RANDOM VARIABLES EXPECTATION & TRANSFORMATIONS

Jerzy Szulga

Department of Mathematics and Statistics
Auburn University

STAT 3600 SUMMER 2019

September 20, 2019

Outline

Basics

Variance

Percentiles

Moment generating function

For independent RVs

- sum
- mgf
- covariance

Definition

Given a RV X, the **expectation** is

$$\mu = \mu_X = \mathrm{E}\,X = \begin{cases} \displaystyle\sum_x x\,f(x), & \text{if } X \text{ is discrete with pmf } f(x) \\ \displaystyle\int_{\mathbb{R}} x\,f(x)\,dx, & \text{if } X \text{ is continuous with pdf } f(x) \end{cases}$$

Synonyms (some come from physics):

mean, expectation, expected value, expectancy (archaic, but still in use in special cases), the first moment, average, ... static moment, center of mass, ...

transformed RV

Let $Y = \phi(X)$, a transformation of a given X. According to the definition, it seems that the pmf or pdf $f_Y(y)$ of Y is needed to find $\mathrm{E}\,Y$. Not necessarily!

Theorem

If X is transformed by a function, e.g., $Y = \phi(X)$, then

$$\mu_Y = \begin{cases} \displaystyle\sum_x \phi(x)\, f(x), & \textit{discrete}, \\ \displaystyle\int_{\mathbb{R}} \phi(x)\, f(x)\, dx, & \textit{continuous}. \end{cases}$$

Example

1. The **kth moment** is

$$\mathrm{E}\,(X^k) = \begin{cases} \sum_x x^k\, f(x), & \text{discrete,} \\ \int_{\mathbb{R}} x^k\, f(x)\, dx, & \text{continuous.} \end{cases}$$

2. The **moment generating function, mgf** is

$$M(t) = M_X(t) = \mathrm{E}\, e^{tX} = \begin{cases} \sum_x e^{tx}\, f(x), & \text{discrete,} \\ \int_{\mathbb{R}} e^{tx}\, f(x)\, dx, & \text{continuous.} \end{cases}$$

The mean might not exist

Let $f(x) = \dfrac{c}{x^p}$, $x = 1, 2, \ldots$.

Q: For what values of p does the mean $\operatorname{E} X$ exist?

A: The series $\displaystyle\sum_x x\,\frac{c}{x^p} = c\sum_x \frac{1}{x^{p-1}}$ must converge.

It happens iff $p > 2$.

Q: For what values of p does the second moment $\operatorname{E} X^2$ exist?

A: The series $\displaystyle\sum_x x^2\,\frac{c}{x^p} = c\sum_x \frac{1}{x^{p-2}}$ must converge.

It happens iff $p > 3$.

brackets or no brackets?

Ross writes $\mathrm{E}\,[X]$.

Here you see $\mathrm{E}\,X$, without parentheses. Why?

> The formula involves multiplication and summation.
>
> The old rule: the first before the second.
>
> So, parentheses are unnecessary (when there's no ambiguity)

Although $\mathrm{E}\,X^k$ could suffice (powers before products before sums) yet the chances of ambiguity increase. Now, parentheses help:

$$\mathrm{E}\,(X^k) \neq (\mathrm{E}\,X)^k.$$

Some simple algebra

The arithmetic distributive property entail the following

- $\mathsf{E}\,(X+Y) = \mathsf{E}\,X + \mathsf{E}\,Y$ (when both means exist);
- $\mathsf{E}\,c = c$ (when the RV is constant);
- Hence $\mathsf{E}\,(aX+bY) = a\,\mathsf{E}\,X + b\,\mathsf{E}\,Y$ (so called "linearity");

An immediate extension to finite sums (subject to existence):

$$\mathsf{E}\left(\sum_n c_n X_n\right) = \sum_n c_n\,\mathsf{E}\,X_n \tag{1.1}$$

The case of an infinite sum is more advanced, formula (1.1) may fail.

Moment of inertia

The second moment $\mathrm{E}\,X^2 = \mathrm{E}\,(X^2)$ in physics is called also the **moment of inertia** with respect to origin (or 0).

The static moment stems from the quantity **distance** $\times$ **mass**, the moment of inertia stems from **distance squared** $\times$ **mass**.

It's proportional to the energy released when a mass revolving around a center point is stopped (the sling or atlatl, or catapults).

Definition

Consider the second moment $\mathrm{E}\,(X-c)^2$ relative to a point c.

Q: What value of c does minimize the quantity?

A: $c = \mathrm{E}\,X$. The minimum moment of inertia occurs when the mass is revolved about its center.

Indeed, by simple algebra $\mathrm{E}\,(X-c)^2 = \mathrm{E}\,(X^2) - 2c\,\mathrm{E}\,X + c^2(\mathrm{E}\,X)^2$. Then the answer follows either by HS Algebra (minimum of a quadratic function) or Calculus 1 (minimum of a function).

Def. $\sigma^2 = \mathrm{E}\,(X - \mathrm{E}\,X)^2 = \mathrm{E}\,(X^2) - (\mathrm{E}\,X)^2$ is called the **variance** while σ is called the **standard deviation**.

location and spread

The mean $\mu_X = \mathsf{E}\, X$ indicates the **location** of the probability mass.

The variance (or standard deviation) indicates the **spread** or **scatter** of the probability mass.

Why bother?

A brief pair of numbers replaces the entire pmf.

The full yet complex information is forfeited for transparency.

Definition

Q. Sometimes the mean or variance do not exist. This happens quite often in socio-economic models (distribution of wealth, salaries, education level, etc.). Does it mean that we cannot describe the location and the spread?

A. We can switch to other measures, insensitive to existence restrictions, such as **percentiles** (Ross: only in Exercises).

Roughly speaking, a $100p^{\text{th}}$ percentile is a number π_p such that $\mathrm{P}(X \le \pi_p) \approx p$ and $\mathrm{P}(X \ge \pi_p) \approx 1 - p$. The most common are **deciles** ($p = 0.1, 0.2, ..., 0.8, 0.9$) and **quartiles** ($p = 0.25, 0.50, 0.75$. The 2^{nd} quartile - a.k.a. **median**.

Trouble with comparison

Often (if not always in practice) a RV X is a physical quantity and as such it has a physical unit (kg, m, sec, $, etc.), and the mean and standard deviation has the same unit.

Q Let $\mu_X = 100$ and $\sigma_X = 50$, and $\mu_Y = 0.1$, $\sigma_Y = .5$

Does it mean that values of X are large and greatly scattered while values of Y are small and concentrated about the mean?

Trouble with comparison

Often (if not always in practice) a RV X is a physical quantity and as such it has a physical unit (kg, m, sec, $, etc.), and the mean and standard deviation has the same unit.

Q Let $\mu_X = 100$ and $\sigma_X = 50$, and $\mu_Y = 0.1$, $\sigma_Y = .5$

Does it mean that values of X are large and greatly scattered while values of Y are small and concentrated about the mean?

A: No, because “large” or “small” are relative terms. No such claims should be posed without additional information.

Normalization

The invertible ($a \neq 0$) affine transformation

$$\widetilde{X} = \frac{X - a}{b}, \quad X = a\widetilde{X} + b,$$

may serve as a **normalization**, a.k.a. **standardization**, when $a = \mu,\ b = \sigma$, yielding a unit-less RV with mean 0 and variance 1.

Such transformations are common in sciences, e.g., temperature in Fahrenheit (F) vs. Celsius (C) scale:

$$C = \frac{5}{9}(F - 32), \quad \text{or} \quad F = \frac{9}{5}C + 32.$$

What are the higher moments for?

The third moment $\mathsf{E}\,X^3$. If X is symmetric, i.e. $f(x) = f(-x)$, then not only $\mathsf{E}\,X = 0$ but $\mathsf{E}\,X^3 = 0$ as well. Hence

skewness $= \mathsf{E}\,\widetilde{X}^3$ shows asymmetry about the mean.

The fourth moment $\mathsf{E}\,X^4$. The heavier the tails (extreme values with relatively high probability), the larger the fourth moment. Hence

kurtosis $= \mathsf{E}\,\widetilde{X}^4$ shows how heavy the tails are.

Note that the normalization is necessary in both cases.

mgf

Definition. The **moment generating function** of a RV X is

$$M(t) = M_X(t) = \mathrm{E}\, e^{tX}, \qquad \underline{\text{subject to existence}}.$$

A desired domain should include an interval $[0, b)$ or $(-a, 0]$.

The name stems from the following property:

$$M^{(n)}(t) = \frac{d^n M}{dt^n} = \mathrm{E}\, \left(X^n\, e^{tX}\right) \quad \text{entails} \quad M^{(n)}(0) = \mathrm{E}\, X^n.$$

Misleading name

However, generation of moments is only a minor feature of the mgf.

More importantly:

- An mgf identifies the distribution and parameters of a RV .
- A distribution formula may be complex while the mgf is simple.
- Mgf allows powerful calculus and algebra.
- It simplifies the concept of limits of probability distributions.

Alternatives

Laplace transform. $L(t) = L_X(t) = \mathrm{E}\, e^{-tX}$.

Small disadvantage: $L^{(n)}(0) = (-1)^n \mathrm{E}\, X^n$.

Big advantage: It always exists for $X \geq 0$ (frequent).

Fourier transform. $\varphi(t) = \varphi_X(t) = \mathrm{E}\, e^{itX}$.

Small disadvantage: Complex numbers. $\varphi^{(n)}(0) = i^n \mathrm{E}\, X^n$.

Big advantage: Complex numbers. It always exists.

Example: mgf may not exist

RVs with **power tails** (a.k.a. **fat tails** or **heavy tails**) often occur in engineering, science, or economics:

$$\mathrm{P}(|X| > x) \approx \frac{1}{x^p}, \quad \text{pmf or pdf } f(x) \approx \frac{c}{|x|^{p+1}}.$$

Then the mgf doesn't exist, in general.

If $X \geq 0$, then $M(t)$ exists for $t \leq 0$.

No moments exist for $p \leq 1$. No second moment exists for $p \leq 2$. Then the mgf is not "mgf".

Product of independent RVs

Theorem

Let X and Y be independent with finite means. Then

$$\mathsf{E}\left(X\,Y\right) = \left(\mathsf{E}\,X\right)\left(\mathsf{E}\,Y\right).$$

The extension to finitely many independent RVs is immediate:

$$\mathsf{E}\left(X_1 \cdots X_n\right) = \left(\mathsf{E}\,X_1\right) \cdots \left(\mathsf{E}\,X_n\right).$$

Proof

We present the proof for discrete RVs.

The continuous case requires a bit of Calculus 3.

Let $f(x)$ and $g(y)$ be pmf's of X and Y. By independence

$$\mathrm{P}(X = x,\ Y = y) = f(x)g(y).$$

Hence, by the arithmetic distributive property

$$\begin{aligned} \mathrm{E}\left(XY\right) = \sum_x \sum_y xy\, f(x)g(y) \quad &= \left(\sum_x x\, f(x)\right) \left(\sum_y y\, g(y)\right) \\ &= \left(\mathrm{E}\, X\right) \left(\mathrm{E}\, Y\right) \end{aligned}$$

The sum

Corollary

Let X and Y be independent with finite variances. Then

$$\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y).$$

The extension to finitely many independent RVs and to linear combinations is immediate:

$$\mathrm{Var}(c_1\,X_1+\cdot+c_n\,X_n)=c_1^2\,\mathrm{Var}(X_1)+\cdots c_n^2\,\mathrm{Var}(X_n)$$

Proof

W.l.o.g. we may and do assume that $\mathrm{E}\, X = \mathrm{E}\, Y = 0$.
(If not, replace X and Y by centered $X - \mathrm{E}\, X$ and $Y - \mathrm{E}\, Y$.)

By this simplification, then by linearity of expectation and independence, which implies $\mathrm{E}\,(XY) = 0$,

$$\begin{aligned} \mathrm{Var}(X+Y) &= \mathrm{E}\,(X+Y)^2 = \mathrm{E}\left(X^2 + 2XY + Y^2\right) \\ &= \mathrm{E}\left(X^2\right) + 2\mathrm{E}\left(XY\right) + \mathrm{E}\left(Y^2\right) = \mathrm{Var}(X) + \mathrm{Var}(Y) \end{aligned}$$

Corollary

Let X and Y be independent with finite mgf's $M_X(t) = \mathrm{E}\, e^{tX}$ and $M_Y(t) = \mathrm{E}\, e^{tY}$. Then

$$M_{X+Y}(t) = \mathrm{E}\, e^{t(X+Y)} = M_X(t) \cdot M_Y(t).$$

The extension to finitely many independent RVs is immediate:

$$M_{X_1+\cdots+X_n} = M_{X_1}(t) \cdots M_{X_n}(t).$$

If X_k are iid with the mgf $M(t)$, then

$$M_{X_1+\cdots+X_n} = M^n(t).$$

Definition

Exercise. $|\mathrm{E}\,(XY)|^2 \le (\mathrm{E}\,X^2)(\mathrm{E}\,Y^2)$.

Hint: The quadratic function $\mathrm{E}\,(X+tY)^2 \ge 0$ for every t.

Apply algebra.

Let X, Y posses second moments. Hence $\mathrm{E}\,(XY)$ exists.

The **covariance**

$$\mathrm{Cov}(X,Y) = \mathrm{E}\,(X-\mathrm{E}\,X)(Y-\mathrm{E}\,Y) = \mathrm{E}\,(XY) - (\mathrm{E}\,X)(\mathrm{E}\,Y)$$

generalizes the variance: $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.

Polarization formulas

Also by simple algebra

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \frac{1}{2}\Big(\mathrm{Var}(X+Y) - \mathrm{Var}(X) - \mathrm{Var}(Y)\Big)\\ &= \frac{1}{4}\Big(\mathrm{Var}(X+Y) - \mathrm{Var}(X-Y)\Big)\end{aligned}$$

Correlation coefficient

The **correlation coefficient** is just the covariance after normalization:

$$\rho = \mathrm{Cov}(\widetilde{X}, \widetilde{Y}) = \frac{\mathrm{E}\,(XY) - \mu_X \mu_Y}{\sigma_X \sigma_Y}$$

From the previous exercise it follows that

$$-1 \le \rho \le 1,$$

and the extreme values are attained when one variable is an affine transformation of the other, e.g., $Y = aX + b$.

Correlation vs. independence

So we call X and Y **uncorrelated** if $\rho = 0$ (i.e., $\mathrm{Cov}(X, Y) = 0$. Among RVs with finite variances it's a weaker property than independence. That is, it is implied by independence.

The property is fragile. E.g., for uncorrelated X and Y, even their simple transformation such as X^2 vs. Y^2 may destroy it.

In contrast, when X and Y are independent, so are their every transformations.

However, the property is much easier to detect. The correlation coefficient can be simply approximated from a sample while independence is hard to confirm.