Clinical Biostatistics

Suggested Answers: Summarising Data

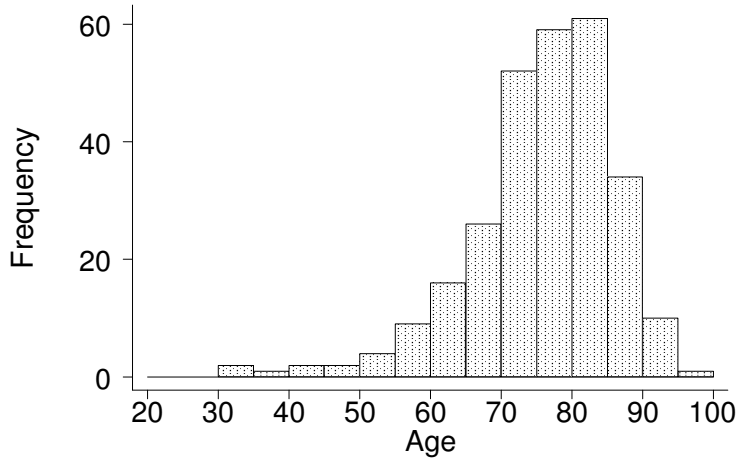
(a) *What kind of diagram is this?* We would call this a histogram

(b) *How would you describe the shape of the distribution?* The distribution is has a slight positive skew or is skew to the right, the longer tail being on the right.

(c) *The class intervals are 5mm Hg wide . In what interval would a diastolic blood pressure of 70 be put?* The convention is to include the lower limit in the interval, thus the interval which starts at 65 and finishes at 70 does not actually include 70, which is put into the interval 70 to 75.

(d) *What is the shape of the distribution?* The long tail is on the left, so the distribution is skew to the left or negatively skew .

(e) *Could the presentation of the figure be improved?* Since age is continuous, there is no need to put gaps between the vertical bars . These gaps suggest that there are gaps in the distribution, which there are not . The distribution could be shown as a conventional histogram:



(f) *What is meant by `inter-quartile range'? Why are the median and inter-quartile range' used here, rather than mean and standard deviation?* T he median is the central point in a set of observations when they are ordered from the smallest to the largest . Hence half of the observations lie above the median and half lie below it . Because it is the central point in the ordered data, it is not affected by extreme values and so may be a better descriptive summary measure for skewed data such as alcohol consumption . The lower and upper quartiles are the points such that a quarter of the observations lie below the lower quartile and a quarter lie above the upper quartile . The interval between the lower and upper quartile is known as the inter-quartile range . Hence the inter-quartile range gives a range of values which includes half of the observations . It, too, is unaffected by the most extreme observations.

(g) *What is the shape of the distribution of alcohol consumption?* The medians are much closer to the first quartiles than to the third quartiles . The distribution of alcohol consumption is therefore positively skew for both groups.

(h) *What is meant by mean and standard deviation?* The mean or average is the sum of all the observations divided by the number of observations and gives a measure of the centrality of the data . The standard deviation is a measure of the scatter of the data about the mean value and is measured in the same units as the data itself . It is the square root of the variance where the variance of the sample is given by the sum of the squared deviations from the mean divided by the total sample size minus one.

(i) *What do they tell us here about the shape of the distribution?* If the data were symmetrically distributed, we would expect about 2.5% of the observations to be less than the mean minus 2 standard deviations . For a variable which can only take positive values, we can deduce positive skewness where the standard deviation is bigger than half the mean . However, here the variable is excess insulation where negative values are possible and so we cannot deduce the shape of the distribution from these values.

(j) *What is wrong with this statement?* The three quartiles are the values which divide the distribution into four equal parts . They are points, not groups of people . They therefore do not have averages . The word is sometimes incorrectly used to mean one of the four groups into which the three quartiles divide the population, more correctly called `fourths' . In this sense there is no `middle quartile' because there are four of them and each fourth contains 25% of the population . (The newspaper in question is notorious for misprints, so this may not be quite what the author wrote.)

(k) *What is the more usual name for the 'middle quartile'?* The middle quartile is the median.

(l) *Which Honourable Member is correct, if any, and why?* Mr Lloyd is correct that the mean is not the same as the median but his definition of the median is wrong . It is the middle number when the numbers are arranged in ascending order and not the difference between the minimum and maximum values . So using his first example (2, 2, 5, 6, 7), the median here would be 5 and not 3.5 as asserted . He is correct that altering the extreme values will not change the median but his argument is confused.

Mr Carrington is correct that the median is not the mid-point between the first number and the last but is wrong in his definition . His definition appears to be referring to the mode of a distribution, i.e . the category which has the highest frequency, although his explanation is not very clear . His definition of the average is wrong (the sample multiplied by the number of items) . He is wrong in stating that the median will change if the number at the bottom of the scale changes . This will not change the median but will alter the mean.

(m) *What would be the effect on the skewness of the earnings distribution if the minimum wage were made a fixed proportion of the median, assuming that this figure was then higher than the current wage of some members of the population?* The distribution would be become more positively skew . The smallest wages would be increased and the short, left-hand tail of the distribution would be made even shorter.