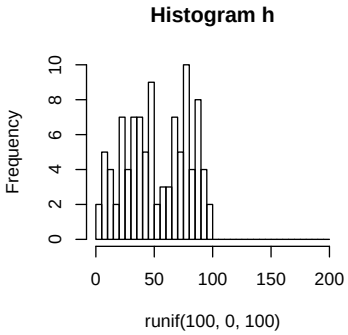
HistogramTools 0.1: Quick Reference Guide

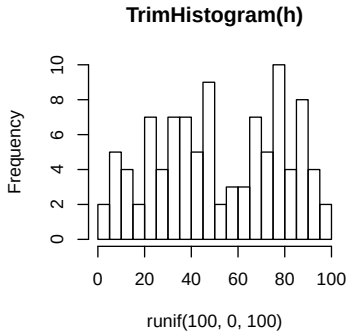
Murray Stokely

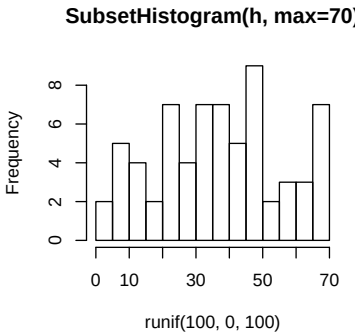
September 13, 2013

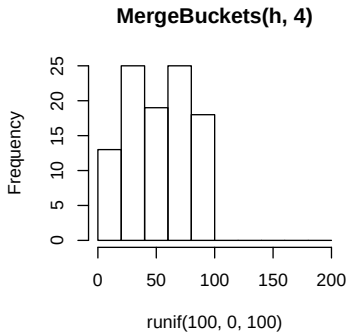
Histogram Manipulation This package includes a number of basic functions for subsetting, trimming, merging, adding, and otherwise manipulating basic R histogram objects.

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









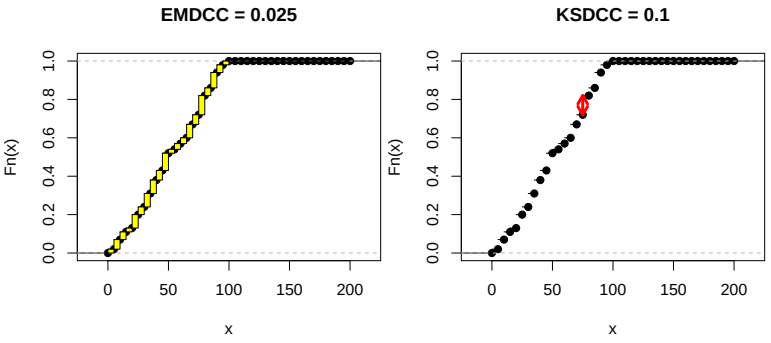
Information Loss The introduction of binning a dataset into a histogram introduces information loss. The Kolmogorov-Smirnov Distance of the Cumulative Curves (KSDCC) and Earth Mover's Distance of the Cumulative Curves (EMDCC) are two provided error metrics for histograms. The plots here show a visual 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))
```

```
NULL
```

Common HistogramTools Methods

`HistToEcdf`	Returns cumulative distribution function of histogram
`AddHistograms`	Aggregate two or more histograms for different data sets with identical bucket boundaries
`MergeBuckets`	Merge adjacent bucket boundaries to return a histogram with fewer buckets
`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
`SubsetHistogram`	Returns a new histogram with a subset of the buckets
`TrimHistogram`	Returns a new histogram with empty buckets at the left or right of distribution removed
`PlotKSDCC`	Plot ECDF of hist with annotation at point of KS distance of the cumulative curves
`PlotEMDCC`	Plot ECDF of hist with annotation showing earth mover's distance of the cumulative curves
`KSDCC`	Return 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)
`as.histogram`	Parse a HistogramState protocol buffer and return an R histogram.
`as.Message`	Serialize an R histogram as a HistogramState protocol buffer