Statistics: A look at what they are and how they can be used by the Teradata Database

There are lies, damned lies and … statistics. Statistics are “the science that deals with the collection, classification, analysis, and interpretation of numerical facts or data, and that, by use of mathematical theories of probability, imposes order and regularity on aggregates of more or less disparate elements.” At least that is according to the definition provided by the Merriam-Webster Dictionary. Many other more philosophical definitions have been provided for statistics over time, as well as many observations regarding the value of statistics.

*Statistics are the triumph of the quantitative method, and the quantitative method is the victory of sterility and death.*
- Hilaire Belloc

*Statistics are no substitute for judgment.*
- Henry Clay

*Our scientific age demands that we provide definitions, measurements, and statistics in order to be taken seriously. Yet most of the important things in life cannot be precisely defined or measured. Can we define or measure love, beauty, friendship, or decency, for example?*
- Dennis Prager

*Definition of Statistics: The science of producing unreliable facts from reliable figures.*
- Evan Esar

*There are two ways of lying. One, not telling the truth and the other, making up statistics.*
- Josefina Vazquez Mota

*Statistics: the mathematical theory of ignorance.*
- Morris Kline

*Statistics is the grammar of science.*
- Karl Pearson

The opening quote of this article, “There are lies, damned lies and … statistics”, is usually attributed to Mark Twain. This did not originate with Twain, however, and appears to be connected to Twain because he included it in one of his autobiographical chapters published in the North American Review. In this publication Twain mistakenly attributed the quote’s origination to the British statesman, Benjamin Disraeli. However, Disraeli is not the source of the quote, either, and it’s origins are dubious and disputable. Apparently the only connection Mark Twain has with the quotation is the following passage in Mark Twain’s Autobiography, Volume I, Berkeley, Los Angeles and London: University of California Press 2010, p. 228 (this particular passage was dictated in Florence in 1904):

*I was deducing from the above that I have been slowing down steadily in these thirty-six years, but I perceive that my statistics have a defect: three thousand words in the spring of 1868, when I was working seven or eight or nine hours at a sitting, has little or no advantage over the sitting...*
of to-day, covering half the time and producing half the output. Figures often beguile me, particularly when I have the arranging of them myself; in which case the remark attributed to Disraeli would often apply with justice and force: “There are three kinds of lies: lies, damned lies, and statistics.”

The manipulating of statistics has become a fact of modern existence. Indeed, especially in an election year, as we are now have in the United States, the opening statement of this article has a corollary:

There are Five kinds of lies:
Lies
Damned Lies,
Statistics,
Politicians quoting statistics
and Novelists quoting Politicians on Statistics.
- Stephen K. Tagg

The Scientific Application of Statistics
Before we pronounce all statistics to be evil, we should revisit to the definition of our impartial friend, the dictionary. “The science that deals with the collection, classification, analysis, and interpretation of numerical facts or data, and that, by use of mathematical theories of probability, imposes order and regularity on aggregates of more or less disparate elements.”

So, there is also a science associated with statistics, and they are not just fodder for the spin doctors. Certainly, one of the reasons statistics get a bad rap is that we think that statistics provide a measure of proof that something is true, but they actually do no such thing. Instead, statistics provide a measure of the probability of observing a certain result.

One account of the origination of the field of statistics dates back to 1654 when Antoine Gombaud, a French writer who also liked to gamble, asked his buddy, the noted mathematician, physicist, and philosopher Blaise Pascal, about how one should divide the stakes among players when a game of chance is interrupted prematurely. Pascal then posed the question to his buddy, the lawyer and mathematician Pierre de Fermat. In correspondence through a series of letters they wrote to each other, Pascal and Fermat devised a mathematical system that not only answered Gombaud’s original question about gambling stakes, but laid the foundations of modern probability theory and statistics.

From its roots in gambling, statistics has grown into a field of study that relies on the development of methods and tests that can be utilized to quantitatively define the variability inherent in data, the probability of certain outcomes, and the error and uncertainty associated with those outcomes. Statistical methods are used extensively throughout a variety of scientific processes, from the design of research questions through data analysis and to the final interpretation of data. The specific statistical methods used vary widely between different scientific disciplines. But, the reasons that these tests and techniques are used are similar across disciplines.
Often, statements of likelihood and probability are misinterpreted as a sign of a weak argument or indefinite scientific results. However, the use of statistical methods and probability tests in research is a vital aspect of science that adds credibility and certainty to scientific conclusions.

Statistics and the Teradata Database

If you were to survey Teradata experts on the most important factors in obtaining Teradata Database performance, they would undoubtedly have good primary index selection and timely and appropriate statistics collection at, or near, the top of their lists. Over the last two decades, Teradata software releases have consistently provided improvements and enhancements in the way statistics are collected, and then utilized by the cost-based Teradata Optimizer. The Optimizer doesn’t perform a detailed evaluation of every possible query plan (multiple joins could produce billions of possibilities). Instead, it uses sophisticated algorithms to identify and select the most promising candidates for detailed evaluation, then picks what it perceives as the best plan among those. The essential task of the optimizer is to produce the optimal execution plan (the one with the lowest cost) from many possible plans. The basis on which different plans are compared with each other is the cost which is derived from the estimation of cardinalities of the temporary or intermediate relations, after an operation such as selections, joins and projections. The estimations in Teradata are derived primarily from statistics and random AMP samples. Accurate estimations are crucial to get optimal plans. Providing statistical information for performance optimization is critical to optimal query plans, but collecting statistics can prove difficult due to the demands of time and system resources. Without full or all-AMP sampled statistics, query optimization must rely on extrapolation and dynamic AMP sample estimates of table cardinality, which does not collect all of the statistics that a COLLECT STATISTICS request does. Besides estimated cardinalities, dynamic AMP samples also collect a few other statistics, but far fewer than are collected by a COLLECT STATISTICS request. Statistics and demographics provide the Optimizer with information it uses to reformulate queries in ways that permit it to produce the least costly access and join plans. The critical issues you must evaluate when deciding whether to collect statistics are not whether query optimization can or cannot occur in the face of inaccurate statistics, but the following pair of probing questions.

• How accurate must the available statistics be in order to generate the best possible query plan?
• How poor a query plan are you willing to accept?

Different strategies can be used to attain the right balance between the need for statistics and the demands of time and resources. The main strategies for collecting statistics are:

Random AMP Sampling.

The optimizer builds an execution plan for each SQL statement that enters the parsing engine. When no statistics have been collected, the system default is for the optimizer to make a rough estimate of a table’s demographics by using dynamic samples from one or more AMPs (one being the default). These samples are collected automatically each time the table header is accessed from disk and are embedded in the table header when it is placed in the dictionary cache.

By default, the optimizer does the single AMP sampling to produce random AMP sample demographics with some exceptions (volatile, sparse single table join indexes and aggregate join
indexes). By changing an internal field in the dbscontrol record called RandomAMPSampling, it can be requested that sampling be performed on 2 AMPs, 5 AMPs, all AMPs on a node, or all AMPs in the system. When using these options, random sampling uses the same techniques as single-AMP random AMP sampling, but more AMPs participate. Touching more AMPs may improve the quality of the statistical information available during plan generation, particularly if rows are not evenly distributed.

In Teradata Database 12.0 and higher releases, all-AMP sampling was enhanced to use an efficient technique using “Last done Channel mechanism” which considerably reduces the messaging overhead. This is used when all-AMP sampling is enabled in the dbscontrol or cost profile but dbscontrol internal flag RowsSampling5is set to 0 (which is the default). If set to greater than 0, this flag causes the sampling logic to read the specified percentage of rows to determine the number of distinct values for primary index.

Pros and cons of Random AMP sampling:

Pros:

- Provides row count information of all indexes including the Primary Index.
- The row count of Primary Index is the total table rows.
- The row count of NUSI subtable is the number of distinct values of the NUSI columns.
- The estimated number of distinct values is used for single-table equality predicates, join cardinality, aggregate estimations, costing, etc.
- Can potentially eliminate the need to collect statistics on the indexes.
- Up-To-Date information – usually most fresh
- This operation is automatically performed

Cons:

- Works only with indexed columns.
- The single-AMP sampling may not be good enough for small tables and tables with non-uniform distribution on the primary index.
- Does not provide the following information.
  - Number of nulls.
  - Skew Info.
  - Value Range.
- For NUSIs, the estimated number of distinct values on a single-AMP is assumed to be the total distinct values. This is true for highly non-unique columns but can cause distinct value underestimation for fairly unique columns. On the other hand, it can cause overestimation for highly nonunique columns because of rowid spill over.
- Can not estimate the number of distinct values for non-unique primary indexes.
- Single table estimations can use this information only for equality conditions assuming uniform distribution.

It is strongly recommended to contact Teradata Global Support Center (GSC) to assess the impact of enabling all-AMP sampling on your configuration and to help change the internal dbscontrol settings.

Full Statistics Collection.

Generically defined, a histogram is a count of the number of occurrences, or cardinality, of a
particular category of data that fall into defined disjunct value range categories. These categories are typically referred to as bins or buckets.

Issuing a COLLECT STATISTICS statement is the most complete method of gathering demographic information about a column or an index. Teradata Database uses equal-height, high-biased, and history interval histograms (a representation of a frequency distribution) to represent the cardinalities and other statistical values and demographics of columns and indexes for all-AMPs sampled statistics and for full-table statistics.

The greater the number of intervals in a histogram, the more accurately it can describe the distribution of data by characterizing a smaller percentage of its composition per each interval. Each interval histogram in the system is composed of a number of intervals (the default is 250 and the maximum is 500) intervals. A 500 interval histogram permits each interval to characterize roughly 0.25% of the data.

Because these statistics are kept in a persistent state, it is up to the administrator to keep collected statistics fresh. It is common for many Teradata Warehouse sites to re-collect statistics on the majority of their tables weekly, and on particularly volatile tables daily, if deemed necessary.

**Collection with the USING SAMPLE option**

Collecting full statistics involves scanning the base table and performing a sort, sometimes a sort on a large volume of data, to compute the number of occurrences for each distinct value. The time and resources required to adequately collect statistics and keep them fresh can be problematic, particularly with large data volumes.

Collecting statistics on a sample of the data reduces the resources required and the time to perform statistics collection. However, the USING SAMPLE alternative was certainly not designed to replace full statistics collection. It requires some careful analysis and planning to determine under which conditions it will add benefit.

The quality of the statistics collected with full-table sampling is not guaranteed to be as good as the quality of statistics collected on an entire table without sampling. Do not think of sampled statistics as an alternative to collecting full-table statistics, but as an alternative to never, or rarely, collecting statistics.

When you use sampled statistics rather than full-table statistics, you are trading time in exchange for what are likely to be less accurate statistics. The underlying premise for using sampled statistics is usually that sampled statistics are better than no statistics.

Do not confuse statistical sampling with the dynamic AMP samples (system default) that the Optimizer collects when it has no statistics on which to base a query plan. Statistical samples taken across all AMPs are likely to be much more accurate than dynamic AMP samples.

Sampled statistics are different from dynamic AMP samples in that you specify the percentage of rows you want to sample explicitly in a COLLECT STATISTICS (Optimizer Form) request to collect sampled statistics, while the number of AMPs from which dynamic AMP samples are collected and the time when those samples are collected is determined by Teradata Database, not by user choice. Sampled statistics produce a full set of collected statistics, while dynamic AMP samples collect only a subset of the statistics that are stored in interval histograms.
What you can do to understand what Teradata statistic collection options are best for you

The subject of Teradata database statistics is far too complex and detailed to be summarily defined or exhausted in this article. There are many new statistics collection options with Teradata Release 14.0, and also improvements to existing options. For example, one of the new options in 14.0 is called SUMMARY. This is used to collect only the table-level statistical information such as row count, average block size, average row size, etc. without the histogram detail. This option can be used to provide up-to-date summary information to the optimizer in a quick and efficient way. When SUMMARY option is specified in a collect statistics statement, no column or index specification is allowed.

The following resources are recommended reading to further your knowledge of statistics as they pertain to the Teradata Database.

*SQL Request and Transaction Processing Release 14.0* manual. Excellent, technically detailed information on different statistic collection strategies is provided in chapter 2. Also, great explanations of how the optimizer uses statistics.

*Teradata Statistics Wizard User Guide Release 14.00*. You don’t have to do it all yourself. Teradata Statistics Wizard automates the process of collecting statistics for a particular workload or selecting arbitrary indexes or columns for collection/re-collection purposes. Additionally, you can validate the proposed statistics on a production system which enables you to verify the performance of the proposed statistics before applying the recommendations. Check it out and use it.


The following Teradata Orange Books:
- *Optimizer Cardinality Estimation Improvements Teradata Database 12.0* by Rama Korlapati
- *Teradata 14.0 Statistics Enhancements* by Rama Korlapati
- *Statistics Extrapolations* by Rama Korlapati
- *Collecting Statistics* by Carrie Ballinger (written for Teradata release V2R6.2, but still a valuable resource)

Anything written by Carrie Ballinger on the subject of Teradata statistics. Check out her contributions to the Teradata Developers Exchange at: http://developer.teradata.com/

Including - *New opportunities for statistics collection in Teradata 14.0* on Carrie’s Blog.


Also on http://developer.teradata.com/
- *When is the right time to refresh statistics? - Part I* (and Part II) by Marcio Moura
- Others
Summary: Teradata Statistics Collection

The decision between full-table and all-AMPs sampled statistics seems to be a simple one: always collect full-table statistics, because they provide the best opportunity for producing optimal query plans.

While the above statement may be true, the decision is not so easily made in a production environment. Other factors must be taken into consideration, including the length of time required to collect the statistics and the resource consumption the collection of full-table statistics incurs while running other workloads on the system.

To resolve this, the benefits and drawbacks of each method must be considered.

An excellent information table comparing the three methods (Full Statistics, Sampled Statistics, Dynamic AMP Samples) is provided in Chapter 2 of the SQL Request and Transaction Processing Release 14.0 manual, under the heading *Relative Benefits of Collecting Full-Table and Sampled Statistics*. Please view this table for comparison details. The information from this table is listed below:

**Method Characteristics Best Use**

**Full statistics**

**Characteristics**
- Collects all statistics for the data.
- Time consuming.
- Most accurate of the three methods of collecting statistics.
- Stored in interval histograms in the Data Dictionary.

**Best Use**
- Best choice for columns or indexes with highly skewed data values.
- Recommended for tables with fewer than 1,000 rows per AMP.
- Recommended for selection columns having a moderate to low number of distinct values.
- Recommended for most NUSIs, PARTITION columns, and other selection columns because collection time on NUSIs is very fast.
- Recommended for all column sets or indexes where full statistics add value, and where sampling does not provide satisfactory statistical estimates.

**Sampled Statistics**

**Characteristics**
- Collects all statistics for the data, but not by accessing all rows in the table.
- Significantly faster collection time than full statistics.
- Stored in interval histograms in the Data Dictionary.
**Best Use**
- Acceptable for columns or indexes that are highly singular; meaning that their number of distinct values approaches the cardinality of the table.
- Recommended for unique columns, unique indexes, and for columns or indexes that are highly singular. Experience suggests that sampled statistics are useful for very large tables; meaning tables with tens of billions of rows.
- Not recommended for tables whose cardinality is less than 20 times the number of AMPs in the system.

**DynamicAMP Sample Characteristics**
- Estimates fewer statistics than COLLECT STATISTICS does.
  Statistics estimated include the following for all columns.
  - Cardinalities
  - Average rows per value
  For indexes only, the following additional statistics are estimated.
  - Average rows per index
  - Average size of the index per AMP
  - Number of distinct values
- Extremely fast collection time, so is not detectable.
- Stored in the file system data block descriptor for the table, not in interval histograms in the Data Dictionary.
- Occurs automatically. Cannot be invoked by user.
- Automatically refreshed when batch table INSERT … DELETE operations exceed a threshold of 10% of table cardinality.
  Cardinality is not refreshed by individual INSERT or DELETE requests even if the sum of their updates exceed the 10% threshold.
- Cached with the data block descriptor.
- Not used for non-indexed selection criteria or indexed selection with non-equality conditions.

**Best Use**
- Good for cardinality estimates when there is little or no skew and the table has significantly more rows than the number of AMPs in the system.
- Collects reliable statistics for NUSI columns when there is limited skew and the table has significantly more rows than the number of AMPs in the system.
- Useful as a temporary fallback measure for columns and indexes on which you have not yet decided whether to collect statistics or not. Dynamic AMP sampling provides a reasonable fallback mechanism for supporting the optimization of newly devised ad hoc queries until you understand where collected statistics are needed to support query plans for them. Teradata Database stores cardinality estimates from dynamic AMP samples in the interval histogram for estimating table growth even when complete, fresh statistics are available.

**Some Closing Thoughts on the Subject of Statistics**
Statistics are vital to our existence, whether they are being twisted by politicians, scientifically utilized, or created for consideration by the Teradata Optimizer. This broad use of statistics for so many different purposes leads to some varying views on the existence and usefulness of statistics.
Modern statisticians are familiar with the notion that any finite body of data contains only a limited amount of information on any point under examination; that this limit is set by the nature of the data themselves, and cannot be increased by any amount of ingenuity expended in their statistical examination: that the statistician's task, in fact, is limited to the extraction of the whole of the available information on any particular issue.
- R. A. Fisher

The statistics on sanity are that one out of every four Americans is suffering from some form of mental illness. Think of your three best friends. If they're okay, then it's you.
- Rita Mae Brown

In ancient times they had no statistics so they had to fall back on lies.
- Stephen Leacock

The science of statistics is the chief instrumentality through which the progress of civilization is now measured, and by which its development hereafter will be largely controlled.
- S. N. D. North

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.
- H.G.Wells

To understand God's thoughts we must study statistics, for these are the measure of His purpose.
- Florence Nightingale

Statistics are just a way for the mathematician to evangelize his faith.
- Hunter Brinkmeier

Now go, write it before them in a table, and note it in a book.
- Book of Isaiah, 30:8

Do not put your faith in what statistics say until you have carefully considered what they do not say.
- William W. Watt

There are two kinds of statistics, the kind you look up and the kind you make up.
- Rex Stout, Death of a Doxy

The theory of probabilities is at bottom nothing but common sense reduced to calculus.
- Laplace, Théorie analytique des probabilités, 1820

These days the statistician is often asked such questions as "Are you a Bayesian?" "Are you a frequentist?" "Are you a data analyst?" "Are you a designer of experiments?". I will argue that the appropriate answer to ALL of these questions can be (and preferably should be) "yes", and that we can see why this is so if we consider the scientific context for what statisticians do.
- G.E.P. Box
The manipulation of statistical formulas is no substitute for knowing what one is doing.
- Hubert M. Blalock, Jr., Social Statistics

The same set of statistics can produce opposite conclusions at different levels of aggregation.
- Thomas Sowell

Whether we like it or not, quantification in history is here to stay for reasons which the quantifiers themselves might not actively approve. We are becoming a numerate society: almost instinctively there seems now to be a greater degree of truth in evidence expressed numerically than in any literary evidence, no matter how shaky the statistical evidence, or acute the observing eye.
- John Harold Plumb

Having given the number of instances respectively in which things are thus and so, in which they are thus and not so, in which they are so and not thus, and in which they are neither thus nor so, it is required to eliminate the general quantitative relativity inhering in the mere thingness of the things, and to determine the special quantitative relativity subsisting between the thusness and the soness of the things.
- M.H. Doolittle, 1887

I never keep a scorecard or the batting averages. I hate statistics. What I got to know, I keep in my head.
- Dizzy Dean