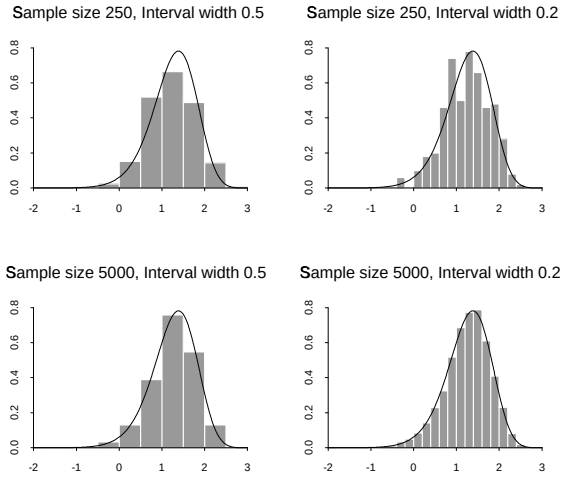
How Probability Density Histograms Approach the Probability Density Curve

A histogram based on a *large sample* using *narrow intervals* will be a good approximation to the population's *probability density curve*.



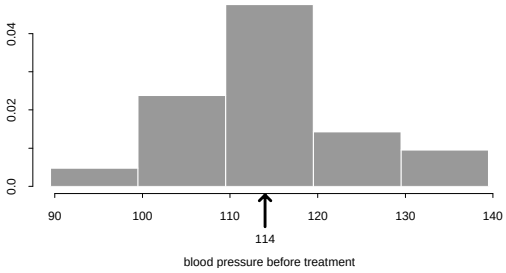
The Mean of a Data Set

The *mean* of a set of numbers, $x_1, \ldots, x_n$, is their arithmetic average:

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n}$$

If the distribution of the numbers is symmetrical, the mean will be at its centre.

If the distribution isn't symmetrical, the mean is the point where the histogram could be balanced:



The Median of a Data Set

The *median* of $x_1, \ldots, x_n$ is a number such that half the x_i are above it and half are below it (approximately).

From the experiment on calcium and blood pressure, here are the blood pressures after treatment of the 11 men who took a placebo:

124 97 113 105 95 119 114 114 121 118 133

Arranged in increasing order, these are:

95 97 105 113 114 114 118 119 121 124 133

The median is the number in the middle, 114.

If there are an even number of observations, the median is the average of the middle two.

The blood pressures after treatment of the 10 men who took calcium pills are, in order:

100 102 104 105 106 112 114 115 116 125

The median is $(106 + 112)/2 = 109$.

The Mean Versus the Median

The mean and the median are both measures of "location". How do they differ?

If the distribution is symmetrical, they don't differ.

But for a "skewed" distribution, they can be quite different. Consider this data:

2 8 15 3 29 5 8 1 20 17 6 5 31 44 10 12 23 62

Here are the stem plot, mean, and median:

```
0 | 12355688
1 | 0257
2 | 039              mean = 16.72
3 | 1
4 | 4                median = 11
5 |
6 | 2
```

Also, the median is a *resistant* measure of location — it doesn't change much if you change just a few data points. The mean is not resistant.

The Mean and Median of a Population

We can talk about the mean and median for the whole population, as well as for a sample from the population — even if the population is infinite.

The population mean is the "average" value, the point where the probability density curve for the population would balance.

The population median is the value where half the population is above and half is below.

If we look at *larger and larger samples* from a population, we will see that

> The sample mean gets closer and closer to the population mean.
>
> The sample median gets closer and closer to the population median.

Quartiles of a Distribution

The mean and median can't tell us how "spread out" a distribution is.

One way of doing this is to find the *quartiles*: the points 1/4 and 3/4 of the way along the sorted list of numbers (approximately).

Here again, in order, are the blood pressures after taking a placebo of the men in the control group:

95 97 105 113 114 114 118 119 121 124 133
105 114 121

The first quartile is 105, the third quartile is 121. (The second quartile is the median, 114.)

The exact definition of quartiles varies. Some software thinks the 1st and 3rd quartiles of this data are 109 and 120.

The Five-Number Summary

One way of summarizing the distribution of a set of observations is with the following five numbers:

Minimum 1st-quartile Median 3rd-quartile Maximum

For the blood pressures of the 10 men in the treatment group after taking calcium:

100 102 104 105 106 112 114 115 116 125

the five-number summary is

100 104 109 115 125

The Boxplot

The boxplot is a graphical display of the five-number summary. (Though later we'll see that boxplots often show "outliers" separately.)

Here are side-by-side boxplots of blood pressure after treatment for the two groups in the calcium and blood pressure experiment: