Unpacking "comparisons"

Q1. A researcher implements a new sorting algorithm and runs it on 30 random inputs. The average runtime is 142ms. The theoretical model predicts the algorithm should run in 150ms on average. Is the measured runtime *different* from the theoretical prediction?

Q2. Three sorting algorithms (QuickSort, MergeSort, HeapSort) are each run on 30 *different* randomly generated inputs. Do the algorithms differ in runtime?

Q3. Two groups of users were recruited to test two user interfaces (UI) (A vs. B). Each group tested one UI and rated its usability on a scale of 1–5 (Terrible / Bad / Neutral / Good / Excellent). Is the usability rating of the design A *higher* than that of the design B?

Q4. A company is developing a new version of the software and has asked a group of users to rate the current version on a 1–5 scale. Then, the same users tested the new version and rated it. Did the new version's rating improve?