Central tendency measures for grouped univariate data

BEA140 Quantitative Methods - Module 2

Central Tendency

In these slides we will look at a number of central tendency measures for grouped univariate data.

Note: While it is usually preferable to work with ungrouped (raw) data, you may come across grouped data:

(i) that was never collected in ungrouped (raw) form to begin with; or

(ii) where the corresponding ungrouped (raw) data has been lost or is simply unavailable.

Consequently it is important for us to also know how to work with such data sets.

Central Tendency - Travel Time Example

We previously looked at the following sample of ungrouped (raw) univariate data consisting of 9 travel times (in minutes):

15	29	8	42	35	21	18	42	26

If we group our data in to 10 minute intervals, we obtain:

time	frequency (f_i)
$0 \leq x_i < 10$	1
$10 \leq x_i < 20$	2
$20 \leq x_i < 30$	3
$30 \leq x_i < 40$	1
$40 \leq x_i < 50$	2

Central Tendency - Class Marks

Recall that in statistics **central tendency measures** attempt to give us an idea of where the *middle* of some data points is.

To obtain central tendency measures for grouped data, we make use of the **class mark** of a group/interval, which is the mean of the minimum and maximum number in the group/interval.

Note: Measures obtained using group data are likely to differ from those obtained using the same corresponding ungrouped (raw) data, since class marks may not be representative of how data is distributed within each group/interval.

Central Tendency - Class Marks

Example: Consider again our data set on travel times, since travel times have been recorded in whole minutes:

(i) the minimum possible number in the group '$10 \leq x_i < 20$' is 10 minutes; and

(ii) the maximum number in the group '$10 \leq x_i < 20$' is 19 minutes.

Therefore the class mark of '$10 \leq x_i < 20$' is $\frac{10+19}{2} = 14.5$ minutes.

Filling out the rest of the class marks gives us:

time	frequency (f_i)	class mark (X_i)
$0 \leq x_i < 10$	1	4.5
$10 \leq x_i < 20$	2	14.5
$20 \leq x_i < 30$	3	24.5
$30 \leq x_i < 40$	1	34.5
$40 \leq x_i < 50$	2	44.5

Central Tendency - Discrete vs. Continuous

Note that the data for the example on the previous slide has been rounded to every whole number (minute).

(i) if the data was rounded to every half minute rather than every minute, then the maximum number in '$10 \leq x_i < 20$' would be 19.5 minutes, which gives a class mark of $\frac{10+19.5}{2} = 14.75$ minutes; and

(ii) if the data was continuous, then the maximum number in '$10 \leq x_i < 20$' is *essentially* 20 minutes, which gives a class mark of $\frac{10+20}{2} = 15$ minutes.

Hence when working with a data set it is important to consider whether your data is discrete or continuous, and if it is discrete to also consider *how* discrete it is (e.g. can the data only take whole numbers or increments of 0.5, 0.25, etc.).

Note: Sometimes it may not be obvious whether data is discrete or continuous. If in doubt, make a best-judgement call and state the assumption that you have made.

Central Tendency - Mean

To obtain the **mean** of a grouped data set we:

(i) take the weighted sum of the class marks (where the weights are the class frequencies); and

(ii) divide it by the number of data points (which is equal to the sum of the class frequencies, i.e. $n = \Sigma f_i$).

$$\text{I.e.} \quad \mu_X = \frac{\Sigma f_i X_i}{N} \qquad \text{(population)}$$

$$\overline{X} = \frac{\Sigma f_i X_i}{n} \qquad \text{(sample)}$$

Central Tendency - Example Mean Calculation

Going back to our grouped travel time data:

time	frequency (f_i)	class mark (X_i)	$f_i X_i$
$0 \leq x_i < 10$	1	4.5	4.5
$10 \leq x_i < 20$	2	14.5	29
$20 \leq x_i < 30$	3	24.5	73.5
$30 \leq x_i < 40$	1	34.5	34.5
$40 \leq x_i < 50$	2	44.5	89
			$\Sigma f_i X_i = 230.5$

$$\overline{X} = \frac{\Sigma f_i X_i}{n} = \frac{230.5}{9} = 25.61 \text{ (to 2 dp)}.$$

I.e. the mean travel time for the grouped data is 25.61 minutes, whereas we obtained 26.22 minutes with the ungrouped (raw) data (that these two means are reasonably close is a good sanity check!).

Note: Histograms can also be used as a sanity check.

Central Tendency - Median

There are several ways in which the **median** of grouped data can be interpolated within the **median class** (i.e. the first class with a cumulative frequency $\geq \frac{n+1}{2}$).

Note: We will use the easiest of these methods, and this method should only really be used if the data is truly continuous (which the data for the example later on is not!).

Central Tendency - Median

Let:

(i) L be the lower-class-limit of the median class;

(ii) w be the width/range of the median class;

(iii) h be how far in to the median class that the median falls (i.e. the median position - the cumulative frequency up to the start of the median class); and

(iv) f be the frequency of the median class.

Then the interpolated estimate of the median for grouped data is:

$$M_d = L + w\frac{h}{f}$$

Central Tendency - Example Median Calculation

Going back to our grouped travel time data:

time	frequency (f_i)	cumulative frequency
$0 \leq x_i < 10$	1	1
$10 \leq x_i < 20$	2	3
$20 \leq x_i < 30$	3	6
$30 \leq x_i < 40$	1	7
$40 \leq x_i < 50$	2	9

The median position is $\frac{n+1}{2} = 5$, so the median class is $20 \leq x_i < 30$. Furthermore:

$$L = 20; w = 9; h = 5 - 3 = 2; \text{ and } f = 3.$$

Hence $M_d = 20 + 9\frac{2}{3} = 26$, which coincidentally is the same as the ungrouped (raw) median.

Note: Closeness of the grouped and ungrouped/raw medians can be used as a sanity check!).

Central Tendency - Mode

With grouped data it is more common to hear the terms **modal class(es)** (i.e. the class(es) with the highest frequency) rather than the **mode**.

If asked for a mode, the best we can do is use the class mark of the modal class.

Example: Going back to our grouped travel time data:

time	frequency (f_i)
$0 \leq x_i < 10$	1
$10 \leq x_i < 20$	2
$20 \leq x_i < 30$	3
$30 \leq x_i < 40$	1
$40 \leq x_i < 50$	2

The modal class is $20 \leq x_i < 30$, and the mode is 24.5 minutes.

. . . that's it for now, thanks for watching!

Don't forget that you can ask questions via:

(i) face-to-face lectures;

(ii) workshops or tutorials;

(iii) consultation hours; or

(iv) email.