Variations on the histogram

Lorraine Denby and Colin Mallows

It is usual to choose to make the bars in a histogram all have the same width. One could also choose to make them all have the same area. These two options have complementary strengths and weaknesses; the equal-width histogram oversmooths in regions of high density, and is poor at identifying sharp peaks; the equal-area histogram oversmooths in regions of low density, and so does not identify outliers. We describe a compromise approach which avoids both of these defects. We regard the histogram as an exploratory device, rather than as an estimate of a density. We argue that relying on the asymptotics of Integrated Mean Square Error leads to inappropriate recommendations for choosing bin-widths.

Key Words: Diagonally-cut histogram; equal-area histogram; asymptotics; IMSE.

Supplementary Materials

The following files are all contained in the archive `supplement_07-004.tar`.

1. Data Sets

`fig2.txt` This data is from a Six Sigma quality improvement project. The purpose of the study was to reduce the time from trouble report to restored service. The particular repairs of interest were those that needed a technician dispatched and could not be fixed remotely. This file contains 506 durations of one step of the process from trouble report to service restoration.

`fig3.txt` This data is from the Boston Housing data from Harrison and Rubinfeld (1978). This dataset contains information collected by the U.S Census Service concerning housing in the area of Boston Massachusetts. It contains 506 observations and 14 variables. This file contains the PT variable (pupil-teacher ratio by town).

Lorraine Denby is a Research Scientist and Colin Mallows is a Consultant in the Data Analysis Research Department at Avaya Labs

This document describes supplementary material to an article published in the Journal of Computational and Graphical Statistics.

`fig4.txt` This is an artificial example, exhibiting both a spike and outliers. It is a mixture of three distributions: 775 points drawn from a standard Normal distribution, 150 points of value 7, and 75 points distributed uniformly over the interval (0,15).

`fig5.txt` Our final example is drawn from a study of round trip time of Voice over IP (VoIP) packets between two devices on the data network. Each observation is the median round trip time of 100 packets sent 20 milliseconds apart from a particular source and destination device on the network. One thousand sets of packets were sent, of which 69 were lost, so 931 remain.

2. Computer Code

`dhist.r` This is the R function that produces a dhist. The arguements are explained in the code.

`read.data.R` This reads the 4 data files into R and names them appropriately for use in the 4 R source files that produce Figures 2 to 5.

`fig.2.R` This R source file produces Figure 2. Note that these source files do not get the character sizes identical to the figures in the paper since the figures were produced with S-Plus.

`fig.3.R` This R source file produces Figure 3.

`fig.4.R` This R source file produces Figure 4.

`fig.5.R` This R source file produces Figure 5.