Probability and Statistics

MA-202

Moments and Percentiles

Reference

- Rohatgi, V. K., & Saleh, A. M. E. (2015). An Introduction to Probability and Statistics. John Wiley & Sons.

Moments

The study of probability distributions of a random variable is essentially the study of some numerical characteristics associated with them. These characteristics, including moments, their functions, and order parameters, are fundamental in the field of mathematical statistics.

Expectation

Let X be a discrete random variable probability mass function (PMF)

$$P_k = P(X = x_k), \quad k = 1, 2, \ldots.$$

If

$$\sum_{k=1}^{\infty} |x_k| p_k < \infty,$$

we say that the expected value (or the mean or the mathematical expectation) of X exists and write

$$E[X] = \sum_{k=1}^{\infty} x_k p_k.$$

Note that the series $\sum_{k=1}^{\infty} x_k p_k$ may converge but the series $\sum_{k=1}^{\infty} |x_k| p_k$ may not. In that case we say that $E[X]$ does not exist.

If X is of the continuous type and has PDF $f(\cdot)$, we say that $E[X]$ exists and equals $\int xf(x)dx$ provided that

$$\int |x|f(x)dx < \infty.$$

We denote $E[X]$ by μ.

Remark 1. *Note that the condition $\int |x|f(x)dx < \infty$ must be checked before it can be concluded that EX exists and equals $\int xf(x)dx$. The same holds for discrete random variable. Further, the integral $\int_{-\infty}^{\infty} f(x)dx$ exists, provided that the limit $lim_{b\to\infty}^{a\to\infty} \int_{-b}^{a} f(x)dx$ exists. It is quite possible for the limit*

$$lim_{a\to\infty} \int_{-a}^{a} f(x)dx$$

to exist without the existence of $\int_{-\infty}^{\infty} f(x)dx$. As an example, consider the following pdf:

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}, -\infty < x < \infty.$$

Clearly,

$$lim_{a\to\infty} \int_{-a}^{a} \frac{x}{\pi}\frac{1}{1+x^2}dx = 0.$$

However, $E[X]$ does not exist since the integral $(1/\pi)\int_{-\infty}^{\infty} |x|/(1+x^2)dx$ diverges.

A similar definition is given for the mean of any function $g(X)$ of X.[1] Thus, we have the following

- If X is of discrete type, we say that $E[g(X)]$ exists and equals $\sum_{k=1}^{\infty} g(x_k)p_k$, provided that $\sum_{k=1}^{\infty} |g(x_k)|p_k < \infty$.
- If X is of continuous type, we say that $E[g(X)]$ exists and equals $\int g(x)f(x)dx$, provided that

$$\int |g(x)|f(x)dx < \infty.$$

Some results:

[1] Borel-measurable function

a) Let $X(\omega) = I_A(\omega), \quad \omega \in \Omega$, where

$$I_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \text{otherwise} \end{cases},$$

for some $A \in \mathcal{F}$, the sigma-field. Then $E[X] = P(A)$.

b) $E[X]$ exists if and only if $E[|X|]$ does.

c) If a and b are constants and X is an RV with $E[|X|] < \infty$, then $E[|aX + b|] < \infty$ and $E[aX + b] = aE[X] + b$.

d) If X is bounded, that is, if $P\{|X| < M\} = 1, \quad 0 < M < \infty$, then $E[X]$ exists.

e) If $P\{X \geq 0\} = 1$ and $E[X]$ exists, then $E[X] \geq 0$.

Higher Order Moments

a) Consider the functions $g(x) = x^n$, where n is a positive integer. If $E[X^n]$ exists for some positive integer n, we call $E[X^n]$ the nth moment of X about the origin. We use the following notation,

$$m_n = E[X^n]$$

b) Consider the functions $g(x) = x^\alpha$, where α is a positive real number. If $E[|X|^\alpha]$ exists for some positive real number α, we call $E[|X|^\alpha]$ the αth absolute moment of X about the origin. We use the following notation,

$$\beta_\alpha = E[|X|^\alpha]$$

c) Let k be a positive integer and c be a constant. If $E[(X - c)^k]$ exists, we call it the moment of order k about the point c. If we take $c = E[X] = \mu$, which exists since $E[|X|] < \infty$, we call $E[(X - \mu)^k]$ the central moment of order k or the moment of order k about the mean. We shall write

$$\mu_k = E[(X - \mu)^k].$$

Remark 2. *If the moment of order t exists for an RV X, moments of order $0 < s < t$ exist.*

Remark 3. *If we know $m_1, m_2, \ldots, m_k$, we can compute $\mu_1, \mu_2, ..., \mu_k$, and conversely. We have*

$$\mu_k = E[(X-\mu)^k] = m_k - \binom{k}{1}\mu m_{k-1} + \binom{k}{2}\mu^2 m_{k-2} - \ldots + (-1)^k \mu^k$$

and

$$m_k = E[(X-\mu+\mu)^k] = \mu_k + \binom{k}{1}\mu\mu_{k-1} + \binom{k}{2}\mu^2\mu_{k-2} + \ldots + \mu^k.$$

Variance

If $E[X^2]$ exists, we call $E(X-\mu)^2$ the variance of X, and we write $\sigma^2 = Var(X) = E(X-\mu)^2$. Note that σ is called the standard deviation (SD) of X.

Important Properties

a) $Var(X) \geq 0$

b) $\sigma^2 = \mu_2 = E[X^2](E[X])^2$.

c) $Var(X) = 0$ if and only if X is degenerate (a constant random variable).

d) $Var(X) < E[(X-c)^2]$ for any $c \neq E[X]$.

Proof. We have

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

Hence the result. $\square$

e)

$$\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X).$$

Remark 4. *Note the following*

- ***Standardized RV***: *Let* $E[|X|^2] < \infty$. *Then define* Z *as follows:*

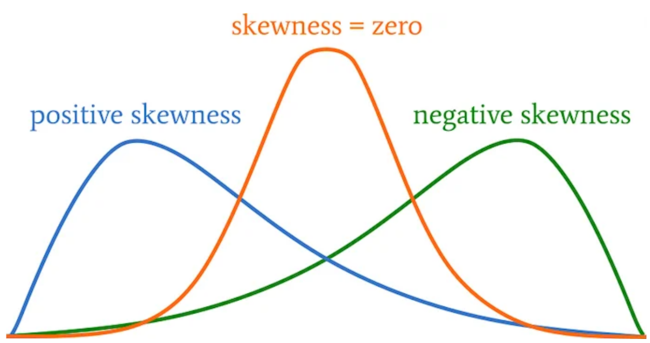
$$Z = \frac{X - E[X]}{\sqrt{Var(X)}} = \frac{X - \mu}{\sigma}$$

We call Z *a standardized random variable. Note that it satisfies* $E[Z] = 0$ *and* $Var(Z) = 1$.

- *Significance of first four moments:*
  - *Expectation: measures central tendency*
  - *Variance: measures dispersion (spread)*
  - *Skewness: symmetry. The coefficient of skewness is given by*

$$\alpha_3 = E[Z^3] = E[\frac{(X-\mu)^3}{\sigma^3}] = \frac{E[(X-\mu)^3]}{\sigma^3} = \frac{\mu_3}{\mu_2^{3/2}}.$$

If $\alpha_3 = 0$, *the distribution is symmetric around mean. If* $\alpha_3 > 0 (< 0)$, *the distribution is asymmetric and positively (negatively) skewed.*



  - *Kurtosis: peakedness (tail behavior). The coefficient of kurtosis is given by*

$$\alpha_4 = E[Z^4] = E[\frac{(X-\mu)^4}{\sigma^4}] = \frac{E[(X-\mu)^4]}{\sigma^4} = \frac{\mu_4}{\mu_2^2}.$$

If $\alpha_4 = 3$, *the distribution is called mesokurtic (corresponds to normal distribution, will discuss in upcoming lectures). If* $\alpha_4 > 3 (\alpha_4 < 3)$, *the distribution is called leptokurtic (platykurtic).*

- ***Symmetric random variable:*** *We say that an RV* X *is symmetric about a point* α *if*

$$P(X \geq \alpha + x) = P(X \leq \alpha - x), \quad \forall x$$

which is same as

$$F(\alpha - x) = 1 - F(\alpha + x) + P\{X = \alpha + x\}$$

in which case we say that the DF F (or the RV X) is symmetric with α as the center of symmetry.

Percentiles

Note that in certain distributions, the mean may not be defined. Moving forward, we'll examine certain parameters known as order parameters, which always exists.

A number x is called a quantile of order p [or $(100p)$th percentile] for the RV X if it satisfies the following (refer Figure 1)

$$P\{X \leq x\} \geq p, \quad P\{X \geq x\} \geq 1 - p, \quad 0 < p < 1$$

One can write the above conditions as

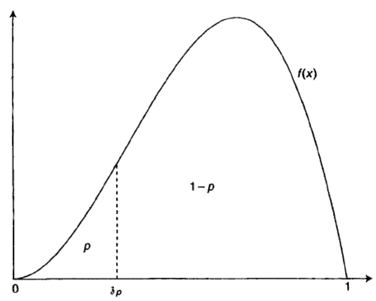


Figure 1: Quantile of order p

$$p \leq F(x) \leq p + P\{X = x\}.$$

In case of a continuous random variable, we know that $P\{X = x\} = 0$ for all x, a quantile of order p is a solution of the equation

$$F(x) = p$$

Note that there may be many (even uncountably many) solutions of $F(x) = p$, each of which is then called a quantile of order p.

Remark 5. *For $p = \frac{1}{2}$, the pth quantile is also called median.*