Section 6.4: cont'd

Wednesday, December 03, 2014
8:38 AM

So, once you've made your measurements (on population or sample), you can summarize it either

→ graphically (pie charts, bar graphs, line graphs, histograms)

→ numerically (giving average, highest/lowest value)

interpreting histograms

symmetry - see handout

number of peaks:

unimodal (one peak)
bimodal (two peaks)
multimodal (many peaks)

measures of centre: (where the "middle" of the data set lies)

mean (aka average)

median - middle value when data set is written as an ordered list

(if there are an even number of points, average the two middle ones)

example: Four former CST students were surveyed after graduation and their starting salaries at new jobs were found to be:

$46,000

$52,000

$38,000

$500,000

Calculate the mean and median

mean = $159,000

median = $49,000 (average of $46,000 and $52,000)

note: the mean is strongly influenced by "outliers" (unusually small or large data points)

→ median is a better measure of

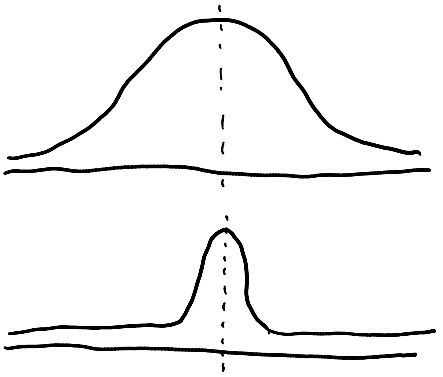
centre for asymmetric distributions or data sets with outliers

two possible reasons for an outlier:

1. error, which you should go back and correct
2. it is simply an unusual data point

measures of variability (measures of spread):

measure the "width" of the data set



two distributions with same mean but different "spread"

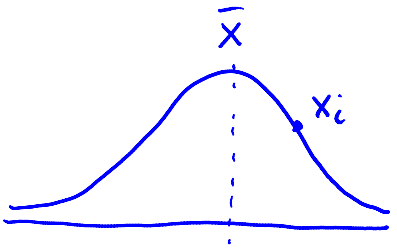
range ≡ maximum value − minimum value

↑ ↑

heavily influenced by outliers so its pretty useless

standard deviation - measure of how far on average each data point

standard deviation – measure of how far on average each data point is from the mean



$(x_i - \bar{x})$

IMPORTANT: I will not ask you to calculate standard deviations

but I will ask you to interpret them

example:

which data set has a higher standard deviation? or are they the same?

a) Set 1: 31, 33, 34, 35, 37

Set 2: 31, 31, 34, 37, 37

31 and 37 are further away from the middle value

$\therefore$ Set 2

b) Set 1: 31, 33, 34, 35, 37

Set 2: 41, 43, 44, 45, 47

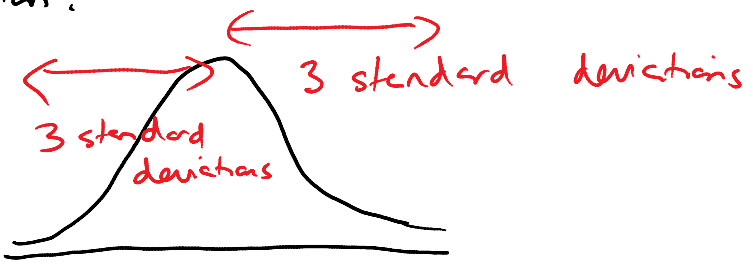
same! shifting the data set doesn't change the spread

c) Set 1: 1, 2, 3, 4, 5
Set 2: 2, 4, 6, 8, 10

set 2 (the points are farther away from each other)

in fact doubling the values increased the standard deviation by a factor of 2

how can you estimate the standard deviation from a distribution?



for most data sets, almost all of the data lies within 3 standard deviations of the mean

so standard deviation $\approx \frac{\text{range}}{6}$ (ish)