Probability & Statistics – Histograms

Histograms are a useful way of representing the *distribution* of a set of measurements of a particular quantity.. In other words, how the quantity being measured varies, over what range, and where is it most likely to occur. A histogram contains much more information than simply an average like the mean or median, and even if a measure of spread such as a standard deviation is also calculated.

To avoid *visual bias* (which is possible for bar charts), the *area* under each bar of a histogram corresponds to the number of measurements, or **frequency**. *The total area under the histogram is therefore the total frequency.* To make this happen, the height of a histogram bar is the **frequency / variable range.** This is called the **frequency density.**

If we scale a histogram such that its total area is one (unity), then the histogram approximates a *probability distribution*. We can *integrate* this (i.e. find areas) to determine the *chance* of the measurement being between particular ranges. In the example below, the probability of x being between 20 and 30 is approximately 30/85 = **0.353.**

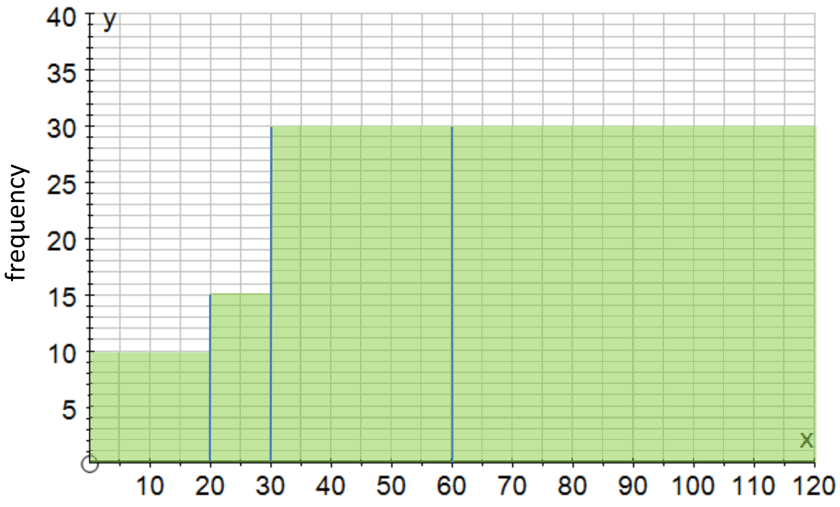
Variable range	Frequency	Frequency density = frequency / variable range
$0 \le x < 20$	10	10/20 = 0.5
$20 \le x < 30$	15	15/10 = 1.5
$30 \le x < 60$	30	30/30 = 1
$60 \le x < 120$	30	30/60 = 0.5

Consider a data set of 85 measurements, grouped into a *frequency table* as shown on the left.

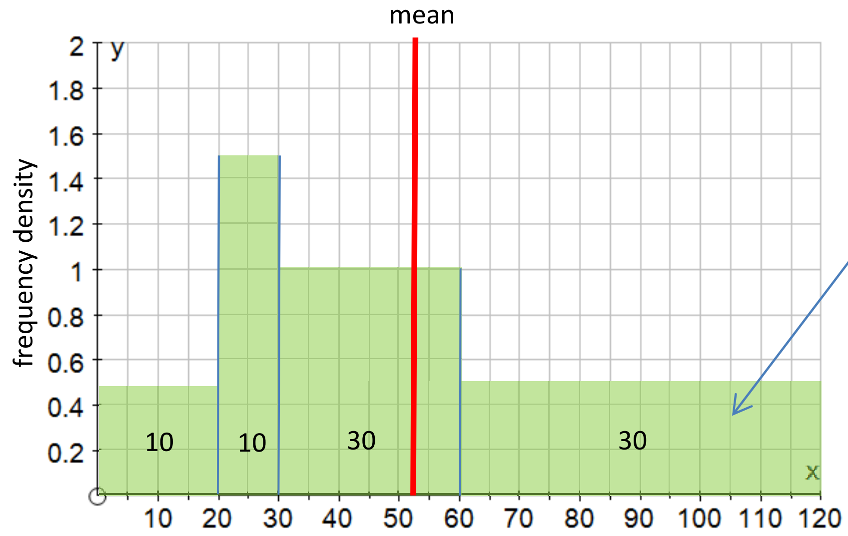
The *mean* for this data set can be *estimated* by assuming we have the same x value in each variable range, and taking the value to be the middle of the range.

Therefore the mean of x is approximately:

((10)(10) + (25)(15) + (45)(30) + (90)(30)) / (10 + 15 + 30 + 30) = **53.24**



Bar chart – it *looks like* most of the x values are towards the higher values, i.e. we might expect an *average* of around 80 to 90. **However, we would be wrong!**



Total area of bars is the total frequency i.e. 85

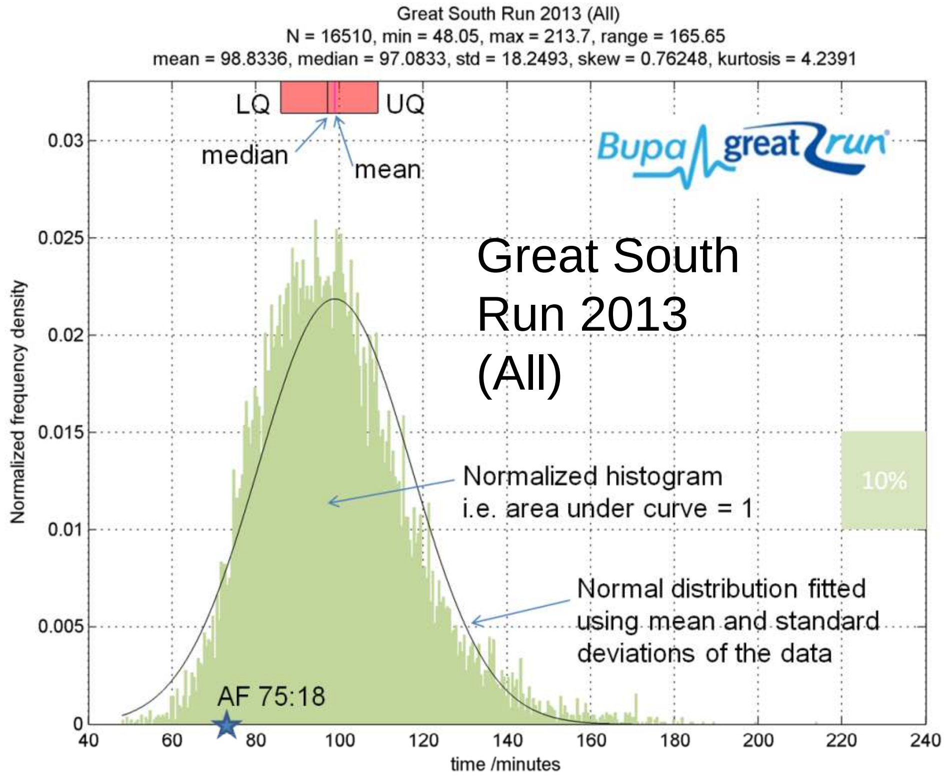
Areas of each bar are given, which correspond to the frequency of measurements of the quantity x in the range associated with each bar.

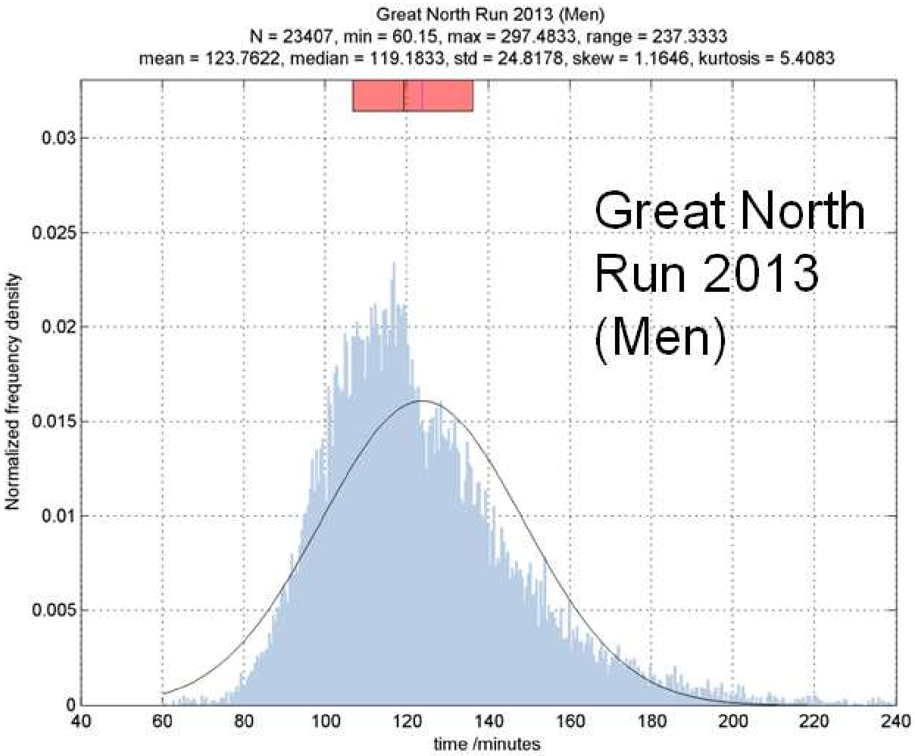
Histogram – Since the *bar areas correspond to frequency*, we can see that most of the measurements of x are actually around 50. This is a much more realistic representation of the distribution of the data. The bar chart is visually biased, which may cause us to misinterpret its meaning.

The shape of histograms, and approximating the *underlying probability distribution*

If many measurements are taken of a particular quantity, and this quantity has a *random* element, the resulting histogram often tends to have a symmetric 'bell curve' shape.
If this is the case then the *underlying probability distribution* of the quantity is said to be *Gaussian* or *'Normal.'* Note we can *approximately* measure a probability distribution by constructing a histogram and then *normalizing* it, i.e. scaling the frequency density such that the area of the histogram is unity.

A histogram which is *not* Normal in shape might be described as *skewed* (i.e. one tail is 'fatter' than the other either side of a peak) or possibly *multi-modal,* i.e. has several peaks. For the latter, one might model the distribution to be the sum of *different* normal distributions, each with a different peak position (mean) and width (standard deviation).





A *Normal distribution* is *symmetric* about the **mean** and has 'width' proportional to *the* **standard deviation** of the data

The *normalized histogram* (30s time intervals between 40 and 240 minutes) of times for male runners in the Great North Run of 2013 appears to be *bi-modal.* A fit of a single Normal distribution does not represent the data very well. A better model is for a Normal distribution of more elite runners (mean around 110 minutes) + another, wider distribution, with mean around 130 minutes, for everyone else.

The normalized histogram of a *Weibull distribution* can be set to be *skewed* to larger values of the random quantity x beyond what would be expected for a symmetric, Normal distribution.

PDF stands for 'probability density function.' It is effectively the frequency density of a normalized histogram.