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

Vladik Kreinovich and Christian Servin

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 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 20 and 40, etc.

Resulting Computational Challenge

• We want to estimate the value of a statistic.
• 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 \( t_\alpha \).
• In particular, for different \( x_i, x_j \in \mathbb{R} \), one can uniquely identify the patient.
• We have such thresholds \( t_\alpha \) for each sample.
• One cannot compute the range for each \( x_i, x_j \) and is, thus, feasible.

Intuitive Idea

• All expressions for \( t \) have the form \( \frac{x_i - x_j}{s} \).
• The smallest value \( t \) is attained when \( x_i - x_j \) is the largest.
• So, for each \( i \), we select \( x_i \) which are as close from the mean as possible.

Towards Algorithm for t

A function \( f(x) \) attains its maximum on \([a, b] \) if:
• at every interval of the range \( [a, b] \), the function has a maximum.
• At the boundaries of the interval, the function value is equal to the boundary value.

Towards Algorithm for \( t \) (cont-d)

• For privacy data, intervals \([L, U] \) can be sorted so that \( L \leq L' \) and \( U \geq U' \).
• Let us show that \( t \) is in attained when \( x_i \leq x_{i+1} \).
• Indeed, the only possibility for \( x_i \leq x_{i+1} \) is when both intervals contain \( x_i \) and \( x_{i+1} \) which is symmetric w.r.t. all \( x_i \).

This Algorithm Is Feasible and Can Be Further Improved

• The algorithm takes time \( O(n^2) \) or \( O(n^3) \) and is, thus, feasible.
• When we change from \( k \to k+1 \), only one value changes \( x_i \) of \( x_i \leq x_{i+1} \) to \( x_{i+1} \).
• Thus, we can change \( t' \) and \( t'' \) in \( O(1) \) steps.

References


Towards Algorithm for \( t \)

• For every \( i \), when the minimum \( t \) is attained:
• either \( x_i \leq x_{i+1} \) and \( x_{i+1} \leq x_{i+2} \), then \( t = 0 \).
• or \( x_i \leq x_{i+1} \) and \( t_{i+1} \leq t_{i+2} \), then \( t = 0 \).
• So, for every \( i \), when the minimum \( t \) is attained:
• either \( x_i \leq x_{i+1} \) and \( x_{i+1} \leq x_{i+2} \), then \( t = 0 \).
• or \( x_i \leq x_{i+1} \) and \( t_{i+1} \leq t_{i+2} \), then \( t = 0 \).

Towards Algorithm for \( t \) (cont-d)

• For privacy data, intervals \([L, U] \) can be sorted so that \( L \leq L' \) and \( U \geq U' \).
• Let us show that \( t \) is attained when \( x_i \leq x_{i+1} \).
• Indeed, the only possibility for \( x_i \leq x_{i+1} \) is when both intervals contain \( x_i \) and \( x_{i+1} \) which is symmetric w.r.t. all \( x_i \).

In this case, since \( t \) is symmetric w.r.t. all \( x_i \), we can swap these values and take \( x_i \leq x_{i+1} \) and \( x_{i+1} \leq x_i \).

This work was supported in part by the National Science Foundation grants:
• HRD-0734823 and HRD-1234212
• (CyberShARE Center of Excellence), and
• DUE-0967281.