Expectation and Moments

Ben Gravell

benjamin.gravell@utdallas.edu
The Erik Jonsson School of Engineering and Computer Science
The University of Texas at Dallas
800 W. Campbell Rd.
Richardson, TX 75080

Outline

Expectation and moments

Expectation

The **expectation** or **mean** of a random variable X is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx \tag{1}$$

The expectation of a function of a random variable $g(X)$ is

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx \tag{2}$$

If the RV is discrete, these integrals become simple sums:

$$\mathbb{E}[X] = \sum_i x_i P_X(x_i) \tag{3}$$

$$\mathbb{E}[g(X)] = \sum_i g(x_i) P_X(x_i) \tag{4}$$

Linearity of expectation

Expectation is a **linear operator** - follows from linearity of integration

$$
\begin{aligned}
&\mathbb{E}[X+Y] && (5)\\
&= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x+y) f_{XY}(x,y) dx dy && (6)\\
&= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x f_{XY}(x,y) dx dy + \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} y f_{XY}(x,y) dx dy && (7)\\
&= \int_{-\infty}^{+\infty} x \left(\int_{-\infty}^{+\infty} f_{XY}(x,y) dy \right) dx + \int_{-\infty}^{+\infty} y \left(\int_{-\infty}^{+\infty} f_{XY}(x,y) dx \right) dy && (8)\\
&= \int_{-\infty}^{+\infty} x f_X(x) dx + \int_{-\infty}^{+\infty} y f_Y(x) dy && (9)\\
&= \mathbb{E}[X] + \mathbb{E}[Y] && (10)
\end{aligned}
$$

Use induction to conclude the linearity property

$$
\mathbb{E}\left[\sum_{i=1}^{N} X_i\right] = \sum_{i=1}^{N} \mathbb{E}\left[X_i\right] \tag{11}
$$

Expectation of a Gaussian

Recall the Gaussian random variable $X \sim \mathcal{N}(\mu, \sigma^2)$.

Let's show the mean is μ using the change of variable $z = \frac{x-\mu}{\sigma}$

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x \cdot f_X(x) dx && (12) \\
&= \int_{-\infty}^{\infty} x \cdot \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx && (13) \\
&= \int_{-\infty}^{\infty} (\sigma z + \mu) \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2\right) dz && (14) \\
&= \frac{\sigma}{\sqrt{2\pi}} \underbrace{\int_{-\infty}^{\infty} z \cdot \exp\left(-\frac{1}{2} z^2\right) dz}_{=0 \text{ because integrand odd}} + \mu \underbrace{\left[\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2\right) dz\right]}_{=1 \text{ because } P[Z \leq \infty] = 1} && (15) \\
&= \mu && (16)
\end{aligned}$$

Conditional expectation

The **conditional expectation** of random variable Y given event B has occurred is

$$\mathbb{E}[Y|B] = \int_{-\infty}^{\infty} y f_{Y|B}(y|B) dy \tag{17}$$

The **conditional expectation** of random variable Y conditioned on random variable X is

$$\mathbb{E}[Y|X=x] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy \tag{18}$$

We have a **law of total expectation** (like law of total probability)

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} \mathbb{E}[Y|X=x] f_X(x) dx \tag{19}$$

Moments are expectations of monomials of (shifted and scaled) RVs

Moments

The k^{th} **(raw) moment** of X is

$$m_k = \mathbb{E}[X^k] \tag{20}$$

The k^{th} **central moment** of X is

$$c_k = \mathbb{E}[(X - \mathbb{E}[X])^k] \tag{21}$$

The k^{th} **standardized moment** of X is

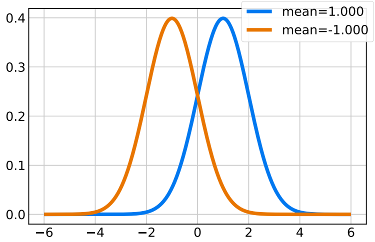
$$s_k = \frac{\mathbb{E}[(X - \mathbb{E}[X])^k]}{\mathbb{E}[(X - \mathbb{E}[X])^2]^{k/2}} = \frac{c_k}{c_2^{k/2}} \tag{22}$$

Moments summarize different aspects of the **shape** of a distribution

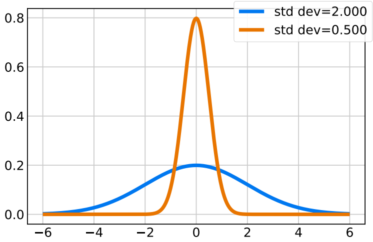
Name	Definition	Intuition
Mean	$\mu = m_1$	Location or center
Variance	$\sigma^2 = c_2$	Dispersion or spread
Std deviation	$\sigma = \sqrt{\sigma^2}$	Dispersion or spread
Skewness	s_3	Asymmetry or tilt
Kurtosis	s_4	Heaviness of tails

See `moments.py`

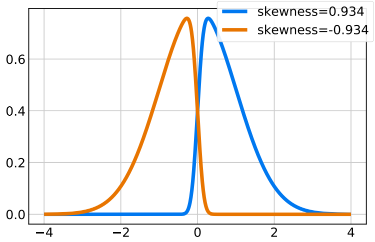
Comparison of pdfs with different moments



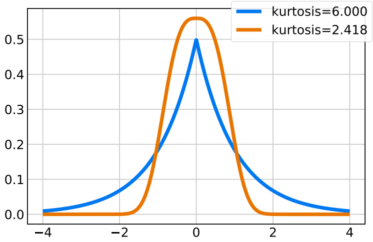
(a) Mean



(b) Standard deviation



(c) Skewness



(d) Kurtosis

We can convert between raw and central moments

Example: Second moment

$$
\begin{aligned}
c_2 &= \mathbb{E}[(X - \mathbb{E}[X])^2] && (23) \\
&= \mathbb{E}[X^2 - 2\mathbb{E}[X]X + \mathbb{E}[X]^2] && (24) \\
&= \mathbb{E}[X^2] - 2\mathbb{E}[X]^2 + \mathbb{E}[X]^2 && (\text{linearity of } \mathbb{E}[\cdot]) \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2 && (25) \\
&= m_2 - m_1^2 && (26)
\end{aligned}
$$

This relation generalizes to higher-order moments as

$$
c_k = \sum_{i=0}^{k} \binom{k}{i} (-1)^i \mu^i m_{k-i} \tag{27}
$$

Homework P3-1:
Verify the expression for the variance of a Gaussian.
Hint: See Example 4.1-7 in [1]

Optional Exercise:
Find expressions for all moments of a Gaussian.
Hint: See e.g. https://arxiv.org/abs/1209.4340

Basic probability bounds

Often we want to bound the probability of certain events or random variables without having to specify/compute their distribution

c.f. the first several pages of Wainwright's book [2]

Tail bounds

Markov inequality

Given a non-negative random variable X with finite mean, we have

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t} \quad \text{for all } t > 0 \tag{28}$$

"X is probably small when its mean is small"

The most basic tail bound.

Basis for several "classical" concentration inequalities.

Chebyshev inequality

Given a random variable X with finite mean μ and variance σ^2, we have

$$\mathbb{P}[|X - \mu| \geq t] \leq \frac{\sigma^2}{t^2} \quad \text{for all } t > 0 \tag{29}$$

"X is probably close to its mean whenever its variance is small"

The most basic concentration inequality.

Proof: Follows by applying Markov inequality to the non-negative random variable $(X - \mu)^2$.

Moment bound

Given a non-negative random variable X with finite moments up to order k, we have

$$\mathbb{P}[|X-\mu| \geq t] \leq \frac{\mathbb{E}\left[|X-\mu|^k\right]}{t^k} \quad \text{for all } t > 0 \tag{30}$$

Proof: Follows by applying Markov inequality to the random variable $|X-\mu|^k$

Chernoff bound

Given a non-negative random variable X with a moment generating function in a neighborhood of zero, we have

$$\mathbb{P}[X \geq 0] \leq \inf_{\theta > 0} \mathbb{E}\left[e^{\theta X}\right] \tag{31}$$

Proof: Follows by applying Markov inequality to the random variable $e^{\theta(X-\mu)}$ and optimizing over θ.

The moment bound with an optimal choice of k is never worse than the Chernoff bound.

Nonetheless, the Chernoff bound is most widely used in practice, possibly due to the ease of manipulating moment generating functions.

Homework P3-2:
Compare the Markov inequality bound with the exact tail probability from the exponential cdf with parameter $\lambda = 1$; compute the probability bounds at the level $t = 2$. How bad is the Markov bound compared with the exact tail probability?
Hint: The mean of an exponential random variable is $\mu = 1/\lambda$.

Homework P3-3:
Compare the Chebyshev inequality bound with the exact tail bound from the standard normal cdf; compute the probability bounds at the level $t = 2$. How bad is the Chebyshev bound compared with the exact concentration probability?
Hint: The standard normal cdf does not have a closed-form expression, so either use the `cdf()` method of `scipy.stats.norm` or a table of the standard normal cdf to get the exact value. In case you run into issues, $\Phi(2) = 1 - \Phi(-2) = 0.9772$.

Joint moments summarize different aspects of the shape of a joint distribution

Joint moments

The **ijth (raw) joint moment** of random variables X and Y is

$$m_{ij} = \mathbb{E}[X^i Y^j] \tag{32}$$

The **ijth central joint moment** of random variables X and Y is

$$c_{ij} = \mathbb{E}[(X - \mathbb{E}[X])^i (Y - \mathbb{E}[Y])^j] \tag{33}$$

Some joint moments have special, confusing names

The **correlation** is

$$m_{11} = \mathbb{E}[XY] \tag{34}$$

The **covariance** is

$$c_{11} = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \tag{35}$$

The **correlation coefficient** is

$$\rho = \frac{c_{11}}{\sqrt{c_{02} c_{20}}} \tag{36}$$

Homework P3-4:
Prove the relation

$$m_{11} = c_{11} + \mathbb{E}[X]\mathbb{E}[Y]$$

Hint: It is similar to the earlier second moment relation $m_2 = c_2 + m_1^2$

Homework P3-5:
When are the correlation and covariance equal?
Hint: Use the relation $m_{11} = c_{11} + \mathbb{E}[X]\mathbb{E}[Y]$ *you just proved.*

Homework P3-6:
Prove that $\rho \in [-1, 1]$
Hint: See Ch. 4.3 of [1]

Uncorrelated and orthogonal random variables

Uncorrelated random variables

Two random variables are **uncorrelated** if their **covariance** is zero.

Orthogonal random variables

Two random variables are **orthogonal** if their **correlation** is zero.

- Yes I know the terminology is confusing :/

Homework: Uncorrelated random variables

Homework P3-7:
Prove that if X and Y are uncorrelated, then $\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$
i.e. "the variance of the sum is the sum of the variances."
Hint: Use linearity of expectation.

Homework P3-8:
Prove that if X and Y are independent, then they are uncorrelated.
Remark: The converse does not hold unless X and Y are both Gaussian.

Homework P3-9:
Under what condition(s) can a pair of uncorrelated random variables be orthogonal?
Hint: This is a special case of one of the earlier exercises.

Random vectors

Random vector

A **random vector** is a vector of random variables.

The **cdf** of a random vector is defined as

$$F_X(x) = \mathbb{P}[X_1 \leq x_1 \text{ and } X_2 \leq x_2 \text{ and } \ldots X_n \leq x_n] \tag{37}$$

The **pdf** is defined as

$$f_X(x) = \frac{\partial^n F_X(x)}{\partial x_1 \partial x_2 \cdots \partial x_n} \tag{38}$$

Similar definitions for joint, marginal, and conditional distributions

- See Ch. 5.1 of [1]

The **expectation** of a random vector X is the vector μ_X with entries

$$[\mu_X]_i = \mathbb{E}[X]_i = \int_{-\infty}^{\infty} x_i f_{X_i}(x_i) dx_i \tag{39}$$

where $f_{X_i}(x_i)$ is the ith marginal pdf.

Moments are defined similarly as with random variables.

(Auto)-covariance matrix of X

$$K_X = \mathbb{E}[(X - \mu_X)(X - \mu_X)^\intercal] \tag{40}$$

(Cross)-covariance matrix between X and Y

$$C_{XY} = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)^\intercal] \tag{41}$$

We can gather these up into the block covariance matrix

$$D_{XY} = \begin{bmatrix} K_X & C_{XY} \\ C_{XY}^\intercal & K_Y \end{bmatrix} = \mathbb{E}\left[\begin{bmatrix} X - \mu_X \\ Y - \mu_Y \end{bmatrix} \begin{bmatrix} X - \mu_X \\ Y - \mu_Y \end{bmatrix}^\intercal\right] \tag{42}$$

(Auto)-correlation matrix of X

$$R_X = \mathbb{E}[XX^\intercal] \succeq 0 \tag{43}$$

(Cross)-correlation matrix between X and Y

$$S_{XY} = \mathbb{E}[XY^\intercal] \tag{44}$$

We can gather these up into the block correlation matrix

$$B_{XY} = \begin{bmatrix} R_X & S_{XY} \\ S_{XY}^\intercal & R_Y \end{bmatrix} = \mathbb{E}\left[\begin{bmatrix} X \\ Y \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix}^\intercal\right] \tag{45}$$

Homework 3-10:
Prove the identity between covariance and correlation matrices

$$R = K + \mu\mu^\intercal \tag{46}$$

Hint: Use linearity of expectation.

Homework 3-11:
Write an expression for D in terms of B, μ_X, μ_Y.
Hint: It follows immediately from $R = K + \mu\mu^\intercal$ *by stacking* X *and* Y*.*

Homework 3-12:
Prove that $R \succeq K \succeq 0$ and $B \succeq D \succeq 0$ where $A \succeq B$ means $A - B$ is symmetric positive semidefinite.
Hint: It follows by the above relations and the property of outer product matrices $AA^\intercal \succeq 0$ *for any matrix* A*, and taking* $A = \mu$*.*

A random vector X is **uncorrelated** with itself if K is diagonal.

A random vector X is **orthogonal** with itself if R is diagonal.

Two random vectors X and Y are **uncorrelated** if $C = 0$.

Two random vectors X and Y are **orthogonal** if $S = 0$.

Optional Exercise:
Think about how these expressions can be summarized in terms of the block matrices C and D.

Optional Exercise:
Under what condition(s) can a pair of uncorrelated random vectors be orthogonal?
Hint: You already solved this in the scalar case.

Whitening transformation

Sometimes we need to get a standardized version of a random variable

In the scalar case we used the standardizing transform

$$Z = \frac{X - \mu}{\sigma} \tag{47}$$

- Subtract out the mean and normalize by the standard deviation, so Z has zero mean and variance one
- Need to assume $\sigma > 0$ for non-degeneracy

Whitening transformation

The **whitening transformation** is the multivariate generalization of the scalar standardizing transform

- Based on the eigen-decomposition of the covariance matrix

The **whitening transformation** is

$$Z = \Lambda_X^{-1/2} U_X^\intercal (X - \mu) \tag{48}$$

- Subtract the mean out and normalize, so Z has zero mean and identity auto-covariance
- Λ_X is a diagonal matrix whose entries are the n eigenvalues of K_X
  - The eigenvalues λ_i are real numbers since K_X is symmetric
  - Need to assume $\lambda_i > 0$ for $i = 1, \ldots, n$ for non-degeneracy
    - Equivalent to assuming K_X full rank
  - $\Lambda_X^{-1/2}$ is diagonal with entries $\lambda_i^{-1/2}$
- U_X is an orthogonal matrix whose columns are n eigenvectors of K_X

Sometimes we need to get a a random vector Y with nonzero mean μ_Y and non-identity covariance K_Y from a white random vector

- Inverse operation of the whitening transformation

The **coloring transformation** is

$$Y = U_Y \Lambda_Y^{1/2} X + \mu \tag{49}$$

- Λ_Y is a diagonal matrix whose entries are the n eigenvalues of K_Y
- U_Y is an orthogonal matrix whose columns are n eigenvectors of K_Y

The n-**dimensional multivariate Gaussian** pdf is

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n \det(K)}} \exp\left[-\frac{1}{2}(x-\mu)^\intercal K^{-1}(x-\mu)\right] \qquad (50)$$

- Mean is $\mu \in \mathbb{R}^n$
- Covariance is $K \in \mathbb{R}_+^{n \times n}$

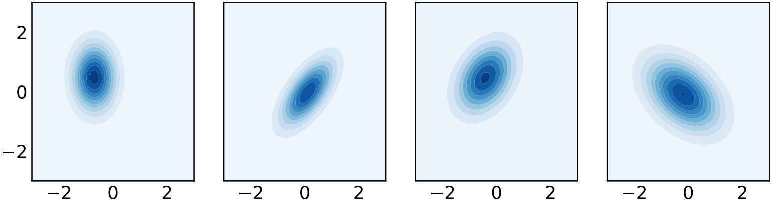


Figure 2: Various multivariate Gaussian pdfs for $n = 2$.
See `multivariate_gaussian.py`

Properties of multivariate Gaussians

Gaussians are extremely special distributions with nice properties

- Marginals of a Gaussian are Gaussian
- Gaussians conditioned on Gaussians are Gaussian
- Any affine transformation of a Gaussian is Gaussian
- All pertinent information about a Gaussian is encoded in the mean and covariance
- Sums of random vectors tend towards a Gaussian
  (central limit theorem, coming up)

Homework 3-13:
What is the pdf of a white (zero mean and identity covariance) multivariate Gaussian random vector X? Can it be expressed in terms of the marginal densities of each component of X? If so, write the expression. Are the components of X statistically independent?

[1] John Woods and Henry Stark.
Probability, Statistics, and Random Processes for Engineers.
Pearson Higher Ed, 4 edition, 2011.

[2] Martin J Wainwright.
High-dimensional statistics: A non-asymptotic viewpoint, volume 48.
Cambridge University Press, 2019.
https://people.eecs.berkeley.edu/~wainwrig/BibPapers/Wai19.pdf.