Preliminary Statistics September 2013

Answers to class exercise

Problem Set 1: Descriptive Statistics

1. (a) (i) $\sum_{i=1}^{4}(i+4) = (1+4)+(2+4)+(3+4)+(4+4) = 26$

(ii) $\sum_{i=1}^{n} 2 = 2n$

(iii) $\sum_{i=1}^{3} 3^i = 3^1 + 3^2 + 3^3 = 39$

(b) (i) $X_1 + 2X_2 + 3X_3 + 4X_4 + 5X_5 = \sum_{i=1}^{5}(iX_i)$

(ii) $(X_1 + Y_1) + (X_2 + Y_2) + \cdots + (X_k + Y_k) = \sum_{i=1}^{k}(X_i + Y_i) = \sum_{i=1}^{k} X_i + \sum_{i=1}^{k} Y_i$

2. Mean, median and mode for the two cohorts are given in the table below.

	Cohort 1	Cohort 2
Mean	40	38
Median	40	40
Mode	20	50
N	100	100

For Cohort 1 the mean is calculated as such:

$$\frac{\sum n_i X_i}{\sum n_i} = \frac{(70 \times 20) + (50 \times 20) + (40 \times 20) + (20 \times 40)}{(20 + 20 + 20 + 40)} = \frac{4000}{100} = 40$$

For Cohort 2 the mean is calculated as such:

$$\frac{\sum n_i X_i}{\sum n_i} = \frac{(50 \times 40) + (40 \times 20) + (30 \times 20) + (20 \times 20)}{(40 + 20 + 20 + 20)} = \frac{3800}{100} = 38$$

Cohort 1 data are skewed to the right, so mode ≤ median ≤ mean.

Cohort 2 data are skewed to the left, so mean ≤ median ≤ mode.

It is not clear that the reorganisation was an improvement. It reduced the variance, cutting the number of fails, but also cutting the number of firsts.

3. (a) The mean is calculated as: $\bar{X}_{mean} = \frac{27+31+24+33+29+34+30+26}{8} = 29.25\ miles/gallon$; though the median is $\bar{X}_{median} = \frac{(29+30)}{2} = 29.5\ miles/gallon.$

Nr.	1.	2.	3.	4.	5.	6.	7.	8.
X_i	24	26	27	29	30	31	33	34

There is no mode in the sample.

(b) The sample variance is given by:

$Var(X_i) = \frac{\sum(x_i - \bar{x})^2}{(N-1)} =$

$\frac{(27-29.25)^2+(31-29.25)^2+(24-29.25)^2+(33-29.25)^2+(29-29.25)^2+(34-29.25)^2+(30-29.25)^2+(26-29.25)^2}{(8-1)} =$

11.93. The standard deviation is just the square root of the sample variance, so it is equal to 3.45.

4. The summary statistics for the stratified[1] sample for income are provided in the following table.

	n_i	t_i	$\bar{\bar{X}}_i$	N_i	w_i	$w_i \times \bar{\bar{X}}_i$
Males	50	1,400,000	28,000	800	0.8	22,400
Females	50	1,100,000	22,000	200	0.2	4,400
	n=100			N=1,000		$\bar{X} = 26{,}800$

Subscript i refers to whether being male or female, n_i is the sample used from each of the two genders, t_i is the total income reported by the sample males and females respectively. $\bar{X}_i$ shows the average income in each sample $[\bar{X}_i = \frac{t_i}{n_i}]$. N and N_i refer to the total, and total male and female population. The overall average income is estimated using a weighted average $\bar{X}_w = \sum w_i \bar{X}_i = (0.8 \times 28{,}000) + (0.2 \times 22{,}000) = 26{,}800$, with $w_M = \frac{800}{1{,}000} = 0.8$ and $w_M = \frac{200}{1{,}000} = 0.2$ being the respective weights for the two subpopulations male (M) and female (F).

[1] Stratification is the process of dividing members of the population into homogeneous subgroups before sampling so that certain subpopulations can be analysed.