Privacy

1. **How can potentially identifying attributes be chosen to achieve k-anonymity?**
   - To achieve k-anonymity, one must identify and modify the quasi-identifiers, which are attributes that could potentially be linked with external information to re-identify individuals within a dataset. This typically involves generalizing these attributes to a less specific level (e.g., changing a precise birthdate to just a year) or suppressing them (removing the attribute from the dataset). The goal is to ensure that in the modified dataset, each combination of quasi-identifiers appears at least k times, making it ambiguous as to which individual corresponds to which record.
2. **How should the parameters of the two ideas, ε-DP and k-anonymity be chosen?**
   - The parameter ε in ε-differential privacy (ε-DP) controls the trade-off between privacy and utility: a smaller ε provides stronger privacy guarantees but at the cost of potentially less accurate or useful data. The choice of ε is context-dependent; it requires careful consideration of the privacy risks and the utility needs of the dataset.
   - The parameter k in k-anonymity dictates the size of the anonymity set: a larger k value increases privacy by making each person indistinguishable from at least $k - 1$ others. However, higher values of k can overly generalize the data, which may reduce its utility for certain analyses.
3. **Does having more data available make it easier to achieve privacy?**
   - Having more data can, in some instances, make it easier to achieve privacy. With a larger dataset, it's more likely that there are enough similar records to create anonymity sets that satisfy the k-anonymity criterion. However, this also depends on the diversity and uniqueness of the records; a larger dataset composed of very unique records may still struggle with achieving high levels of k-anonymity. For differential privacy, a larger dataset can dilute the relative impact of noise added to the data, maintaining utility while satisfying ε-DP. Nevertheless, the ease of achieving privacy also hinges on the nature of the data, such as its variability and sensitivity, and the intended use of the data.

Reproducibility

1. **What is a good use case for cross-validation over hold-out sets?**
   Cross-validation is especially useful when you have a limited amount of data and want to maximize its use for training while also ensuring that you have a robust estimate of your model's performance. Unlike a single hold-out set, which might not be representative of the overall population, cross-validation allows the model to be trained and validated on multiple splits of the data. This technique is good for reducing the variance of the model performance estimate, making it a preferred method over a hold-out set in many cases.
2. **When is it a good idea to use bootstrapping?**
   Bootstrapping is a good idea when you need to estimate the sampling distribution of a statistic (such as mean, median, or variance) by resampling with replacement from the data set, especially when the theoretical distribution of that statistic is unknown or when the sample size is too small for reliable asymptotic approximations. It's also useful for constructing confidence intervals and for hypothesis testing when the standard methods assume a normal distribution or other conditions that do not hold for the data.
3. **How can we use the above techniques to avoid the false discovery problem?**
   Cross-validation and bootstrapping can help avoid the false discovery problem by providing a more accurate assessment of a model's ability to generalize to new data. By repeatedly assessing model performance on different subsets of the data (cross-validation) or by resampling the data (bootstrapping), we can reduce the risk of overfitting to the idiosyncrasies of a single dataset, which is often a source of false discoveries. Additionally, when performing multiple hypothesis tests, these techniques can be used in conjunction with methods like the Bonferroni correction to adjust for multiple comparisons and further control for false discovery rate.
4. **Can these techniques fully replace independent replication?**
   While cross-validation and bootstrapping can provide a strong indication of a model's reliability and the stability of scientific findings, they do not fully replace the need for independent replication. Independent replication involves collecting new data under the same or different conditions to confirm the findings. This is crucial for establishing the external validity of a study because it ensures that the results are not due to the peculiarities of a particular dataset or the specific conditions under which the original experiment was conducted. These statistical techniques are tools for internal validation and help to ensure that the model or statistical method is not overfitting the data it was trained on, but they cannot address all the issues that independent replication can.