Economics 361
Expectation

Jun Ishii *
Department of Economics
Amherst College

Fall 2023

1 Overview

The expected value (*expectation*) of a random variable is essentially the weighted sum of all the possible realization of the random variable, with the PMF/PDF value of the realization chosen as weights. Using standard notation, $E[X]$ represents the expected value of X

$$E[X] = \begin{cases} \sum_{x=-\infty}^{\infty} x\ f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x\ f(x)\ dx & \text{if } X \text{ is continuous} \end{cases}$$

Recall that $f(x) = 0$ for any $x \notin \mathcal{X}$

We can also take the expectation of functions of random variables

$$E[g(X)] = \begin{cases} \sum_{x=-\infty}^{\infty} g(x)\ f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x)\ f(x)\ dx & \text{if } X \text{ is continuous} \end{cases}$$

Similarly, you can take expectations with respect to the joint distribution

$$E[g(X,Y)] = \begin{cases} \sum_{x=-\infty}^{\infty} \sum_{y=-\infty}^{\infty} g(x,y)\ f(x,y) & \text{if } (X,Y) \text{ is discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)\ f(x,y)\ dx\ dy & \text{if } (X,Y) \text{ is continuous} \end{cases}$$

Again, recall that $f(x,y) = 0$ if $x \notin \mathcal{X}$ or $y \notin \mathcal{Y}$

And, similarly, you can take expectations with respect to the conditional distribution

$$E[g(X,Y) \mid X = x] = \begin{cases} \sum_{y=-\infty}^{\infty} g(x,y)\ f(y|x) & \text{if } (X,Y) \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x,y)\ f(y|x)\ dy & \text{if } (X,Y) \text{ is continuous} \end{cases}$$

*Office: Converse Hall 315 Phone: (413) 542-2901 E-mail: jishii@amherst.edu

2 Moments

The expected value of the random variable and of the functions of the random variable are often referred to as **moments**. For random variable X

- $E[X^k]$ is the k^{th} (raw) moment of X
- $E[(X-\mu_X)^k]$, where $\mu_X = E[X]$, is the k^{th} cental moment of X

The moments correspond to some important (and familiar) statistical terms

$$\begin{array}{rcccl} \text{Mean of X} & = & \mu_X & = & E[X] \\ \text{Variance of X} & = & \sigma_X^2 & = & E[\ (X-E(X))^2\ ] \\ \text{Covariance of X,Y} & = & \sigma_{XY} & = & E[\ (X-E[X])\ (Y-E[Y])\ ] \end{array}$$

The mean is the first (raw) moment and the variance the second *central* moment.

The first two (mean, variance) can be calculated using the marginal distribution of X but the third (covariance) requires the joint distribution of X, Y. Of course, all three can be calculated from the joint distribution as the marginal distribution can be obtained from the joint distribution

$$\begin{aligned} E[g(X)] &= \int_{-\infty}^{\infty} g(x)\ f(x)\ dx \\ &= \int_{-\infty}^{\infty} g(x) \underbrace{\left(\int_{-\infty}^{\infty} f(x,y)\ dy \right)}_{f(x)}\ dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x) f(x,y)\ dy\ dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x) f(x,y)\ dy\ dx \qquad \text{(order of integration does not matter)} \end{aligned}$$

From above, it should be clear that

$$\underbrace{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x) f(x,y)\ dy\ dx}_{E[g(X)]} \neq \underbrace{g\left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\ dy\ dx \right)}_{g(E[X])} \qquad \text{in general}$$

$E[g(x)] = g(E[X])$ only for certain functions $g(\cdot)$ and/or certain distributions $f(x)$. This result is an application of the more general **Jensen's Inequality**.

2.1 Conditional Distribution

Moments can also be calculated using the conditional distribution. Recall that the conditional distribution of $\{Y|X = x\}$ is just the distribution of Y *after* the sample space has been restricted such that $X = x$ always.

- $E[Y|X = x]$ is the mean of Y given $X = x$ ("conditional mean")
- $E[(Y - E[Y])^2 \mid X = x]$ is the variance of Y given $X = x$ ("conditional variance")

The following statements can be proven for X and Y distributed **independently** of each other

- $E[Y|X = x] = E[Y]$
- $E[(X - E[X])^2 \mid Y = y] = E[(X - E[X])^2]$
- $E[(X - E[X])\ (Y - E[Y])\ ] = 0$

Consider the proof for the first statement

$$\begin{aligned} E[Y|X = x] &= \int_{-\infty}^{\infty} y \underbrace{\left(\frac{f(x,y)}{f(x)}\right)}_{f(y|x)} dy \qquad \text{using definition of conditional probability/distribution} \\ &= \int_{-\infty}^{\infty} y \left(\frac{f(x)f(y)}{f(x)}\right) dy \qquad \text{as } f(x,y) = f(x)f(y) \text{ due to independence} \\ &= \int_{-\infty}^{\infty} y\ f(y)\ dy \quad = \quad E[Y] \end{aligned}$$

The other two statements can be similarly proven.

So, if X and Y are independently distributed of each other, the conditional mean equals the (unconditional) mean, the conditional variance equals the (unconditional) variance, and the covariance between the two is zero.

Note that the above three statements, even together, do not necessarily imply that X and Y are independently distributed. The first two moments may not differ with conditioning but the later moments may.

2.2 Law of Iterated Expectations

Note that $f(x,y) = f(y|x)\ f(x)$. This result can be used to show that the "unconditional" expectation of $g(Y)$ can be obtained by iterative application of the conditional and marginal expectation

$$E[g(Y)] \quad = \quad E_X[\ E_{Y|X}[\ g(X,Y)\ ]\ ]$$

where $E_X[\cdot]$ is the expected value with respect to the marginal distribution and $E_{Y|X}[\cdot]$ is the expected value with respect to the conditional distribution.

First take the expected value of $g(X,Y)$ with respect to conditional $f(Y|X)$

$$E_{Y|X}[g(X,Y)] \quad = \quad \underbrace{\int_{-\infty}^{\infty} g(X,y)\ f(y|X)\ dy}_{\text{will result in some function of } X}$$

Then take the expected value of $E_{Y|X}[g(X,Y)]$ with respect to marginal distribution $f(X)$

$$\begin{aligned}
E_X[\ E_{Y|X}[g(X,Y)]\ ] &= \int_{-\infty}^{\infty} E_{Y|X}[g(X,Y)]\ f(x)\ dx \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} g(x,y)\ f(y|x)\ dy \right)\ f(x)\ dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)\ f(y|x)\ f(x)\ dy\ dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)\ f(x,y)\ dy\ dx \\
&= E[g(X,Y)]
\end{aligned}$$

Taking the expected value while conditioning on X essentially makes X a constant for that calculation; X is fixed to take the value x for that conditional expectation calculation.

The Law of Iterated Expectations is a useful property. Often, we will have already calculated the conditional expectation of some random function $g(X,Y)$. The Law of Iterated Expectations allows us to calculate the unconditional expectation using just the marginal distribution (less tedious/messy than going back to the joint distribution)

2.3 Other Useful Properties

The following statements involving three known and fixed constants (a, b, c) and two random variables (X, Y) can be proven. For notational simplicity, let $Z \equiv aX + bY + c$

$$\begin{aligned} E[Z] &= a\ E[X] + b\ E[Y] + c \\ \sigma_X^2 = E[\ (X - E[X])^2\ ] &= E[X^2] - (E[X])^2 \\ \sigma_{XY} = E[\ (X - E[X])\ (Y - E[Y])\ ] &= E[XY] - E[X]E[Y] \\ \sigma_Z^2 = E[\ (Z - E[Z])^2\ ] &= a^2\ \sigma_X^2 + b^2\ \sigma_Y^2 + 2ab\ \sigma_{XY} \\ E[Z|X = x] &= ax + b\ E[Y|X = x] + c \end{aligned}$$

The statements can be extended to the multivariate case involving k random variables.

Consider the proof for the first statement

$$\begin{aligned} E[Z] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (ax + by + c)\ f(x, y)\ dx\ dy \\ &= a\left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\ f(x, y)\ dx\ dy\right) + b\left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\ f(x, y)\ dx\ dy\right) + c\left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\ dx\ dy\right) \\ &= a\int_{-\infty}^{\infty} x \underbrace{\left(\int_{-\infty}^{\infty} f(x, y)\ dy\right)}_{f(x)} dx + b\int_{-\infty}^{\infty} y \underbrace{\left(\int_{-\infty}^{\infty} f(x, y)\ dx\right)}_{f(y)} dy + c\underbrace{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\ dx\ dy}_{\text{integrates to 1}} \\ &= a\ E[X] + b\ E[Y] + c \end{aligned}$$

The other statements can be similarly proven.

Again, do **not** make the common error of assuming that $E(g(X)) = g(E(X))$. This does not **not** hold in general. It holds only for special cases such as when $g(\cdot)$ is a linear (affine) transformation.

Suppose that $Y \equiv g(X)$. The expected value of $h(Y) = h(g(X))$ can be calculated using distributions involving Y or those involving X

$$\underbrace{\int_{-\infty}^{\infty} h(y)\ f(y)\ dy}_{E[h(y)]} = \underbrace{\int_{-\infty}^{\infty} h(g(x))\ f(x)\ dx}_{E[h(g(x))]}$$

as $\{s_j \in S : h(Y(s_j)) = h(y)\}$ is the same set as $\{s_j \in S : h(g(X(s_j))) = h(g(x))\}$

This is useful as, sometimes, $E[h(Y)]$ is easier to calculate than $E[h(g(X))]$. For example, suppose $Y \equiv X^2$ and X is distributed N(0,1). Instead of calculating $E[X^2]$, note that

- $E[X^2] = E[Y]$ and Y is distributed χ_1^2
- So $E[Y]$ is the mean of a random variable distributed χ_1^2, which is 1

3 Prediction

One of the main activities associated with statistical inference is the following:

> Suppose you are interested in some random experiment involving two random variables X and Y. Specifically, you want to estimate/predict the value of Y given the value of X. Let $\hat{Y}(X)$ be your estimate/prediction of Y given X. How does one choose $\hat{Y}(X)$?

Suppose further that you know the exact joint distribution of X and Y, $f(x, y)$. How might this information help you choose $\hat{Y}(X)$? Note: $\hat{Y}(X)$ is a function solely of X (does not involve Y)

Statisticians "solve" this problem using an approach borrowed from decision theory. They first choose a **loss function** that embodies their statistical objective and then use the related **risk function** to help them choose $\hat{Y}(X)$.

For the specific problem above,

- A loss function, $LF(\hat{Y}(X))$, evaluates how far off $\hat{Y}(X)$ is from the "ideal"
  - The greater the value of the loss function, the less ideal the estimator/predictor
  - The loss function is usually some increasing function of the distance between the prediction/estimation and the actual (unobserved) realization, $\hat{Y}(X) - Y$, as the ideal is usually $\hat{Y}(X) = Y$ ... perfect prediction
- The risk function is simply the expected value of the loss function
  - $R(\hat{Y}(X)) = E[\ LF(\hat{Y}(X))\ ]$
  - The risk function is the mean loss of the chosen $\hat{Y}(X)$

3.1 Absolute Deviation and Mean-squared Error

Why this distinction between loss and risk functions?

Suppose your statistical objective was to minimize the distance between your prediction and the actual realized value of Y: $\min_{\hat{Y}(X)} \mid\mid \hat{Y}(X) - Y \mid\mid$. Then the loss function is simply

$$LF(\hat{Y}(X)) \quad = \quad \mid\mid \hat{Y}(X) - Y \mid\mid$$

This is the **absolute deviation (AD)** loss function.

But $Y = y$ is unobserved. This means that the above loss function cannot be evaluated. The actual loss associated with $\hat{Y}(X)$ is unknown. But the expected loss can be calculated using $f(x, y)$

$$\begin{aligned} R(LF(\hat{Y}(X))) \quad = \quad E[\ LF(\hat{Y}(X))\ ] \quad &= \quad E[\ ||\hat{Y}(X) - Y||\ ] \\ &= \quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} ||\hat{Y}(x) - y||\ f(x, y)\ dx\ dy \end{aligned}$$

This risk function for $\hat{Y}(X)$ can be evaluated if we know $f(x, y)$

So, one can choose $\hat{Y}(X)$ to minimize the *risk* function but not the *loss* function, in general.

In the case of the absolute deviation loss function, we can try to solve for

$$\tilde{Y}(X) \quad = \quad \text{argmin}_{\hat{Y}(X)} \; E[\; ||\hat{Y}(X) - Y|| \;]$$

which is, in words, the function of X that minimizes the *mean* absolute deviation loss.

But the above risk function is difficult to evaluate. Absolute values complicate calculus. Consequently, a more popular loss function is the **mean squared error (MSE)** loss function

$$LF(\hat{Y}(X)) \quad = \quad \Big(\; \hat{Y}(X) - Y \; \Big)^2$$

Instead of examining the absolute difference between $\hat{Y}(X)$ and Y, the mean squared error loss function examines the *squared* difference.

The risk function for the mean-squared error is

$$R(LF(\hat{Y}(X))) \quad = \quad E[\; \Big(\hat{Y}(X) - Y\Big)^2 \;] \quad = \quad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Big(\; \hat{Y}(x) - y \; \Big)^2 \; f(x,y) \; dx \; dy$$

This function is easier to evaluate using calculus.

Note the similarities and differences between the absolute deviation (AD) and mean squared error (MSE) loss functions

- Both loss functions do not distinguish between positive and negative differences of the same magnitude (symmetric around 0)
- Both loss functions penalize more for larger magnitude differences
- But the MSE penalizes larger magnitude differences more, as it is the square of the difference (not the absolute value) that is added to the loss function

3.2 Best Predictor (BP)

Consider now the function $\hat{Y}(X)$ that minimizes the risk function $R(LF(\hat{Y}(X)))$.

> **DEFINITION**: the **best predictor (BP)** of Y given X for a given loss function is the predictor that minimizes the risk function associated with that loss function
>
> $$BP(Y|X) \quad = \quad \text{argmin}_{\hat{Y}(X)} \; E[\; LF(\hat{Y}(X)) \;]$$

$BP(Y|X)$ for the mean squared error (MSE) loss function is the function $\hat{Y}(X)$ that minimizes

$$\begin{aligned} E[\ LF(\hat{Y}(X))\ ] &= E[\ \left(\hat{Y}(X) - Y\right)^2\ ] \\ &= E[\ \left(\hat{Y}(X) - E[Y|X] + E[Y|X] - Y\right)^2\ ] \\ &= \underbrace{E[\ \left(\hat{Y}(X) - E[Y|X]\right)^2\ ]}_{\text{(A)}} + \underbrace{E[\ (E(Y|X) - Y)^2\ ]}_{\text{(B)}} \\ &\quad + 2\underbrace{E[\ \left(\hat{Y}(X) - E[Y|X]\right)\ (E[Y|X] - Y)\ ]}_{\text{(C)}} \end{aligned}$$

where the conditional mean of Y given X, denoted as $E[Y|X]$ instead of $E_{Y|X}[Y]$, is "added and subtracted" in the middle of the squared term.

(B) does not vary with $\hat{Y}(X)$. So the choice of $\hat{Y}(X)$ should not depend on (B). But neither should $\hat{Y}(X)$ depend on (C) as

$$\begin{aligned} \underbrace{E[\ \left(\hat{Y}(X) - E[Y|X]\right)\ (E[Y|X] - Y)\ ]}_{(C)} &= E_X[\ E_{Y|X}[\ \left(\hat{Y}(X) - E[Y|X]\right)\ (E[Y|X] - Y)\ ]\ ] \\ &= E_X[\ \left(\hat{Y}(X) - E[Y|X]\right)\ \left(E[Y|X] - E_{Y|X}[Y]\right)\ ] \\ &= E_X[\ \left(\hat{Y}(X) - E[Y|X]\right)\ \underbrace{(E[Y|X] - E[Y|X])}_{=0}\ ] \\ &= 0 \end{aligned}$$

The first step above is an application of the Law of Iterated Expectations. The second step uses the fact that $\hat{Y}(X)$ and $E[Y|X]$, as functions solely of X, are constants with respect to the conditional distribution $f(y|x)$. And the third step uses the identity $E_{Y|X}[Y] = E[Y|X]$.

So the $BP(Y|X)$ for the MSE loss function is the function $\hat{Y}(X)$ that minimizes (A).

Note that (A) is the expected value of a squared term. This means that the minimum value that (A) could possibly take is zero. And (A) = 0 when $\hat{Y}(X) = E[Y|X]$!!!

> **The Best Predictor of Y given X under the mean squared error (MSE) loss function is the conditional mean of Y:** $BP_{MSE}(Y|X) = E[Y|X]$

This result means that if (1) the appropriate loss function is MSE and (2) the joint distribution of X and Y, $f(x, y)$, is known then the best prediction of Y one can make given $X = x$ is

$$E[Y|X] = E_{Y|X}[Y] = \int_{-\infty}^{\infty} y\ f(y|x)\ dy$$

This is one of the main reasons why many empirical economists consider $E[Y|X]$ the "Holy Grail."

Interestingly enough, we can use similar steps to show that

> **The Best Predictor of Y given X under the absolute deviation (AD) loss function is the conditional *median* of Y**

where the conditional median of Y given X is the value c such that

$$P_{Y|X}(Y \leq c) \quad = \quad \int_{-\infty}^{c} f(y|x)\ dy \quad = \quad \frac{1}{2}$$

3.3 Best Linear Predictor (BLP)

Consider a related problem. Find a *linear* predictor of Y given X that minimizes some risk function. The linear constraint implies that $\hat{Y}(X) = a + bX$.

> **DEFINITION**: the **best linear predictor (BLP)** of Y given X for a given loss function is the linear predictor $a + bX$ that minimizes the risk function associated with that loss function
>
> $$BLP(Y|X) \quad = \quad \text{argmin}_{a,b}\ E[\ LF(a + bX)\ ]$$

For the MSE loss function, we can rewrite the risk function for the linear predictor as

$$\begin{aligned} R(LF(a+bX)) &= E[\ (a + bX - Y)^2\ ] \\ &= E[\ (a+bX)^2 - 2Y(a+bX) + Y^2\ ] \\ &= a^2 + 2ab\ E[X] + b^2\ E[X^2] - 2a\ E[Y] - 2b\ E[XY] + E[Y^2] \end{aligned}$$

To find the values of a, b that minimize $R(LF(a+bX))$, consider the first order conditions

$$\underbrace{2a + 2b\ E[X] - 2\ E[Y]}_{\frac{d}{da}\ R(LF(a+bX))} \quad = \quad 0 \qquad\qquad \underbrace{2a\ E[X] + 2b\ E[X^2] - 2\ E[XY]}_{\frac{d}{db}\ R(LF(a+bX))} \quad = \quad 0$$

It can be shown that these first order conditions implicitly give us the values of a, b that minimizes $R(LF(a+bX))$. Solving the above two equations for a, b

$$\begin{aligned} a^* &= E[Y] - b^* E[X] \quad = \quad \mu_Y - b^* \mu_X \\ b^* &= \frac{E[XY] - E[X]\ E[Y]}{E[X^2] - (E[X])^2} \quad = \quad \frac{\sigma_{XY}}{\sigma_X^2} \end{aligned}$$

So $BLP_{MSE}(Y|X) \quad = \quad a^* + b^* X \quad$ where (a^*, b^*) are defined as above.

Note that the $BLP_{MSE}(Y|X)$ can be calculated even if we do not know $f(x, y)$. All we need is a few moments associated with $f(x, y)$ – namely the mean of X, mean of Y, the covariance of X and Y, and the variance of X.

We will revisit this result when we discuss the ordinary least squares (OLS) model – also known as the linear regression model.