

Ec140 - Core Concepts from Statistics

Fernando Hoces la Guardia

06/23/2022

Housekeeping

- Updated Syllabus
- Unofficial Course Capture!
- What is the weirdest concept you remember from yesterday?
- Switch to finish yesterday's slides

This Lecture

- Introduction to Data
- Mean and Expectation
- Variance and Standard Deviation

What Defines a Data Set?

- Data Set is the collection of any type information (of multiple *Datum*)
- In quantitative analysis we focus on *structured* data sets (unlike, for example, unstructured field notes).
- In econometrics the most common way to structure data is in tabular, or rectangular, form.
- A tabular data set is a collection of variables that with information for one or more entities.
- Entities can represent multiple individuals, one individual over time, firms, countries, etc.
- Variables are represented in columns, and observations are represented by rows. (for more on variables [The Effect, Ch3](#))

Data

Sample of US workers (Current Population Survey, 1976)

	Wage ↕	Education ↕	Tenure ↕	Female? ↕	Non-white? ↕
1	3.1	11	0	1	0
2	3.24	12	2	1	0
3	3	11	0	0	0
4	6	8	28	0	0
5	5.3	12	2	0	0
6	8.75	16	8	0	0

Showing 1 to 6 of 526 entries

But What Can We Do With Data?

- We summarized it! (see the great [short story by J.L. Borges](#) on why summarizing is essential)
- One of the first thing we do when summarizing data is to look at *some type of average*.
 - Wait? *Type* of average? Isn't there just one average? called *the mean*?

But What Can We Do With Data?

- We summarized it! (see the great [short story by J.L. Borges](#) on why summarizing is essential)
- One of the first thing we do when summarizing data is to look at *some type of average*.
 - Wait? *Type* of average? Isn't there just one average? called *the mean*?
- These is also referred as measure of central tendency.
- In this course, we will focus primarily on the mean. **From now on in this course**

average noun



av·er·age | \ 'a-v(ə-)rij \ 

Definition of *average* (Entry 1 of 3)

- 1 a** : a single value (such as a mean, mode, or median) that summarizes or represents the general significance of a set of unequal values

b : [MEAN sense 1b](#)
- 2 a** : an estimation of or approximation to an [arithmetic mean](#)

b : a level (as of intelligence) typical of a group, class, or series

// above the *average*

Mean

- The mean is defined by the sum of a set of values divided by the number of values.

Let's look at the mean from the "hang out with a friend" exercise.

- Total over N

$$\textit{Average}(X) = \frac{1 \times 10 + 2 \times 9 + 3 \times 11}{30} = 2.03$$

- One number, **highly informative** for a variable of interest.
- Always important to keep an eye on the units and magnitude (relevant for PS1).

Mean: Notation

$$\textit{Average}(X) = \frac{1 \times 10 + 2 \times 9 + 3 \times 11}{30} = 2.03$$

$$\textit{Ave}(X) = 1 \times \frac{10}{30} + 2 \times \frac{9}{30} + 3 \times \frac{11}{30} = 2.03$$

$$\overline{X}_n = x_1 \times \textit{proportion}(x_1) + x_2 \times \textit{proportion}(x_2) + x_3 \times \textit{proportion}(x_3)$$

$$\overline{X}_n = \text{summing across all } x (x \times \textit{proportion}_n(x))$$

$$\overline{X}_n = \sum_x x \times \textit{prop}_n(x)$$

Expected Value

- Let's look at the histogram for the exercise above (drawn in the board) and pretend it is not a sample but the entire population. How can we move from frequencies into probabilities?
- Replace frequencies by probabilities
- The population version of the sample mean is the **expected value**.

Expected Value: Definition (Discrete)

The expected value of a discrete random variable X is the weighted average of its k values $\{x_1, \dots, x_k\}$ and their associated probabilities:

$$\begin{aligned}\mathbb{E}(X) &= x_1 \mathbb{P}(X = x_1) + x_2 \mathbb{P}(X = x_2) + \dots + x_k \mathbb{P}(X = x_N) \\ &= \sum_x x \mathbb{P}(X = x)\end{aligned}$$

- Also known as the population mean.

Expected Value: Definition (Discrete)

The expected value of a discrete random variable X is the weighted average of its k values $\{x_1, \dots, x_k\}$ and their associated probabilities:

$$\begin{aligned}\mathbb{E}(X) &= x_1 \mathbb{P}(X = x_1) + x_2 \mathbb{P}(X = x_2) + \dots + x_k \mathbb{P}(X = x_k) \\ &= \sum_x \textcolor{brown}{x} \mathbb{P}(\textcolor{teal}{X} = \textcolor{teal}{x}) = \sum_x \textcolor{brown}{x} \textcolor{teal}{f}(x)\end{aligned}$$

- Also known as the population mean. Compare it to the sample mean:

$$\overline{X}_n = \sum_x \textcolor{brown}{x} \times \textcolor{teal}{prop}_n(x_1)$$

Expected Value

Example

Rolling a six-sided die once can take values $\{1, 2, 3, 4, 5, 6\}$, each with equal probability. What is the expected value of a roll?

$$\mathbb{E}(\text{Roll}) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5.$$

- **Note:** The expected value can be a number that isn't a possible outcome of X .

Expected Value. Definition (Continuous)

- Compare it to the discrete version
- Continuous

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx.$$

- Discrete

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

Expected Value. Definition (Continuous)

- Compare it to the discrete version
- Continuous

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx.$$

- Discrete

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

This explanation was inspired by
[this lecture from Eddie Woo](#)

Expected Value. Definition. One Last Thing 1/2

Let's go back to the mean of our exercise:

$$\overline{X}_n = 1 \times \frac{10}{30} + 2 \times \frac{9}{30} + 3 \times \frac{11}{30} = 2.03$$

But now let's switch the values of the random variables to: 10, 20, 30. How should we compute the mean?

$$\overline{g(X)}_n = 10 \times \frac{10}{30} + 20 \times \frac{9}{30} + 30 \times \frac{11}{30} = 20.33$$

Expected Value. Definition. One Last Thing 2/2

Hence, we can conclude, that for a random variable X , any transformation $g(X)$ has a sample average:

$$\overline{X}_n = \sum_x g(x) \times \text{prop}_n(x_1)$$

And an expectation:

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

The same idea applies in the case of a continuous random variable

Expected Value: Rules (or Properties)

Rule 1

For any constant c , $\mathbb{E}(c) = c$.

Not-so-exciting examples

$$\mathbb{E}(5) = 5.$$

$$\mathbb{E}(1) = 1.$$

$$\mathbb{E}(4700) = 4700.$$

Expected Value

Rule 2

For any constants a and b , $\mathbb{E}(aX + b) = a \mathbb{E}(X) + b$.

Example

Suppose X is the high temperature in degrees Celsius in Eugene during August. The long-run average is $\mathbb{E}(X) = 28$. If Y is the temperature in degrees Fahrenheit, then $Y = 32 + \frac{9}{5}X$. What is $\mathbb{E}(Y)$?

- $\mathbb{E}(Y) = 32 + \frac{9}{5} \mathbb{E}(X) = 32 + \frac{9}{5} \times 28 = 82.4$.

Expected Value

Rule 3: Linearity

If $\{a_1, a_2, \dots, a_n\}$ are constants and $\{X_1, X_2, \dots, X_n\}$ are random variables, then

$$\mathbb{E}(a_1 X_1 + a_2 X_2 + \dots + a_n X_n) = a_1 \mathbb{E}(X_1) + a_2 \mathbb{E}(X_2) + \dots + a_n \mathbb{E}(X_n).$$

In English, the expected value of the sum = the sum of expected values.

Expected Value

Rule 3

The expected value of the sum = the sum of expected values.

Example

Suppose that a coffee shop sells X_1 small, X_2 medium, and X_3 large caffeinated beverages in a day. The quantities sold are random with expected values $\mathbb{E}(X_1) = 43$, $\mathbb{E}(X_2) = 56$, and $\mathbb{E}(X_3) = 21$. The prices of small, medium, and large beverages are 1.75, 2.50, and 3.25 dollars. What is expected revenue?

$$\begin{aligned}\mathbb{E}(1.75X_1 + 2.50X_2 + 3.25X_3) &= 1.75 \mathbb{E}(X_1) + 2.50 \mathbb{E}(X_2) + 3.25 \mathbb{E}(X_3) \\ &= 1.75(43) + 2.50(56) + 3.25(21) \\ &= 283.5\end{aligned}$$

Expected Value

Caution

Previously, we found that the expected value of rolling a six-sided die is $\mathbb{E}(\text{Roll}) = 3.5$.

- If we square this number, we get $[\mathbb{E}(\text{Roll})]^2 = 12.25$.

Is $[\mathbb{E}(\text{Roll})]^2$ the same as $\mathbb{E}(\text{Roll}^2)$?

No!

Expected Value

Caution

Except in special cases, the transformation of an expected value **is not** the expected value of a transformed random variable.

For some function $g(\cdot)$, it is typically the case that

$$g(\mathbb{E}(X)) \neq \mathbb{E}(g(X)).$$

Activity 1

- Let's watch [another Stat 110's video](#). Then get together in groups of 3 and discuss:
 - Don't worry about the law of large numbers yet
 - How does the random variables becomes continuous?
 - How does linearity help with computations?