Session 8: Introduction to Statistics

Module BUSN9690

Business Statistics with Python

Outline

- Data type
- Sampling techniques
- Basic statistics

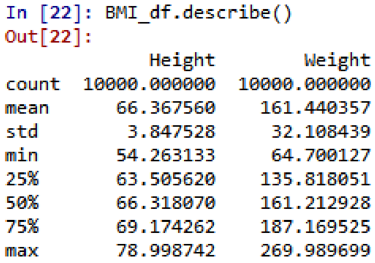
Population vs Sample; Descriptive vs Inferential

- **Population:** The collection of all individuals or items under consideration in a statistical study.
- **Sample:** That part of the population from which information is obtained.

- Statistics has these two broad classes
  - *Descriptive statistics* are used to organize and summarise your data. It focuses on the main characteristics of your data.
    - For example, mean, variance, skewness, etc
  - *Inferential statistics* make generalisations or inferences from your data to a larger set of data, based on probability theory.
    - For example, parameter estimation, hypothesis testing

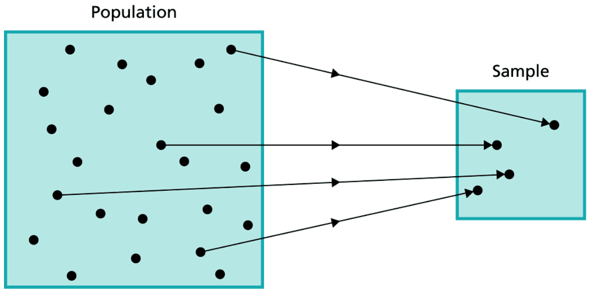
Examples

- *Example 1* (**Descriptive statistics)***:* Suppose we want to describe the variables in the income dataset. We calculate the summary statistics and produce graphs.



```
In [22]: BMI_df.describe()
Out[22]:
             Height        Weight
count  10000.000000  10000.000000
mean      66.367560    161.440357
std        3.847528     32.108439
min       54.263133     64.700127
25%       63.505620    135.818051
50%       66.318070    161.212928
75%       69.174262    187.169525
max       78.998742    269.989699
```

- *Example 2* (**Inferential statistics)**: Political polling provides an example of **inferential statistics**. Interviewing everyone of voting age in Scotland on their voting preferences would be expensive and unrealistic. Statisticians who want to gauge the sentiment of the entire **population** of Scottish voters can afford to interview only a carefully chosen group of a few thousand voters. This group is called a **sample** of the **population**



Sources of data

- *Primary Data*: Data which must be collected.
- *Secondary Data*: Data which are already available.
  - Advantage: less expensive. Disadvantage: may not satisfy your needs.

- **Methods of Collecting Primary Data**
  - *Experimental*
  - *Non-experimental (Observational)*
    - Collect data from database systems
    - Survey as a common type:
      - Focus Group; Telephone Interview; Mail Questionnaires; Door-to-Door Survey; Mall Intercept; New Product Registration; Personal Interview

An example of the experimental study

- ***A question:*** *does folic acid have an effect on birth defect?*

- ***Folic Acid and Birth Defects*** For the study, the doctors enrolled 4,753 women prior to conception, and divided them randomly into two groups.
  - One group took daily multivitamins containing 0.8 mg of folic acid, whereas
  - the other group received only trace elements.

- Statistical methods are then used to compare the effect on birth defect of the two groups

Data collection

- To collect data from your company's database systems;
- To collect data through survey
  - Focus Group; Telephone Interview; Mail Questionnaires; Door-to-Door Survey; Mall Intercept; New Product Registration; Personal Interview
- To buy data, for example, from www.experian.co.uk/
- To collect data from open sources
  - Eurostat: http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home
  - European Central Bank: http://www.ecb.int/home/html/index.en.html
  - UK Statistics Authority: http://www.statistics.gov.uk/
  - **UCI databank**: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing
- Question
  - In the above four types of data, which one is primary and which one is secondary data?

Non-experimental method: Sampling techniques

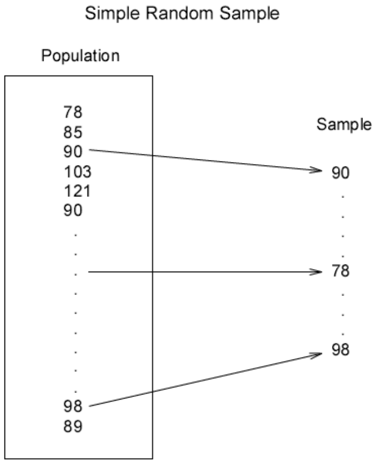
- Non-experimental method
  - Statistical sampling
    - Simple random sampling
    - Systematic sampling
    - Stratified sampling
    - Cluster sampling
  - Non-statistical sampling
    - Judgement Sampling
    - Convenience Sampling
    - Quota Sampling

Simple Random Sampling

Simple random sampling: A sampling procedure for which each possible sample of a given size is equally likely to be the one obtained.
Simple random sample: A sample obtained by simple random sampling.

There are two types of **simple random sampling.** One is simple random sampling **with replacement**, whereby a member of the population can be selected more than once; the other is simple random sampling **without replacement**, whereby a member of the population can be selected at most once.

- Questions: which of the following examples are Sampling with Replacement?
  - toss a coin
  - throw a die
  - assess the average house price in London



Python Codes

- Syntax: DataFrame.sample(n=None, frac=None, replace=False, weights=None, random_state=None, axis=None)
- Parameters:
  - n: int value, Number of random rows to generate.
  - frac: Float value, Returns (float value * length of data frame values). Either frac or n can be used, but not both.
  - replace: Boolean value, return sample with replacement if True.
  - random_state: int value or numpy.random.RandomState, optional. if set to a particular integer, will return same rows as sample in every iteration.
  - axis: 0 or ‘row’ for Rows and 1 or ‘column’ for Columns.

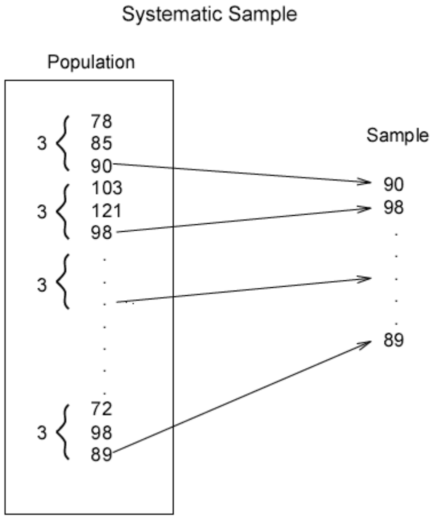
- Examples

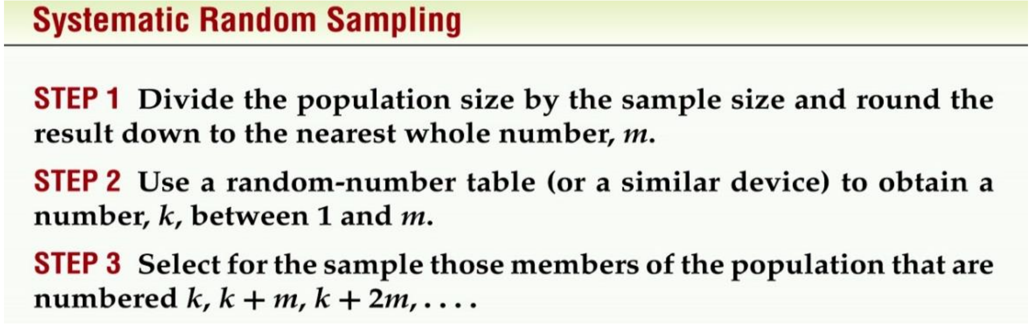
```
>>>dd1=BMI_df.sample(n=4000,replace=False) #sample 4000 observations, without replacement
>>>dd2=BMI_df.sample(frac=0.4,replace=True) #sample 40% of the observations, with replacement
```

Systematic random sampling

Systematic random sampling *is useful when* the researcher is unsure how many individuals will eventually be in the population.

Example: Suppose Mr Smith is interested in customers' preference on a new brand of milk sold in ASDA. He will design a couple of questions, select a couple of dates, stand in front of ASDA, and select every 10 customers to answer his questions.





Systematic Random Sampling

STEP 1 Divide the population size by the sample size and round the result down to the nearest whole number, m.

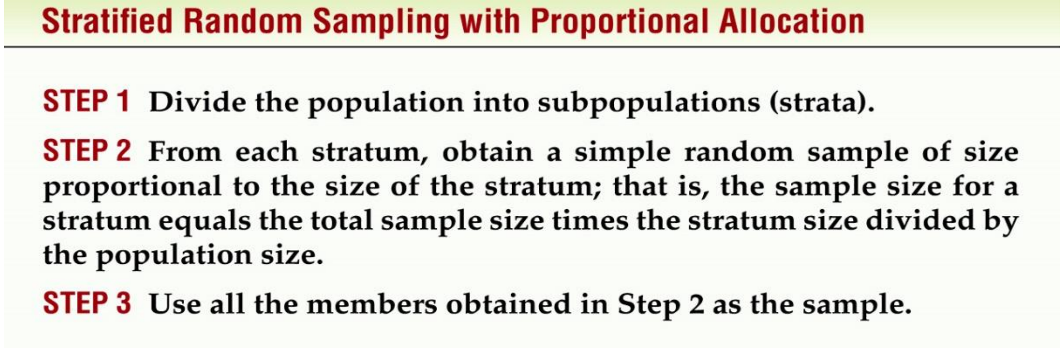
STEP 2 Use a random-number table (or a similar device) to obtain a number, k, between 1 and m.

STEP 3 Select for the sample those members of the population that are numbered $k, k + m, k + 2m, \ldots$.

Stratified Random Sampling

Stratified sampling *is useful when* a research question focuses on a stratum or on strata that make up a small proportion of the population

Example: There is a random sample of 3,000 college grads, within which 2,034 are European, 832 are African, and 134 are Asian. You wonder if there are any differences in income one year after graduation between the different subgroups and if the demographics of your sample are truly representative of the demographics of UK college graduates. One way of examining these questions is by using a *stratified random sampling*.



Stratified Random Sampling with Proportional Allocation

STEP 1 Divide the population into subpopulations (strata).

STEP 2 From each stratum, obtain a simple random sample of size proportional to the size of the stratum; that is, the sample size for a stratum equals the total sample size times the stratum size divided by the population size.

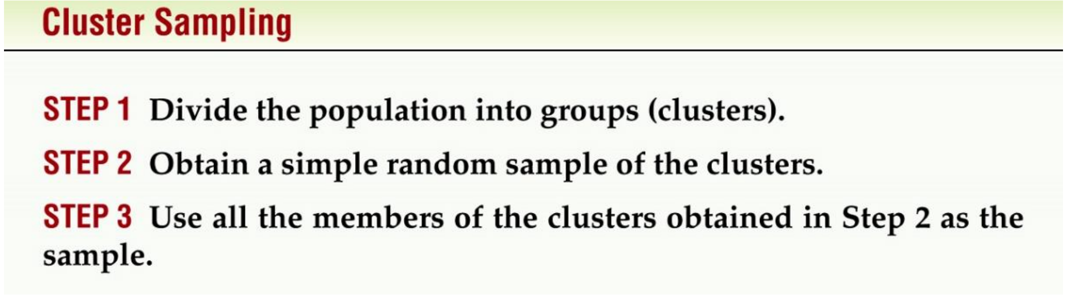
STEP 3 Use all the members obtained in Step 2 as the sample.

Cluster Sampling

Cluster sampling is useful when it is difficult or costly to generate a simple random sample.

Example: You aim to estimate the average annual household income in a large city. A less expensive way is to let each block within the city represent a cluster.

- A sample of clusters could then be randomly selected, and
- every household within these clusters could be interviewed to find the average annual household income.



Cluster Sampling

STEP 1 Divide the population into groups (clusters).

STEP 2 Obtain a simple random sample of the clusters.

STEP 3 Use all the members of the clusters obtained in Step 2 as the sample.

Non-statistical sampling

1. Judgement Sampling (Selecting what seems like a good enough sample.)

   Example: A TV researcher wants a quick sample of opinions about a political announcement. They stop what seems like a reasonable cross-section of people in the street to get their views.

2. Convenience Sampling (Use those available)

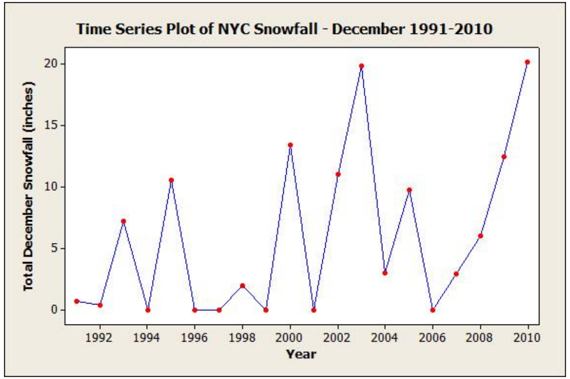
   Example: A group of students in a high school do a study about teacher attitudes. They interview teachers at the school, a couple of teachers in the family and few others who are known to their parents.

3. Quota Sampling (Keep going until the sample size is reached)

   Example: A researcher in the high street wants 100 opinions about a new style of cheese. She sets up a stall and canvasses passers-by until she has got 100 people to taste the cheese and complete the questionnaire.

Cross-sectional vs. Time-series

- cross-sectional data are collected at the same or approximately the same point in time
- time-series data are data collected over several time periods



- Panel data are time-series cross-sectional (TSCS) data

person	year	income	age	sex
1	2001	1600	23	1
1	2002	1500	24	1
2	2001	1900	41	2
2	2002	2000	42	2
2	2003	2100	43	2
3	2002	3300	34	1

Uni-variate data analysis: Basic statistics

- Number of observations (**sample size**), missing data
- Central tendency: mode, mean, median
- Location: Minimum, maximum, quartiles
- Dispersion (variability): range, variance, standard deviation, interquartile range

Measures of central tendency: Mode, Median, Mean

Mode: Most frequently occurring value (peak).

- Useful for qualitative rather than quantitative purposes.
- May not be very descriptive of distribution.

- Example
  - A bag of 10 balls: 4 red, 3 blue, 2 yellow, 1 green
  - What is the mode of this dataset?

Median: The middle data value

- 2 situations:
  - An odd number of data values: **number of obs =middle value**
  - An even number of data values: **number of obs = average of two middle values**
- Not as sensitive to extreme values, eg. House price, salary

- Example
  - Calculate the median of the following datasets, respectively
    - 23, 30, 60, 45, 90, 990
    - 23, 30, 60, 45, 90, 990, 51

- **Mean:** The average
  - Sensitive to extreme values
  - The sum of the differences of each data value from the mean is always 0.

- Examples
  - Calculate the median of the following dataset, respectively
    - 23, 30, 60, 45, 90, 990
    - 23, 30, 60, 45, 90, 990, 51
  - The answer is 206.333 and 184.14, respectively

Location: Min, Max, Quartiles

- **Min, Max**
  - Min and max are useful in risk analysis, for example, the max value a river can contain; the max magnitude of earthquake a house can stand

- **Quartiles**: Arrange the data in increasing order and determine the median.
  - The first quartile is the median of the part of the entire data set that lies at or below the median of the entire data set.
  - The second quartile is the median of the entire data set.
  - The third quartile is the median of the part of the entire data set that lies at or above the median of the entire data set.

Example: to find 1st, 2nd, 3rd quartiles

- Dataset
  - 1955, 2260, 2020, 2040, 2060, 2040, 2050, 2070, 2165, 2075, 2125

- Quartiles:
  1) arrange the data in ascending order

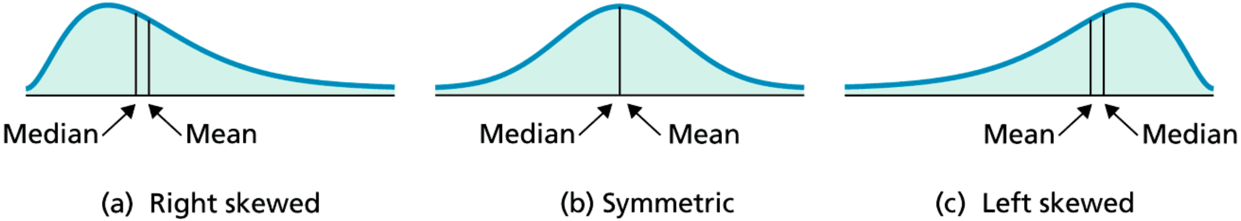
     (e.g.) 1955 2020 2040 2040 2050 2060 2070 2075 2125 2165 2260
  2) The median is the middle of the dataset, which is 2060. That is, the 2nd quartile is 2060.

     1955 2020 2040 2040 2050 2060 2070 2075 2125 2165 2260
  3) The first quartile: the middle of the 1st part (red and 2060), and it‘s (2040 +2040)/2=2040
  4) The third quartile: the middle of the 2nd part (blue and 2060), and it‘s (2075+2125)/2=2100

- Note: there are different methods of calculating quartiles

Relative positions of the mean and median

- This figure shows the relative positions of the mean and median for right-skewed, symmetric, and left-skewed distributions.

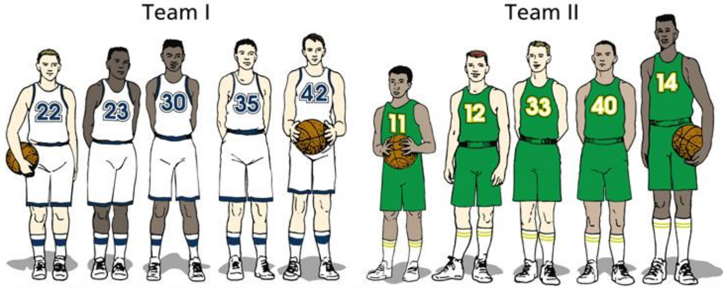


- Question: in the above three figures, provide the mode for each of them.

- Examples:
  - Left skewed: household income in the UK is left skewed with a very long left tail. (https://www.theguardian.com/society/datablog/2012/jun/22/household-incomes-compare); human longevity (https://understandinguncertainty.org/why-life-expectancy-misleading-summary-survival)
  - Right skewed: age-of-companies (http://theodi.org/diverse-uk-companies-open-data); time-to failure distribution

Variability: range, variance, standard deviation, Interquartile Range

- Variability measures risk
- The “data sets” have the same Mean, Median, and Mode, yet clearly differ!

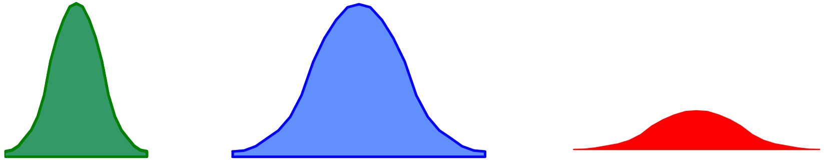


Feet and inches	6'	6'1"	6'4"	6'4"	6'6"	5'7 "	6'	6'4"	6'4"	7'
Inches	72	73	76	76	78	67	72	76	76	84

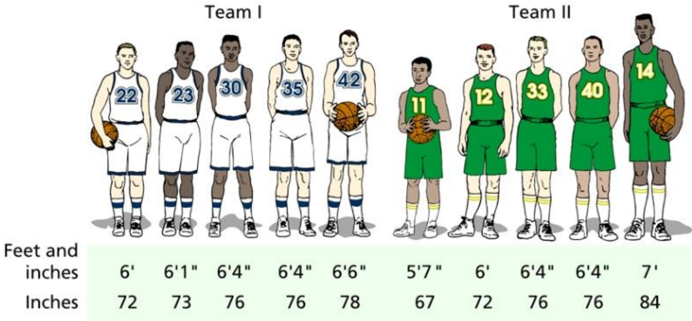
- Measures of Variation or Measures of Spread:
  - range, variance, standard deviation, Interquartile Range
- The two teams have the same
  - Mean (375/5=75)
  - Median (=76)
  - Mode (=76)

Measures of dispersion I – Range

- Purpose of characterising variation and dispersal in a distribution – the wider the dispersal, the less representative the central measure will be.



- Range----Interval between minimum and maximum values. The range: range = largest value - smallest value
  - is easy to calculate and understand
  - is influenced by extreme data values
  - tells you nothing about the variability of your data between these two extreme values.
- Example: Ranges for the 2 teams
  - Range of team I=78-72=6
  - Range of team II=84-67=17



Measures of dispersion II – Variance

- Sample mean (the mean of n observations),

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

- Variance, s^2:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Team I
Team II
22
23
30
35
42
11
12
33
40
14

Feet and inches	6'	6'1"	6'4"	6'4"	6'6"	5'7"	6'	6'4"	6'4"	7'
Inches	72	73	76	76	78	67	72	76	76	84

  - Here, n -1 in the divisor makes this sample variance formula an unbiased estimate of the population variance. $n-1$ **degrees of freedom**: s^2 is the sum of the n values $(x_i - \bar{x})^2$, but they have a constraint: $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, hence, there are $n-1$ degrees of freedom
  - Sensitive to extreme values, which might mislead
  - One of most popular measures of dispersion

- Example
  - For team I: Variance=6
  - For team II: Variance=39

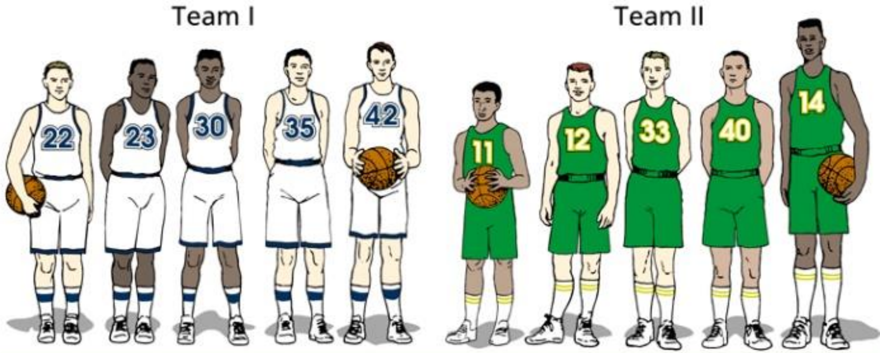
Measures of dispersion III – Standard deviation

- Standard deviation, s (sample), σ (population), SD (either):

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

  - the standard deviation is a measure of how X deviates from its expected value

- Example
  - For team I: Standard deviation=2.45
  - For team II: Standard deviation=6.24

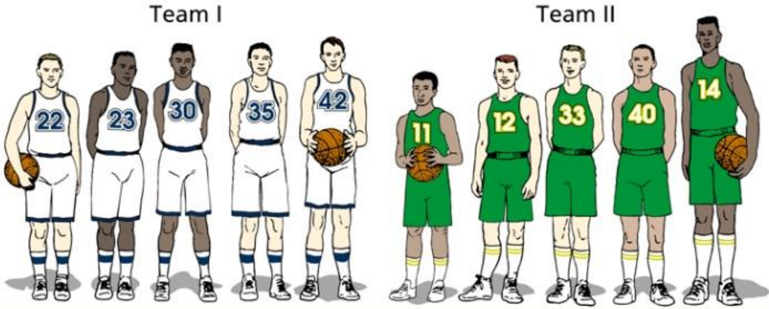


Feet and inches	6'	6'1"	6'4"	6'4"	6'6"	5'7"	6'	6'4"	6'4"	7'
Inches	72	73	76	76	78	67	72	76	76	84

Measures of dispersion IV: Interquartile Range

- **Interquartile Range:** The **interquartile range,** or **IQR,** is the difference between
  - the first and third quartiles; that is, IQR = $Q_3 - Q_1$.

- Example
  - For team I: IQR=76-73=3
  - For team II: IQR=76-72=4



Feet and inches	6'	6'1"	6'4"	6'4"	6'6"	5'7"	6'	6'4"	6'4"	7'
Inches	72	73	76	76	78	67	72	76	76	84

Python pandas built-in functions

```
import numpy as np
arr = np.array([10, 20, 30, 40, 50, 60])
```

Function	Description
np.mean(arr)	mean of array arr
np.std(arr)	standard deviation of array arr
np.median(arr)	median
np.quantile(arr,probs)	quantiles where array arr is the numeric vector whose quantiles are desired and probs is a numeric vector with probabilities in [0,1]. # 30th percentile of arr y <- np.quantile(arr,0.30)
np.ptp(arr,axis=0)	range
np.sum(arr)	sum
np.min(arr)	minimum
np.max(arr)	maximum