Describing data with numeric value summary

Dr. Abbas Adigun (PhD)

Biostatistician

19th May 2017

Describing data with numeric value summary

Summary measures of location

A summary measure of location is a value around which most of the data values tend to centre or congregate.

We will be discussing three measures of location: the mode; the median; and the mean.

The mode

The *mode* is that category or value in the data that has the highest frequency (i.e. occurs the most often). In this sense, the mode is a measure of *commonness* or *typicalness*.

The mode is not particularly useful with metric continuous data where no two values may be the same.

The other shortcoming of this measure is that there may be more than one mode in a set of data.

Example

The modal Apgar score in Table 3.1 is 8, this being the category with the highest frequency (of 9 infants), i.e. is the most commonly occurring.

Apgar score	No of infants
4	2
5	3
6	6
7	7
8	9
9	3

Stata command

```
use "C:\Users\ADIGUM\Desktop\abuja\infant.dta", clear
tab apgar_score
table apgar_score
```

Exercise

Determine the modal value of parity from the infant data.

The Median

If we arrange the data in ascending order of size, the *median* is the middle value. Thus, half of the values will be equal to or less than the median value, and half equal to or above it. The median is thus a measure of *centralness*.

An advantage of the median is that it is not much affected by skewness in the distribution, or by the presence of outliers.

However, it discards a lot of information, because it ignores most of the values, apart from those in the centre of the distribution.

Example

Suppose you had the following data on age (in ascending order of years), for five individuals: 30 31 **32** 33 35. The middle value is 32, so the median age for these five people is 32 years.

If you have an *even* number of values, the median is the average of the two values either side of the 'middle'.

There is a quite easy way, of determining the value of the median. If you have *n* values arranged in ascending order, then;

$$\text{Median} = {}^{1}/_{2}(n + 1)^{\text{th}} \text{ value}$$

So, for instance, if the ages of six people are: 30 31 32 33 35 36, then $n = 6$, therefore:

$^1/_2(n + 1) = {}^1/_2 \times (6 + 1) = {}^1/_2 \times 7 = 3.5.$

Therefore the median is the 3.5th value. That is, it is the value half way between the 3rd value of 32, and the 4th value of 33, which is 32.5 years.

Exercise: Determine the median percentage mortality using the information within the table 3.2.

Table 3.2

ICU	1	2	3	4	5	6	7	8	9	10	11	12	13
% mortality	15.2	31.3	14.9	16.3	19.3	18.2	20.2	12.8	14.7	29.4	21.1	20.4	13.6
ICU	14	15	16	17	18	19	20	21	22	23	24	25	26
% mortality	22.4	14.0	14.3	22.8	26.7	18.9	13.7	17.7	27.2	19.3	16.1	13.5	11.2

Median= $^{1}/_{2}(n + 1)^{th}$ value

= $^{1}/_{2}(26 + 1)^{th}$ value

= 13.5th value, which is the average of 13 and 14 values

= (17.7+18.2)/2

=17.95%

`

Arithmetic mean

- The *arithmetic mean* is more commonly known as the average. One advantage of the mean over the median is that it uses all of the information in the data set. However, it is affected by skewness in the distribution, and by the presence of outliers in the data. This may, on occasion, produce a mean that is not very representative of the general mass of the data. Moreover, it cannot be used with ordinal data.

Mean is calculated using the formula

$$\frac{\sum X_i}{n} \qquad i = 1,2 \ldots, n$$

Example

Determine the mean ICU percentage mortality from the data given in table 5.6

Mean=
15.2+31.3+14.9+16.3+19.3+18.2+20.2+12.8+14.7+29.4
+21.1+20.4+13.6+22.4+ 14.0+ 14.3+ 22.8+ 26.7+ 18.9
+ 13.7+ 17.7+ 27.2+ 19.3+16.1+13.5+11.2
=485.2/26
=18.66

Stata commands

```
use "C:\Users\ADIGUM\Desktop\abuja\ICU.dta", clear
summ
```

Percentiles

Percentiles are the values which divide an ordered set of data into 100 equal-sized groups.

As an illustration, suppose you have birth weights for 1200 infants, which you've put in ascending, order.

If you identify the birthweight that has 1 per cent (i.e. 12) of the birthweight values below it, and 99 per cent (1188) above it, then this value is the *1st percentile*.

Similarly, the birthweight which has 2 per cent of the birthweight values below it, and 98 per cent above it is the 2nd percentile. You could repeat this process until you reached the 99th percentile, which would have 99 per cent (1188) of birthweight values below it and only 1 per cent above.

Notice that this makes the median the *50th percentile*, since it divides the data values into two equal halves, 50 per cent above the median and 50 per cent below.

Calculating percentiles

The birthweight data of 30 infants but now in ascending order and their position is reproduced below.

Birthweight	2860	2994	3193	3266	3287	3303	3388	3399	3400	3421	3447	3508	3541	3594	3613
position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Birthweight	3615	3650	3666	3710	3798	3800	3886	3896	4006	4010	4090	4094	4200	4206	4490
position	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30

The *p*th percentile is the value in the

p/100(n **+** 1)th position.

For instance the 20th percentile is (n+1) th position (30+1) = 0.20×31th value

= 6.2th value

The 6th value is 3303 g and the 7th value is 3388g, a difference of 85g,

so the 20th percentile is 3303g plus 0.2 of 85g, which is 3303g **+** 0.2 **×** 85g

= 3303g **+** 17g

= 3320g.

Stata commands

```
use "C:\Users\ADIGUM\Desktop\abuja\infant.dta", clear
centile birthweight, centile(20)
```

Guide to choosing appropriate measure of location

	Summary measure of location		
Type of variable	**Mode**	**Median**	**Mean**
Nominal	Yes	no	no
Ordinal	Yes	yes	no
Metric discrete	Yes	Yes if distribution is markedly skewed	yes
Metric continuous	No	Yes if distribution is markedly skewed	yes

Summary measure of spread

As important as a summary measure of location, a summary measure of spread or dispersion can also be very useful.

There are three main measures in common use. These are range, inter-quartile range, and Standard deviation.

The range

The *range* is the distance from the smallest value to the largest. The range is not affected by skewness, but is sensitive to the addition or removal of an outlier value.

Example; The birthweight data values ranges between 2860 and 4490.

The Inter-quartile range

The **inter-quartile range** is a measure of where the "middle fifty" is in a data set.

Where a range is a measure of where the beginning and end are in a set, an inter-quartile range is a measure of where the bulk of the values lie.

The inter-quartile range formula is the first quartile subtracted from the third quartile.

That is IQR= Q3-Q1

Q3 is the value which has 75percent of the data below it

Q1 is the value which has 25percent of the data below it

Example: Find the 25th and 75th percentile value of the infant birthweight data

25th percentile = 25/100 × (n+1)th value.

n=30

= 25/100 × (30+1)th

= 7.75th value

the 7th value is 3388 and the 8th value is 3399

Difference between 7th and 8th value is 11.

0.75 × 11=8.25

so 7.75th value=3388+8.25=3396.25

Also, the 75th percentile is calculated likewise

75/100 × (30+1)th value

=0.75 × 31

=23.25th value

the 23rd value is 3896 and the 24th value is 4006

The difference between 23rd and the 24th value = 110

0.25*110=27.5

so 23.25th value=3896+27.5=3923.5

The inter-quartile range is then written as (Q1 to Q3).

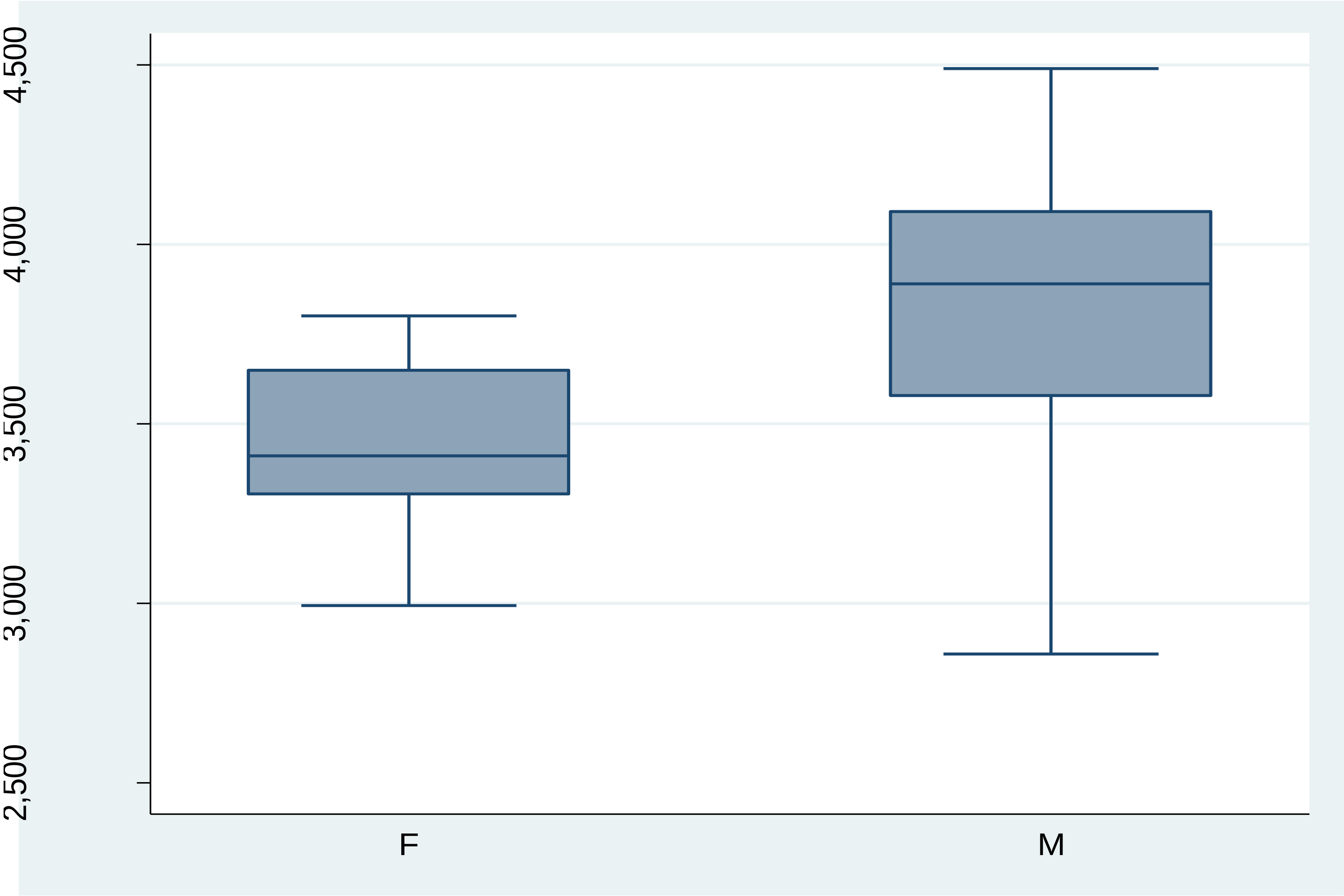
With the birthweight data: Q1 *=* 3396.25 g, and Q3 *=* 3923.50 g. Therefore: inter-quartile range *=* (3396.25 to 3923.50) g.

This result informed that the middle 50 per cent of infants (by weight) weighed between 3396.25 g and 3923.50 g.

Box-plot

Boxplots provide a graphical summary of the three quartile values, the minimum and maximum values, and any outliers. They are usually plotted with value on the vertical axis. Like the pie chart, the boxplot can only represent one variable at a time, but a number of boxplots can be set alongside each other.

Example: Boxplot for weight of infants boys and girls.



Note:

The bottom end of the lower ‘whisker’ (the line sticking out of the bottom of the box), corresponds to the minimum value

The bottom of the box is the 1st quartile value, Q1

The line across the inside of the box is the median, Q2.

The more asymmetric (skewed) the distributional shape, the further away from the middle of the box will be the median line, closer to the top of the box is indicative of negative skew, closer to the bottom of the box – positive skew.

The top of the box is the third quartile Q3.

The top end of the upper whisker is the maximum.

Stata commands

```
use "C:\Users\ADIGUM\Desktop\abuja\infant.dta", clear
graph box birthweight
graph box birthweight, over(gender)
```

Standard deviation

An approach which summarized the spread by measuring the mean (average) distance of all the data values from the *overall* mean of all of the values is known as standard deviation. The smaller this mean distance is, the narrower the spread of values must be, and vice versa. One advantage of the standard deviation is that, unlike the inter-quartile range, it uses all of the information in the data.

It is calculated using the formula $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Exercise:

Calculate the standard deviation for birthweight and breast cancer data

Stata commands

```
use "C:\Users\ADIGUM\Desktop\abuja\infant.dta", clear
summ btw
```

```
use "C:\Users\ADIGUM\Desktop\abuja\breast_cancer.dta", clear
summ parity
```

To sum up summary measures of spread

With ordinal data use either the range or the interquartile range. The standard deviation is not appropriate because of the non-numeric nature of ordinal data.

With metric data use either the standard deviation, which uses all of the information in the data, or the interquartile range.

The latter if the distribution is skewed, and/or you have already selected the median as your preferred measure of location.

Guide to choose appropriate measure of spread

	Summary measure of spread		
Type of variable	**Range**	**Inter-quartile range**	**Standard deviation**
Nominal	no	no	no
Ordinal	yes	yes	no
Metric	yes	yes, if skewed	yes

CORBIS/Brian Leng (05065)