Histograms

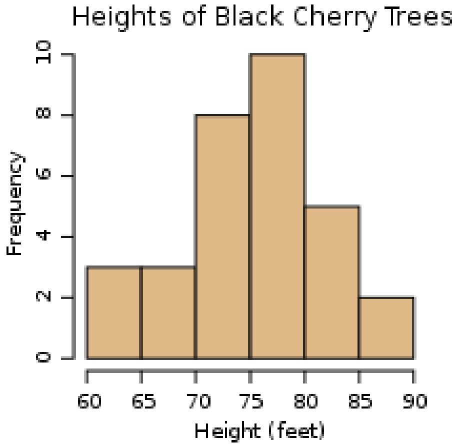
MUDS
Herramientas de análisis y visualización de datos
La Salle - Universitat Ramon Llull

1. Definition and description

A histogram consists of tabular frequencies of data collected and classified over discrete intervals (bins), with an area equal to the frequency of the observations in the interval. It is often shown as a graphical representation, shown as adjacent rectangles, giving a visual impression of the distribution of data. It was first introduced by Karl Pearson.

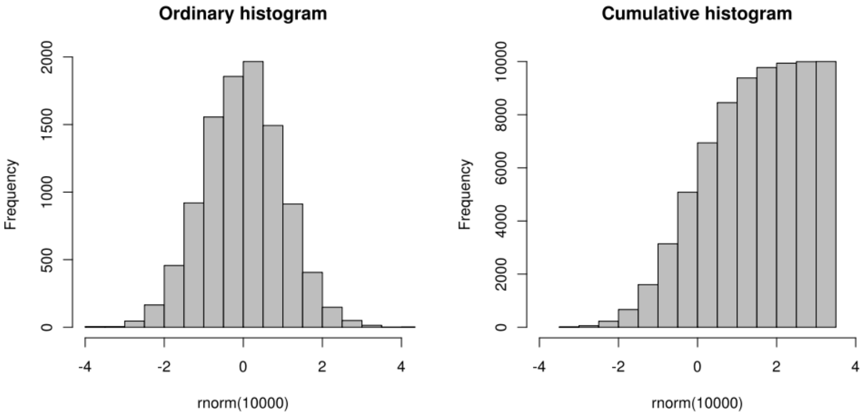
If working with continous distributions, it is an estimate of the probability distribution.

We can extend the concept to discrete or categorical data by just keeping the representation of the relative frequencies of each category.



One also defines cumulative histograms. A cumulative histogram is a mapping that counts the cumulative number of observations in all of the bins up to the specified bin. That is, the cumulative histogram M_i of a histogram m_j is defined as:

$$M_i = \sum_{j=1}^{i} m_j$$



2. How an histogram is filled

Assume a discrete random variable or a categorical variable. The probability for each value n_k is given by $f(n_k)$

$$f(n_k) = p_k = P[N = n_k]$$

We keep repeating the experiment independently while we count how many times we obtain each of the possible values. We will have,

$\mathcal{N}$ Total number of experiments *i.e.* number of times we repeat our experiment.

N_1 number of occurences of the first value of N, n_1,

N_2 number of occurences of the second value of N, n_2,

$\ldots\ldots$,

N_n number of occurences of the last value of N, n_n,

p_k is the probability to obtain the k^{th} value of N, $p_k = f(n_k)$.
We have

$$N_1 + N_2 + \ldots + N_n = \mathcal{N}$$

The set of numbers $(N_1, N_2, \ldots, N_n)$ follows a MULTINOMIAL DISTRIBUTION

$$f(N_1, N_2, \ldots, N_n) = \frac{\mathcal{N}!}{N_1! N_2! \ldots N_n!} p_1^{N_1} \ldots p_n^{N_n}$$

The marginal distribution for each N_k is a binomial,

$$f(N_k) = \frac{\mathcal{N}!}{N_k!(\mathcal{N} - N_k)!} p_k^{N_k} (1 - p_k)^{\mathcal{N} - N_k}$$

while the average is,

$$\nu_k = \mathcal{N} p_k,$$

and the variance,

$$\sigma_k^2 = \mathcal{N} p_k (1 - p_k).$$

Now, by Chebishev inequality,

$$P\left[\mid N_k - \nu_k \mid \geq \mathcal{N}\epsilon\right] \leq \frac{\sigma_k^2}{\mathcal{N}^2 \epsilon^2} = \frac{p_k(1 - p_k)}{\mathcal{N}\epsilon^2},$$

which we can recast into,

$$P\left[\mid \frac{N_k}{\mathcal{N}} - p_k \mid \geq \epsilon\right] \leq \frac{p_k(1 - p_k)}{\mathcal{N}\epsilon^2},$$

Remark that ϵ is a measure of the error precision of the error with which we want to find p_k from $N_k/\mathcal{N}$. Since $1 - p_k$ shall be established with the same error, we define the relative error,

$$\epsilon_r = \frac{\epsilon}{p_k}$$

so

$$P\left[\mid \frac{N_k}{\mathcal{N}} - p_k \mid \geq \epsilon \right] \leq \frac{1}{\mathcal{N}\epsilon_r^2},$$

In the worst case, $\mathcal{N}\epsilon_r^2 = 1$ that is,

$$\epsilon_r = \frac{1}{\sqrt{\mathcal{N}}}.$$

Still this is a very bad limit, since it comes from Chebishev inequality, which is very crude. Instead we may use *De Moivre Laplace* theorem and define,

$$x = \frac{N_k - \mathcal{N}p_k}{\sqrt{\mathcal{N}p_k(1-p_k)}}$$

In the limit $\mathcal{N} \to \infty$, x behaves like a Gaussian variable. In fact, the x behaves like a Gaussian random variable in a good enough approximation, provided,

$$\mathcal{N}p_k >> 5$$

Example. With this approximation, for instance,

$$P[\mid x \mid \leq 3] = 0.997$$

Since

$$\mid x \mid \leq 3 \Longrightarrow \frac{\mid N_k - \mathcal{N}p_k \mid}{\sqrt{\mathcal{N}p_k(1-p_k)}} \leq 3$$

so

$$\epsilon = \left\| \frac{N_k}{\mathcal{N}} - p_k \right\| \leq 3\sqrt{\frac{p_k(1-p_k)}{\mathcal{N}}}$$

with probability 0.997.

If the total number of data elements $\mathcal{N}$ is not fixed but data appear in an independent way, $\mathcal{N}$ should follow a Poisson distribution. We call its average ν. If this is so, each category or value of the histogram shall also follow a Poisson distribution. This Poisson distribution shall have average ν_k,

$$f_k(N_k) = \frac{1}{N_k!}\nu_k^{N_k}e^{-N_k}$$

The expected probability for the corresponding category is

$$p_k = \frac{\nu_k}{\nu}$$

while the probability we evaluate is

$$\frac{N_k}{\mathcal{N}}$$

For large values of ν (say $\nu >> 5$, the Poisson variable also can be approximated by a Gaussian.

3. Bins

If the variable is not discrete or even if discrete, has a too large number of possibilites, we build our histogram out of bins. In a bin, we count all events in a given range.

$$\mathrm{BIN}_i, \quad x_i < x < x_{i+1}, \quad p_i = \int_{x_i}^{x_{i+1}} dx\ f(x)$$

One key element is the choice of the number of bins and its size. In order to guarantee a good gaussian approximation for the number of counts in each bin, we need to guarantee that

$$\nu_k = \mathcal{N} p_k >> 5.$$

Two typical choices are

Evenly spaced. The bin width is constant and equal to h.

$$x_k = x_0 + kh.$$

This is a simple choice but we need to guarantee the above condition.

Equiprobable. Bins are chosen so that they all have the same expected probability. This condition is harder to build but still, is the best to guarantee optimal conditions for some tests.

For equally spaced bins, the bin size is a delicate matter.

If the goal of our histogram is to estimate the probability density function we face the following dilemma.

- If you choose too small bin size, a bar height at each bin suffers significant statistical fluctuation due to paucity of samples in each bin.
- If you choose too large bin size, a histogram can not represent shape of the underlying distribution because the resolution isn't good enough.

Determining how many histogram bins should be used for estimating distributions is a problem in non-parametric statistics, although histogram-based methods are not the only form of distribution estimation. However, histogram-based methods are the most practical as other methods usually involve too much computational overhead to be useful for this problem.

Scott shows that the optimal histogram bin size, which provides the most efficient, unbiased estimation of the probability density function, assuming an underlying gaussian distribution, is achieved when

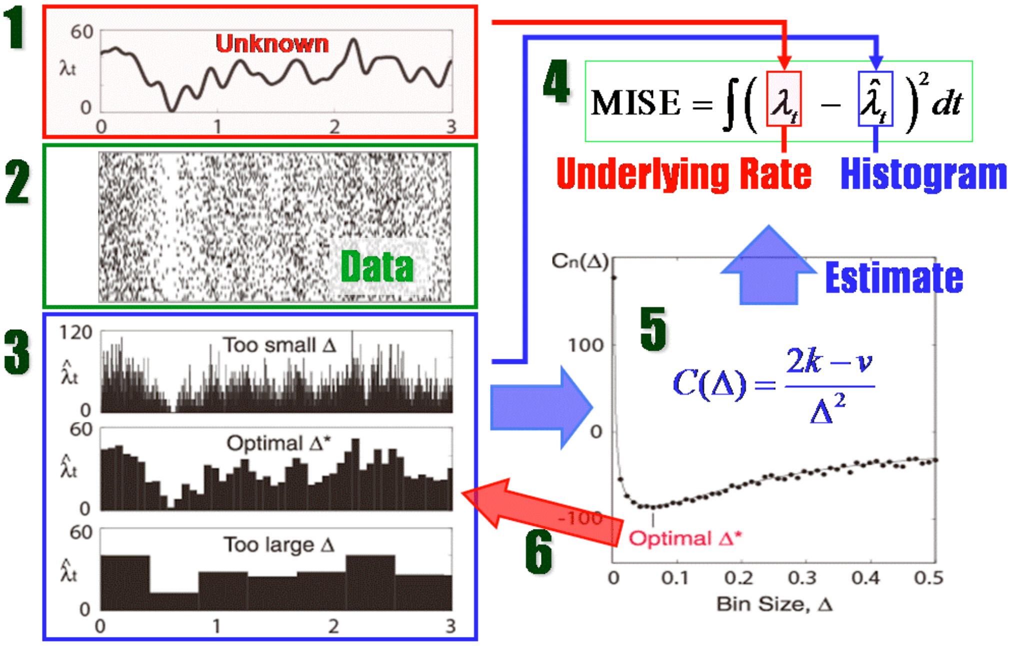
$$h = \frac{3.49\sigma}{\mathcal{N}^{1/3}}$$

where σ is the standard deviation of the distribution, though in practice, the estimated standard deviation. must be used. A similar, but more robust, result was also obtained by Freedman and Diaconis, which gives the bin width as:

$$h = \frac{2(IQR)}{\mathcal{N}^{1/3}}$$

where IQR is the interquartile range (the 75th percentile minus the 25th percentile).

More modern methods, propose,



- An unknown underlying rate that generates observable events.

- Event sequences generated from the unknown rate.

- Histograms constructed from the events. It is not obvious which bin size should be used.

- Mean Integrated Squared Error (MISE), a measure of the goodness-of-the-fit of a histogram to the unknown rate. A histogram with the bin size that minimizes the MISE is optimal.

- Note that we can not directly compute the MISE since we do not know the underlying rate. However, the MISE can be estimated from the data.

- The optimal bin size can be estimated as the one that minimizes the estimated MISE.