x|$, $m_U$ is median, and $V_U$ is average absolute deviation $V_U = \int \rho(x) \cdot |x - m_U| \, dx$. 

8. What Are the Reasonable Measures of Dependence for Heavy-Tailed Distributions?

- In the traditional statistics, a reasonable measure of dependence is the correlation

\[ \rho_{xy} = \frac{E[(x - E(x)) \cdot (y - E(y))]}{\sqrt{V_x \cdot V_y}}. \]

- For heavy-tailed distributions, variances are infinite, so this formula cannot be applied.

- **Possibility**: Kendall’s tau, the proportion of pairs \((x, y)\) and \((x', y')\) s.t. \(x\) and \(y\) change in the same direction:

  either \((x \leq x' \& y \leq y')\) or \((x' \leq x \& y' \leq y)\).

- **Remaining problem**: what if we are interested only in linear dependencies?
9. Proposed Definition

- **Idea:** \( c \) describes how much disutility decreases when we use \( x \) to help predict \( y \):

\[
c \overset{\text{def}}{=} \frac{V_U(y) - V_{U,F}(y|x)}{V_U(y)},
\]

where

\[
V_U(y) \overset{\text{def}}{=} \min_m \int \rho(x,y) \cdot U(y - m) \, dx \, dy
\]

and

\[
V_{U,F}(y|x) \overset{\text{def}}{=} \min_{f \in F} \int \rho(x,y) \cdot U(y - f(x)) \, dx \, dy.
\]

- The function \( f \) at which the minimum is attained is called \( F \)-regression.

- When \( U(d) = d^2 \) and \( F \) is the class of all linear functions, \( c = \rho^2 \).
10. Discussion

- For normal distributions and linear functions, correlation is symmetric:
  - if we can reconstruct $y$ from $x$,
  - then we can reconstruct $x$ from $y$.
- Our definition is, in general, not symmetric.
- This asymmetry makes perfect sense.
- For example, suppose that $y = x^2$:
  - then, if we know $x$, then we can uniquely reconstruct $y$;
  - however, if we know $y$, we can only reconstruct $x$ modulo sign.
11. How to Estimate the New Characteristics from Observations

- In the above text: we defined the desired characteristics in terms of the probability density function (pdf) $\rho(x)$.
- In practice: we often do not know the distribution.
- Instead: we know the sample values $x_1, \ldots, x_n$.
- A natural idea: use the “histogram” distribution, in which each $x_i$ appears with equal probability $\frac{1}{n}$.
- Example: for $\rho(x) = \frac{1}{n} \cdot \sum_{i=1}^{n} \delta(x - x_i)$, the mean

$$E = \int \rho(x) \cdot x \, dx \text{ turns into } \hat{E} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i.$$

- Similarly: we get

$$\hat{V} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left( x_i - \hat{V} \right)^2.$$
12. Resulting Estimates for $m_U$ and $V_U$

- For each sample $x_1, \ldots, x_n$, by a $U$-estimate, we mean the value $\hat{m}_U$ that minimizes $\frac{1}{n} \cdot \sum_{i=1}^{n} U(x_i - m)$.

- By an estimate for $U$-deviation, we mean

$$\hat{V}_U \overset{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} U(x_i - \hat{m}_U).$$

- When $U(x) = x^2$, $\hat{m}_U$ is arithmetic mean, and $\hat{V}_U$ is sample variance.

- When $U(x) = |x|$, $\hat{m}_U$ is sample median, and $\hat{V}_U$ is average absolute deviation $\hat{V}_U = \frac{1}{n} \cdot \sum_{i=1}^{n} |x_i - \hat{m}_U|$. 
13. How to Estimate $m_U$ and $V_U$

- Once we compute $\hat{m}_U$, the computation of
  \[ \hat{V}_U = \frac{1}{n} \cdot \sum_{i=1}^{n} U(x_i - \hat{m}_U) \] is straightforward.

- Estimating $\hat{m}_U$ means optimizing a function of a single
  variable $\frac{1}{n} \cdot \sum_{i=1}^{n} U(x_i - m) \rightarrow \text{min}$.

- This optimization problem is equivalent to the Maximum Likelihood (ML): for $U(x) = -\ln(\rho_0(x))$,
  \[ L = \rho_0(x_1 - m) \cdot \ldots \cdot \rho_0(x_n - m) \rightarrow \max \Leftrightarrow \]
  \[ \psi \overset{\text{def}}{=} -\ln(L) = \sum_{i=1}^{n} U(x_i - m) \rightarrow \min. \]

- Similar algorithms are used in robust statistics, as $M$-methods, which are mathematically equivalent to ML.
14. Estimates for $U$-Correlation

- **Idea:** $\hat{c}$ describes how much disutility decreases when we use $x_i$ to help predict $y_i$:

$$\hat{c} \overset{\text{def}}{=} \frac{\hat{V}_U(y) - \hat{V}_{U,F}(y|x)}{\hat{V}_U(y)},$$

where

$$\hat{V}_U(y) \overset{\text{def}}{=} \min_m \frac{1}{n} \cdot \sum_{i=1}^{n} U(y_i - m)$$

and

$$\hat{V}_{U,F}(y|x) \overset{\text{def}}{=} \min_{f \in F} \frac{1}{n} \cdot \sum_{i=1}^{n} U(y_i - f(x_i)).$$

- The function $\hat{f}$ at which the minimum is attained is called *sample $\mathcal{F}$-regression*.

- When $U(d) = d^2$ and $\mathcal{F}$ is the class of all linear functions, $\hat{c} = \hat{\rho}^2$. 
15. Need to Take into Account Interval Uncertainty

- In practice, we often know approximate values \( \tilde{x}_i \approx x_i \).
- Sometimes, we know the probabilities of different values of the approximation error \( \Delta x_i \defeq \tilde{x}_i - x_i \).
- Often, we only know the upper bound \( \Delta_i : |\Delta x_i| \leq \Delta_i \).
- So, we only know that \( x_i \in x_i = [\tilde{x}_i - \Delta_i, \tilde{x}_i + \Delta_i] \).
- For each estimator \( C(x_1, \ldots, x_n) \), different \( x_i \in x_i \) lead, in general, to different values \( C(x_1, \ldots, x_n) \).
- Thus, we must find the range:

\[
C = [\underline{C}, \overline{C}] = \{ C(x_1, \ldots, x_n) : x_1 \in x_1, \ldots, x_n \in x_n \}.
\]

- This *interval computations* problem is, in general, NP-hard.
16. Estimating the Heavy-Tailed-Related Deviation Characteristics under Interval Uncertainty

- When we know the exact values of $x_i$, we know how to compute $\hat{V}_U = \min_m \frac{1}{n} \sum_{i=1}^{n} U(x_i - m)$.

- In practice, the values $x_i$ are often only known with interval uncertainty.

- We only know the intervals $x_i = [x_i, \bar{x}_i]$ that contain the unknown values $x_i$.

- In this case, it is desirable to compute the range $[\hat{V}_U, \bar{V}_U]$ of possible values of $\hat{V}_U$ when $x_i \in x_i$. Here:
  - The value $\underline{\hat{V}_U}$ is the minimum of the function $\hat{V}_U(x_1, \ldots, x_n)$ when $x_i \in x_i$.
  - The value $\overline{\hat{V}_U}$ is the maximum of the function $\hat{V}_U(x_1, \ldots, x_n)$ when $x_i \in x_i$. 
17. Algorithm for Computing $\widehat{V}_U$

- First, sort all $2n$ endpoints $x_i$ and $\overline{x}_i$ into an increasing sequence $x(1) \leq x(2) \leq \ldots \leq x(2n)$.

- These values, with $x(0) \overset{\text{def}}{=} -\infty$ and $x(2n+1) \overset{\text{def}}{=} +\infty$, divide the real line into zones $[x(k), x(k+1)]$, $k = 0, 1, \ldots, 2n$.

- For each zone $z$, we select the values $x_1, \ldots, x_n$ as follows: for some value $m$ (to be determined),
  - if $\overline{x}_i \leq r(k)$, then we select $x_i = \overline{x}_i$;
  - if $r(k+1) \leq x_i$, then we select $x_i = \overline{x}_i$;
  - for all other $i$, we select $x_i = m$.

- Then, we take only the values for which $x_i \neq m$, and find their $U$-estimate $\widehat{m}_U$; if $m_U \in z$, we compute $\widehat{V}_U$.

- The smallest of thus computed $U$-deviations is the desired value $\widehat{V}_U$. 
18. Computation Time for This Algorithm

- Sorting takes $O(n \cdot \log(n))$ steps.
- After that, for each of $2n = O(n)$ zones, we need:
  - $O(n)$ steps to perform the computations and
  - the time – that we will denote by $T_{\text{exact}}$ – to compute the $U$-estimate and $U$-deviation.
- Thus, the total computation time is equal to
  \[ O(n \cdot \log(n)) + O(n^2) + O(n) \cdot T_{\text{exact}} = O(n^2) + O(n) \cdot T_{\text{exact}}. \]
- Conclusion:
  - if we can compute $\hat{V}_U$ for exactly known $x_i$ in polynomial time (e.g., linear), then
  - we can compute $\hat{V}_U$ under interval uncertainty also in polynomial time (e.g., quadratic).
19. Computing $\widehat{V}_U$: Analysis of the Problem

- **Fact:** the maximum $\widehat{V}_U$ is attained:
  - if $\overline{x}_i \leq m$, for $x_i = \overline{x}_i$;
  - if $m \leq \underline{x}_i$, for $x_i = \overline{x}_i$;
  - if $\underline{x}_i \leq m \leq \overline{x}_i$, for $x_i = \overline{x}_i$ or $x_i = \underline{x}_i$.

- **Resulting algorithm:**
  - try all possible combinations of endpoints that satisfy the above conditions, and
  - select the largest of the resulting values $\widehat{V}_U$.

- **Problem:** we may need $2^n$ combinations, too long already for $n \approx 300$.

- **Explanation:** even for $U(d) = d^2$, the problem of computing $\widehat{V}_U$ is NP-hard.
20. Case when a Feasible Algorithm Is Possible

- **Reminder:** we consider cases where:
  - if \( \bar{x}_i \leq m \), for \( x_i = \bar{x}_i \);
  - if \( m \leq \underline{x}_i \), for \( x_i = \underline{x}_i \);
  - if \( \underline{x}_i \leq m \leq \bar{x}_i \), for \( x_i = \underline{x}_i \) or \( x_i = \bar{x}_i \).

- **Situation.** For some \( C \), every group of \( > C \) intervals has an empty intersection.

- **Algorithm:** for each zone \( z \), we consider case \( m \in z \).

- For each zone, there are \( \leq C \) intervals for which
  \[ \underline{x}_i \leq m \leq \bar{x}_i. \]

- So we need to check \( \leq 2^C \) combinations for each zone.

- Since \( C \) is a constant, \( 2^C = O(1) \).
21. Resulting Algorithm for Computing $\hat{V}_U$

• First, we sort all endpoints $x_i$ and $\overline{x}_i$ into an increasing sequence, and add $x(0) = -\infty$ and $x(2n+1) = +\infty$:

$$-\infty = x(0) \leq x(1) \leq x(2) \leq \ldots \leq x(2n) \leq x(2n+1) = +\infty.$$  

• For each zone $[x(k), x(k+1)]$, we do the following:
  - if $\overline{x}_i \leq r(k)$, then we select $x_i = \overline{x}_i$;
  - if $r(k+1) \leq \overline{x}_i$, then we select $x_i = x_i$;
  - for all other $i$, we select either $x_i = x_i$ or $x_i = \overline{x}_i$.

• For each zone, we have $\leq C$ indices $i$ that allow two selections, so we thus get $\leq 2^C$ selections.

• For each of these selections, we compute the $U$-deviation.

• The largest of these $\hat{V}_U$ is the desired value $\overline{V}_U$.

• This algorithm requires time $O(n^2) + O(n) \cdot T_{\text{exact}}$. 

22. When a Feasible Algorithm Is Possible

• 2nd case: no interval is a proper subinterval of another: 
  \([x_i, \bar{x}_i] \not\subset (x_j, \bar{x}_j)\) for all \(i\) and \(j\).

• Example: measurements made by the same instrument.

• Under this property, lexicographic order
  \([x_i, \bar{x}_i] \leq [x_j, \bar{x}_j] \iff ((x_i < x_j) \lor (x_i = x_j \land \bar{x}_i < \bar{x}_j))\)
  sorts the intervals by both endpoints:
  \(\bar{x}_1 \leq \bar{x}_2 \leq \ldots \leq \bar{x}_n; \ \bar{x}_1 \leq \bar{x}_2 \leq \ldots \leq \bar{x}_n\).

• One can prove that, for some \(k\), the maximum is attained at a tuple \((x_1, \ldots, x_k, \bar{x}_k+1, \ldots, \bar{x}_n)\).

• There are \(n + 1\) such tuples, so we have a polynomial-time algorithm.

• Similar arguments can be made when the intervals can be divided into \(m\) groups with this property.
23. Resulting Algorithms for Computing $\hat{V}_U$

- **Applicable:** when $[x_i, \bar{x}_i] \not\subset (x_j, \bar{x}_j)$ for all $i$ and $j$.

- First, we sort all the intervals in lexicographic order
  $$[x_i, \bar{x}_i] \leq [x_j, \bar{x}_j] \iff ((x_i < x_j) \lor (x_i = x_j \land \bar{x}_i < \bar{x}_j)).$$

- Then, we compute $V_U$ for all $n + 1$ tuples of the form $(x_1, \ldots, x_k, \bar{x}_{k+1}, \ldots, \bar{x}_n)$, with $k = 0, 1, \ldots, n$.

- The largest of thus computed $U$-deviations is the desired value $\hat{V}_U$.

- This algorithm requires time
  $$O(n \cdot \log(n)) + O(n) \cdot T_{exact}.$$
24. Algorithms for Computing $\hat{V}_U$ (cont-d)

- **Applicable:** all intervals can be divided into $m$ groups each of which satisfies the no-subinterval property.

- We sort all intervals within each group in lexicographic order.

- For each group $j = 1, \ldots, m$, with $n_j \leq n$ elements, we consider $n_j + 1 \leq n + 1$ tuples of the form
  
  $$(x_1, \ldots, x_{k_j}, \bar{x}_{k_j+1}, \ldots, \bar{x}_n).$$

- We consider all possible combinations of such tuples corresponding to all possible vectors $(k_1, \ldots, k_m)$.

- For each of these $\leq n^m$ vectors, we compute $\hat{V}_U$.

- The largest of these $\hat{V}_U$ is the desired value $\hat{V}_U$.

- This algorithm requires time
  
  $$O(n \cdot \log(n)) + O(n^m) \cdot T_{\text{exact}}.$$
25. **Conclusion**

- Uncertainty is usually gauged by using standard statistical characteristics: mean, variance, correlation, etc.
- Then, we use the known values of these characteristics to select a decision.
- Sometimes, we only know bounds, then we use these bounds in decision making.
- Sometimes, it becomes clear that the selected characteristics do not always describe a situation well.
- Then other known (or new) characteristics are proposed.
- A good example is description of volatility in finance:
  - it started with variance, and
  - now many descriptions are competing, all with their own advantages and limitations.
26. Conclusion (cont-d)

- **Reminder**: sometimes, the traditional statistical characteristics do not work well.

- In such situations, a natural idea is to come up with characteristics tailored to specific application areas.

- E.g., a characteristic that maximizes the expected utility of the resulting risk-informed decision making.

- How to estimate these characteristics when the sample values are only known with interval uncertainty?

- We show that:
  - algorithms originally developed for estimating traditional characteristics
  - can often be modified to cover new characteristics.
27. Acknowledgments

- This work was supported in part:
  - by the National Science Foundation grants HRD-0734825 and DUE-0926721, and
  - by Grant 1 T36 GM078000-01 from the National Institutes of Health.

- The work of N. Buntao was supported by a grant from Office of the Higher Education Commission, Thailand.

- The authors are thankful:
  - to Hung T. Nguyen,
  - to Sa-aat Niwitpong,
  - to Tony Wang, and
  - to the anonymous referees for valuable suggestions.