Main

Suppose you get some data from the experiment, say:

$$X_1, X_2, X_3, \ldots, X_n$$

We can easily define

$$mean\ of\ the\ data := \frac{1}{n}\sum_{i=1}^{n} X_i$$

The above equation is without any arguments.

But when you want to define a thing to capture the deviance of the data, what should you do?

There are 2 candidates

$$deviance\ of\ the\ data\ version\ 1 := \frac{1}{n}\sum_{i=1}^{n}(X_i - mean\ of\ the\ data)^2$$
$$deviance\ of\ the\ data\ version\ 2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - mean\ of\ the\ data)^2$$

Q: Which is better?

A:

First let us use some new notations

$$\begin{aligned} M &:= mean\ of\ the\ data \\ D_1 &:= deviance\ of\ the\ data\ version\ 1 \\ D_2 &:= deviance\ of\ the\ data\ version\ 2 \end{aligned}$$

Obviously, M, D_1, D_2 are the functions of $\vec{X}$

$$\begin{aligned} M &= M(\vec{X}) \\ D_1 &= D_1(\vec{X}) \\ D_2 &= D_2(\vec{X}) \end{aligned}$$

If we assume that $X_i \sim (\mu, \sigma^2) i.i.d.$

i.e.

$$\begin{aligned} E(X_i) &= \mu \\ Var(X_i) &= \sigma^2 \\ X_i\ &i.i.d. \end{aligned}$$

Then we can get (proof see below)

$$\begin{aligned} E(M) &= \mu \\ E(D_1) &= \frac{n-1}{n}\sigma^2 \\ E(D_2) &= \sigma^2 \end{aligned}$$

So, D_2 is better than D_1

Proof

$$\begin{aligned} E(D_1) &= E(\frac{1}{n}\sum_{i=1}^{n}(X_i - M)^2) \\ &= \frac{1}{n}\sum_{i=1}^{n}E((X_i - M)^2) \\ &= \frac{1}{n}\sum_{i=1}^{n}E(X_i^2 + M^2 - 2X_iM) \\ &= \frac{1}{n}\sum_{i=1}^{n}E(X_i^2) + E(M^2) - 2E(X_iM) \end{aligned}$$

From

$$E(X_i) = \mu$$
$$Var(X_i) = \sigma^2$$
$$X_i \; i.i.d.$$

We get

$$\begin{aligned} E(X_i^2) &= \mu^2 + \sigma^2 \\ E(M) &= \mu \\ Var(M) &= \frac{\sigma^2}{n} \\ E(M^2) &= \mu^2 + \frac{\sigma^2}{n} \end{aligned}$$

and

$$E(X_iX_j) = \begin{cases} \mu^2 + \sigma^2, i = j \\ \mu^2, i \neq j \end{cases}$$

So

$$E(X_iM) = \frac{1}{n}\sum_{j=1}^{n}E(X_iX_j) = \mu^2 + \frac{\sigma^2}{n}$$

So

$$\begin{aligned} E(D_1) &= \frac{1}{n}\sum_{i=1}^{n}((\mu^2 + \sigma^2) + (\mu^2 + \frac{\sigma^2}{n}) - 2(\mu^2 + \frac{\sigma^2}{n})) \\ &= \frac{1}{n}\sum_{i=1}^{n}(\frac{n-1}{n}\sigma^2) \\ &= \frac{1}{n}(\frac{n-1}{n}\sigma^2)\sum_{i=1}^{n}1 \\ &= \frac{n-1}{n}\sigma^2 \end{aligned}$$

And obviously,

$$E(D_2) = \sigma^2$$

QA

Q: Can't I just use D_1

A: In fact, you can, I just say that D_2 is better than D_1.

Q: What if not i.i.d.?

A:

If not, it really doesn't matter too much to use D_1 or D_2.

It is recommended to find a new random variable from your data which is likely to be $i.i.d$.

Another similar method is that, if the D_1 or D_2 of a variable is too big, then you must find a new random variable with smaller D_1 or D_2.

Q: In your taxis project?

A:

In my project:

- runs are not $i.i.d.$.
- tracks/worms are $i.i.d.$!

So, it doesn't matter I use D_1 or D_2 when handling runs.

When handling tracks/worms, I will use D_2

Q: How do you know if a variable is i.i.d. or not? You are not the god!

A:

Yes, you are right, only the God know if a variable is i.i.d or not.

But, we can guess.

For example, $X_1, X_2, X_3, \dots, X_n$ can be the height of a human. If you choose X_1 from Jilin and X_2 from Anhui, then it is likely that they are not i.i.d. But if you choose X_1 from Changchun and X_2 from Songyuan, it is likely that they are i.i.d..

For another example, $X_1, X_2, X_3, \dots, X_n$ can be the run speed of a worm. If you choose X_1 from N_2 and X_2 from $RIA - twk18$, then it is likely that they are not i.i.d. But if you choose X_1 from N_2 and X_2 from N_2, it is likely that they are i.i.d..