Ec140 - Variance and Sampling

Fernando Hoces la Guardia
06/27/2022

Housekeeping

- Updated Syllabus
  - Fixed dates on PS1. Due this Friday 5pm on gradescope.
- Unofficial Course Capture! (second attempt!)
- Finish Ch 1 of MM by the end of the week.

Variance and Standard Deviation 1/N (Sample)

- Random variables -> probabilities -> distributions -> data -> mean/expectation
- Let's look at another data set:

Rotten Tomatos Scores

i	**Harry Potter Movies (X)**	**Game of Thrones Seasons (Y)**
1	81	90
2	83	96
3	90	96
4	88	97
5	78	93
6	83	94
7	77	93
8	96	55

Variance and Standard Deviation 2/N (Sample)

$$\overline{X} = 84.5$$
$$\overline{Y} = 89.2$$

$$\frac{\sum_{1:8}\left(x - \overline{X}\right)}{8} = 0$$

$$\frac{\sum_{1:8}\left(y - \overline{Y}\right)}{8} = 0$$

Rotten Tomatos Scores

i	X	$X - \overline{X}$	Y	$Y - \overline{Y}$
1	81	-3.5	90	0.75
2	83	-1.5	96	6.75
3	90	5.5	96	6.75
4	88	3.5	97	7.75
5	78	-6.5	93	3.75
6	83	-1.5	94	4.75
7	77	-7.5	93	3.75
8	96	11.5	55	-34.25

Variance and Standard Deviation 2/N (Sample)

$$\overline{X} = 84.5$$
$$\overline{Y} = 89.2$$

$$\frac{\sum_{1:8}\left(x - \overline{X}\right)}{8} = 0$$

$$\frac{\sum_{1:8}\left(y - \overline{Y}\right)}{8} = 0$$

Rotten Tomatos Scores

i	X	$X - \overline{X}$	$(X - \overline{X})^2$	Y	$Y - \overline{Y}$	$(X - \overline{X})^2$
1	81	-3.5	12.25	90	0.75	0.5625
2	83	-1.5	2.25	96	6.75	45.5625
3	90	5.5	30.25	96	6.75	45.5625
4	88	3.5	12.25	97	7.75	60.0625
5	78	-6.5	42.25	93	3.75	14.0625
6	83	-1.5	2.25	94	4.75	22.5625
7	77	-7.5	56.25	93	3.75	14.0625
8	96	11.5	132.25	55	-34.25	1173.0625

Variance and Standard Deviation 2/N (Sample)

$$\overline{X} = 84.5$$
$$\overline{Y} = 89.2$$

$$\frac{\sum_{1:8}\left(x - \overline{X}\right)}{8} = 0$$

$$\frac{\sum_{1:8}\left(y - \overline{Y}\right)}{8} = 0$$

$$\frac{\sum_{1:8}\left(x - \overline{X}\right)^2}{8} = 36.2$$

$$\frac{\sum_{1:8}\left(y - \overline{Y}\right)^2}{8} = 171.9$$

Rotten Tomatos Scores

i	X	$X - \overline{X}$	$(X - \overline{X})^2$	Y	$Y - \overline{Y}$	$(X - \overline{X})^2$
1	81	-3.5	12.25	90	0.75	0.5625
2	83	-1.5	2.25	96	6.75	45.5625
3	90	5.5	30.25	96	6.75	45.5625
4	88	3.5	12.25	97	7.75	60.0625
5	78	-6.5	42.25	93	3.75	14.0625
6	83	-1.5	2.25	94	4.75	22.5625
7	77	-7.5	56.25	93	3.75	14.0625
8	96	11.5	132.25	55	-34.25	1173.0625

- These represent the sample variances of HP and GoT ratings
- But what about the units?

Variance and Standard Deviation 2/N (Sample)

$$\overline{X} = 84.5$$
$$\overline{Y} = 89.2$$

$$s_X^2 = \frac{\sum_{1:8}\left(x - \overline{X}\right)^2}{8-1} = 41.4$$

$$s_Y^2 = \frac{\sum_{1:8}\left(y - \overline{Y}\right)^2}{8-1} = 196.5$$

Rotten Tomatos Scores

i	X	$X - \overline{X}$	$(X - \overline{X})^2$	Y	$Y - \overline{Y}$	$(X - \overline{X})^2$
1	81	-3.5	12.25	90	0.75	0.5625
2	83	-1.5	2.25	96	6.75	45.5625
3	90	5.5	30.25	96	6.75	45.5625
4	88	3.5	12.25	97	7.75	60.0625
5	78	-6.5	42.25	93	3.75	14.0625
6	83	-1.5	2.25	94	4.75	22.5625
7	77	-7.5	56.25	93	3.75	14.0625
8	96	11.5	132.25	55	-34.25	1173.0625

- Due to a minor technicality we divide by $N - 1$ instead of N (not relevant for the course).
- s_X^2 and s_X correspond to the sample variance and standard deviation.

Variance and Standard Deviation 3/N (Population)

Let's focus on the formula for mean and sample variance of Harry Potter only. And for now, I will continue use N (8) in the denominator for the variane to illustrate the following concept.

$$\overline{X} = \frac{\sum_{1:8} x}{8} = 84.5$$

$$s_X^2 = \frac{\sum_{1:8} \left(x - \overline{X}\right)^2}{8} = 36.2$$

Variance and Standard Deviation 4/N (Population)

Sample

Population

$$\overline{X} = \frac{\sum_{1:8} x}{8} = 84.5$$

$$s_X^2 = \frac{\sum_{1:8} \left(x - \overline{X}\right)^2}{8} = 36.2$$

Variance and Standard Deviation 4/N (Population)

Sample

Population

$$\overline{X} = \frac{\sum_{1:8} x}{8} = \sum_{1:8} x \frac{1}{8} =$$

$$\sum_{1:8} x \times prop(x) = 84.5$$

$$s_X^2 = \frac{\sum_{1:8} \left(x - \overline{X}\right)^2}{8} = 36.2$$

Variance and Standard Deviation 4/N (Population)

Sample

$$\overline{X} = \frac{\sum_{1:8} x}{8} = \sum_{1:8} x \frac{1}{8} =$$

$$\sum_{1:8} x \times prop(x) = 84.5$$

$$s_X^2 = \frac{\sum_{1:8} \left(x - \overline{X}\right)^2}{8} = 36.2$$

Population

$$\mathbb{E}(X) = \sum_x x f(x)$$

Variance and Standard Deviation 4/N (Population)

Sample

$$\overline{X} = \frac{\sum_{1:8} x}{8} = \sum_{1:8} x \frac{1}{8} = \sum_{1:8} x \times prop(x) = 84.5$$

$$s_X^2 = \frac{\sum_{1:8} \left(x - \overline{X}\right)^2}{8} = 36.2$$

Population

$$\mathbb{E}(X) = \sum_{x} x f(x)$$

Variance and Standard Deviation 4/N (Population)

Sample

$$\overline{X} = \frac{\sum_{1:8} x}{8} = \sum_{1:8} x \frac{1}{8} =$$

$$\sum_{1:8} x \times prop(x) = 84.5$$

$$s_X^2 = \frac{\sum_{1:8} g(x)}{8} = 36.2$$

Population

$$\mathbb{E}(X) = \sum_x x f(x)$$

Variance and Standard Deviation 4/N (Population)

Sample

$$\overline{X} = \frac{\sum_{1:8} x}{8} = \sum_{1:8} x \frac{1}{8} =$$

$$\sum_{1:8} x \times prop(x)$$

$$s_X^2 = \frac{\sum_{1:8} g(x)}{8} = \sum_{1:8} g(x) \frac{1}{8} =$$

$$\sum_{1:8} g(x) \times prop(x)$$

Population

$$\mathbb{E}(X) = \sum_x x f(x)$$

$$\mathbb{E}(g(x)) =$$

$$\mathbb{E}\Big((X - \overline{X})^2\Big) = \sum_x (x - E(X))^2 f(x)$$

Variance and Standard Deviation 4/N (Population)

Sample

$$\overline{X} = \frac{\sum_{1:8} x}{8} = \sum_{1:8} x \frac{1}{8} =$$

$$\sum_{1:8} x \times prop(x)$$

$$s_X^2 = \frac{\sum_{1:8} g(x)}{8} = \sum_{1:8} g(x) \frac{1}{8} =$$

$$\sum_{1:8} g(x) \times prop(x)$$

Population

$$\mathbb{E}(X) = \sum_x x f(x)$$

$$\mathbb{E}(g(x)) =$$

$$\mathbb{E}\left((X - E(X))^2\right) = \sum_x (x - E(X))^2 f(x)$$

Usually $E(X)$ is defined as μ, so you might see:

Variance and Standard Deviation 5/N (Done!)

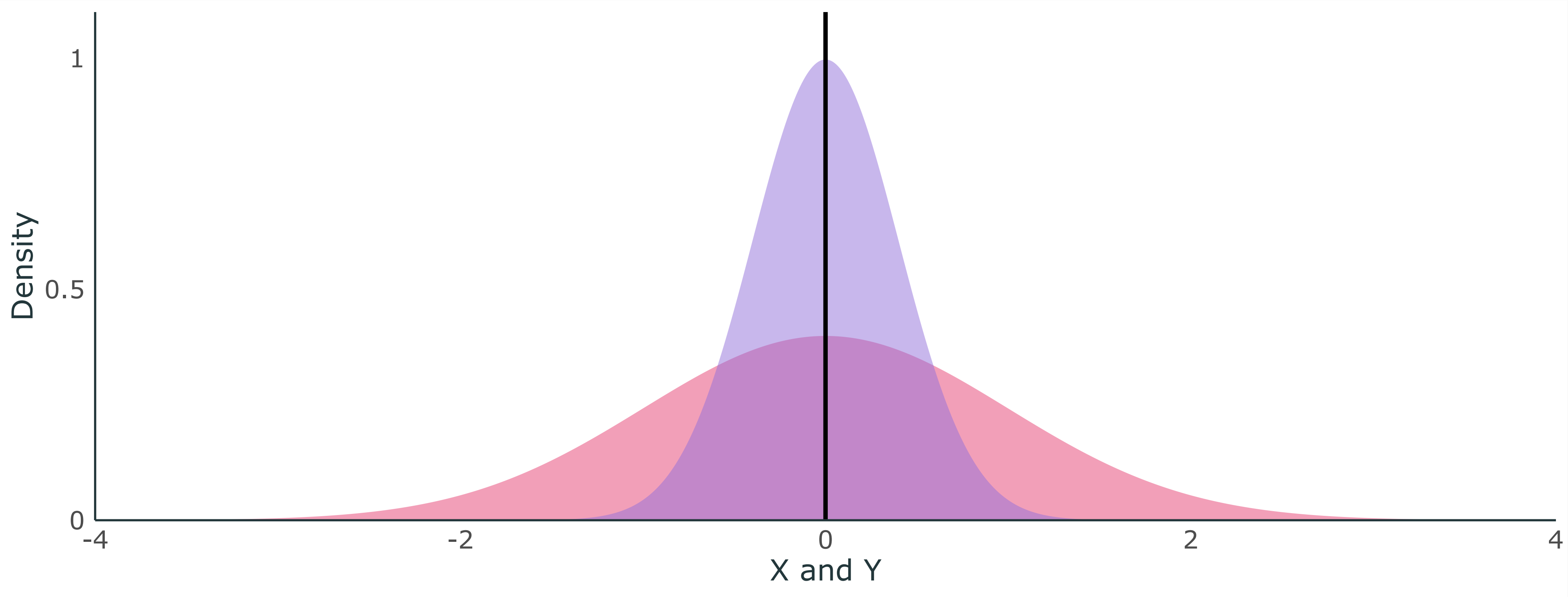
You now know what are the variance and standard deviation and where do they come from!

$$Var(X) = \sigma^2 = \mathbb{E}\left((X - \mu)^2\right)$$

$$SD(X) = \sigma = \sqrt{\mathbb{E}\left((X - \mu)^2\right)}$$

Variance

Random variables X and Y share the same population mean, but are distributed differently.



Variance

Rule 1

$\mathbf{Var}(X) = \mathbf{0} \iff X$ is a constant.

- If a random variable never deviates from its mean, then it has zero variance.
- If a random variable is always equal to its mean, then it's a (not-so-random) constant.

Variance

Rule 2

For any constants a and b, $\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X)$.

Example

Suppose X is the high temperature in degrees Celsius in Eugene during August. If Y is the temperature in degrees Fahrenheit, then $Y = 32 + \frac{9}{5}X$. What is $\mathrm{Var}(Y)$?

- $\mathrm{Var}(Y) = (\frac{9}{5})^2\,\mathrm{Var}(X) = \frac{81}{25}\,\mathrm{Var}(X)$.

Variance

Variance Rule 3

For constants a and b,

$$\mathrm{Var}(aX + bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X, Y).$$

- If X and Y are uncorrelated, then $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$
- If X and Y are uncorrelated, then $\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Expectation and Variance of the Sample Mean

- Time for a subtle, but very important change of focus.
- Until now we have been talking about the expectation and variance of a random variable. Now we are going to focus on the expectation and variance of the **mean of a collection of random variables**.
  - Wait? We talk last class that the expectation is like the mean. So basically you want to focus on the mean of the mean? What do that we even mean (!)?
- A combination of random variables is also a random variable (e.g., remember how a Binomial random variable was a summation of Bernoullis?). In particular, a summation of random variables $Y_1, Y_2, Y_3 \ldots, Y_n$ is also a random variable, and the sample size is a constant. Hence, $\overline{Y} = \frac{\sum_n Y}{n}$ is also a random variable.

Expectation and Variance of the Sample Mean

- This potentially cofusing, as before we would have one random variable X, from which we would sample a collection of values $\{x_1, x_2, \ldots, x_n\}$, and with this we could compute the mean $\overline{X}$.
- But now we will have to imagine that we do this sampling multiple times. To help with the transition (and because it will also help with future notation), I will use the letter $Y_{\text{number } i}$ to denote random variable number i (where i is used to represent any given number) or Y_i for short.
- Hard to imagine if one sample corresponds to one survey that cost millions of dollars and took months or years to carry out, but think about it as a thought exercise. Believing in the multiverse in this case helps with the thought exercise :)

Expectation and Variance of the Sample Mean

- Before we start combining random variables, we need to make two important assumptions: **independence** and **identically distributed**.

- **Independence:** Two (or more) random variables are independent when knowing one random variable provides no information about the value of the other. A bit more formally, if two random variables X and Y are independent, then $P(X = x \& Y = y) = P(X = x)P(Y = y)$. A nice shorthand is to think of "independence as multiplication".

- **Identically Distributed:** Two (or more) random variables are identically distributed if they have the same probability distribution (or density) function. As a consequence these random variables have the same expected value, let's call it μ_Y, and standard deviation σ_Y

Expectation of the Sample Mean

- The expected value of the sample mean $(\overline{Y})$ is, at first glance, nothing too surprising:

$$\mathbb{E}(\overline{Y}) = \frac{1}{n}\sum \mathbb{E}(Y_i)$$

$$\mathbb{E}(\overline{Y}) = \frac{1}{n}\sum \mu_Y = \frac{n\mu_Y}{n}$$

$$\mathbb{E}(\overline{Y}) = \mu_Y$$

(The first equality comes from Rule 2 and 3 of expectation. The second equality comes from identical means, and the third from summing n times the same constant)

The Standard Deviation of the Sample Mean

- The formula for variance and standard deviation of the sample mean $(\overline{Y})$ is less straight forward:

$$Var(\overline{Y}) = \frac{\sigma_Y^2}{n}$$

$$SD(\overline{Y}) = \frac{\sigma_Y}{\sqrt{n}}$$

- Unlike the expectation of the mean its the standard deviation is not the same as the standard deviation of a single random variable. Moreover, it shrinks (to zero) as the sample size increases.

Exact v. Approximate Approches

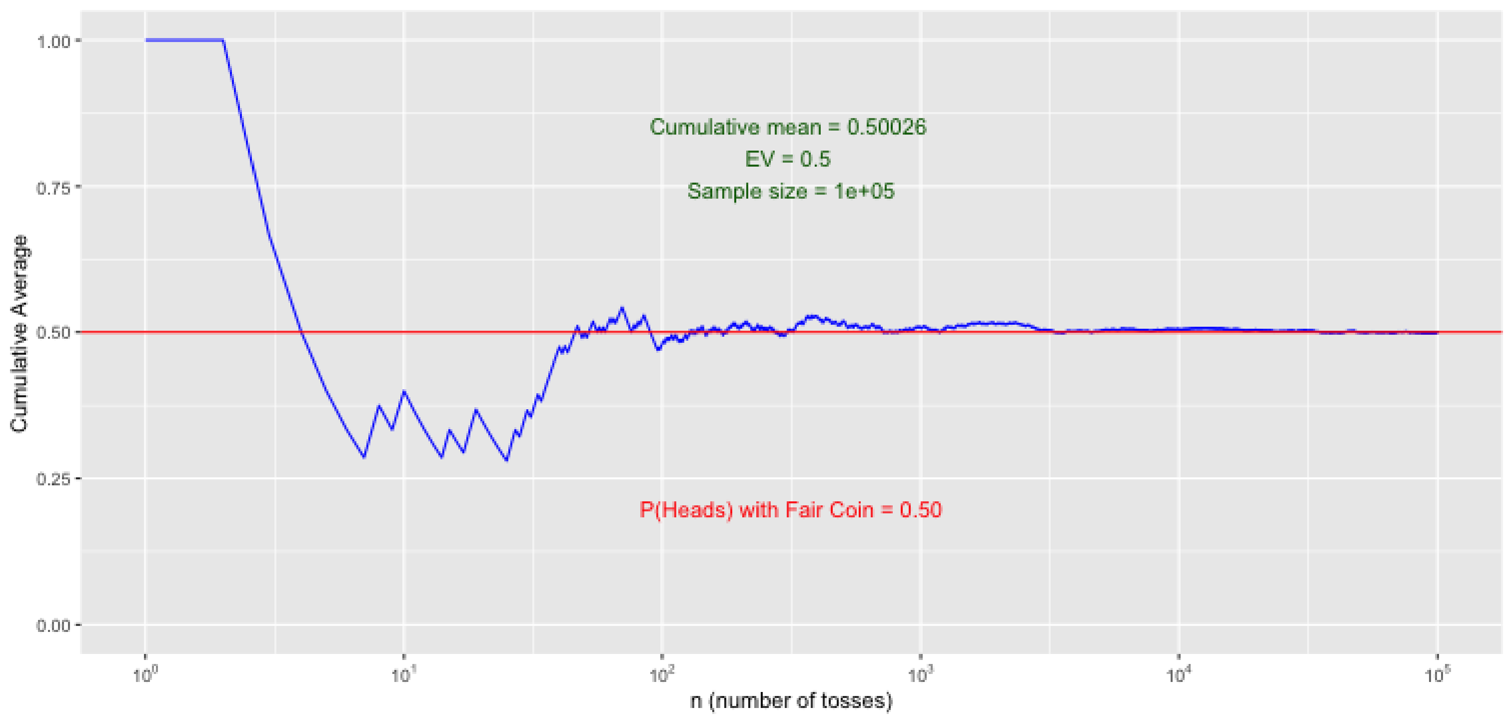
- We just examine the expectation and variance for the sampling mean $(\overline{Y})$ using theoretical properties of $E()$ and $Var()$ this results hold true *regardless* of the sample size n. But at the same time answer to a highly hypothetical question (what is the population mean of the sample mean?).

- In addition to this "exact" derivation. We can also ask what happens with $\overline{Y}$ when its sample size (n) increases. This "approximate" approach is refer to as the asymptotic properties $\overline{Y}$ (but either term is fine).

- In econometrics we make extensive use of the two following approximations:

Law of Large Numbers (LLN)

- Under general conditions, of independence (and finite variance), $\overline{Y}$ will be near its expected value (μ_Y) with arbitrary high probability as n is large $(\overline{Y} \xrightarrow{p} \mu_Y)$



Law of Large Numbers (LLN): Observations

- In practical terms n doesn't have to be too large. $n = \mathbf{25 - 35}$ tends to be enought. In social sciences we tend to work with much more that.

- As n grows the standard deviation of the sample mean drops to zero. In the example above: $SD(\overline{Y_{10}}) = 0.15$, $SD(\overline{Y_{100}}) = 0.04$, $SD(\overline{Y_{1000}}) = 0.01$, $SD(\overline{Y_{10000}}) = 0$.

Central Limit Theorem (CLT)

- Under general conditions, of independence (and finite variance), the **distribution** of $\overline{Y}$ is approximately $N(\mu_Y, \frac{\sigma_Y^2}{n})$ as n is large.
- This is true **for any** type of distribution (not only normal) of the underlying Y_i.
- This is very hard to believe, so we are going to spend some significant time in Seeing Theory simulating different scenarios (and probably over session too).
- In real life the key assumption is that of independence. If observations are obtained at random, a procedure called *random sampling*, then independence achieved.
- Random sampling is necessary so the LLN and CLT can be used.