Bits & Bytes: Bias in Data Analytics

|  |  |  |
| --- | --- | --- |
| When I wrote about the analysis of “Task Planner” data last week, many of you advised me about biases in the analysis.  Thank you for sharing all sorts of historical stories!  I learnt a lot of facts that were not shown in the data and I really appreciate that I have had an opportunity to learn from that indirect experience.  When we analyse data, there is always a great chance of bias involvement, and I would like to talk about that. | | |
| BIAS is defined as “the [action](https://dictionary.cambridge.org/dictionary/english/action) of [supporting](https://dictionary.cambridge.org/dictionary/english/supporting) or [opposing](https://dictionary.cambridge.org/dictionary/english/oppose) a [particular](https://dictionary.cambridge.org/dictionary/english/particular) [person](https://dictionary.cambridge.org/dictionary/english/person) or thing in an [unfair](https://dictionary.cambridge.org/dictionary/english/unfair) way, because of [allowing](https://dictionary.cambridge.org/dictionary/english/allow) [personal](https://dictionary.cambridge.org/dictionary/english/personal) [opinions](https://dictionary.cambridge.org/dictionary/english/opinion) to [influence](https://dictionary.cambridge.org/dictionary/english/influence) [your](https://dictionary.cambridge.org/dictionary/english/your) [judgment](https://dictionary.cambridge.org/dictionary/english/judgment)” in Cambridge Dictionary.  I assume that you already understand the meaning very well and use it correctly.  However, how much do you know about the mathematical meaning of BIAS.  BIAS refers to “the tendency of a measurement process to over- or under-estimate the value of a population [parameter](https://stattrek.com/Help/Glossary.aspx?Target=Parameter)” in Statistics.  This definition has been extended to data analytics too.  To understand the definition, we need to learn of the “Population Parameter”.  **Population vs. Sample**  A population is the whole of a group, and a sample is a fraction of the population. The first figure represents a population of fish in a pond, and there are three different samples that follow. As shown in the figure, a number of samples can be collected from a population. | | |
|  | | |
| A. | B. | C. |
| Now, we consider the relationship between the population and samples in terms of characteristics of data.  For example, we are concerned about fish types and average size of fish in the pond above.  Which do you think to be the most representative sample of those three samples?  I believe that all of you will pick the right one and say “B”, without any hesitation.  **Parameter vs. Statistic**  Actual fish types and average size of the population in the pond are called parameters. On the other hand, fish types and average size of each sample are called statistics.  From the parameter of population and the statistics of sample, we can make a simple formula of bias:  **Bias = parameter - Statistic**  In real life, we hardly have a chance to collect data for the population for a number of reasons. Typically, because of time, budget and labour. The census is one of the examples of population data collection. In our business, we do have quite a large number of population data collections. Such as sales and customer’s information.  BIAS can be introduced at any stage of data analysis, although population data are less susceptible to be exposed to bias compared to samples.  When I read some literature about BIAS I was really surprised to learn that there are at least 30 different bias types. Most names have faded from my brain but I still remember some of them.  **Selection Bias**  When we run a survey via a mobile phone application we probably miss a portion of elderly customers due to technological difficulties. As a result, we cannot collect a representative sample that we intend.  **Confirmation Bias**  Assume that we make the options for our ticket quality survey -Good, Very good and Excellent. We have predetermined that our service result will be Great in the survey.  **Outliers**  The mean of the set of (1, 2, 3, 4, 5) is 3. When we put an extra value in the set, say 15, the number of elements of the new set increases to 6, and the mean jumps to 5. If the actual value is 1 (not 15), the new false value twists the actual value. In this example, an outlier affects the statistic greatly and can betray the actual figure.  **Interpretation Bias**  Interpretation bias is the tendency to read ambiguous results in a positive or negative fashion, especially when the outcome of data analytics is presented. When we have half of water in a glass, some people say “Still have a half” and others “Only have half”.  **Availability Bias**  Sometimes we do not have sufficient data to analyse, but the situation forces us to analyse them. This can result wrong outcomes and lead to a failure to come to a correct decision. | | |
| Though an analytics job is not for everyone, all of us are exposed a certain amount of analytical outcomes in daily life. Understanding the possible bias and having a critical eye on the figures can help us to find the true stories.  When we consider bias, there is a combo-term with BIAS – it is called VARIANCE.  Both terms are used for judging how the given data are good or bad in a different sense.  Recall that the bias is the difference between a parameter of a population and a statistic of a sample.  VARIANCE is the aggregated differences of values in a sample. So it represents how the data are scattered.  With both measures larger is worse. | | |
| There are four different cases in the plot below:   * Case A: Large Bias and Small Variance – data fail to represent to the actual value but are consistent. * Case B: Large Bias and Large Variance – data fail to represent the actual value and inconsistent. * Case C: Small Bias and Small Variance – data represent the actual value well and are consistent. * Case D: Small Bias and Large Variance – data represent the actual value well but are inconsistent. | | |
|  | | |
| I hope we are all aware of the Bias and Variance when we collect, analyse and interpret data so that we make the data and their analytical results more valuable. | | |