# How to Test Hypotheses When Exact Values are Replaced by Intervals to Protect Privacy: Case of t-Tests

[1]Vladik Kreinovich and [2]Christian Servin

[1]Department of Computer Science
University of Texas at El Paso
El Paso, Texas 79968
vladik@utep.edu

[2]Information Technology Department
El Paso Community College
919 Hunter, El Paso, TX 79915
cservin@gmail.com

## Need for t-Tests

- Biomedical researchers continuously look for possible relations between relevant quantities.
- Such relations may help in preventing and curing diseases.
- Once a hypothesis is made about such a relation, it is necessary to test whether it is confirmed by the data.
- For such hypothesis testing, t-tests are most widely used.
- A t-test can check, whether two samples come from distributions with the same mean.
- Example: checking whether the average blood pressure decreases after a proposed treatment.

## Need to Preserve Privacy

- In traditional statistics, we assume that we know the exact values of the corresponding quantities.
- In biomedical research, however, it is important to preserve patients' privacy and confidentiality.
- Knowing the exact values of age, height, weight, etc., one can uniquely identify the patient.
- One of the most efficient ways to preserve privacy is thus to replace the exact values with intervals containing such values.
- Example: instead of the exact age, we only store an interval containing this age:
  - between 20 and 30, or
  - between 30 and 40, etc.

## Resulting Computational Challenge

- We want to estimate the value of a statistic $s$.
- We know how the statistic depends on the sample values $x_1, \ldots, x_n$.
- For example, for the t-test, we estimate a statistic $t$.
- The hypothesis is confirmed, with given confidence $\alpha$, if this value is below a certain threshold $t_\alpha$: $t \in [0, t_\alpha]$
- Example: the mean is $s = \dfrac{1}{n} \cdot \sum\limits_{i=1}^{n} x_i$.
- For privacy-protected data, instead of the exact values $x_i$, we only know the intervals $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$.
- Different values $x_i \in \mathbf{x}_i$ lead, in general, to different values of the corresponding statistic $s$.
- In particular, for different $x_i \in \mathbf{x}_i$ and $y_i \in \mathbf{y}_i$, we have different values $t(x_1, \ldots, x_n, y_1, \ldots, y_n)$.
- To confirm the hypothesis, we need to check that $t(x_1, \ldots, y_1, \ldots) \le t_\alpha$ for all $x_i \in \mathbf{x}_i$ and $y_i \in \mathbf{y}_i$.
- This is equivalent to $\overline{t} \le t_\alpha$, where
$$\overline{t} \stackrel{\text{def}}{=} \max\{t(x_1, \ldots, y_1, \ldots) : x_i \in \mathbf{x}_i, y_i \in \mathbf{y}_i\}.$$
- To reject the hypothesis, we need to check that $t(x_1, \ldots, y_1, \ldots) > t_\alpha$ for all $x_i \in \mathbf{x}_i$ and $y_i \in \mathbf{y}_i$.
- This is equivalent to $\underline{t} > t_\alpha$, where
$$\underline{t} \stackrel{\text{def}}{=} \min\{t(x_1, \ldots, y_1, \ldots) : x_i \in \mathbf{x}_i, y_i \in \mathbf{y}_i\}.$$
- Thus, we need to compute the range
$$[\underline{t}, \overline{t}] = \{t(x_1, \ldots, y_1, \ldots) : x_i \in \mathbf{x}_i, y_i \in \mathbf{y}_i\}.$$

## Interval Computations

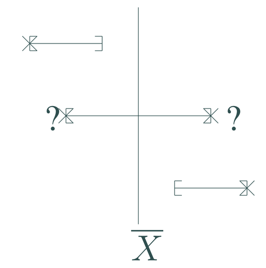
- Computation under interval uncertainty about inputs is known as *interval computations*.
- In general, computing the range is NP-hard.
- This means, crudely speaking, that no feasible algorithm can solve all instances of this problem.
- In some cases, feasible algorithms are possible.
- For example, it is easy to compute the range of the mean $s = \dfrac{1}{n} \cdot \sum\limits_{i=1}^{n} x_i$.
- Since this function in monotonic in all $x_i$, the range is
$$[\underline{s}, \overline{s}] = \left[\frac{1}{n} \cdot \sum_{i=1}^{n} \underline{x}_i, \frac{1}{n} \cdot \sum_{i=1}^{n} \overline{x}_i\right].$$
- We provide efficient algorithms for computing t-tests under privacy-motivated interval uncertainty.

## Versions of t-Test: Reminder

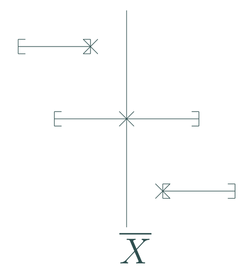
- General statistics: sample mean $\overline{X} = \dfrac{1}{n} \cdot \sum\limits_{i=1}^{n} x_i$ and sample variance $s^2 = \dfrac{1}{n-1} \cdot \sum\limits_{i=1}^{n} (x_i - \overline{x})^2$.
- For testing that the actual mean $\mu$ is $\mu_0$: $t = \dfrac{\overline{X} - \mu_0}{s/\sqrt{n}}$.
- For testing that the means are equal ($\mu_1 = \mu_2$), case of equal sample sizes $n_1 = n_2$ and equal variance: $t = \dfrac{\overline{X}_1 - \overline{X}_2}{s_{X_1X_2} \cdot \sqrt{2/n}}$, where $s_{X_1X_2} = \sqrt{\dfrac{1}{2} \cdot (s^2_{X_1} + s^2_{X_2})}$.
- Case of unequal sample sizes $n_1 \ne n_2$, equal variance:
$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{X_1X_2} \cdot \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}, s_{X_1X_2} \stackrel{\text{def}}{=} \sqrt{\frac{(n_1 - 1)s^2_{X_1} + (n_2 - 1)s^2_{X_2}}{n_1 + n_2 - 2}}.$$
- Case of unequal variance: $t = \dfrac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}}$, where $s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\dfrac{s^2_{X_1}}{n_1} + \dfrac{s^2_{X_2}}{n_2}}$.

## Intuitive Idea

- All expressions for $t$ have the form $\dfrac{\overline{X}}{s}$.
- Thus, the smallest value $\underline{t}$ of $t$ is attained when $s$ is the largest.
- So, for each $i$, we select $x_i$ which are as far from the mean as possible.




- For intervals $[\underline{x}_i, \overline{x}_i]$ containing $\overline{X}$, we have two options: $x_i = \underline{x}_i$ and $x_i = \overline{x}_i$.
- For all other intervals $[\underline{x}_i, \overline{x}_i]$, we have only one option.
- Similarly, the largest value $\overline{t}$ of $t$ is attained when $s$ is the smallest.
- This means that for each $i$, we select $x_i$ which are as close from the mean as possible.




## Towards Algorithm for $\overline{t}$

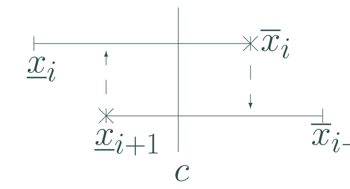
- A function $f(x)$ attains its maximum on $[\underline{x}, \overline{x}]$:
  - either inside the interval, then $\dfrac{df}{dx} = 0$;
  - or for $x_i^M = \underline{x}$, then $\dfrac{df}{dx} \le 0$;
  - or for $x^M = \overline{x}$, then $\dfrac{df}{dx} \ge 0$.
- So, for every $i$, when the maximum $t = \overline{t}$ is attained:
  - either when $\underline{x}_i < x_i^M < \overline{x}_i$ and $\dfrac{\partial t}{\partial x_i} = 0$;
  - or when $x_i^M = \underline{x}_i$ and $\dfrac{\partial t}{\partial x_i} \le 0$;
  - or when $x_i^M = \overline{x}_i$ and $\dfrac{\partial t}{\partial x_i} \ge 0$.
- Here, $\dfrac{\partial t}{\partial x_i} \sim x_i - c$ for some quadratic expression $c$ which is independent on $i$.
- When $\underline{x}_i \le c \le \overline{x}_i$, we cannot have $x_i^M = \underline{x}_i$ and $x_i^M = \overline{x}_i$, so $x_i^M$ is in between, so $\dfrac{\partial t}{\partial x_i} = 0$ and $x_i^M = c$.
- Similarly, when $\overline{x}_i \le c$, we have $x_i^M = \overline{x}_i$.
- When $c \le \underline{x}_i$, we have $x_i^M = \underline{x}_i$.
- In all three cases, $x_i^M$ is the point closest to $c$.
- Let's sort all endpoints of the intervals: $x_{(1)} \le x_{(2)} \le \ldots$
- The value $c$ is in one of the zones $[x_{(k)}, x_{(k+1)}]$.
- For each zone $k$, for each $i$, we either know $x_i^M$, or we know that $x_i^M = c$.
- Substituting $x_i = x_i^M$ into the quadratic expression $c(x_1, \ldots, x_n)$, we get a quadratic equations for $c$. After solving the quadratic equation, we find $c$.
- Based on these values, we compute the value $t$ corresponding to the $k$-th zone.
- We repeat this for each pair of $X_1$- and $X_2$-zone.
- The largest of the computed values $t$ is the desired maximum $\overline{t}$.
- For a sample of size $n$, we have $2n$ bounds, so we have $2n + 1 = O(n)$ zones.
- Thus, we have $O(n) \cdot O(n) = O(n^2)$ pairs of zones.
- For each pair of zone, we need $O(n)$.
- Thus, overall, we need $O(n^2) \cdot O(n) = O(n^3)$ steps.
- So, our algorithm is indeed feasible.

## Towards Algorithm for $\underline{t}$

- A function $f(x)$ attains its minimum on an interval $[\underline{x}, \overline{x}]$:
  - either inside the interval, then $\dfrac{df}{dx} = 0$;
  - or for $x^m = \underline{x}$, then $\dfrac{df}{dx} \ge 0$;
  - or for $x^m = \overline{x}$, then $\dfrac{df}{dx} \le 0$.
- So, for every $i$, when the minimum $t = \overline{t}$ is attained:
  - either when $\underline{x}_i < x_i^m < \overline{x}_i$ and $\dfrac{\partial t}{\partial x_i} = 0$;
  - or when $x_i^m = \underline{x}_i$ and $\dfrac{\partial t}{\partial x_i} \ge 0$;
  - or when $x_i^m = \overline{x}_i$ and $\dfrac{\partial t}{\partial x_i} \le 0$.
- Here, $\dfrac{\partial t}{\partial x_i} \sim x_i - c$ for some quadratic expression $c$ which is independent on $i$.
- When $c < \underline{x}_i$, we cannot have $x_i^m = \underline{x}_i$ and $\underline{x}_i < x_i^m < \overline{x}_i$, so $x_i^m = \overline{x}_i$.
- Similarly, when $\overline{x}_i < c$, we have $x_i^m = \underline{x}_i$.
- When $\underline{x}_i \le c \le \overline{x}_i$, we can have both $x_i^m = \underline{x}_i$ and $x_i^m = \overline{x}_i$.

## Towards Algorithm for $\underline{t}$ (cont-d)

- For privacy data, intervals $[\underline{x}_i, \overline{x}_i]$ can be sorted so that $\underline{x}_i \le \underline{x}_{i+1}$ and $\overline{x}_i \le \overline{x}_{i+1}$.
- Let us show that min is attained when $x_i^m \le x_i^{m+1}$.
- Indeed, the only possibility for $x_i^m \le x_i^{m+1}$ is when both intervals contain $c$, $x_i^m = \overline{x}_i$, and $x_{i+1}^m = \underline{x}_{i+1}$.
- In this case, since $t$ is symmetric w.r.t. all $x_i$ we can swap these values and take $x_i^m = \underline{x}_{i+1}$, and $x_{i+1}^m = \overline{x}_i$.




- We see that the resulting tuple is not minimizing.
- Thus, there exists $k$ for which the minimizing sequence $x_i^m$ has the form
$$(\underline{x}_1, \ldots, \underline{x}_k, \overline{x}_{k+1}, \ldots, \overline{x}_n).$$
- We have such thresholds $k_1$ and $k_2$ for both samples.
- There are $n^2$ pairs of such thresholds.
- For each pair, we know the values $x_i$ and thus, we can compute $t$ by using time $O(n)$.
- The smallest of these values $t$ is the desired value $\underline{t}$.

## This Algorithm Is Feasible and Can Be Further Improved

- The algorithm takes time $O(n^2) \cdot O(n) = O(n^3)$ and is, thus, feasible.
- When we change from $k$ to $k + 1$, only one value changes $x_{k+1}^m$, from $\underline{x}_{k+1}$ to $\overline{x}_{k+1}$.
- Thus, we can change $\overline{X}_i$ and $S_{X_i}$ is $O(1)$ steps.
- With this improvement, we can compute $\underline{t}$ is time $O(n^2)$.

## References

- S. Ferson, V. Kreinovich, J. Hajagos, W. Oberkampf, and L. Ginzburg, *Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty*, Sandia National Laboratories, Report SAND2007-0939, May 2007.
- H. T. Nguyen, V. Kreinovich, B. Wu, and G. Xiang, *Computing Statistics under Interval and Fuzzy Uncertainty*, Springer Verlag, Berlin, Heidelberg, 2012.
- C. Servin and V. Kreinovich, *Propagation of Interval and Probabilistic Uncertainty in Cyberinfrastructure-Related Data Processing and Data Fusion*, Springer Verlag, Berlin, Heidelberg, to appear.

## Acknowledgments

This work was supported in part by the National Science Foundation grants:
- HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence), and
- DUE-0926721.