Q4

Upon first sight we can see that the percentage success rates for each variation are as follows:

Baseline: 5.4%
Variation 1: 5.0%
Variation 2: 2.9%
Variation 3: 8.4%
Variation 4: 6.6%

The problem in comparing these values is of course, the varying sample sizes of each case. In that case, I would consider using a Student's t-test to determine if there are differences in the average number of quotes made. This is because it will consider the standard error of the estimates of the means for each group. My t-test will tell me whether two variations are statistically different from each other or not, and I will assume a p-value greater than 5% will confirm whether this difference is statistically significant.

I think it's incredibly important to consider the external factors that may be contributing to the variation in quotes made. It is necessary to assume that the experiment was constructed such that the forms were available at the exact same times during the week, via the same user flows. Otherwise, if for instance, variation 1 was available on Monday morning and not Friday night, the difference in traffic may cause this spike in quotes made. To avoid these biases, I would ask whether the forms were available throughout the week for the same periods.

Furthermore, a question I would want to ask is whether securing the most number of quotes is the aim. Perhaps a filtered set of quality quotes that have a high probability of being accepted is more desirable than a flood of useless quotes. In that case, when analysing the above data set we need understand where the value lies. To understand this, perhaps we need consider the acceptance rate of quotes sent via each form. For instance, what proportion of submitted quotes, sent in via variation 2 throughout the week, were accepted by the consumer?

In conclusion, there are methods to compare the five data sets despite the variation in sample size. However, before making any conclusions we need understand exactly how the data was collected over the week for each variation i.e. were they provided the same exposure throughout the week? It is necessary to understand the goals of this study better, to question whether the value lies in the greatest number of quotes made. Perhaps we need better appreciate that successful matches between a consumer and service provider is best for the company. Hence, less quotes with a high probability of being accepted may provide more value and the experiment should be altered to account for this.