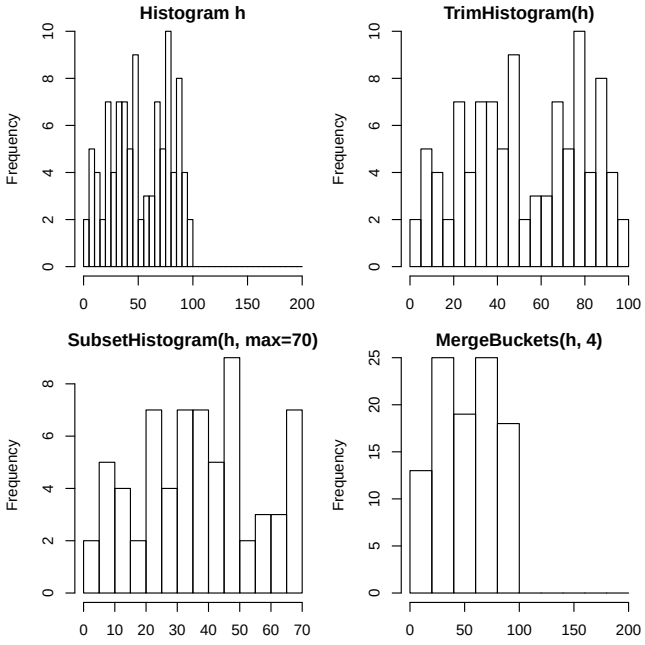
HistogramTools 0.3.1 Quick Reference Guide

Murray Stokely **August 26, 2014**

Histogram Manipulation This package includes a number of basic functions for subsetting, trimming, merging, adding, and otherwise manipulating basic R histogram objects.

```
> h <- hist(runif(100, 0, 100),
+           breaks=seq(from=0,to=200,by=5))
> plot(TrimHistogram(h))
> plot(SubsetHistogram(h, maxbreak=70))
> plot(MergeBuckets(h, adj.buckets=2))
```



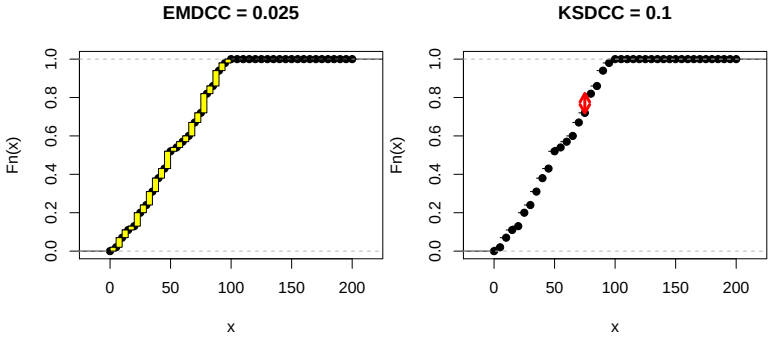
Information Loss The introduction of binning a dataset into a histogram introduces information loss. The package provides standard Histogram distance measures as well as distance functions on the possibile ECDFs that could have been constructed from the original data: the Kolmogorov-Smirnov Distance of the Cumulative Curves (KSDCC) and Earth Mover's Distance of the Cumulative Curves (EMDCC). The plots here show a visual representation of the returned value. EMDCC is the area of the yellow boxes and KSDCC is the distance of the red arrow.

```
> par(mfrow=c(1,2), par(mar=c(5,4,4,0)+0.1))
> PlotEMDCC(h)
> PlotKSDCC(h)
> EMDCC(h)

[1] 0.025

> KSDCC(h)

[1] 0.1
```



Serialize a Histogram This package includes functions for reading and writing Histograms from other tools. Most notably, it can encode or decode any arbitrary R histogram into a portable protocol buffer format to send to other programs written in other languages.

```
> hist.msg <- as.Message(h)
> length(hist.msg$serialize(NULL))

[1] 469
```

Common HistogramTools Functions

Bin Manipulation	
`TrimHistogram`	Remove empty consecutive buckets from ends
`AddHistograms`	Aggregate two or more histograms
`MergeBuckets`	Merge adjacent bucket boundaries
`SubsetHistogram`	Return histogram with subset of buckets
`IntersectHistograms`	Return a intersection of two histograms
`ScaleHistogram`	Scale the counts of a histogram by a factor
Quantiles and Empirical CDFs	
`HistToEcdf`	Return the ECDF of histogram
`Count`	Return the number of data points in hist
`ApproxMean`	Return an approximate mean of the binned data
`ApproxQuantile`	Return an approximate quantile of the binned data
Distance Measures of Two Histograms	
`minkowski.dist`	The Minkowski distance of order p.
`intersect.dist`	The intersection distance.
`kl.divergence`	Kullback-Leibler Divergence.
`jeffrey.divergence`	Jeffrey Divergence.
Binned CDF Distance Measures	
`PlotKSDCC`	Plot ECDF with annotation at point of KS distance of the cumulative curves
`PlotEMDCC`	Plot ECDF with annotation showing EMD of the cumulative curves
`KSDCC`	Return the Kolmogorov-Smirnov distance of the cumulative curves (btwn 0 and 1)
`EMDCC`	Return the Earth Mover's distance of the cumulative curves (btwn 0 and 1)
Misc	
`PlotLog2ByteEcdf`	Plot ECDF of hist with power of two bucket boundaries
`PlotLogTimeDurationEcdf`	Plot ECDF of hist with log-scaled time duration bucket boundaries
`as.histogram`	Parse a HistogramState protocol buffer and return an R histogram
`as.Message`	Serialize an R histogram as a HistogramState protocol buffer
`ReadHistogramsFromDtraceOutputFile`	Read histograms from DTrace output