Basic Statistics Analysis

version 1b, 15-Nov-2011
author: Craig Shallahamer, craig@orapub.com / orapub.general@gmail.com

Introduction

It’s difficult for us humans to aborb and grasp more than just a few values. Statistics provides us with a way to grasp complexity by simplifying or abstracting. As Oracle performance specialist, we typically have a large number of numeric samples, comprising our sample set.

Sample values

Enter your sample values; seperated by commas and the entire set enclosed in braces. Usually you can simply copy and paste the values directly into the notebook; which usually removes any line feeds or returns. It is common to not consider all sample vaues. Below you can specify threshholds for the minimum and maximum values to run through the analysis.

```
cutoffMin = 0.0;
cutoffMax = 9999.0;
sampleSetRaw = {0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
   0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
   0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
   0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
   0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
   0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
   0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
   0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
   0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
   0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
   0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
   0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
   0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
   0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
   0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
   0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
   0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
   0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
   0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
   0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
   0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1,
   1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
   1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
   1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
   1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
   1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
   1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
   1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
   2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
   2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
   2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3,
   3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
   3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
   3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
   5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 8, 8};
```

Basic Numeric Statistics

The basic numeric statistics shown are:

First five values simply lists the first five values of your sample set. Use this to check the data being used is what you think.
Number of samples is in fact the number of samples. Use this to check the data has been entered correctly.
Average is the statistical mean.
Median is the middle value after all the values are sorted. This is also the 50-percentile value.
Standard deviation is a measure of dispersion and is particularly valuable when the distribution is normal.
P-Value is a measure of contrast. In this instance, we are constrasting the sample set to the normal distribution. Loosely speaking, if the P-Value is greater than 0.05 then our sample set is likely to be normally distributed.

```
sampleSetGood = {};
sampleSetBad = {};
Table[
  If[((sampleSetRaw[[i]] ≤ cutoffMax) && (sampleSetRaw[[i]] ≥ cutoffMin)),
   AppendTo[sampleSetGood, sampleSetRaw[[i]] ],
   AppendTo[sampleSetBad, sampleSetRaw[[i]] ]
  ]
  , {i, 1, Length[sampleSetRaw]}
 ];

countRaw = Length[sampleSetRaw];
countGood = Length[sampleSetGood];
countBad = Length[sampleSetBad];

firstFiveRaw = Take[sampleSetRaw, 5];
firstFiveGood = Take[sampleSetGood, 5];
firstFiveBad = Take[sampleSetBad, 5];

avg = Round[N[Mean[sampleSetGood]], 0.00010];
med = Round[N[Median[sampleSetGood]], 0.00010];
std = Round[N[StandardDeviation[sampleSetGood]], 0.0010];
pValue = Round[N[DistributionFitTest[sampleSetGood]], 0.0000010];
pct90 = Round[N[Quantile[sampleSetGood, 0.90]], 0.000010];
pct95 = Round[N[Quantile[sampleSetGood, 0.95]], 0.000010];
pct99 = Round[N[Quantile[sampleSetGood, 0.99]], 0.000010];
maxV = Max[sampleSetGood];

Grid[{
  {"Number of total samples", countRaw},
  {"Number of good samples", countGood},
  {"Number of bad samples", countBad},
  {"First five raw samples", firstFiveRaw},
  {"First five good samples", firstFiveGood},
  {"First five bad samples", firstFiveBad},
  {"Good Sample Details", "---"},
  {"  Average", avg},
  {"  Median (50%-tile)", med},
  {"  Maximum", maxV},
  {"  Percentiles (90,95,99)", {pct90, pct95, pct99}},
  {"  Standard deviation", std},
  {"  P-Value", pValue}
 },
 {Alignment → {Left},
  Frame → None}
]
```

```
Take::take : Cannot take positions 1 through 5 in {}. ≫
```

```
Number of total samples  1024
Number of good samples   1024
Number of bad samples    0
First five raw samples   {0, 0, 0, 0, 0}
First five good samples  {0, 0, 0, 0, 0}
First five bad samples   Take[{}, 5]
Good Sample Details      ---
  Average                0.8389
  Median (50%-tile)      0.
  Maximum                8
  Percentiles (90,95,99) {3., 4., 6.}
  Standard deviation     1.337
  P-Value                0.
```

Basic Visual “Statistics”

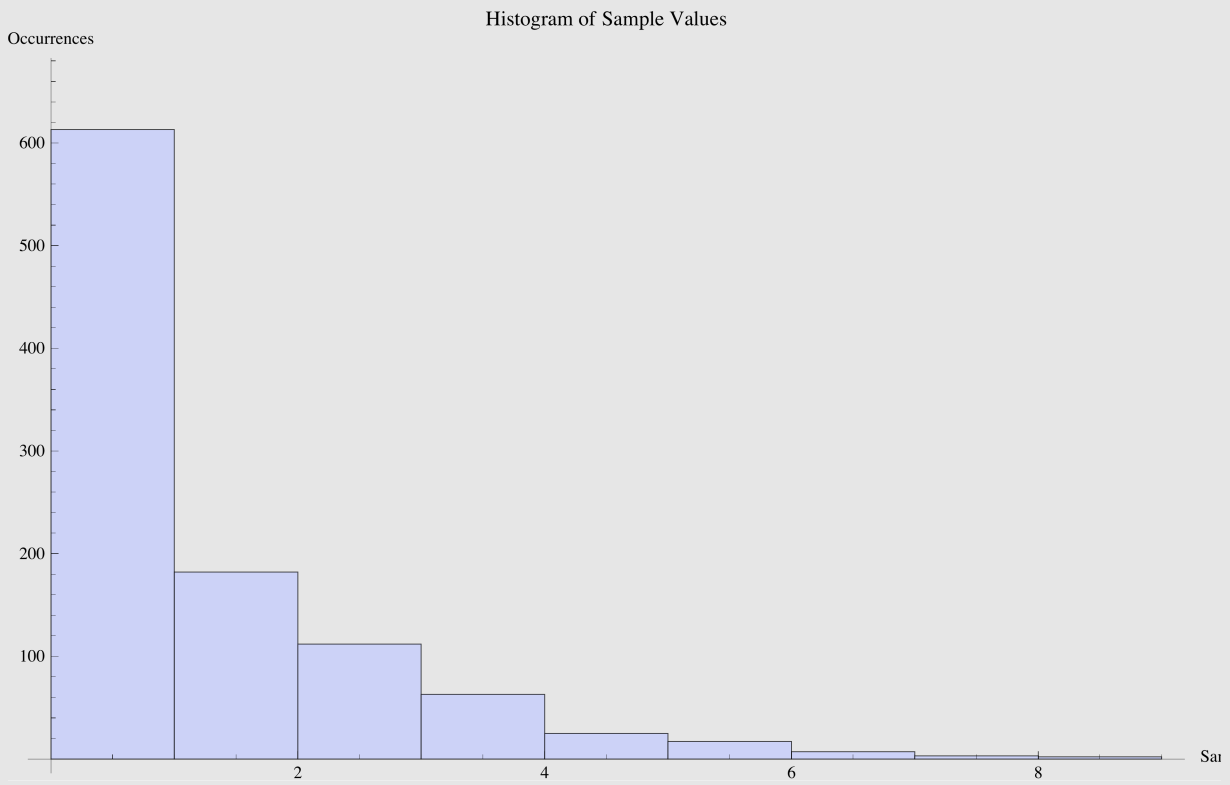
Histograms are a fantastic way to get a quick grasp of a large number of samples. Below are a select number of histogram, each focusing on a specific numeric quality.

```
hLabel = "Sample Values";
vLabel = "Occurrences";

histStnd = Histogram[sampleSetGood,
   PlotLabel → "Histogram of Sample Values", AxesLabel → {hLabel, vLabel}];
histStndSmooth = SmoothHistogram[sampleSetGood, PlotLabel →
   "Smoothed Histogram of Sample Values\n(Probability Distribution Function)",
  AxesLabel → {hLabel, ""}]; histCC = Histogram[sampleSetGood, Automatic,
  "CumulativeCount", PlotLabel → "Histogram of Sample Values\nCumulative Count",
  AxesLabel → {hLabel, vLabel}];
histProb = Histogram[sampleSetGood, Automatic, "Probability", PlotLabel →
    "Histogram of Sample Values\nProbability", AxesLabel → {hLabel, "% Occurs"}];
histStndSmallBin = Histogram[sampleSetGood, {0.250}, PlotLabel →
    "Histogram of Sample Values\nbin size 0.250", AxesLabel → {hLabel, vLabel}];
Print[
  "
    "];
```

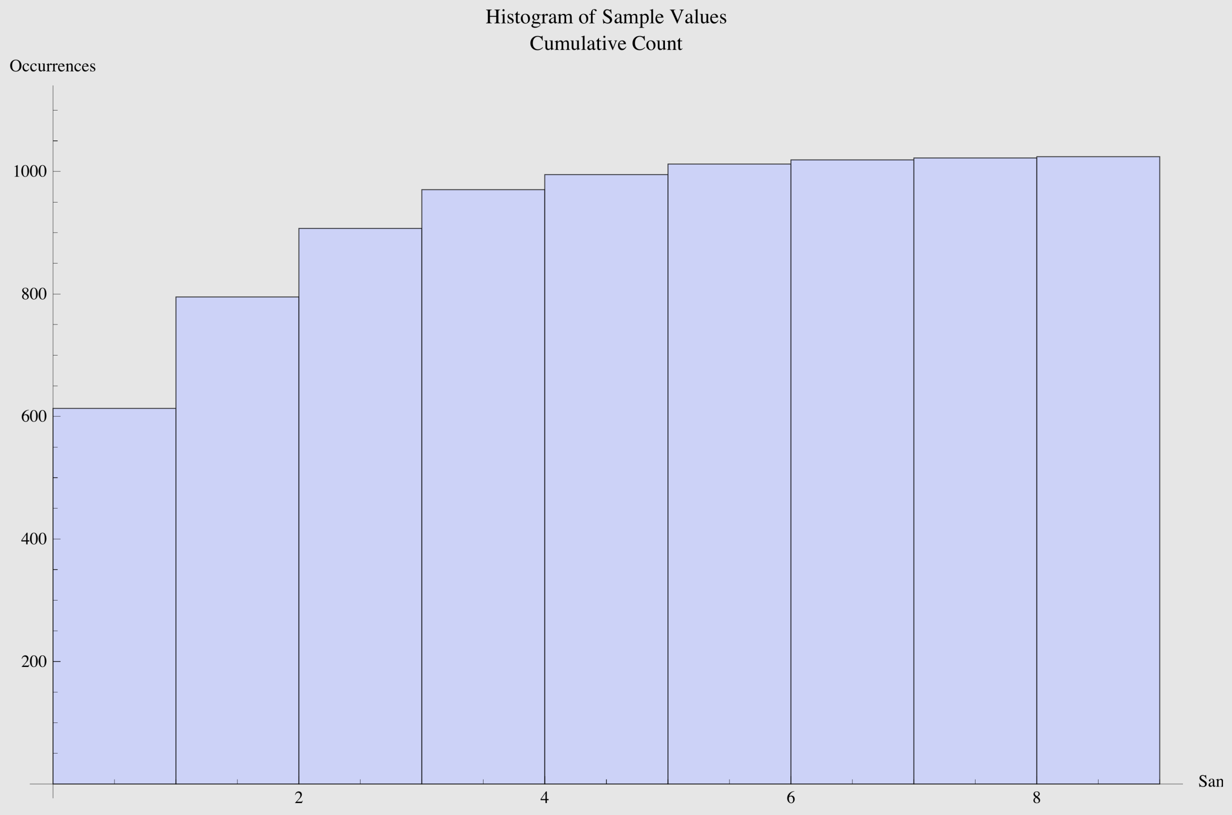
Below is a standard histogram, where each sample is shown as a single block placed on the vertical axis based on its value. Common sample values (i.e., blocks) show as high stacks.

```
histStnd
```



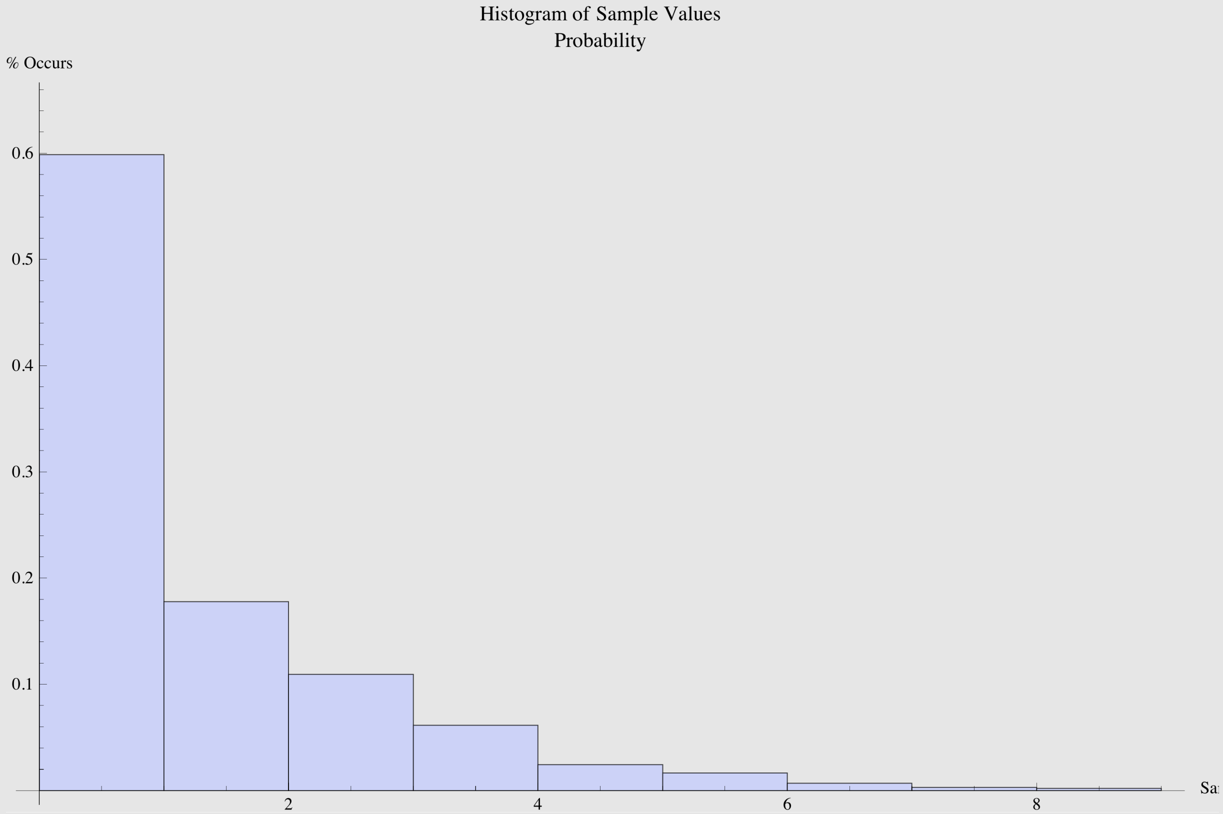
Below is known as a Cummlative Count histogram. Each vertical bar represents the total number of samples values that are less than or equal to the bin.

```
histCC
```



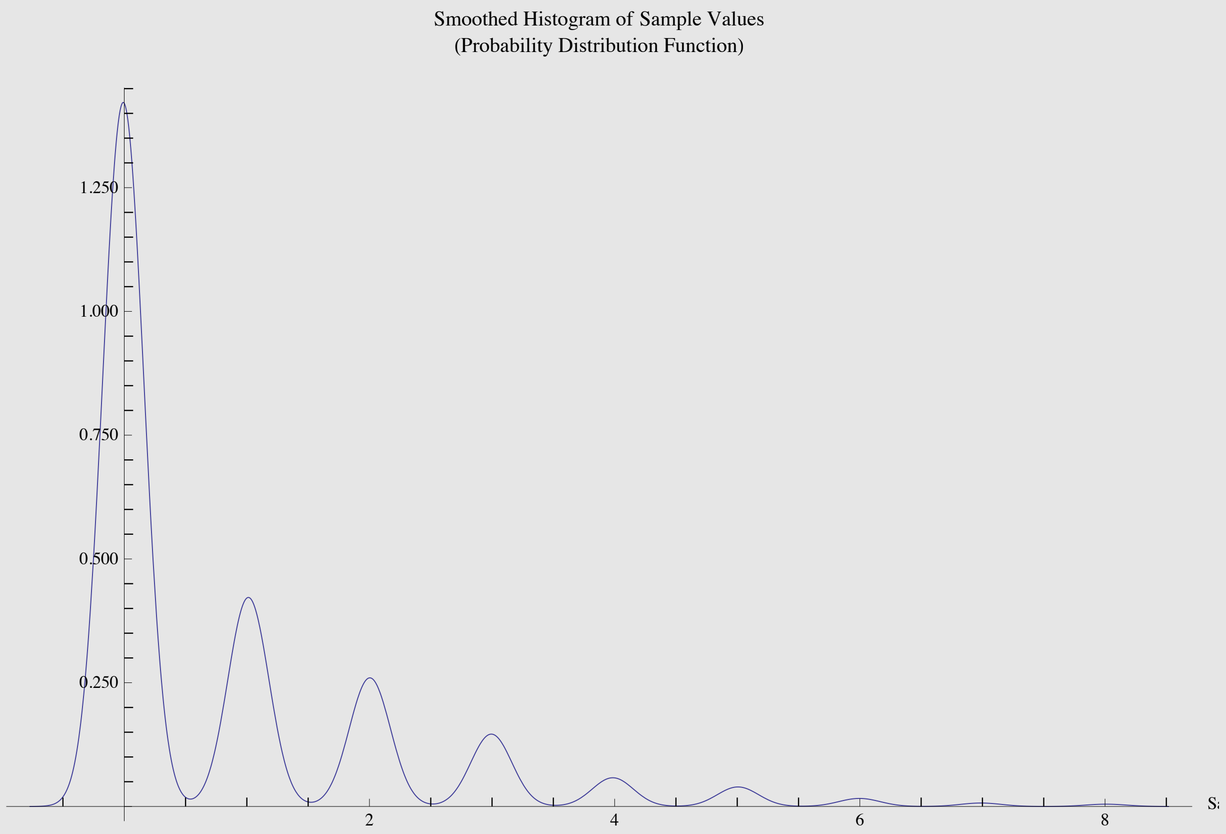
Below is a Probability Historgram. It will visually look exactly like the standard histogram but the vertical axis is a percentage value. Each vertical bar’s height represents the percentage of values that is contains. In contrast, the standard histogram hight is the actual number of sample occurrences.

```
histProb
```



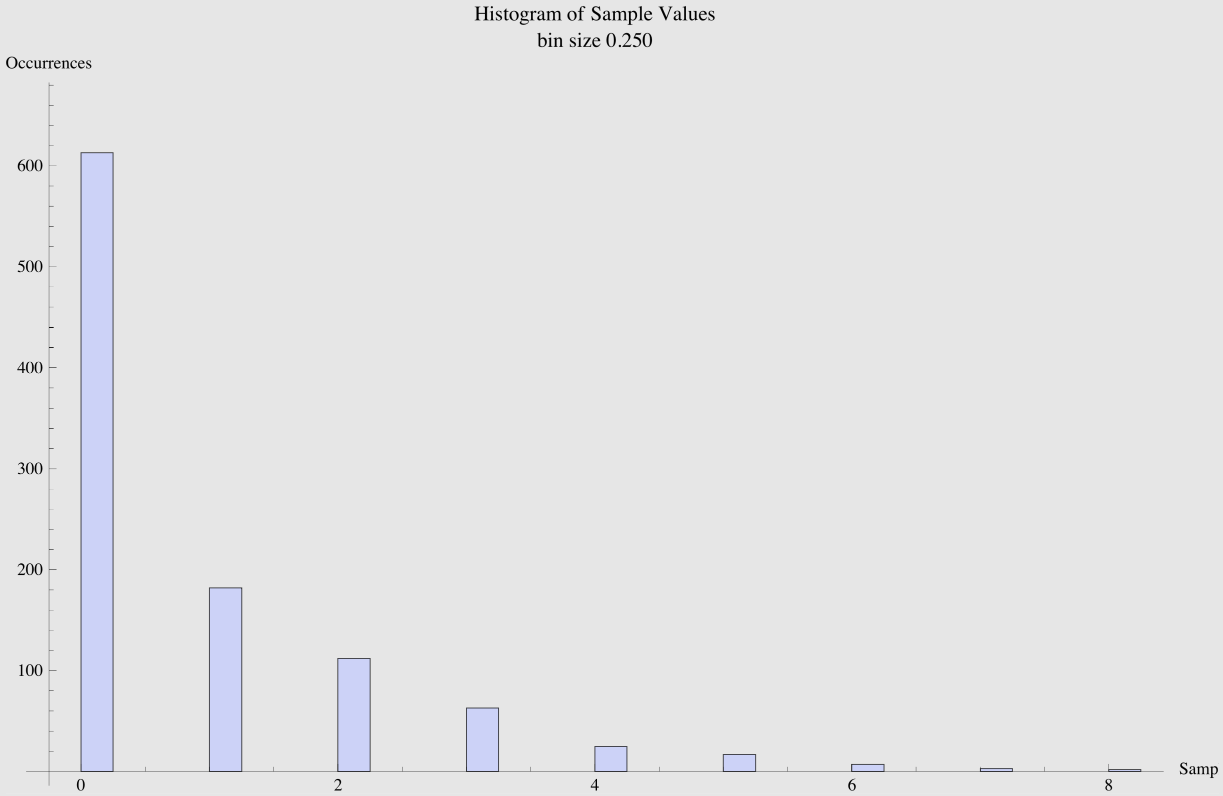
Below is a Smoothed Histogram. It will have a similar shape to the Standard Histogram, but will be mathematically smoothed. Sometimes this is a much more pleasent and informative visual, but not always. Remember, it is smoothed so it does not consist of the actual values. For example, you may see the line go negative, even though there are no negative values. In reality, the plot is the probability distribution function (PDF).

```
histStndSmooth
```



Below is a standard histogram but with the bin size set to 0.250. This is only useful when the sample values range below 1.0, such as when sampling SQL statements (we all hope). Sample sets with large sample values will likely not result in a plot.

```
histStndSmallBin
```


Histogram of Sample Values
bin size 0.250
Occurrences
Samp