Moments and Deviations

Piyush Patil

February 7, 2017

In this chapter we'll look at methods for bounding *tail distributions* of random variables, ie the probabiltiy that a random variable deviates from its expected value. We'll quantify the probability that a random variable assumes values far from its expectation, which will be useful in the analysis of the worst-case runtime of randomized algorithms. The two main bounds we'll look at in this chapter are Markov's inequality Chebyshev's inequality.

1 Markov's Inequality

Markov's inequality is a useful upper bound on the tail of a random variable; that is, it provides a cieling for the probability that a random variable, under certain conditions, assumes values beyond some threshold.

(Theorem) Markov's inequality: *Let X be a non-negative random variable. Then*

$$\forall a \in \mathbb{R} : \Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

- *Proof*: Fix $a > 0$ and define the indicator variable I to be 1 when $X \geq a$ and 0 otherwise. It follows that since $X \geq 0$,

$$I \leq \frac{X}{a}$$

Moreover, $\mathbb{E}[I] = \Pr(I = 1) = \Pr(X \geq a)$. It follows that

$$\Pr(X \geq a) = \mathbb{E}[I] = \mathbb{E}\left[\frac{X}{a}\right] = \frac{\mathbb{E}[X]}{a}$$

□

2 Variance and Moments

If all we know about a random variable is its expectation (and that it's non-negative), then Markov's inequality gives the best possible tail bound. However, if we have more information, we can obtain better bounds. To precisely define what we mean by "information" about the distribution of a random variable, we'll introduce the idea of *moments*, which are used to quantify the overall "shape" of a set of points, or a distribution. The expectation of the random variable is its first moment, and we define subsequent moments iteratively with the following generalization.

(Definition) Moment: *The k^{th} **moment** of a random variable X is $\mathbb{E}[X^k]$.*

- *Intuition*: The first moment is the mean, and the second is related to the variance (which will be defined later). The first moment thus measures the average point, whereas the second measures the average distance of the points from the mean. The same pattern holds - it turns out the third moment measures the skewness of the distribution, while the fourth measures its peakedness. In general, higher order moments generalize variance - whereas the variance takes the average distance of the points from the mean, higher order moments take the distance between points and the mean and amplify it by raising to a power before taking the average. Intuitively, this reveals differently characteristics of the distribution. Similar to how knowing all the higher order derivatives of an analytic function completely specify the function (through its Taylor series), knowing all the higher order moments of a random variable completely specify its distribution. This can be made more formal by introducing a bit of more advanced theory. First, define the *moment-generating function* of a random variable X to be the function

$$\forall t \in \mathbb{R} : M(t) = \mathbb{E}[e^{tX}]$$

It's known that the moment generating function of the random variable is directly linked to the probability distribution of X. In probability theory and statistics in particular a central question is to find the underlying probability distribution of a set of data, a very difficult question in general, so the link between moment generating functions and probability distributions

if quite significant. The moments of a random variable "ground" the moment-generating function and hence the probability distribution, by completely determining its values, with greater and greater precision as we look at higher order moments. This can be seen by examining the Taylor series:

$$M(t) = \mathbb{E}\left[\sum_{k\in\mathbb{N}} \frac{X^k}{k!}t^k\right] = \sum_{k\in\mathbb{N}} \frac{t^k}{k!}\mathbb{E}[X^k]$$

As we discussed in the section on the intuition behind the definition of a moment, a specific moment of interest is the second moment, better known as the *variance*. Technically speaking, the above moments are more commonly referred to as *raw moments*, to contrast them with *central moments*, which are defined as

$$\mathbb{E}[(X-\mu)^k] \text{ where } \mu = \mathbb{E}[X]$$

In other words, central moments are merely moments about the center of the distribution, ie the moment of the random variable after its been centered (shifted so that the "middle" of the distribition is at the origin). The variance is then just a special name we give to a specific moment.

(Definition) Variance: *The **variance** of a random variable X is the second central moment, ie*

$$\text{Var}(X) = \mathbb{E}[(X-\mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Because the variance involves squaring the values of the random variable, this means that in practice when the variable's values represent some physical quantity, the units are squared as well. To return to the original units, we take the square root, giving a more practical notion of spread that we call the *standard deviation*:

$$\sigma[X] = \sqrt{\text{Var}(X)}$$

We can generalize the idea of variance to a notion of how much two random variables "vary" with each other. This is meant to quantify the answer to the question of how the behavior of one variable predicts or influences the behavior of another - if having information about the distribution of one variable tells us a lot about another variable, then the former variable approximates the latter in some sense, and the two vary together more than a pair of variables for which knowing about one tells us nothing about the other.

(Definition) Covariance: *The **covariance** of two random variables X and Y is defined*

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

• *Intuition*: The definition of covariance is natural in and of itself. All we've done is first center both variables, to remove bias from translational differences between X and Y, and look at the expected value of the product. The product of the centered random variables is a good representation of how much the variables vary with each other because its magnitude directly depends on whether X and Y have the same sign, and moreover is larger whenever large values of one of the variables coincide with large values of the other variable. In fact, the correlation of the two random variables is defined as the normalized covariance - ie the quantity obtained by taking the covariance and dividing by the product of the standard deviations, to further standardize the random variables and account for differences in the absolute quantity of X or Y (this way, the same intuition holds but instead of comparing values of X to values of Y, we compare respective z-scores - distances from the mean).

However, the covariance does have a deeper interpretation. Recall that random variables are really real-valued functions, "under the hood". From a functional analysis perspective, if we consider the space of square integrable functions L_2, the covariance as defined above is an inner product of functions. Specifically, the quantity

$$\langle f, g\rangle := \mathbb{E}[fg] \text{ where } \mathbb{E}[f] := \int_{\mathbb{R}} f(x)d\rho(x) \text{ and } \rho \text{ is the distribution of } f$$

is a well defined inner product. The covariance then is simply the process of taking functions in L_2, projecting them down to the subspace of functions that integrate to zero (ie functions which are "balanced") and taking the inner product. We can thus think of the covariance as the angle between two centered random variables, where identical or directly proportional random variables are parallel and completely independent variables with no influence one another are orthogonal.

With this definition, it's often cleaner to think of the variance as a special case of the covariance - a measure of how much a variable "varies with itself", ie the spread of the variable. Notice that

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$$

because

$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathbb{E}[(X+Y-\mathbb{E}[X+Y])^2] = \mathbb{E}[((X-\mathbb{E}[X])+(Y-\mathbb{E}[Y]))^2] \\ &= \mathbb{E}[(X-\mathbb{E}[X])^2] + \mathbb{E}[(Y-\mathbb{E}[Y])^2] + 2\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)\end{aligned}$$

As we'll see below independent random variables have exactly the expectation and covariance we'd expect.

(Theorem) Expectation is multiplicative for independent variables: *Let X and Y be independent random variables. Then*

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$$

- *Proof*:

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{x\in\mathbb{R}}\sum_{y\in\mathbb{R}} xy\Pr(X=x\cap Y=y) = \sum_{x\in\mathbb{R}}\sum_{y\in\mathbb{R}} xy\Pr(X=x)\Pr(Y=y) = \sum_{x\in\mathbb{R}} x\Pr(X=x)\sum_{y\in\mathbb{R}} y\Pr(Y=y) \\ &= \left(\sum_{x\in\mathbb{R}} x\Pr(X=x)\right)\cdot\left(\sum_{y\in\mathbb{R}} y\Pr(Y=y)\right) = \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

□

(Theorem) Variance is additive for independent variables: *Let X and Y be independent random variables. Then*

$$\mathrm{Cov}(X,Y) = 0 \text{ and } \mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

- *Proof*: The second statement follows from the first and the above statement on the variance of a sum of variables, so let's prove the first. This easily follows from the definition:

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] = \mathbb{E}[X-\mathbb{E}[X]]\mathbb{E}[Y-\mathbb{E}[Y]] = 0$$

□

This result extends by induction to any finite sum of random variables.

3 Chebyshev's Inequality

If we know the variance of a random variable, we can find a much tighter bound on its tail distribution.

(Theorem) Chebyshev's inequality: *Let X be a random variable. Then*

$$\forall a\in\mathbb{R} : \Pr(|X-\mathbb{E}[X]|\geq a) \leq \frac{\mathrm{Var}(X)}{a^2}$$

- *Proof*: We can prove this with Markov's inequality:

$$\Pr(|X-\mathbb{E}[X]|\geq a) = \Pr((X-\mathbb{E}[X])^2\geq a^2) \leq \frac{\mathbb{E}[(X-\mathbb{E}[X])^2]}{a^2} = \frac{\mathrm{Var}(X)}{a^2}$$

if we view $(X-\mathbb{E}[X])^2$ as a random variable. □

- *Intuition*: Intuitively, this statement quantifies the idea that for most probability distributions, most of the values lie near the mean. Formally, the density of points decreases quadratically in z-score - no more than $\frac{1}{k^2}$ of the values of the distribution are more than k standard deviations from the mean.