Statistic Explained

Ang Man Shun

2012-11-25

Summary

- Why Statistics
- Mean and Expectation
- Variance
- What is variance
- Why variance is squared
- Why stadnard deviation is $\sqrt{\text{Variance}}$
- Why $n-1$, not n
- What is random variable
- What is probability distribution function

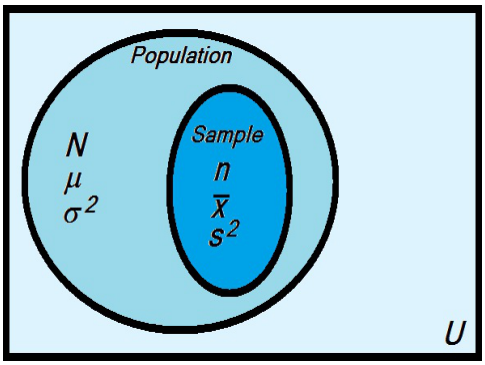
Why statistics

To determine the property of a population, we need to do some test.
Using system as an example, inject some known input, and observe the output, investigate the input/output relationship.

Sometime it is impossible to test all the element in a population
Reason 1. The population size N , is very large ($N \to \infty$)
Reason 2. The experiment is very expensive.

Thus, statistics is a study on how to use finite number of samples to determine the property of the population, with the degree of certainty as high as possible.

Notations



	Population Parameter	Sample Space Statistics
Size	N	n
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard Deviation	σ	s
Proportion of success	p	p_s

Mean , Expectation

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

The term mean has another meaning, the Expectation, in this case, the Expected Value is a weighted sum , for a population that having the element x_i with corresponding occuring frequency $f(x_i)$, the expected value is thus

$$\mu = \sum_{i=1}^{N} x_i f(x_i)$$

The expectation is used in risk evaluation, for example, the chance of a fatal accident is 0.0005, the insurance company will need to pay 1,000,000\$. Then

$$\text{insurnace fee} = \text{Prob(no accident)} \times \begin{matrix}\text{amount the company} \\ \text{had to pay for no accident}\end{matrix} + \text{Prob(accident)} \times \begin{matrix}\text{amount the company} \\ \text{had to pay for accident}\end{matrix}$$

$$= 0.9995 \times 0 + 0.0005 \times 1,000,000\$ = 500\$$$

So the company should charge the client for at least 500\$ (The charge will actually higher due to administrative expenses)

Variance

The sample variance is

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

Expand

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n} \left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x}^2\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2\bar{x}\,(n\bar{x}) + \bar{x}^2\sum_{i=1}^{n} 1\right) \end{aligned}$$

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

What is Variance

Variance, as the name of the term implies, it means the general deviation of all the samples from their central value (the mean).

$$Variance \approx Deviation\ from\ central\ value \propto \frac{1}{central\ tendency}$$

If there is no deviation $\iff$ all data are same $\iff$ $x_i = \bar{x} \forall i$, then

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{x} - \bar{x})^2 = 0$$

i.e. There is no difference between all element, all are the same

If there is some element that are much much different from the central value , for example, let the first one is 10 times to the central value.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\underbrace{(x_1 - \bar{x})^2}_{100} + \frac{1}{n-1}\sum_{i=2}^{n}(x_i - \bar{x})^2$$

Then s^2 is at least 100, then , it means the group has some extreme case.
If all the data are not close to each other, then the variance will be much larger.
As the name implies, larger variance, the data are "less concentrated"

Why variance is squared

If we want to know how diverse the data is , we may just simply add the deviation of them :

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

But this definition is not good, it is always zero :

$$\text{Variance} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}) = \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i - \bar{x}\sum_{i=1}^{n}1\right) = \frac{1}{n-1}(n\bar{x} - n\bar{x}) = 0 \iff \text{alway no variance??}$$

Thus, to avoid this, one method is to take square

$$\text{Variance} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\text{Deviation}^2 \geq 0$$

$$\iff \text{Variance} \begin{cases} > 0 & \text{there is variance} \\ = 0 & \text{iff all data are identical} \\ < 0 & \text{impossible} \end{cases}$$

Why standard deviation is $\sqrt{\text{Variance}}$

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

Let x with physical dimension (i.e. unit) as Q
Then variance has the unit of Q^2 , not Q
Thus, it is better to make a term with same unit as x that derived from variance.

$$\sqrt{Variance} \text{ has the unit of } Q \overset{\text{def}}{=} \text{Standard Deviation}$$

Thus

$$\begin{matrix}\text{Standard}\\ \text{Deviation}\end{matrix} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Why $n-1$, not n

It is related to *degree of freedom*
The simplest definition is

$$\text{Degree of freedom} = \text{Number of values that are free to vary}$$

Example. Four values that have sum equals 100

$$x_1 + x_2 + x_3 + x_4 = 100$$

Once you know x_1 , x_2 , x_3 , then you will know x_4, thus the value of x_4 is not free to vary once you know the others.
For the first 3 values, they are free to vary, the degree of freedom in this case is 3

If x_4 is fixed, for example, $x_4 = 10$, then eqaution becomes

$$x_1 + x_2 + x_3 + 10 = 100 \Rightarrow x_1 + x_2 + x_3 = 90$$

Then once we know x_1 x_2 , x_3 are forced to be the value of $90 - x_1 - x_2$, thus, the degree of freedom this case is 2.

So , for n data, the degree of freedom is $n-1$, since

$$\text{Prob(all event)} = 1 \iff \text{Prob}(x_1) + \text{Prob}(x_2) + ... + \text{Prob}(x_{n-1}) + \text{Prob}(x_n) = 1$$

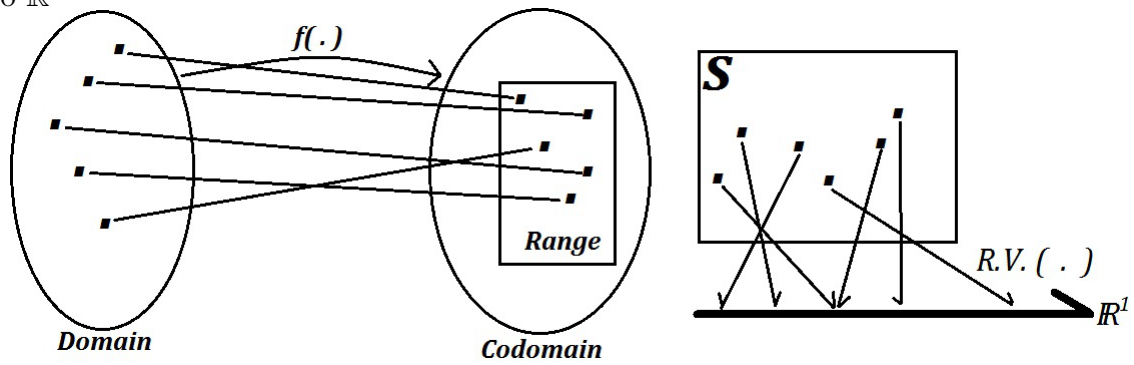
i.e. Once x_1, x_2 , ... , x_{n-1} are found, by $P(A) = 1 - P(\text{not A})$, we can get x_n

What is random variable

Recall that, what is function ?
A *function* , or a *mapping* , is a *realtion* between 2 set. The functions *maps* the first set (input set, *domain*) , into another set (output set, *codomain*)

$$f(.) \text{ is } \begin{cases} \text{Injective} \\ \text{Surjective} \\ \text{Bijective} \end{cases} if \quad \begin{array}{l} \text{one-to-one : no same output from 2 different input} \\ \text{onto : range = codomain} \\ \text{both injective \& surjective} \end{array}$$

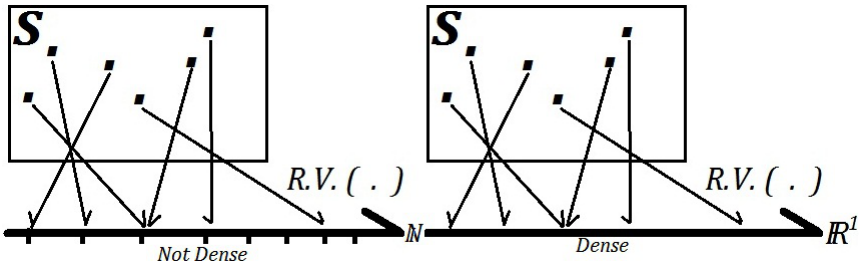
Then , a Random Variable, works like a functions, that *maps* the events in sample space / population into $\mathbb{R}^1$



But for Random Variable $X[.]$, it is not necessary injective nor surjective.

Then, in this way, random variable $x[.]$ can be considered as a mapping that assign a value in $\mathbb{R}^1$
Then it just like function

The random variable can be continuous type and discrete type



What is Probability Distribution Function

For example , roll 2 dice, consider the result of their sum

The reuslt is shown in the following table

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Then the following table show the possible outcomes with corresponding occuring frequency

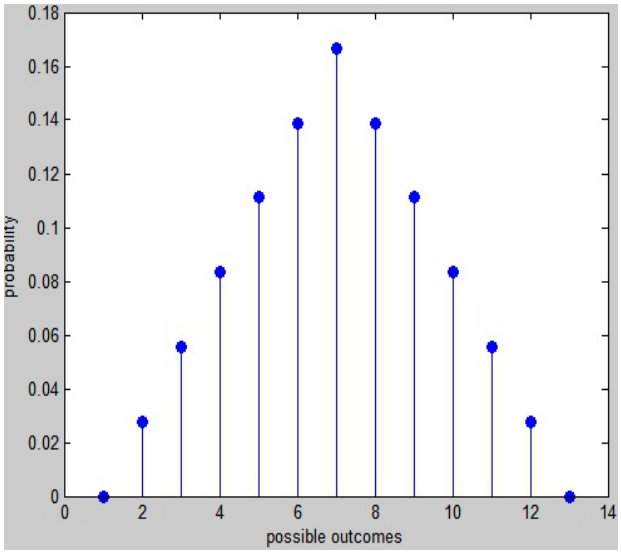
Possible Outcome	2	3	4	5	6	7	8	9	10	11	12
Frequency	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

List out all the probability is Okay, but it is time consuming, why don't we use a function to represent them?

Thus, the probability density function, a short hand notation that *summarize* the distribution behavior for this case is thus

$$f(x) = \begin{cases} \dfrac{x-1}{36} & x \in [2,7] \\ \dfrac{12-x+1}{36} & x \in [8,12] \end{cases}$$

The plot of the PDF (which is a discrete function in this case) is



Thus, the PDF, just like function, it describe the probability of the random variable X , probability that the events that fulfill the requirement can be calculated by

$$P(X \leq a) = \sum_{x_i \leq a} f(x_i)$$

For example, rolling the dice 2 time, the sum is larger than 7 is

$$P(X > 7) = \sum_{x=8}^{12} f(x_i) = \sum_{x=8}^{12} \frac{13-x}{36} = \frac{15}{36}$$

$$-END-$$