EXAM C QUESTIONS OF THE WEEK

S. Broverman, 2007

Week of September 10/07

When data is in interval grouped form, the usual assumption that is made is that within each interval the data points for that interval are uniformly distributed on the interval. This means that for interval $(a, b]$, a data point within that interval would have pdf $f(x) = \frac{1}{b-a}$.
For an interval grouping with interval endpoints $c_0 = 0 < c_1 < c_2 < \cdots < c_{k-1} < c_k$,
and with n_i data points in interval $(c_{i-1}, c_i]$, this results in an empirical estimate of the first moment being $\sum_{i=1}^{k} \frac{n_i}{n} \cdot \frac{c_{i-1}+c_i}{2}$, where $n = \sum_{i=1}^{k} n_i$ (total number of data points).

Suppose that the following pdf is assumed for data in the interval $(a, b]$: $f(x) = \frac{2(x-a)}{(b-a)^2}$.
Find the estimate of the first moment based on the grouped data set.

A) $\sum_{i=1}^{k} \frac{n_i}{n} \cdot \frac{c_{i-1}+c_i}{2}$ B) $\sum_{i=1}^{k} \frac{n_i}{n} \cdot \frac{2c_{i-1}+c_i}{3}$ C) $\sum_{i=1}^{k} \frac{n_i}{n} \cdot \frac{c_{i-1}+2c_i}{3}$

D) $\sum_{i=1}^{k} \frac{n_i}{n} \cdot \frac{3c_{i-1}-c_i}{2}$ E) $\sum_{i=1}^{k} \frac{n_i}{n} \cdot \frac{3c_i-c_{i-1}}{2}$

The solution can be found below.

Week of September 10/07 - Solution

An underlying relationship upon which the estimate is based is

$$E[X] = \sum_{i=1}^{k} E[X \mid c_{i-1} < X \le c_i] \cdot P(c_{i-1} < X \le c_i) \; .$$

When the uniform distribution is assumed for each interval, this results in $E[X \mid c_{i-1} < X \le c_i] = \frac{c_{i-1}+c_i}{2}$. Also, the estimate of $P(c_{i-1} < X \le c_i)$ is $\frac{n_i}{n}$ (fraction of total set of data points that lie in the interval).

With the new pdf, we get $E[X \mid a < X \le b] = \int_a^b x \cdot \frac{2(x-a)}{(b-a)^2}\, dx$.
A somewhat simplified approach is to find
$E[X - a \mid a < X \le b] = \int_a^b (x-a) \cdot \frac{2(x-a)}{(b-a)^2}\, dx = \int_a^b \frac{2(x-a)^2}{(b-a)^2}\, dx = \frac{2(b-a)}{3}$,
so that $E[X \mid a < X \le b] = E[X - a \mid a < X \le b] + a = \frac{2(b-a)}{3} + a = \frac{a+2b}{3}$.

Then, $E[X \mid c_{i-1} < X \le c_i] = \frac{c_{i-1}+2c_i}{3}$, and the estimated mean of X is $\sum_{i=1}^{k} \frac{n_i}{n} \cdot \frac{c_{i-1}+2c_i}{3}$.