Descriptive Statistics

Suppose following data have been collected (heights of 99 five-year-old boys)

117.9	110.2	112.9	115.9	108.0	104.6	107.1	117.9	111.8
106.3	111.0	100.4	112.1	109.2	101.0	105.4	99.4	110.1
103.3	106.9	108.2	119.3	112.0	106.2	105.9	106.9	109.3
105.9	110.0	106.7	108.5	107.7	114.3	108.6	104.6	113.7
116.7	103.5	96.1	110.8	97.2	109.6	110.5	105.9	106.2
107.4	114.9	110.3	104.8	99.2	119.2	111.4	103.0	110.1
105.8	101.5	105.9	107.6	97.1	113.3	109.4	109.4	110.8
106.3	108.1	109.6	102.4	110.4	110.1	115.3	102.9	111.2
99.4	105.7	119.5	109.3	112.8	108.2	117.0	106.8	105.4
108.7	109.2	97.1	103.3	108.8	116.3	115.5	114.9	101.1
104.1	110.8	112.7	105.6	99.9	111.1	109.4	109.1	110.7

Initial impression is of indigestibility

Sorted data

Put data into ascending order

96.1	101.5	105.4	106.3	108.1	109.3	110.3	111.8	115.3
97.1	102.4	105.4	106.3	108.2	109.4	110.4	112.0	115.5
97.1	102.9	105.6	106.7	108.2	109.4	110.5	112.1	115.9
97.2	103.0	105.7	106.8	108.5	109.4	110.7	112.7	116.3
99.2	103.3	105.8	106.9	108.6	109.6	110.8	112.8	116.7
99.4	103.3	105.9	106.9	108.7	109.6	110.8	112.9	117.0
99.4	103.5	105.9	107.1	108.8	110.0	110.8	113.3	117.9
99.9	104.1	105.9	107.4	109.1	110.1	111.0	113.7	117.9
100.4	104.6	105.9	107.6	109.2	110.1	111.1	114.3	119.2
101.0	104.6	106.2	107.7	109.2	110.1	111.2	114.9	119.3
101.1	104.8	106.2	108.0	109.3	110.2	111.4	114.9	119.5

Much more can now be seen but hardly succinct:

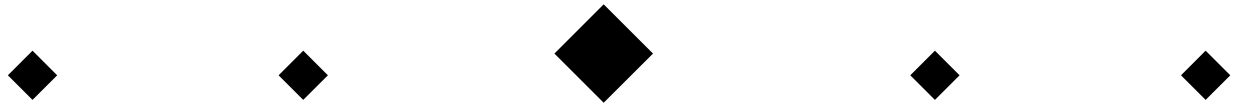
need simple summaries

- all values are between 96.1 cm and 119.5 cm
- ‘middle’ value is 108.7 cm - indication of *location*
- values quarter of way up and down data are 105.6 and 111.1 cm - indication of *spread*
- these comprise the *five number summary* for these data
- ‘Middle’ value is known as the *Median*
- Also upper and lower quartiles
- Definitions need some care

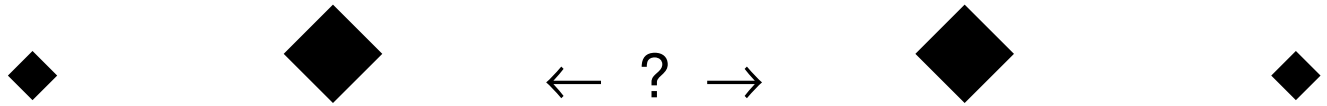
Definition of Median

Median is ‘middle’ value.

For five values there is a definite ‘middle’ one



But not for four values

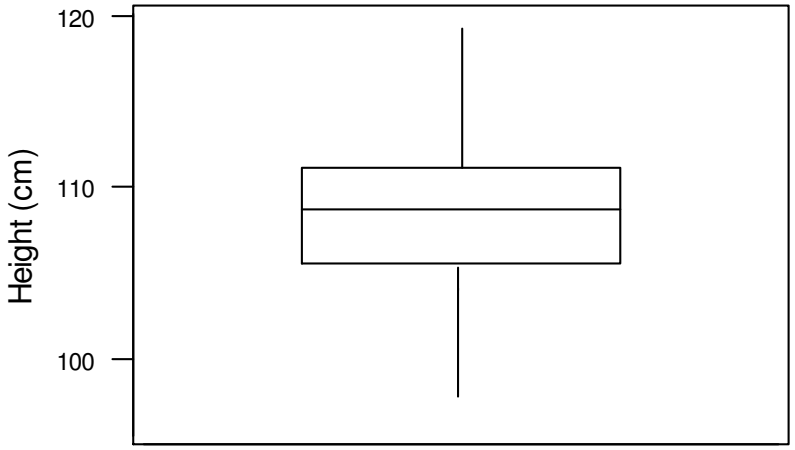


Definition of Median

- Median is the middle value of a sample when the sample size is odd
- Median is the average of the two ‘middle’ values when sample size is even
- Definitions for quartiles are similar and given more precisely in appendix 1 of handout (see web page)

Graphical displays of the data

Display the five figure summary - Box and Whisker plots



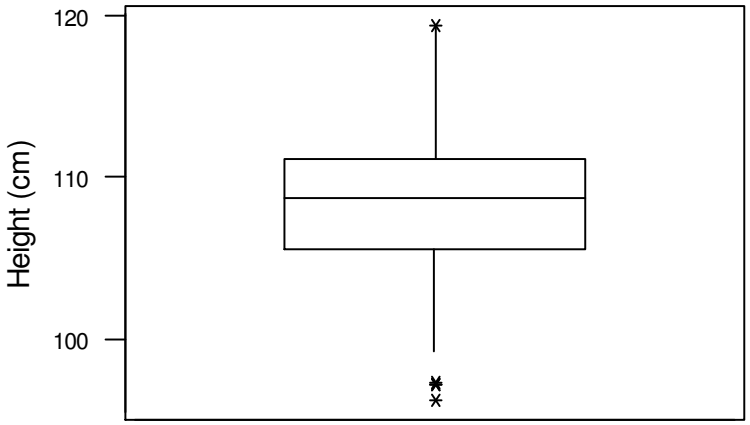
Top of box is upper quartile

Line across box is at level of median

Bottom of box is lower quartile

Whiskers can extend to maximum and minimum but …

Alternative Box and Whisker plots



Whiskers extend to only a given multiple of the box height

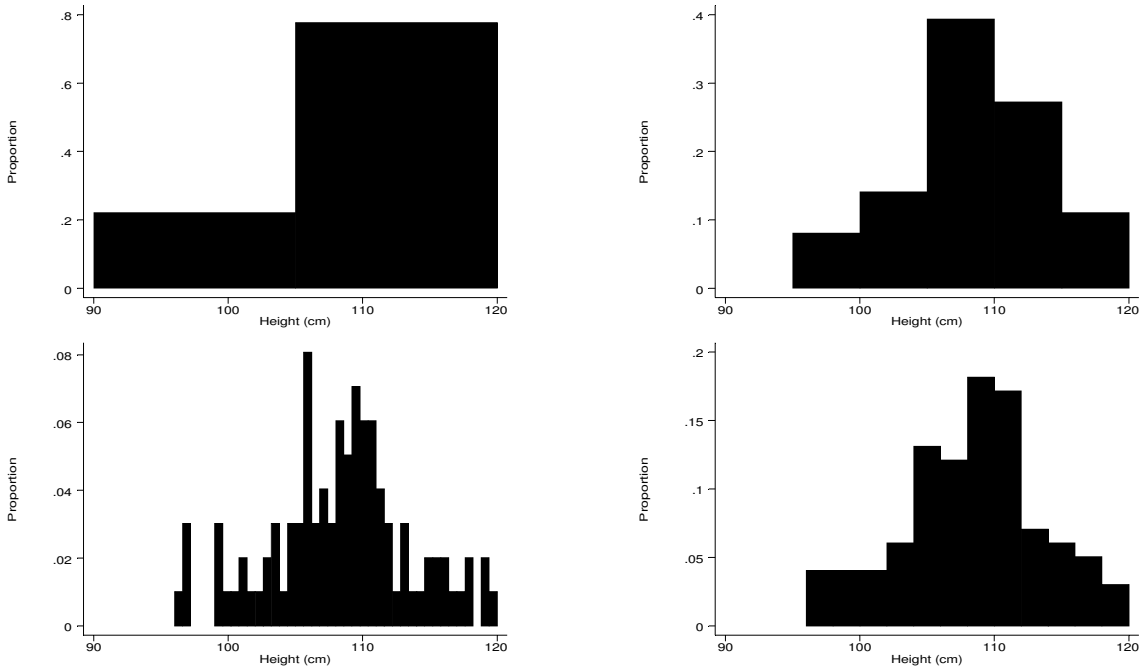
This is intended to allows outliers to be seen

Histograms

A useful way to display all the data in a sample:

Divide range into ‘bins’ and plot numbers/fractions in bins

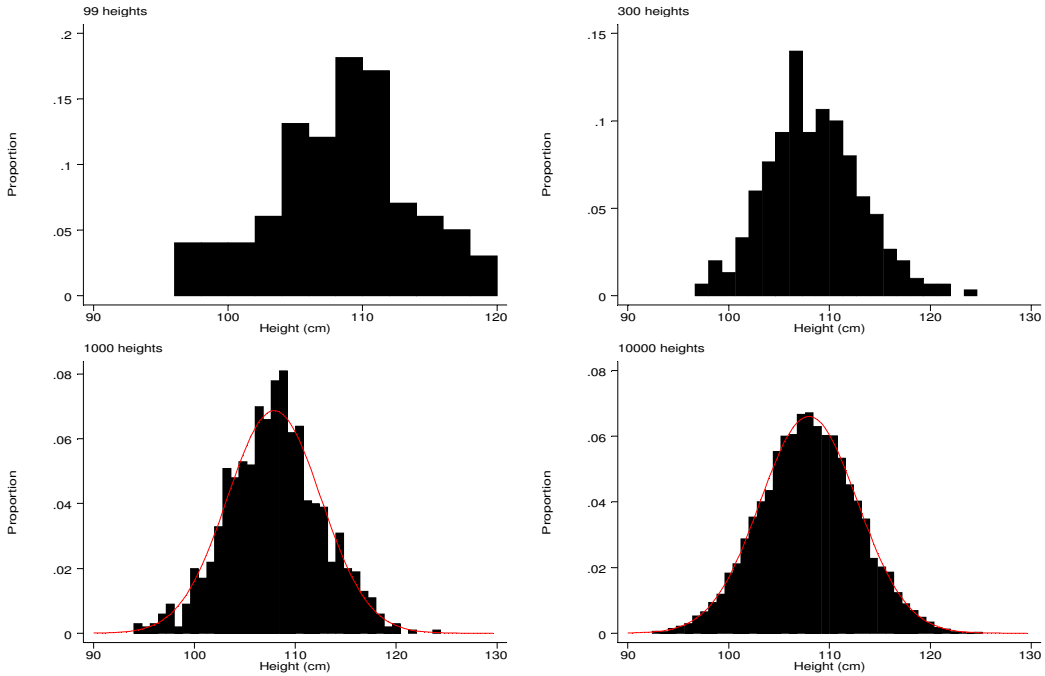
Choose bin width judiciously



(2, 6 15 and 50 bins: clockwise)

Histograms for increasing sample size

Samples of sizes 99, 300, 1000, 10000



Form gets more regular as sample size increases - tends to a limiting form - notion of a population and *inferential statistics*

Populations, Parameters and Inference

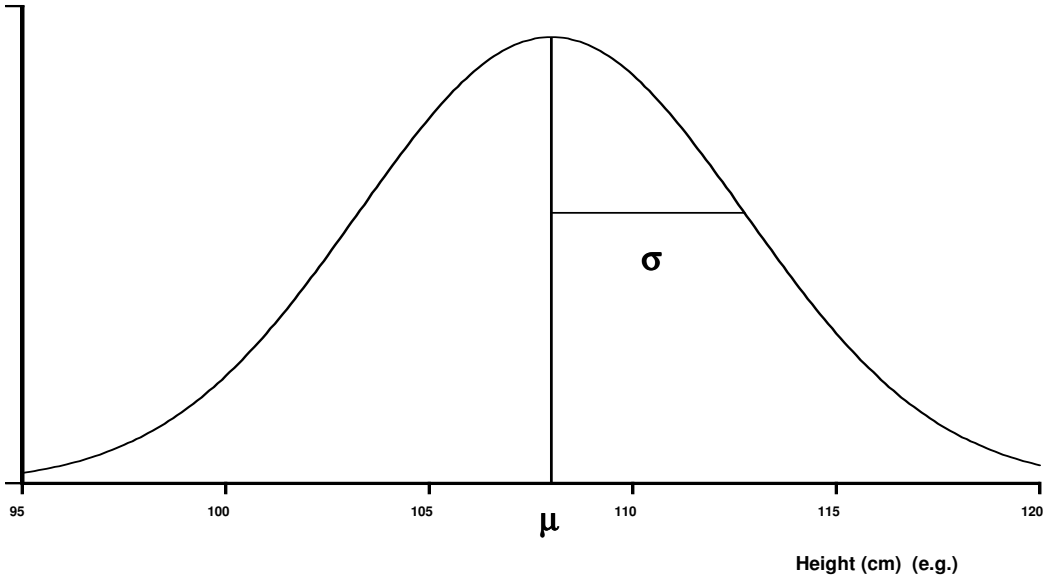
- Sample rarely of interest in its own right
- Value lies in its representativeness of a greater whole, about which we wish to learn
- *Inferential Statistics* seeks to draw conclusions about a population from a sample
- Sample must be representative of population - random selection
- Population often defined in terms of unknown *parameters* μ, σ
- *Estimated* by sample analogues or *statistics*, such as *m* or *s*

Describing Populations

Distributional form: we consider only the Normal distribution: symmetric bell-shaped curve

Associated parameters:

measure of location	μ	(mean)
measure of spread	σ	(SD)



Sample Estimates of μ and σ

Estimate μ by sample mean (arithmetic mean - simple 'average')

$$m = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

where $x_1, x_2, \ldots, x_n$ are the values in the sample (n.b. size n)

Estimate of σ is sample standard deviation (SD):

$$s = \sqrt{\frac{(x_1 - m)^2 + (x_2 - m)^2 + \ldots + (x_n - m)^2}{n - 1}}$$

For sample of 99 boys m = 108.34 cm s = 5.21 cm (note units)

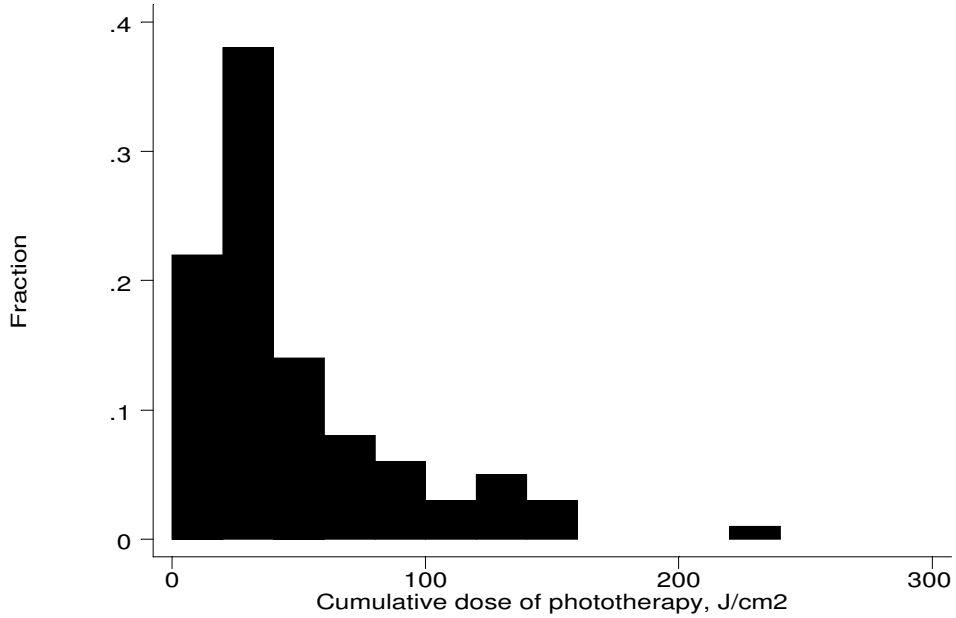
{cf median 108.7 cm quartiles, 105.6 cm, 111.1 cm IQR = 5.5 cm}

- little need these days to work directly with these formulae
- indeed, using a computer is preferable as well as labour-saving
- appendix 2 describes some peculiarities of SD formula for those interested
- mean and SD appropriate summaries for a Normal distribution
- what about non-Normal data?

Skewed data

Medians and quartiles - always legitimate tools

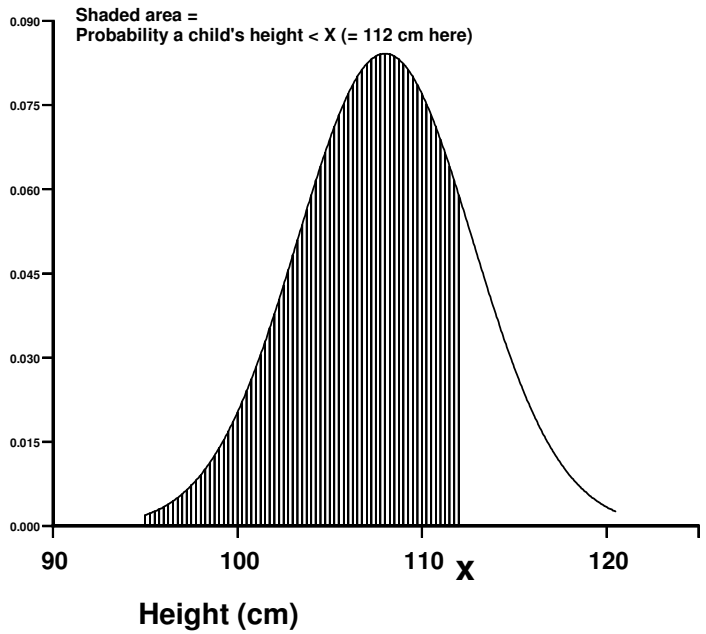
means and SDs - less easy to give general answers



Non-Normal data, especially skewed unduly mean influenced by large values in sample

Quantitative use of the Normal curve

Interpretable feature of curve is area under curve



Area under whole curve is 1: area under curve up to *X* is *P*

Probability randomly chosen member of population has value < *X*

Cumulative probabilities

Given X, and μ, σ the value of P can be found but you need a computer

P is the cumulative probability at X

In example $\mu = 108$ cm $\sigma = 4.7$ cm and $X = 112$ cm

Gives $P = 0.8026$

I.e. about 80% of five year old boy have height < 112 cm

Other probabilities

Some obvious results:

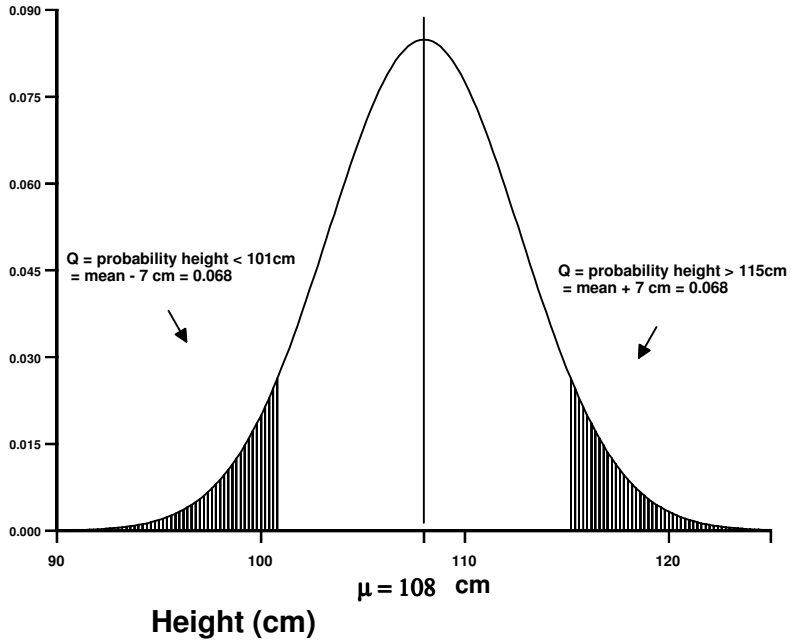
probability value is larger than X is 1-P

I.e. 0.1974 or about 20% of boys have height above 112cm

Symmetry means that it follows that proportion 0.1974 of boys have height below 104 cm

(104 cm is 4 cm below mean, and 112 cm was 4 cm above mean)

Allows virtually all other probabilities to be found



Can also ask inverse question: given *P* what is *X*?

For example, 3% of boys have height less than what?
I.e. $P = 0.03$ and $X = 99.16$ cm
X is the *inverse cumulative probability* of *P*.

Z-scores

P depends on *X* but there is a simpler structure underlying the relationship.

X can be expressed as being 'so many' SDs from the mean

I.e. $X = \mu + Z\sigma$ and P depends on X only through Z

Allows quicker apprehension of the Normal distribution

(e.g. 95% within ± 2 SDs of mean, 68% within ± 1 SD of mean)

Z	0	1	2	1.96	0.675	2.58
P	0.5	0.84	0.977	0.975	0.75	0.995
Proportion within *Z* SDs of mean (=2*P*-1)	0	0.68	0.954	0.95	0.5	0.99

Why is the Normal Distribution Important?

- Empirically important - many variables are Normal
- There are reasons why some variables are Normal e.g. polygenic control - see appendix 3
- Means are ‘more Normal’

Distribution of sample means

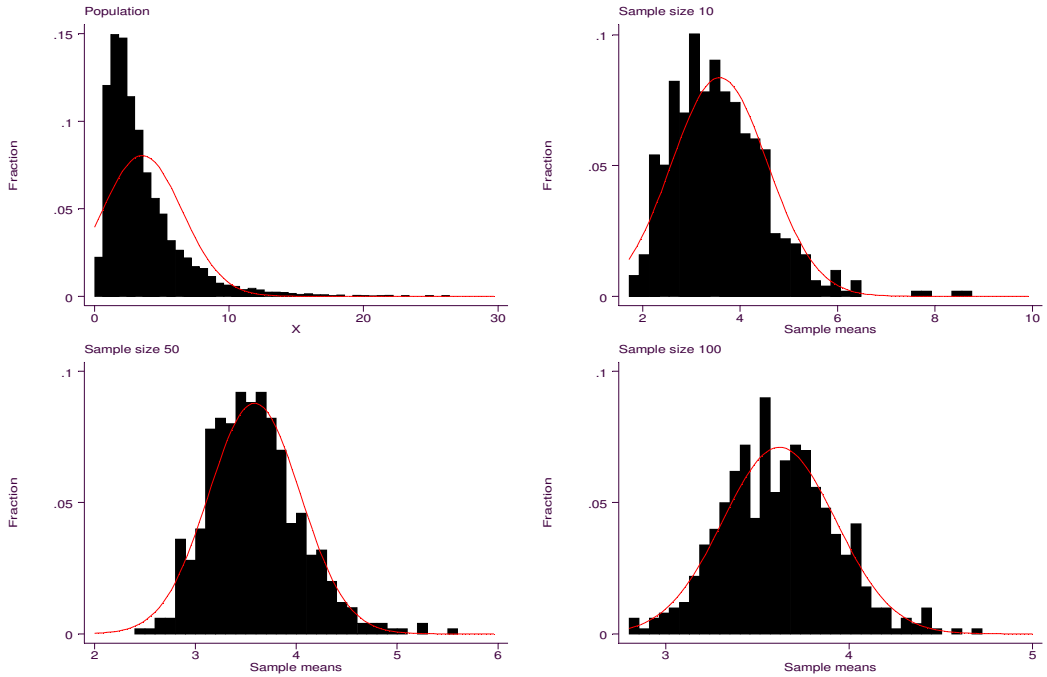


Figure 8: histograms of 10000 observations form a population and of 500 sample means for samples of size 10, 50 and 100.

Assessing Normality

- Quite a preoccupation - not as necessary as might be thought

  cf. Distribution-free or non-parametric methods

- Many methods that assume Normality are robust (or other types of assumption are more important)

- Assessing Normality is not straightforward and its importance depends on purpose of analysis

Assessing Normality - a Quick and Dirty method

About 2½% of population lies below $\mu - 2\sigma$

About 16% of population lies below $\mu - \sigma$

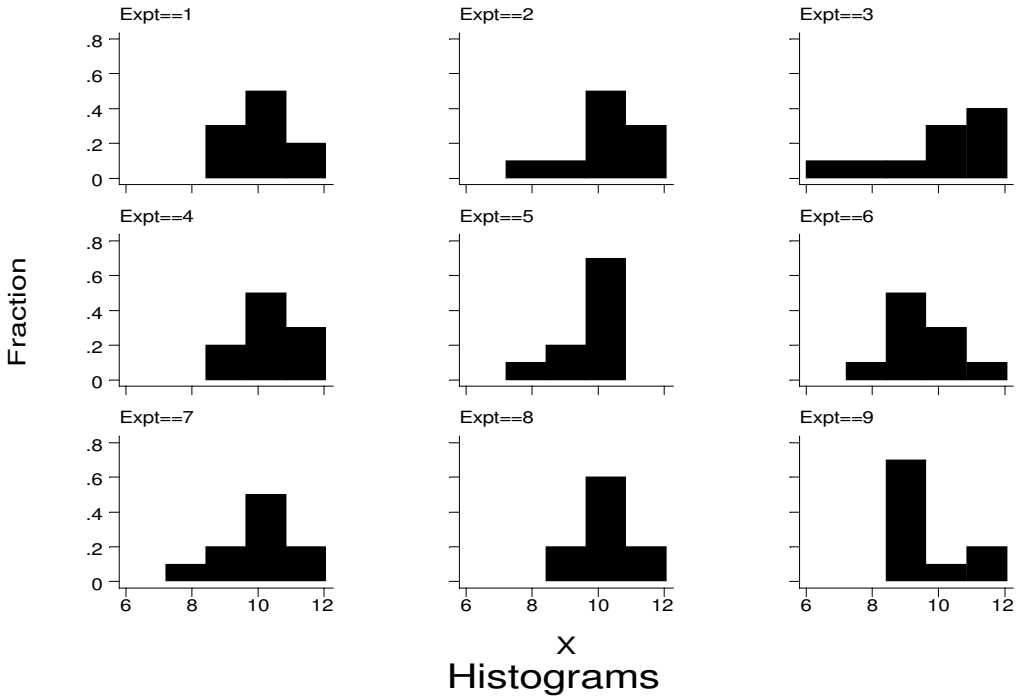
- If variable is positive and $\mu < \sigma$ then Normal distribution highly questionable
- If variable positive and $\mu \approx 2\sigma$ then Normal distribution might be just about OK
- If value lies inbetween then appropriateness should be judged accordingly
- Handy method when reading papers - not at all foolproof

Plot data: I histogram

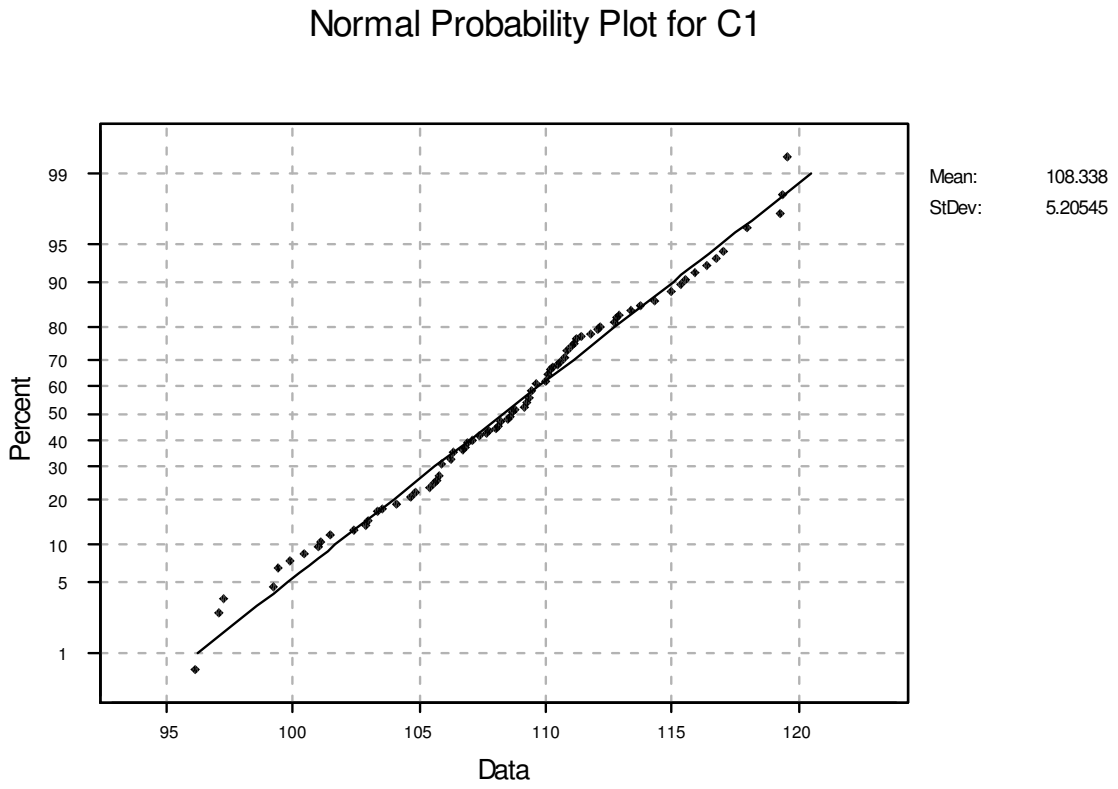
Histogram is obvious way to see if shape of distribution of sample is ‘close’ to that of Normal curve

In small samples this is not easy:

Samples of size 10 from distribution *known* to be Normal



Plot data: II Normal Probability Plot

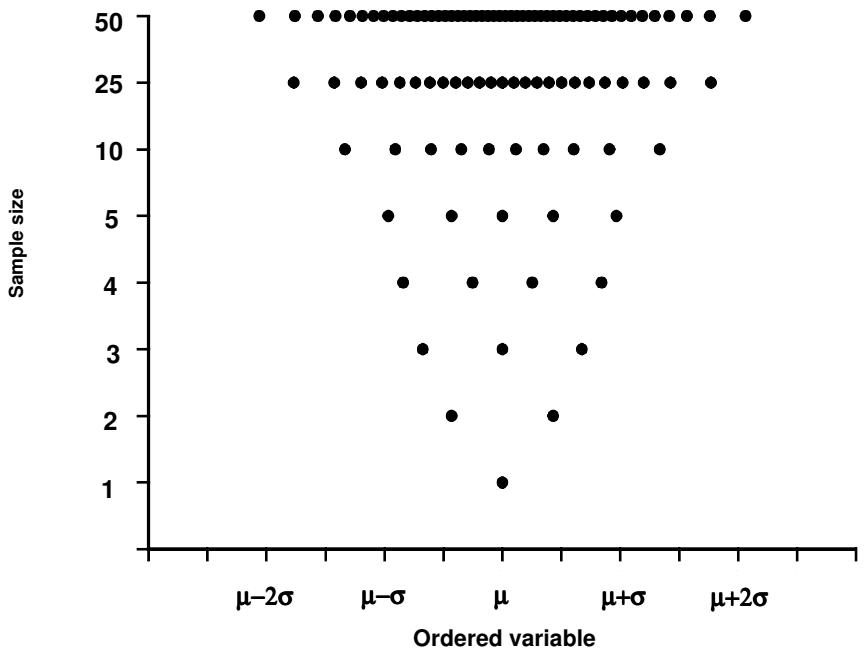


Plot ordered sample values against appropriate y-value

If data are Normal then straight line will result

How does a Normal probability plot work?

Depends on ‘expected’ positions for ordered values in a Normal sample



Normal plot for non-Normal data