Week 1 - Course Introduction and Python Basics

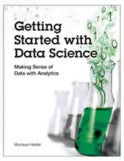
1. Welcome from your Instructors!

Instructor

https://sites.google.com/site/statsr4us/
http://www.youtube.com/regionomics

Murtaza Haider

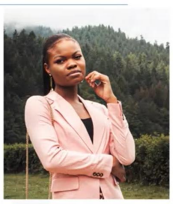
- I am an associate professor at the Ted Rogers School of Management, Ryerson University, in Toronto
- I am the author of *Getting Started with Data Science: Making Sense of Data with Analytics*, which was published by Pearson/IBM Press in 2016
- I am an avid blogger and blog weekly about socio-economics in South Asia for the Dawn newspaper and about urban economics in the Huffington Post
  - http://www.huffingtonpost.ca/murtaza-haider/
- I hold a Masters in transport engineering and planning and a Ph.D. in Civil Engineering (Urban Systems Analysis) from the University of Toronto



Co-Instructor

Aije Egwaikhide

- I am a Senior Data Scientist and Statistician with IBM Canada
- I have experience in supervised and unsupervised Machine learning algorithms
- I have a degree in Economics and Statistics
- On my off days I am a fashion and career blogger on Instagram (@olomohi), and I mentor and encourage young professionals interested in working in STEM



Course Outline

- This course contains five Modules
  1. Introduction and Descriptive Statistics
  2. Data Visualization
  3. Introduction to Probability Distribution
  4. Hypothesis Testing
  5. Regression Analysis
- Each Module will include videos and practical exercises
- Hands-on labs with Jupyter notebooks using Python

2. Python Packages for Data Science

(略。内容同课程 “Data Analysis with Python_IBM“ 中 Week 1 的 3. Python Packages for Data Science 一样)

总结:

Understanding the basics of Descriptive Statistics

3. Welcome to Statistics!

Statistics are all around us

- Will it rain/snow tomorrow?
- Is the housing becoming more expensive over time?
- Has the unemployment rate fallen over the past four months?
- Who is the highest scoring basketball player in NBA?
- Are millennials more likely to rent than the rest?
- Who is the highest paid actress in Hollywood?
- What is the average salary of a starting business analyst?
- Is the average salary of a fresh engineer higher than that of a fresh economist?
- Has crime rate spiked in Chicago in recent years?

The Language of Statistics

- We use Statistics everyday without really being mindful of it
  - Average income, age, height …
  - Highest paid (Maximum) athlete
  - Fastest (Maximum) sprinter
  - Lowest (Minimum) unemployment rate of all OECD countries
  - Percentage of females studying engineering
  - The chance (likelihood) for rain tomorrow
  - How consistent (variance) is a stock performance over the past three months?
  - On average, do men spend more (t-test) on clothes than women?

- So when we say average income, average age, average height, we're relying on average, which is a statistical parameter.
- Highest paid athlete, we're looking at the maximum salary.
- Fastest sprinter, you're looking at the maximum speed.
- Lowest unemployment rate of all the OECD countries, you're looking at a minimum value.
- Percentage of females who study engineering requires us to compute percentages.
- The chance for rain tomorrow is in fact likelihood.
- And how consistent is a stock performance over the past three months? We're concerned about variance, which again is a statistical parameter.
- And then the question of on average, do men spend more on clothes than women? We probably would use a T test to determine this difference, again, relying on statistics.

If you were to recall your conversations in a given day, you probably realize now that you have been using the language of statistics on a daily basis.

We see statistics in news media

Early-Stage (January-June) Favorability Ratings
New York Times / CBS News Polls, Since 1976
Incumbents Shaded in Gray

Year	Candidate	Favorable	Unfavorable	Net	Result
1984	Reagan	54	29	+25	WON
1976	Carter	41	21	+20	WON
1988	Dukakis	34	16	+18	LOST
1976	Ford	52	35	+17	LOST
2008	Obama	43	28	+15	WON
1996	Clinton	48	36	+12	WON
1980	Reagan	42	30	+12	WON
2000	Bush	37	31	+6	WON
2008	McCain	35	30	+5	LOST
1984	Mondale	38	34	+4	LOST
2004	Kerry	28	27	+1	LOST
1988	Bush	34	35	-1	WON
2012	Obama	41	42	-1	???
2004	Bush	40	41	-1	WON
1992	Bush	38	41	-3	LOST
2000	Gore	32	36	-4	LOST
1996	Dole	27	37	-10	LOST
1992	Clinton	19	30	-11	WON
2012	Romney	26	37	-11	???
1980	Carter	33	58	-25	LOST
	AVERAGE	37	34	+3	

***The Economist*'s house-price indices**
% change on a year earlier

	Q3 2004*	Q3 2003	1997-2004
South Africa	35.1	20.9	227
Hong Kong	31.2	-13.6	-49
Spain	17.2	16.5	149
New Zealand	16.4	21.2	56
France	14.7	11.5	76
Britain	13.8	11.0	139
United States	13.0	6.0	65
Ireland	10.8	14.8	187
China	9.9	4.1	*na*
Sweden	9.8	5.5	81
Italy	9.7	10.6	69
Belgium	9.3	5.5	50
Australia	8.2	17.6	112
Denmark	7.3	3.4	50
Canada	6.7	6.5	43
Netherlands	3.3	1.9	76
Switzerland	2.2	2.4	12
Singapore	*nil*	-2.3	*na*
Germany	-1.7†	-4.5	-3
Japan	-6.4	-4.8	-24

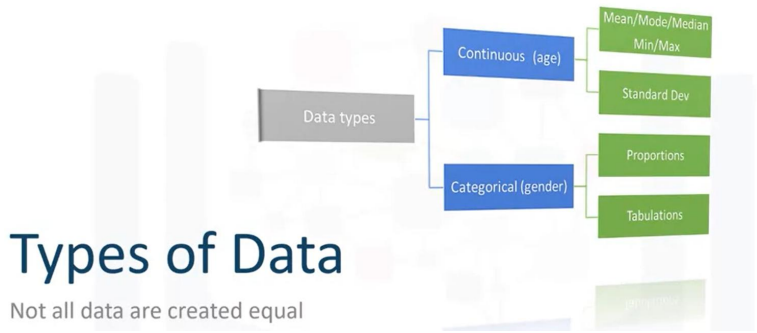
*Or 2004 latest †Second half 2003

Sources: ABSA; Bulwien; ESRI; Japan Real Estate Institute; Nomisma; NVM; ODPM; OFHEO; Quotable Value; Stadim; Swiss National Bank; government offices

At the same time, the news media use statistics all the time to demonstrate how trends are changing. 2016 was the year American presidential elections were held, big surprises there between what the polls forecasted and what the outcome was. But again, you see these numbers portrayed in the newspapers. At the same time, you have other publications that show you how housing prices or other development related statistics vary over countries. In a nutshell, the information we consume and the conversations we have every day involves a lot of statistics, so it pays one to learn some statistics.

总结:

4. Types of Data



The first step in analytics or statistics is to have a good look at your data and before you begin, try to understand what kind of variable you're working with. And based on the type of variable, you will decide what kind of analytics could be performed with it.

Data Types - 1

- Type of Data
  - Cross-Sectional – measurements taken at one time period
    - e.g., students course evaluations in a course
    - Cross sectional Panel
      - Same student's evaluation of different courses in a particular year or in subsequent years
  - Time series – data collected over time
    - E.g., unemployment rate, monthly retail sales
    - Time Series Panel
      - Students' annual satisfaction rating of Ryerson University over 4 years

Data Types - 2

- Number of Variables
  - Univariate– data consisting of a single variable to measure some entity
  - Multivariate– data consisting of two or more variables to measure some entity

Let's have a look at various different types of data that we encounter and is commonly used in our daily lives.

- The most common one would be a **cross-sectional data**, which is basically looking at a measurement taken at one point in time. Census in a given year is a cross-section of the society, as students evaluate course and instructor, that's a cross-section at any given point. Compared to the cross-sectional data, we can have panel or cross-sectional panel data, which is essentially asking the same group of individuals the same questions repeatedly over time. So you may pick a group of people, constitute it as a panel, and then ask the same questions once every year over a given period of time.
- The **time series data** is rather different. You're looking at a particular phenomenon such as unemployment rate, and then you measure it every month and then display that data or analyze that data which is repeated measurements on the same phenomena over time. So you may have monthly data going back to 1940s or climate data going back to hundreds of years.

So based on the type of data cross-sectional panel, time series, we will pick appropriate tools, statistical tools to deal with them.

- If your data set has only one variable, it's called a **univariate data set**. And if you have multiple variables in your data set, then it's a **multivariate data set**.

总结:

Variable Types - 1

- Categorical (nominal) – data sorted into mutually exclusive (an observation cannot belong to more than one category) categories
  - Geographical region, type of employee, gender, state of birth, type of automobile owned
  - Discrete choices
    - Mode of travel (multinomial)
      - Auto drive
      - Auto passenger
      - Public transit
      - Walk/bike
    - Home ownership (binomial)
      - Own
      - Rent
- Properties
  - No quantitative relationships among categories
  - Statistics such as averages are usually meaningless

Variable Types - 2

- Ordinal data – data ordered or ranked according to some relationship to one another
  - Number of cars owned by a household
- Properties
  - Categories can be compared with one another
  - Statistics usually meaningless because of no fixed units of measurement; i.e., differences are meaningless

Variable Types - 3

- Ratio data – data that have a natural zero
  - Sales dollars, length, weight, time from start of a process, most business and economic data
- Properties
  - Strongest form of measurement; both ratios and differences are meaningful

Variable Types - 4

- Interval data – data that are ordered and characterized by a specified measure of distance between observations, but with no natural zero.
  - Temperature scales, time, survey scales that are assumed to be interval
- Properties
  - Ratios are meaningless (50 degrees is not twice as hot as 25 degrees)
  - Differences are meaningful, so statistics such as averages may be compared

Let us now look at variable types.

- Start with **categorical** or **nominal** variables.
  - Let's consider home ownership, for instance. One can either own a home or rent a home, and knowing that there are only two categories here, owning and renting, that is a categorical variable. The tenure status of an individual is essentially a categorical variable. In this particular case, because you only have two choices, own or rent, it's a **binomial** variable.
  - Consider travel choices. You can go to work by driving or by someone drives you there, so you are a passenger, you can take public transit, or you can walk or bike. So in this particular case, you have four choices, more than two, so we call it **multinomial**. So both binomial and multinomial variables are part of categorical variables.
  - You cannot have any quantitative relationships among categories, and for these types of variables, averages are usually meaningless. So if you have a mode of travel and you have four categories, an average category would mean absolutely nothing of use.

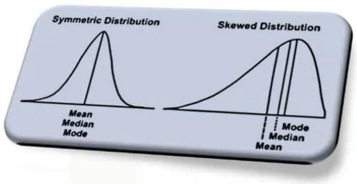
- A particular type of categorical variable is **ordinal** data, which are **ranked or ordered in some particular fashion**.
  - For instance, number of cars owned by household. A household may have zero car, one car, two cars, three or more cars, and that essentially is an ordinal data where zero represents zero, and zero cannot be coded as one and one cannot be coded as zero. So the order in which variable has been recorded matters.
  - Categories can be compared with one another, and you still cannot use regular statistics. The differences are also meaningless in this particular case.

- Another type of data is called **ratio** data, which is the data set that **have a natural zero**.
  - For example, sales dollars, length of distance, or weight of an object. These are all examples of ratio data, and I often would use the term **continuous** data or continuous variable. So a variable such as distance from point A to B could be 8 kilometers, 8.5 kilometers, 6.2 miles. The variable is continuous, and zero makes some logical sense in this particular variable. So for instance, you say I have $0, 0 means something here.
  - It's a **strongest form** of measurement, and you can compute ratios and differences.

- And another type of variable is **interval** data or interval variables that are ordered and characterized by specific measure of distance between observations, and **it may not have a natural zero**.
  - Temperature is a good example. And when you say that it's 0 degrees Celsius, it does not mean that there is no temperature. It's freezing, but it is measuring something that exists.
  - Ratios are also meaningless. So for example, if someone said, well, the temperature in some African countries is 50 degrees, compared to somewhere tropical where it was 25 degrees, it doesn't mean that the temperatures in the African desert is two times or twice as hot as it is in the tropics. But we can say that there is a difference of 25 degrees between the two places.

总结:

5. Measure of Central Tendency

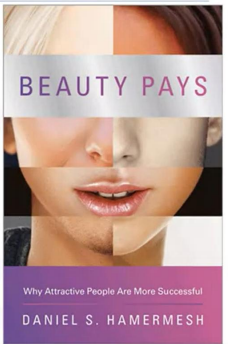


Measures of Central Tendency

Mean, Median, and Mode

Beauty in numbers

- Does beauty pay?
- University of Texas
  - Survey data from 463 courses
  - Teaching evaluations of instructors
- Instructor attributes
  - Gender
  - Fluency in English language
  - Tenure status
  - Beauty score



The measures of central tendency are the most commonly used in statistical analysis. We know them as mean, median, and mode and their use is ubiquitous and statistical analysis. So let us see how it works.

Before we begin, let us take a quick look at our dataset in this course. We have been using the teaching evaluation Data from the University of Texas. The dataset comprises of 463 courses, in which we have information about the teaching evaluation score received by the instructor. We have information about the attributes of the instructor, as well as the characteristics of the course.

Teaching Evaluation Data

```
## get information about each variable
ratings_df.info()
```

```
RangeIndex: 463 entries, 0 to 462
Data columns (total 19 columns):
minority          463 non-null object
age               463 non-null int64
gender            463 non-null object
credits           463 non-null object
beauty            463 non-null float64
eval              463 non-null float64
division          463 non-null object
native            463 non-null object
tenure            463 non-null object
students          463 non-null int64
allstudents       463 non-null int64
prof              463 non-null int64
PrimaryLast       463 non-null int64
vismin            463 non-null int64
female            463 non-null int64
single_credit     463 non-null int64
upper_division    463 non-null int64
English_speaker   463 non-null int64
tenured_prof      463 non-null int64
```

Populations and Samples

- Population – all items of interest for a particular decision or investigation
  - All drivers in the U.S.
  - All individuals who do not own a cell phone
- Sample – a subset of a population
  - All married drivers in the U.S. over age 25
- Why samples are used?
  - To reduce costs of data collection
  - When a full census cannot be taken

Once you have imported a CSV file with a Pandas Python library, the first step in getting to know your data is to discover the different data types it contains. You can display all columns and their data types with dataframe dot info. In this case, we have named our dataframe, ratings_df. It tells you how many rows you have. For the teaching rating data, we have 463 entries from zero to 462 because Python starts counting from zero. Then it also gives you information about the data types. Object represents strings. In 64 represents integer or whole numbers, and float represents real numbers, which could take on decimal points.

Before we begin, let us have a conversation about **population** and **samples**. Essentially, if you have all the information of interest for a particular decision, about every individual that is supposed to be involved in that decision, that is called a population. So if you are interested in looking at some attribute of driving, and we have information about all possible automobile drivers in the US, then we call this, the population. The sample, on the other hand, is a subset of population. So for example, if we have data on all married drivers over the age of twenty five, then that's a subset. Within that subset, if we were to randomly select five percent of those married drivers over the age of 25, that would be our sample. We use samples, especially in cases where we do not want to incur the cost of collecting data for the entire population. Now, let us consider that there are 230 million individuals in the country. A sample size of 330 to 500 individuals randomly selected would suffice. This reduces the cost, especially in cases where you cannot collect information for the entire population. Therefore, using samples, it's really helpful and cost-effective.

Terminology and Notation

- x_i represents the i^{th} observation
- $\sum$ indicates the operation of addition
- N is the size of the population; n is the size of the sample
- f_i is the number of observations in cell i of a frequency distribution

Mean or Average

- Population mean

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

- Sample mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Here you see some Greek symbols on the screen. But don't be afraid. They mostly show the Formula. We will then proceed from here. While they may differ in notation, essentially the **mean** for a population and sample are the same. It is the sum of all the observations, then, divide it by the number of observations to get the mean, which we call **average**s.

总结：

Properties of the Mean

- Meaningful for interval and ratio data (continuous variables)
- Affected by unusually large or small observations (outliers)
  - Hence median is also useful
- The only measure of central tendency where the sum of the deviations of each value from the measure is zero; i.e.,

$$\sum(x_i - \bar{x}) = 0$$

There are several properties of the mean and they are meaningful. But one of the characteristics of a mean is that if you take the difference between the average value for a variable, and subtract from all the observations and sum them up. That sum would be equal to zero.

Median

- Middle value when data are ordered from smallest to largest. This results in an equal number of observations above the median as below it
  - Unique for each set of data
  - Not affected by extremes
  - Meaningful for ratio, interval, and ordinal data

Mode

- Observation that occurs most frequently; for grouped data, the midpoint of the cell with the largest frequency (approximate value)
  - Useful when data consist of a small number of unique values

The **median** is different from the mean. When you order the data from the smallest value to the largest value, the result is in the middle. That is, the value in the middle indicating that there are an equal number of observations, that are above and the equal number of observations are below that family. That value is called the median. So if the median salary in a city is $45 thousand, it means that 50 percent of the people make more than $45 thousand and the other 50 percent make less than $45 thousand.

Mode is essentially the value that occurs most frequently. Therefore, if the most common age and a class of students is 16, then that's the mode.

Let's calculate the average beauty!

```
1  ratings_df[['beauty', 'eval', 'age']].describe()
```

	beauty	eval	age
count	4.630000e+02	463.000000	463.000000
mean	6.271140e-08	3.998272	48.365011
std	7.886477e-01	0.554866	9.802742
min	-1.450494e+00	2.100000	29.000000
25%	-6.562689e-01	3.600000	42.000000
50%	-6.801430e-02	4.000000	48.000000
75%	5.456024e-01	4.400000	57.000000
max	1.970023e+00	5.000000	73.000000

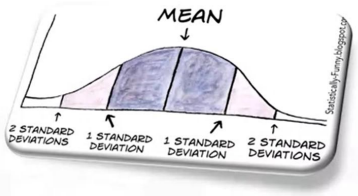
```
1  ratings_df.describe()
```

We will now turn to Python for our hands-on training to estimate the Summary Statistics values for beauty score. Teaching you about evaluation and age.

We will use the DataFrame dot describe function to find the Summary Statistics. This prints out the number of rows, mean, standard deviation, minimum value, 25th, 50th, 75th percentile, and the maximum value. To find the summary statistics for a subset of the variables, you will have to state the column names as we can see here. Otherwise, for the full population, we will call the dot describe function on the DataFrame.

总结:

6. Measure of Dispersion



Measures of Dispersion

The Standard Deviants

Measures of Dispersion

- Dispersion is the degree of variation in the data
  - E.g., the age of instructors {48, 49, 50, 51, 52}
- Range is the difference between the maximum and minimum observations
  - The minimum age of an instructor was 29 and maximum age was 73

Dispersion, which is also called **variability**, **scatter** or **spread**, is the extent to which the data distribution is stretched or squeezed. The common measures of dispersion are **standard deviation** and **variance**.
If you are at a university or college, you may have heard about the bell curve, which looks like this. You will often hear this is within one standard deviation of the mean or within two standard deviations of the mean. Let's see what that means.

Let's look at the age of an instructor. Let's say the average age is 52. This means that the individual ages may differ, some may be 48 or maybe 55 or 75. So the average age is an estimate. But what we also need is an estimate for the dispersion in the dataset. The other thing to note is the **range** in our data-set. For example, the difference of the ranges from a minimum of 29 years of age to a maximum of 73 years. This to you refers to a distance or the difference between the minimum and the maximum.

Variance

- Population $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$
- Sample $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

Standard Deviation

- The standard deviation is the square root of the variance
- The variance is in "square units" so the standard deviation is in the same units as x
- Population $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$
- Sample $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

Unlike the difference between population and sample mean, the difference between a **variance** for the population and a sample is that when you compute the **population variance** denoted as sigma squared, you divide it by the total number of observations. These are the deviations between observation and the mean squared, then added and then divided by the total number of observations. For **sample variance**, which is denoted as S squared, you divide it by n-1. The purpose of using n-1 is so that our estimate is unbiased in the long run. That means if we take a second sample, we'll get a different value of S squared. If we take a third sample, we'll get a third value of S squared and so on. We use n-1 so that the average of all these values of S squared is equal to sigma squared.

We usually talk about squares called **standard deviation** rather than the variance. Standard deviation is essentially the square root of the variance or the variances in square units. It's good to use the standard deviation because it's exactly the same units as the variable. The standard deviation of age will also be measured in years rather than years squared. Here you see that we just took the square root of variance and this becomes standard deviation.

总结:

The aging, beautiful standard deviants

```
ratings_df[['beauty', 'eval', 'age']].describe()
```

	beauty	eval	age
count	4.630000e+02	463.000000	463.000000
mean	6.271140e-08	3.998272	48.365011
std	7.886477e-01	0.554866	9.802742
min	-1.450494e+00	2.100000	29.000000
25%	-6.562689e-01	3.600000	42.000000
50%	-6.801430e-02	4.000000	48.000000
75%	5.456024e-01	4.400000	57.000000
max	1.970023e+00	5.000000	73.000000

Michael Jordan and Wilt Chamberlain

- Jordan and Chamberlain are basketball's most celebrated players
- On average, they both scored almost the same points per game
  - Mean – 30.12 for Jordan
  - Mean – 30.06 for Chamberlain
- However, when we consider standard deviation, Michael Jordan appears much more consistent
  - SD – Jordan: 4.76
  - SD – Chamberlain: 10.59

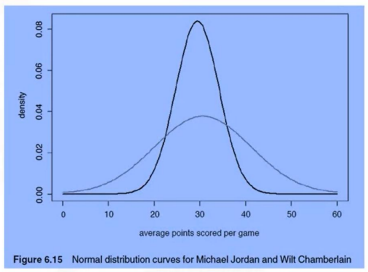


Figure 6.15 Normal distribution curves for Michael Jordan and Wilt Chamberlain

Source: Chapter 6: *Getting Started with Data Science: Making Sense of Data with Analytics*

We'll return to our dataset and we'll look at the variables that we computed before. You can see that the standard deviation for beauty, evaluation scores, and age were computed with the descriptive statistics using the describe function in Python.

Let me explain why mean and standard deviation have to go hand in hand. I will use the example from basketball about the two giants, Michael Jordan and Wilt Chamberlain, who preceded Michael Jordan. If you consider their average score per game, you would notice that they didn't differ much. Their average was around 30 points for both Jordan and Chamberlain. However, when you look at the standard deviation of their performance, Jordan was around 4.76 compared to Chamberlain who was at around 10.59. If you were to plot this distribution to see Michael Jordan scores using the mean and standard deviation. Assuming that their scores are normally distributed. You would notice even though both players had around the same mean, the tighter distribution for Jordan suggests that he was more consistent in his performance than Chamberlain.

Reliability

- Average paints a partial picture
- Average statistics are incomplete without standard deviation/variance
- Risk metrics are all about variance

Mean with SD means something

Case Summaries

minority	eval	
	mean	std
0 no	4.015288	0.545877
1 yes	3.892187	0.601666

The main takeaway(外卖，要点) is that average will only paint a partial picture. If you really want to understand the complete picture about a variable or data-set, it is important to compute both the average and the standard deviation to get insights on what the data is telling us.
So a mean with a standard deviation means something more useful than the mean by itself.

总结:

2. Find the median of the data set. 3 ,8 ,9 ,11, 12, 15

- ○ 9
- ◉ 10
- ○ 11
- ○ 12

✓ **正确**
Correct!

3. The measurements of spread or scatter of the individual values around the central point is called:

- ○ Measures of central tendency
- ○ Measures of central tendency and Measures of dispersion
- ○ Measure of skewness
- ◉ Measures of dispersion

✓ **正确**
Correct!

总结:

1. Which of the following is an example of time series data?

- (●) Annual average housing price in New York
- ( ) Batting average of a baseball player
- ( ) Number of trees in Jardin du Luxemburg in Paris
- ( ) Number of dolphins in the Pacific Ocean

✓ **正确**
A time series data is a sequence taken at successive equally spaced points in time.

2. What is the 25th percentile of the following data set;

1, 3, 3, 4, 5, 6, 6, 7, 8, 8

- (●) 3.5
- ( ) 1
- ( ) 5.5
- ( ) 3

✗ **错误**
The 25th percentile is the score that's greater or equal to 25% of the scores.

2. What is the 75th percentile of the following data set;

1, 3, 3, 4, 5, 6, 6, 7, 8, 8

- (●) 7
- ( ) 8
- ( ) 3
- ( ) 5.5

✓ **正确**
Correct!

4. Which of the following measures of central tendency will always change if a single value in the data changes?

- (●) Mean
- ( ) Mode
- ( ) All of the above
- ( ) Median

✓ **正确**
Correct!

总结:

6. What is meta data?

- ○ The data about metamorphism
- ◉ It's the data about data
- ○ Data about metal fatigue
- ○ The metabolism data in a clinical trial

✓ **正确**
Correct!

8. Median represents a value in the data set where:

- ○ Most observations are positive
- ○ Most observations are negative
- ○ Half of the observations are known and the other half not known
- ◉ Half of the observations are above the median and the other half below it

✓ **正确**
Correct!

9. If the variance of a dataset is correctly computed with the formula using (n – 1) in the denominator, which of the following option is true?

- ○ Data contains other variables with categorical data
- ○ Data is a population
- ○ Data is from an unknown source
- ◉ Data is a sample

✓ **正确**
Correct!

10. Which of the following is NOT a descriptive statistic?

- ◉ t-test
- ○ Mean
- ○ Median
- ○ Standard Deviation

✓ **正确**
Correct!

总结: