

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

Vladik Kreinovich¹ and Christian Servin²

¹Department of Computer Science

University of Texas at El Paso

El Paso, Texas 79968

vladik@utep.edu

²Information 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, \dots, x_n .
- For example, for the t-test, we estimate a statistic t .
- The hypothesis is confirmed, with given confidence α , if this value is below a certain threshold t_α : $t \in [0, t_\alpha]$
- Example: the mean is $s = \frac{1}{n} \cdot \sum_{i=1}^n x_i$.
- For privacy-protected data, instead of the exact values x_i , we only know the intervals $\mathbf{x}_i = [\underline{x}_i, \bar{x}_i]$.
- 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 $t(x_1, \dots, x_n, y_1, \dots, y_n)$.

Computational Challenge (cont-d)

- To confirm the hypothesis, we need to check that $t(x_1, \dots, y_1, \dots) \leq t_\alpha$ for all $x_i \in \mathbf{x}_i$ and $y_i \in \mathbf{y}_i$.

- This is equivalent to $\bar{t} \leq t_\alpha$, where

$$\bar{t} \stackrel{\text{def}}{=} \max\{t(x_1, \dots, y_1, \dots) : x_i \in \mathbf{x}_i, y_i \in \mathbf{y}_i\}.$$

- To reject the hypothesis, we need to check that

$$t(x_1, \dots, y_1, \dots) > t_\alpha \text{ for all } x_i \in \mathbf{x}_i \text{ and } y_i \in \mathbf{y}_i.$$

- This is equivalent to $\underline{t} > t_\alpha$, where

$$\underline{t} \stackrel{\text{def}}{=} \min\{t(x_1, \dots, y_1, \dots) : x_i \in \mathbf{x}_i, y_i \in \mathbf{y}_i\}.$$

- Thus, we need to compute the range

$$[\underline{t}, \bar{t}] = \{t(x_1, \dots, y_1, \dots) : x_i \in \mathbf{x}_i, y_i \in \mathbf{y}_i\}.$$

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 $s = \frac{1}{n} \cdot \sum_{i=1}^n x_i$.
- Since this function is monotonic in all x_i , the range is

$$[\underline{s}, \bar{s}] = \left[\frac{1}{n} \cdot \sum_{i=1}^n \underline{x}_i, \frac{1}{n} \cdot \sum_{i=1}^n \bar{x}_i \right].$$

- We provide efficient algorithms for computing t-tests under privacy-motivated interval uncertainty.

Versions of t-Test: Reminder

- General statistics: sample mean $\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^n x_i$ and sample variance $s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2$.

- For testing that the actual mean μ is μ_0 : $t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$.

- For testing that the means are equal ($\mu_1 = \mu_2$), case of equal sample sizes $n_1 = n_2$ and equal variance: $t = \frac{\bar{X}_1 - \bar{X}_2}{s_{X_1 X_2} \cdot \sqrt{2/n}}$, where $s_{X_1 X_2} = \sqrt{\frac{1}{2} \cdot (s_{X_1}^2 + s_{X_2}^2)}$.

- Case of unequal sample sizes $n_1 \neq n_2$, equal variance:

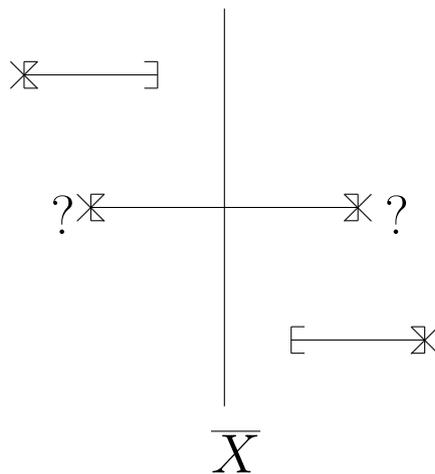
$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{X_1 X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, s_{X_1 X_2} \stackrel{\text{def}}{=} \sqrt{\frac{(n_1 - 1)s_{X_1}^2 + (n_2 - 1)s_{X_2}^2}{n_1 + n_2 - 2}}$$

- Case of unequal variance: $t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}$, where

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_{X_1}^2}{n_1} + \frac{s_{X_2}^2}{n_2}}$$

Intuitive Idea

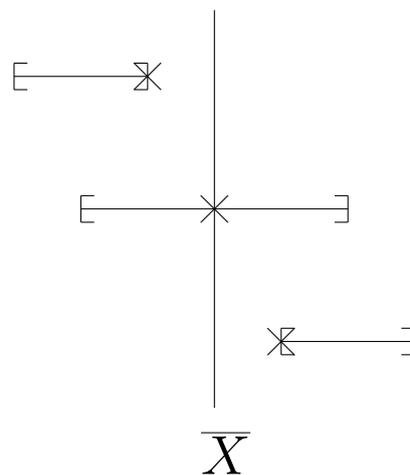
- All expressions for t have the form $\frac{\bar{X}}{s}$.
- Thus, the smallest value \underline{t} of t is attained when s is the largest.
- This means that for each i , we select x_i which are as far from the mean as possible.



- For intervals $[\underline{x}_i, \bar{x}_i]$ containing \bar{X} , we have two options: $x_i = \underline{x}_i$ and $x_i = \bar{x}_i$.
- For all other intervals $[\underline{x}_i, \bar{x}_i]$, we have only one option.

Idea (cont-d)

- Similarly, the largest value \bar{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.



- For each interval $[\underline{x}_i, \bar{x}_i]$, we have only one option:
 - if $\bar{x}_i < \bar{X}$, then $x_i = \bar{x}_i$;
 - if $\bar{X} < \underline{x}_i$, then $x_i = \underline{x}_i$;
 - if $\underline{x}_i \leq \bar{X} \leq \bar{x}_i$, then $x_i = \bar{X}$.

Towards Algorithm for \bar{t}

- A function $f(x)$ attains its maximum on $[\underline{x}, \bar{x}]$:
 - either inside the interval, then $\frac{df}{dx} = 0$;
 - or for $x_i^M = \underline{x}$, then $\frac{df}{dx} \leq 0$;
 - or for $x^M = \bar{x}$, then $\frac{df}{dx} \geq 0$.

- So, for every i , when the maximum $t = \bar{t}$ is attained:
 - either when $\underline{x}_i < x_i^M < \bar{x}_i$ and $\frac{\partial t}{\partial x_i} = 0$;
 - or when $x_i^M = \underline{x}_i$ and $\frac{\partial t}{\partial x_i} \leq 0$;
 - or when $x_i^M = \bar{x}_i$ and $\frac{\partial t}{\partial x_i} \geq 0$.

- Here, $\frac{\partial t}{\partial x_i} \sim x_i - c$ for some quadratic expression c which is independent on i .

Towards Algorithm for \bar{t} (cont-d)

- When $\underline{x}_i \leq c \leq \bar{x}_i$, we cannot have $x_i^M = \underline{x}_i$ and $x_i^M = \bar{x}_i$, so x_i^M is in between, so $\frac{\partial t}{\partial x_i} = 0$ and $x_i^M = c$.
- Similarly, when $\bar{x}_i \leq c$, we have $x_i^M = \bar{x}_i$.
- When $c \leq \underline{x}_i$, 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:

$$x_{(1)} \leq x_{(2)} \leq \dots$$

- 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 $c(x_1, \dots, x_n)$, we get a quadratic equations for c .

Algorithm for \bar{t} (cont-d)

- After solving the quadratic equation, we find c .
- Thus, we know all the values x_i .
- 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 \bar{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.

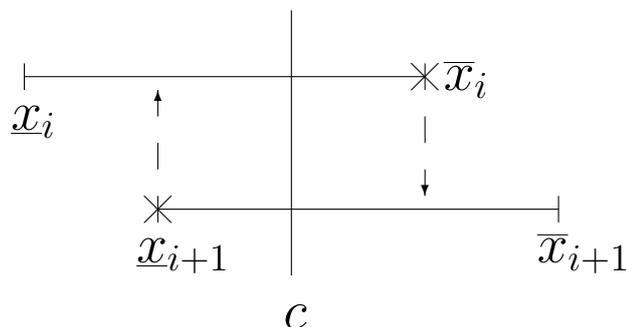
Towards Algorithm for \underline{t}

- A function $f(x)$ attains its minimum on an interval $[\underline{x}, \bar{x}]$:
 - either inside the interval, then $\frac{df}{dx} = 0$;
 - or for $x^m = \underline{x}$, then $\frac{df}{dx} \geq 0$;
 - or for $x^m = \bar{x}$, then $\frac{df}{dx} \leq 0$.
- So, for every i , when the minimum $t = \bar{t}$ is attained:
 - either when $\underline{x}_i < x_i^m < \bar{x}_i$ and $\frac{\partial t}{\partial x_i} = 0$;
 - or when $x_i^m = \underline{x}_i$ and $\frac{\partial t}{\partial x_i} \geq 0$;
 - or when $x_i^m = \bar{x}_i$ and $\frac{\partial t}{\partial x_i} \leq 0$.
- Here, $\frac{\partial t}{\partial x_i} \sim x_i - c$ for some quadratic expression c which is independent on i .

Towards Algorithm for t (cont-d)

- When $c < \underline{x}_i$, we cannot have $x_i^m = \underline{x}_i$ and $\underline{x}_i < x_i^m < \bar{x}_i$, so $x_i^m = \bar{x}_i$.
- Similarly, when $\bar{x}_i < c$, we have $x_i^m = \underline{x}_i$.
- When $\underline{x}_i \leq c \leq \bar{x}_i$, we can have both $x_i^m = \underline{x}_i$ and $x_i^m = \bar{x}_i$.
- For privacy data, intervals $[\underline{x}_i, \bar{x}_i]$ can be sorted so that $\underline{x}_i \leq \underline{x}_{i+1}$ and $\bar{x}_i \leq \bar{x}_{i+1}$.
- Let us show that min is attained when $x_i^m \leq x_i^{m+1}$.
- Indeed, the only possibility for $x_i^m \leq x_i^{m+1}$ is when both intervals contain c , $x_i^m = \bar{x}_i$, 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 $x_{i+1}^m = \bar{x}_i$.

Swap and Resulting Algorithm



- We see that the resulting tuple is not minimizing.
- Thus, there exists k for which the minimizing sequence x_i^m has the form

$$(\underline{x}_1, \dots, \underline{x}_k, \bar{x}_{k+1}, \dots, \bar{x}_n).$$

- 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.
- How can we make computations faster?
- When we change from k to $k + 1$, only one value changes x_{k+1}^m , from \underline{x}_{k+1} to \bar{x}_{k+1} .
- Thus, we can change \bar{X}_i and S_{X_i} is $O(1)$ steps.
- With this improvement, we can compute \underline{t} in 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.