Probability and Statistics

Omid Safarzadeh

February 10, 2022

Table of contents

*Acknowledgement: This slide is prepared based on Casella and Berger, 2002

Section 1

Moments

Moments definition

Definition 1.1

For each of integer n, the n^{th} moment of X is

$$\mu'_n = E[X^n].$$

The n^{th}central moment of X, μ_n, is

$$\mu_n = E[(X - \mu)^n],$$

where $\mu = \mu'_1 = E[X]$.

Expected Value

- Recall that"average" is an arithmetic average where all available observations are weighted equally.
- The expected value, on the other hand, is the average of all possible values a random variable can take, weighted by the probability distribution.
- The question is,which value would we expect the random variable to take on, on average.

Expected Value

Definition 1.2

The expected value or mean of a random variable $g(X)$, denoted by $E[g(X)]$, is

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

If $E[g(X)] = \infty$, we say that $E[g(X)]$ does not exist.

- we are taking the average of $g(x)$ over all of its possible values ($x \in \mathcal{X}$), where these values are weighted by the respective value of the pdf, $f_X(x)$.

Expected Value

Example 1.1

Suppose X has an exponential(λ) distribution, that is, it has pdf given by

$$f_X(x) = \frac{1}{\lambda} e^{-x/\lambda}, \quad 0 \le x < \infty \quad \lambda > 0.$$

Then,

$$E[X] = \int_0^\infty \frac{1}{\lambda} x e^{-x/\lambda} dx = -x e^{-x/\lambda}|_0^\infty + \int_0^\infty e^{-x/\lambda} dx \tag{1}$$

$$= \int_0^\infty e^{-x/\lambda} dx = \lambda. \tag{2}$$

- To obtain this result, we use a method called integration by parts. This is based on

$$\int u dv = uv - \int v du.$$

Expected Value

- A very useful property of the expectation operator is that it is a linear operator.
- take a and b constants:

$$E[a + Xb] = a + E[Xb] = a + bE[x] = a + b\mu.$$

Theorem 1.1

Let X be a random variable and let a, b and c be constants. Then for any functions $g_1(x)$ and $g_2(x)$ whose expectations exist,

- $E[ag_1(X) + bg_2(X) + c] = aE[g_1(X)] + bE[g_2(X)] + c$.
- If $g_1(x) \geq 0$ for all x, then $E[g_1(X)] \geq 0$.
- If $g_1(x) \geq g_2(x)$ for all x, then $E[g_1(X)] \geq E[g_2(X)]$.
- If $a \leq g_1(x) \leq b$ for all x, then $a \leq E[g(X)] \leq b$.

Proof: Exercise!

Expected Value

Example 1.2

Let X have a uniform distribution, such that

$$f_X(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ \\ 0 & \text{if } \text{otherwise} \end{cases}$$

Define $g(X) = -\log X$. Then,

$$E[g(X)] = E[-\log X] = \int_0^1 -\log x dx = (-x \log x + x)|_0^1 = 1,$$

where we use integration by parts.

Variance

- variance measures the variation/dispersion/spread of the random variable around expectation.
- While the expectation is usually denoted by μ, σ^2 is generally used for variance.
- Variance is a second-order moment.

Variance

Definition 1.3

The variance of a random variable X is its second central moment,

$$Var(X) = E[(X - \mu)^2],$$

while $\sqrt{Var(X)}$ is known as the standard deviation of X.

- Importantly,

$$Var(X) = E[(X - \mu)^2] = E[X^2] - \mu^2.$$

Section 2

Covariance and Correlation

covariance

- When it exists, the covariance of two random variables X and Y is defined as

$$Cov(X, Y) = E(\{X - E[X]\}\{Y - E[Y]\}).$$

Covariance and Correlation

- Let X and Y be two random variables. To keep notation concise, we will use the following notation.

$$E[X] = \mu_X, \quad E[Y] = \mu_Y, \quad Var(X) = \sigma_X^2 \quad and \quad Var(Y) = \sigma_Y^2.$$

Definition 2.1

The covariance of X and Y is the number defined by

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)].$$

Definition 2.2

The correlation of X and Y is the number defined by

$$\rho_{XY} = \frac{Cov(X, Y)}{\sigma_x \sigma_y},$$

which is also called the correlation coefficient.

Covariance and Correlation

- If large(small) values of X, tend to be observed with large(small) values of Y, then $Cov(X, Y)$ will be positive.
- Why so? Within the above setting, when $X > \mu_X$ then $Y > \mu_Y$ is likely to be true whereas when $X < \mu_X$ then $Y < \mu_Y$ is likely to be true. Hence

$$E[(X - \mu_X)(Y - \mu_Y)] > 0.$$

- Similarly, if large(small) values of X tend to be observed with small(large) values of Y, then will be negative.

Covariance

- Correlation normalises covariance by the standard deviations and is, therefore, a more informative measure.

Theorem 2.1

For any random variables X and Y,

$$Cov(X, Y) = E[XY] - \mu_X \mu_Y.$$

- **Proof**: Exercise!

Covariance and Correlation

Theorem 2.2

If $X \perp\!\!\!\perp Y$, then Cov$(X, Y)= \rho_{XY} = 0$.

- **Proof**: Exercise!
- Note that although $X \perp\!\!\!\perp Y$ implies that $Cov(X, Y) = \rho_{XY} = 0$, the relationship does not necessarily hold in the reverse direction.

Theorem 2.3

If X and Y are any two random variables and a and b are any two constants, then

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$$

If X and Y are independent random variables, then

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y).$$

- **Proof**: Exercise!

Covariance and Correlation

- Note that if two random variables, X and Y, are positively correlated, then

$$Var(X+Y) > Var(X) + Var(Y),$$

whereas if X and Y are negatively correlated, then

$$Var(X+Y) < Var(X) + Var(Y).$$

- For positively correlated random variables, large values in one tend to be accompanied by large values in the other. Therefore, the total variance is magnified.
- Similarly, for negatively correlated random variables, large values in one tend to be accompanied by small values in the other. Hence, the variance of the sum is dampened.

Variance of Sums of Random Variables

- Let a_i be some constant and X_i be some random variable, where $i = 1, ..., n$.
- Then

$$Var(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i^2 Var(X_i) + \sum_{i \neq j} \sum a_i a_j Cov(X_i, X_j).$$

third and fourth moments

- third and fourth moments are concerned with how symmetric and fat-tailed the underlying distribution is.

Section 3

Moment Generating Functions

Moments and Moment Generating Functions

Definition 3.1

Let X be a random variable with cdf F_X.The moment generating function (mgf)of X (or F_X), denoted by $M_X(t)$, is

$$M_X(t) = E[e^{tX}],$$

provided that the expectation exists for t in some neighbourhood of 0. That is, there is an $h > 0$ such that, for all t in $-h < t < h$, $E[e^{tX}]$ exists.If the expectation does not exist in a neighbourhood of 0, we say that the mgf does not exist.

- We can write the mgf of X as

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)dx \quad \text{if } X \text{ is continuous,}$$

$$M_X(t) = \sum_x e^{tx} P(X = x) \quad \text{if } X \text{ is discrete.}$$

Moment Generating Functions

- moment generating function is used to obtain moments of a random variable.

Theorem 3.1

If X has mgf $M_X(t)$, then

$$E[X^n] = M_X^{(n)}(0),$$

where we define

$$M_X^{(n)}(0) = \frac{d^n}{dt^n} M_X(t)|_{t=0}.$$

That is, the n^{th} moment is equal to the n^{th} derivative of $M_X(t)$ evaluated at t=0.

Normal mgf

- Now consider the pdf for $X \sim N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2], \qquad -\infty < x < \infty.$$

- The mgf is given by

$$M_X(t) = E[e^{Xt}] = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} dx.$$

Normal mgf

- Note that:

$$M_X(t) = exp(\mu t + \frac{\sigma^2 t^2}{2}).$$

- Proof: Exercise!
- Clearly,

$$E[X] = \frac{d}{dt} M_X(t)|_{t=0} = (\mu + \sigma^2 t) exp(\mu t + \frac{\sigma^2 t^2}{2})|_{t=0} = \mu,$$

$$\begin{aligned} E[X^2] = \frac{d^2}{dt^2} M_X(t)|_{t=0} &= \sigma^2 exp(\mu t + \frac{\sigma^2 t^2}{2}|_{t=0} \\ &+ (\mu + \sigma^2 t)^2 exp(\mu t + \frac{\sigma^2 t^2}{2})^2|_{t=0} \\ &= \sigma^2 + \mu^2, \end{aligned}$$

$$Var(X) = E[X^2] - \{E[X]\}^2 = \sigma^2 + \mu^2 - \mu^2 = \sigma^2.$$

Section 4

Matrix Notation for Moments

Matrix Notation for Moments

- Now, let X and Y be $(r * 1)$ and $(c * 1)$ random vectors, respectively. Then

$$Cov(X, Y) = \begin{bmatrix} Cov(X_1, Y_1) & \cdots & Cov(X_1, Y_c) \\ \vdots & \ddots & \vdots \\ Cov(X_r, Y_1) & \cdots & Cov(X_r, Y_c) \end{bmatrix}$$

$$= E \begin{bmatrix} \{X_1 - E[X_1]\}\{Y_1 - E[Y_1]\} & \cdots & \{X_1 - E[X_1]\}\{Y_c - E[Y_c]\} \\ \vdots & \ddots & \vdots \\ \{X_r - E[X_r]\}\{Y_1 - E[Y_1]\} & \cdots & \{X_r - E[X_r]\}\{Y_c - E[Y_c]\} \end{bmatrix}$$

Matrix Notation for Moments

$$= E\left[\begin{pmatrix} X_1 - E[X_1] \\ \vdots \\ X_r - E[X_r] \end{pmatrix} \quad (Y_1 - E[Y_1], \cdots, Y_c - E[Y_c])\right],$$

$$= E(\{X - E[X]\}\{Y - E[Y]\}').$$

Matrix Notation for Moments

- Usually, for a $(c \times 1)$ vector X, one would write $Cov(X)$ for $Cov(X, X)$,
- This is given by

$$Cov(X) = \begin{bmatrix} Var(X_1) & \cdots & Cov(X_1, X_c) \\ \vdots & \ddots & \vdots \\ Cov(X_1, X_c) & \cdots & Var(X_c) \end{bmatrix},$$

which is a $(c \times c)$ symmetric matrix.

Matrix Notation for Moments

- We can also consider block structures. Let

$$X = \begin{pmatrix} Y \\ Z \end{pmatrix},$$

  where Y is $(p \times 1)$ vector and Z is a $(q \times 1)$ vector.

- Then,

$$Cov(X) = \begin{pmatrix} Cov(Y) & Cov(Y,Z) \\ Cov(Z,Y) & Cov(Z) \end{pmatrix},$$

  Exercise: Find dimensions of $Cov(Y)$, $Cov(Y,Z)$, $Cov(Z,Y)$ and $Cov(Z)$ and $Cov(X)$

Matrix Notation for Moments

- Let a and b be $(r \times 1)$ and $(c \times 1)$ non-stochastic vectors.
- Let $E[X_i] = \mu_{X_i}$, $E[Y_i] = \mu_{Y_i}$ and $Cov(X_i, Y_j) = \Sigma_{X_i Y_j}$. Then

$$Cov(a'X, b'Y) = a'\Sigma_{XY} b = a' Cov(X, Y) b$$

Matrix Notation for Moments

- Now, let $\Sigma_{ij} = Cov(X_i, X_j)$ and $\Sigma = Var(X)$. Then,

$$Var(a'X) = E\{[\sum_{i=1}^{r} a_i(X_i - \mu_i)][\sum_{i=1}^{r} a_i(X_i - \mu_i)]\} = a'Var(X)a.$$

Matrix Notation for Moments

- Now, Consider

$$Var(X+Y) = E\{[(X-\mu_X)+(Y-\mu_Y)][(X-\mu_X)+(Y-\mu_Y)]'\}$$

$$= \Sigma_{XX} + \Sigma_{XY} + \Sigma_{YX} + \Sigma_{YY}.$$

- Using this, we get

$$Var[a'(X+Y)] = a'\Sigma_{XX}a + 2a'\Sigma_{XY}a + a'\Sigma_{YY}a,$$

Note that :

$$a'\Sigma_{XY}a = a'\Sigma_{YX}a$$

Matrix Notation for Moments

- These results easily extend to cases where a and b are replaced by matrices.

$$E[RX] = RE[X]$$

$$\begin{aligned} Var(RX) &= E[R(X - \mu_X)(X - \mu_X)' R'] \\ &= RE[(X - \mu_X)(X - \mu_X)'] R' \\ &= R\Sigma_{XX} R'. \end{aligned}$$

Reference

Casella, G., & Berger, R. (2002). *Statistical inference*. Cengage Learning. https://books.google.fr/books?id=FAUVEAAAQBAJ