IIT Madras

ONLINE DEGREE

Statistics for Data Science -1

Lecture 3.5: Describing Numerical Data- Percentiles

Usha Mohan

Indian Institute of Technology Madras

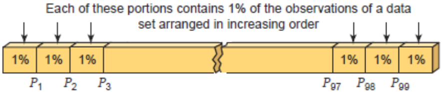
Percentiles

- The sample $100p$ percentile is that data value having the property that at least $100p$ percent of the data are less than or equal to it and at least $100(1-p)$ percent of the data values are greater than or equal to it.

[1] Figure source: Mann, P. S. (2007). Introductory statistics. John Wiley & Sons.

Percentiles

- The sample $100p$ percentile is that data value having the property that at least $100p$ percent of the data are less than or equal to it and at least $100(1-p)$ percent of the data values are greater than or equal to it.

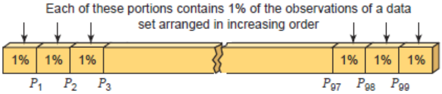


[1]

[1] Figure source: Mann, P. S. (2007). Introductory statistics. John Wiley & Sons.

Percentiles

- The sample $100p$ percentile is that data value having the property that at least $100p$ percent of the data are less than or equal to it and at least $100(1-p)$ percent of the data values are greater than or equal to it.



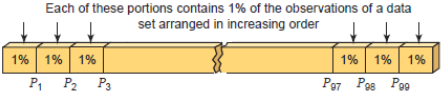
[1]

- If two data values satisfy this condition, then the sample $100p$ percentile is the arithmetic average of these values.

[1] Figure source: Mann, P. S. (2007). Introductory statistics. John Wiley & Sons.

Percentiles

- The sample $100p$ percentile is that data value having the property that at least $100p$ percent of the data are less than or equal to it and at least $100(1-p)$ percent of the data values are greater than or equal to it.



[1]

- If two data values satisfy this condition, then the sample $100p$ percentile is the arithmetic average of these values.
- Median is the 50^{th} percentile.

[1]Figure source: Mann, P. S. (2007). Introductory statistics. John Wiley & Sons.

Computing Percentile

To find the sample 100p percentile of a data set of size n

Computing Percentile

To find the sample 100p percentile of a data set of size n

1. Arrange the data in increasing order.

Computing Percentile

To find the sample 100p percentile of a data set of size n

1. Arrange the data in increasing order.
2. If np is not an integer, determine the smallest integer greater than np. The data value in that position is the sample 100p percentile.

Computing Percentile

To find the sample 100p percentile of a data set of size n

1. Arrange the data in increasing order.
2. If np is not an integer, determine the smallest integer greater than np. The data value in that position is the sample 100p percentile.
3. If np is an integer, then the average of the values in positions np and $np + 1$ is the sample 100p percentile.

Example

Let $n = 10$

Example

Let n= 10

- Arrange data in ascending order 35, 38, 47, 58, 61, 66, 68, 68,70, 79

p	np	

Example

Let n= 10

- Arrange data in ascending order 35, 38, 47, 58, 61, 66, 68, 68,70, 79

p	np	
0.1	1	(35+38)/2=36.5

Example

Let $n = 10$

- Arrange data in ascending order 35, 38, 47, 58, 61, 66, 68, 68,70, 79

p	np	
0.1	1	(35+38)/2=36.5
0.25	2.5	47

Example

Let n= 10

- Arrange data in ascending order 35, 38, 47, 58, 61, 66, 68, 68,70, 79

p	np	
0.1	1	(35+38)/2=36.5
0.25	2.5	47
0.5	5	(61+66)/2=63.5

Example

Let n= 10

- Arrange data in ascending order 35, 38, 47, 58, 61, 66, 68, 68,70, 79

p	np	
0.1	1	(35+38)/2=36.5
0.25	2.5	47
0.5	5	(61+66)/2=63.5
0.75	7.5	68
1	10	79

Computing percentile using googlesheets-PERCENTILE function

Step 1 Paste the dataset in a column.

Step 2 In a blank cell enter PERCENTILE(data, percentile), where data indicates the range of data for which percentile needs to be computed, and percentile is the decimal form of the desired percentile.

- For example if the data is in cell A1:A10, and we are interested in computing the 90^{th} percentile, then enter PERCENTILE(A1:A10,0.9) in a blank cell.

Computing percentile using googlesheets-algorithm

Computing percentile using googlesheets-algorithm

Step 1 Arrange data in increasing order.

Order	1	2	3	4	5	6	7	8	9	10
$x_{[i]}$	$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$	$x_{[6]}$	$x_{[7]}$	$x_{[8]}$	$x_{[9]}$	$x_{[10]}$
Data	35	38	47	58	61	66	68	68	70	79

Computing percentile using googlesheets-algorithm

Step 1 Arrange data in increasing order.

Order	1	2	3	4	5	6	7	8	9	10
$x_{[i]}$	$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$	$x_{[6]}$	$x_{[7]}$	$x_{[8]}$	$x_{[9]}$	$x_{[10]}$
Data	35	38	47	58	61	66	68	68	70	79

Let $x_{[i]}$ denote the i^{th} ordered value of the dataset.

Computing percentile using googlesheets-algorithm

Step 1 Arrange data in increasing order.

Order	1	2	3	4	5	6	7	8	9	10
$x_{[i]}$	$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$	$x_{[6]}$	$x_{[7]}$	$x_{[8]}$	$x_{[9]}$	$x_{[10]}$
Data	35	38	47	58	61	66	68	68	70	79

Let $x_{[i]}$ denote the i^{th} ordered value of the dataset.

Step 2 Find rank using the following formula.
$rank = percentile \times (n - 1) + 1$ where n is total number of observations in the dataset

Computing percentile using googlesheets-algorithm

Step 1 Arrange data in increasing order.

Order	1	2	3	4	5	6	7	8	9	10
$x_{[i]}$	$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$	$x_{[6]}$	$x_{[7]}$	$x_{[8]}$	$x_{[9]}$	$x_{[10]}$
Data	35	38	47	58	61	66	68	68	70	79

Let $x_{[i]}$ denote the i^{th} ordered value of the dataset.

Step 2 Find rank using the following formula.
$rank = percentile \times (n - 1) + 1$ where n is total number of observations in the dataset

- Example: to compute 25 percentile of a set of $n = 10$ observations, $rank = 0.25 \times (10 - 1) + 1 = 3.25$

Computing percentile using googlesheets-algorithm

Step 1 Arrange data in increasing order.

Order	1	2	3	4	5	6	7	8	9	10
$x_{[i]}$	$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$	$x_{[6]}$	$x_{[7]}$	$x_{[8]}$	$x_{[9]}$	$x_{[10]}$
Data	35	38	47	58	61	66	68	68	70	79

Let $x_{[i]}$ denote the i^{th} ordered value of the dataset.

Step 2 Find rank using the following formula.
$rank = percentile \times (n-1) + 1$ where n is total number of observations in the dataset

- Example: to compute 25 percentile of a set of $n = 10$ observations, $rank = 0.25 \times (10-1) + 1 = 3.25$

Step 3 Split the rank into integer part and fractional part.

Computing percentile using googlesheets-algorithm

Step 1 Arrange data in increasing order.

Order	1	2	3	4	5	6	7	8	9	10
$x_{[i]}$	$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$	$x_{[6]}$	$x_{[7]}$	$x_{[8]}$	$x_{[9]}$	$x_{[10]}$
Data	35	38	47	58	61	66	68	68	70	79

Let $x_{[i]}$ denote the i^{th} ordered value of the dataset.

Step 2 Find rank using the following formula.
$rank = percentile \times (n-1) + 1$ where n is total number of observations in the dataset

- Example: to compute 25 percentile of a set of $n = 10$ observations, $rank = 0.25 \times (10-1) + 1 = 3.25$

Step 3 Split the rank into integer part and fractional part.

- Integer part of 3.25 =3; fractional part is 0.25.

Computing percentile using googlesheets-algorithm

Step 1 Arrange data in increasing order.

Order	1	2	3	4	5	6	7	8	9	10
$x_{[i]}$	$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$	$x_{[6]}$	$x_{[7]}$	$x_{[8]}$	$x_{[9]}$	$x_{[10]}$
Data	35	38	47	58	61	66	68	68	70	79

Let $x_{[i]}$ denote the i^{th} ordered value of the dataset.

Step 2 Find rank using the following formula.
$rank = percentile \times (n-1) + 1$ where n is total number of observations in the dataset

- Example: to compute 25 percentile of a set of $n = 10$ observations, $rank = 0.25 \times (10-1) + 1 = 3.25$

Step 3 Split the rank into integer part and fractional part.

- Integer part of 3.25 =3; fractional part is 0.25.

Step 4 Compute the ordered data value $x_{[i]}$ corresponding to the integer part rank.

Computing percentile using googlesheets-algorithm

Step 1 Arrange data in increasing order.

Order	1	2	3	4	5	6	7	8	9	10
$x_{[i]}$	$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$	$x_{[6]}$	$x_{[7]}$	$x_{[8]}$	$x_{[9]}$	$x_{[10]}$
Data	35	38	47	58	61	66	68	68	70	79

Let $x_{[i]}$ denote the i^{th} ordered value of the dataset.

Step 2 Find rank using the following formula.
$rank = percentile \times (n-1) + 1$ where n is total number of observations in the dataset

- Example: to compute 25 percentile of a set of $n = 10$ observations, $rank = 0.25 \times (10-1) + 1 = 3.25$

Step 3 Split the rank into integer part and fractional part.

- Integer part of 3.25 =3; fractional part is 0.25.

Step 4 Compute the ordered data value $x_{[i]}$ corresponding to the integer part rank.

- The ordered data value corresponding to integer part rank of 3, $x_{[3]}$ is 47.

Computing percentile using googlesheets-algorithm-contd

Step 5 The percentile value is given by the formula

$$Percentile = x_{[i]} + fractional\ part \times \left[x_{[i+1]} - x_{[i]}\right]$$

- $Percentile = 47 + 0.25 \times [58 - 47] = 47 + 0.25 \times 11 = 47 + 2.75 = 49.75$

Quartiles

Definition

The sample 25th percentile is called the first quartile. The sample 50th percentile is called the median or the second quartile. The sample 75th percentile is called the third quartile.

Quartiles

Definition

The sample 25th percentile is called the first quartile. The sample 50th percentile is called the median or the second quartile. The sample 75th percentile is called the third quartile.

In other words, the quartiles break up a data set into four parts with about 25 percent of the data values being less than the first(lower) quartile, about 25 percent being between the first and second quartiles, about 25 percent being between the second and third(upper) quartiles, and about 25 percent being larger than the third quartile.

The Five Number Summary

- Minimum
- Q_1: First Quartile or lower quartile
- Q_2: Second Quartile of Median
- Q_3: Third Quartile or upper quartile
- Maximum

The Interquartile Range (IQR)

Definition

The interquartile range, IQR, is the difference between the first and third quartiles; that is,

$$IQR = Q_3 - Q_1$$

.

- IQR for the example

The Interquartile Range (IQR)

Definition

The interquartile range, IQR, is the difference between the first and third quartiles; that is,

$$IQR = Q_3 - Q_1$$

.

- IQR for the example
  - First quartile, $Q_1 = 49.75$

The Interquartile Range (IQR)

Definition

The interquartile range, IQR, is the difference between the first and third quartiles; that is,

$$IQR = Q_3 - Q_1$$

.

- IQR for the example
  - First quartile, $Q_1 = 49.75$
  - Third quartile, $Q_3 = 68$

The Interquartile Range (IQR)

Definition

The interquartile range, IQR, is the difference between the first and third quartiles; that is,

$$IQR = Q_3 - Q_1$$

.

- IQR for the example
  - First quartile, $Q_1 = 49.75$
  - Third quartile, $Q_3 = 68$
  - $IQR = Q_3 - Q_1 = 18.25$

Section summary

- Definition of percentiles.
- How to compute percentiles.
- Definition of quartile.
- Five-number summary.
- Interquartile range as a measure of dispersion.

Summary

1. Frequency tables
    1.1 Frequency table for discrete data.
    1.2 Frequency table for continuous data.
2. Graphical summaries
    2.1 Histograms.
    2.2 Stem-an-leaf plot.
3. Numerical summaries
    3.1 Measures of central tendency
        3.1.1 Mean, Median, Mode
    3.2 Measures of dispersion
        3.2.1 Range, Variance, Standard deviation
    3.3 Percentiles
        3.3.1 Interquartile range as a measure of dispersion.