Chp 2 Describing Distributions w/ Numbers

Two numerical description summaries of data

(A) mean

(1) center

a.k.a. "average" of a set of #s; add all #s up and divide by the number of #s

notation: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ mean

(2) spread

Standard deviation $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

s is square root of variance

variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

measures the average squared distance away from the mean

$x_1, x_2, \ldots, x_n$ are the individual observations (numbers in data set)

Note: For s, we divide by n-1 (not n) because that's the degrees of freedom, since $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ so knowing (n-1) x_i determines x_n.

(B) median

a.k.a. "middle" or "midpoint" of set of #s;

① put all data in increasing order

② find middle value

i.e. if n data values, median is the $\frac{n+1}{2}$ value

(median is denoted as M)

5-number summary: min, Q_1, M, Q_3, max

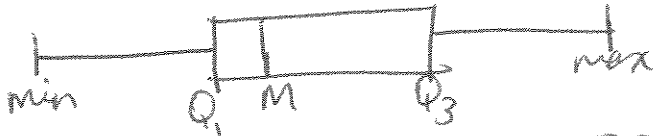
Q_1 = median of lower half of data (called 1st quartile or lower quartile)

Q_3 = median of upper half of data (called 3rd quartile or upper quartile)

min = lowest data value

max = highest data value

☆ can show 5-number summary as box-plot



Ex 1 Weights of defensive linemen (in pounds) for 2007 Dallas Cowboys.

300 300 295 255 298 298 300 310 300

& weights of offensive linemen.

312 305 340 320 366 324 309 315 305 305

(a) find mean for (i) offensive & (ii) defensive

(b) find median for make stemplot & (i) offensive (ii) defensive

(c) give 5-number summary (i) offensive (ii) defensive

Chp 2 (cont)

Ex 1 (cont)

(d) give box plot for (i) offensive (ii) defensive

Ex 2 For these test scores, find (a) mean and (b) standard deviation.

79, 93, 72, 86, 81

Chp 2 (cont)

Other Vocab/Tidbits:

- **IQR** = Interquartile range = distance between Q_1 and Q_3 i.e. $IQR = Q_3 - Q_1$
  - useful for determining outliers
- **outlier**: a data value that is more than 1.5IQR above Q_3 or below Q_1 (mostly used when looking at large sets of data)
- mean is not resistant measure of center, but median is. (i.e. outliers affect mean greatly but don't affect median much)
- for symmetric distribution, $\bar{x} = M$
- In a skewed distribution, $\bar{x}$ is farther out in long tail (usually) than M
- box plots show less detail than histogram or stemplots, so use them side-by-side to compare distributions
- **ALWAYS PLOT DATA** (don't rely only on center and spread as numbers)
- s facts: ① goes along w/ $\bar{x}$ (not M)
  - ② $s \geq 0$ why?
  - ③ s units of measure = same as observation (and mean)
  - ④ s also not resistant; outliers really affect s

M1070

How do I choose a numerical summary?

[B] (5-number summary) is usually better if (and box plot)
describing a skewed distribution or a distribution w/ strong outliers

[A] (mean & s) is useful only for reasonably symmetric distributions that are free of outliers

---

Ex 3 For data in Ex 1, are there any outliers? (calculate IQR first.) If so, what are they?

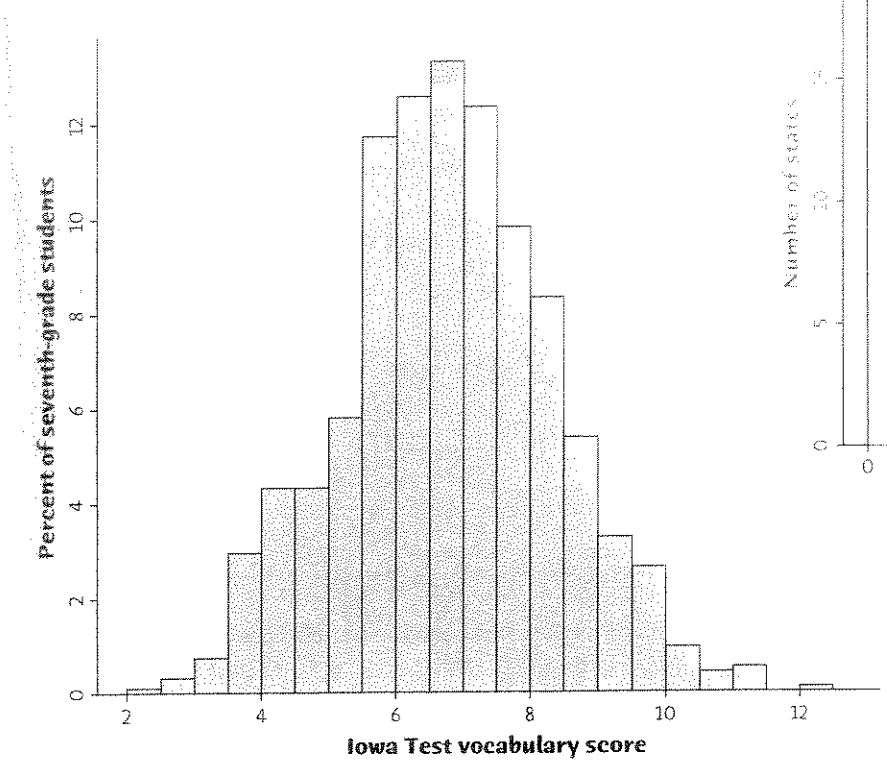
Chp 2 (cont)

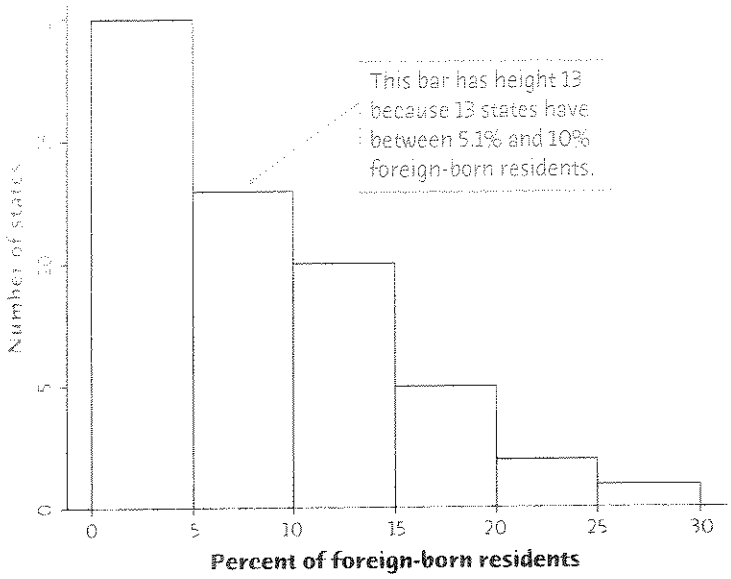
EX 4 For each data set, calculate $\bar{x}$ and s. Then make stemplot & comment on shape of each distribution

Data A	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74
Data B	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.5

Chp 2 (cont)

Ex 5 For each set of data pictured, which numerical summary would you choose to describe data?





23	0
24	1
25	
26	5
27	
28	7
29	
30	2 5 9
31	3 9 9
32	0 3 3 6 7 7
33	0 2 3 7