Caltech SURF Tutorials

Javier Duarte, Cristian Peña
Department of Physics
California Institute of Technology

May 14, 2013

1 Representations of Data

See Cowan, Statstical Data Analysis 1998.

1.1 Histogram

A histogram is a graph of a set of observations. Say we have n observations of a random variable x. Then we can divide the x-axis into m subintervals of width Δx_i, called *bins*. The number of occurences of n_i of x in each subinterval is plotted on the y-axis. In the limit of zero bin width, and infinite observations a histogram exactly replicates a probability density function, or p.d.f. Fig. 1 shows the relation between histograms as we change the number of observations and the p.d.f. from which observations are drawn.

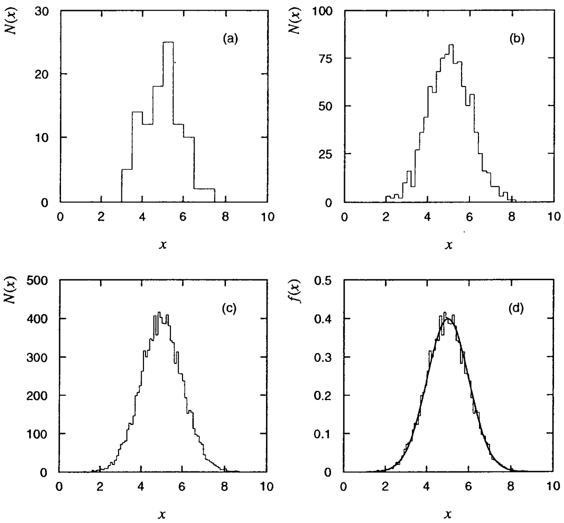

Figure 1: Histograms of various numbers of observations of a random variable x based on the same p.d.f. (a) $n = 100$ observations and a bin width of $\Delta x = 0.5$. (b) $n = 1000$ observations, $\Delta x = 0.2$. (c) $n = 10000$ observations, $\Delta x = 0.1$ and (d) is normalized to unit area. Also shown as a smooth curve is the p.d.f. according to which the observations are distributed.

To create a histogram from input data text file `Day1/basic.dat` simply run the macro named `Day1/histogram1D.C`. To run it, begin at a UNIX prompt in the directory `Day1`, open ROOT, then execute the macro

```
$ root -l
root [0] .x histogram1D.C
x=-1.102279, y=-1.799389, z=4.452822
x=1.867178, y=-0.596622, z=3.842313
x=-0.524181, y=1.868521, z=3.766139
x=-0.380611, y=0.969128, z=1.084074
x=0.552454, y=-0.212309, z=0.350281
root [1]
```

1.2 Tree

A ROOT TTree (just “tree” for short) is structured container of data, specially designed for collider. In a collider, we typically are interested in the distribution of certain quantities over many collision events (or entries). For example, in events with electrons, we may be interested in the distribution of the an electron’s momentum $|\vec{p}|$, maybe after requiring that the electron was within “acceptance” of the detector.

A tree tries to organize this information by effectively linking similar types of information from different events. This link is called a branch. In the tutorial I showed how to create a tree, with a few branches for the variables x, y, and z. Sometimes this is called a “flat ntuple” because none of the information is nested. Here’s a more complicated example, with a nested hierarchy of information. There is a branch of particles, each particle branch has sub-branches
`http://lcg-heppkg.web.cern.ch/lcg-heppkg/ROOT/eventdata.root`

2 Curve Fitting

See PDG Review of Statistics, J. Beringer et al. (Particle Data Group), Phys. Rev. D86, 010001 (2012).
`http://pdg.lbl.gov/2012/reviews/rpp2012-rev-statistics.pdf`

2.1 Least Squares Method

$$\chi^2(\theta) = \sum_{i=1}^{n} \frac{(n_i - F(x_i);\theta)^2}{\sigma_i^2} \tag{1}$$

2.2 Binned Goodness-of-Fit

The chi-square distribution