ELE 382: Probabilistic Systems and Information Processing

Review of Basic Probability Theory: Part 2

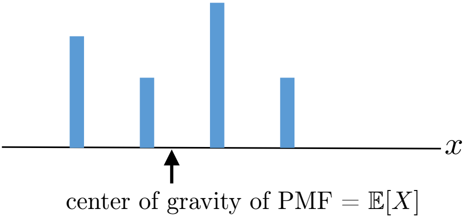
Yuxin Chen

Princeton University, Fall 2018

Outline

- Expectation, variance, covariance, and moments
  - Application: runtime of Quicksort
- Uncorrelatedness and independence
- Conditional expectation and conditional variance

Expectation



- Let X be a discrete r.v. with PMF p_X, and let $g(x)$ be a function of x. The *expectation* (or *mean*) of $g(X)$ is

$$\mathbb{E}[g(X)] \stackrel{\text{def}}{=} \sum_x g(x) p_X(x)$$

- For a continuous r.v. $X \sim f_X(x)$, the expectation of $g(X)$ is

$$\mathbb{E}[g(X)] \stackrel{\text{def}}{=} \int g(x) f_X(x) \mathrm{d}x$$

Expectation

- **Linearity:** expectation is *linear*, i.e. for any constants a_1 and a_2,

$$\mathbb{E}[a_1 g_1(X) + a_2 g_2(X)] = a_1\, \mathbb{E}[g_1(X)] + a_2\, \mathbb{E}[g_2(X)]$$

- Remarks
  - Expectation provides a *summary* of the r.v. — a single number — instead of specifying the entire distribution
  - It is far easier to estimate the expectation of a r.v. from data than to estimate its distribution

Moments

- The *first moment* of $X \sim f_X(x)$ is

$$\mathbb{E}[X] = \int x f_X(x) \mathrm{d}x$$

- The *second moment* of X is

$$\mathbb{E}[X^2] = \int x^2 f_X(x) \mathrm{d}x$$

- The *kth moment* of X is

$$\mathbb{E}[X^k] = \int x^k f_X(x) \mathrm{d}x$$

Variance

- The *variance* of X is

$$\mathsf{Var}(X) \overset{\text{def}}{=} \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

  - can be expressed in terms of moments as

$$\mathsf{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \qquad \text{(also implies } \mathbb{E}[X^2] \geq (\mathbb{E}[X])^2\text{)}$$

  - For any constants a and b, $\mathsf{Var}(aX + b) = a^2 \mathsf{Var}(X)$
  - If $X_1, \cdots, X_k$ are independent, then

$$\mathsf{Var}(X_1 + \cdots + X_k) = \mathsf{Var}(X_1) + \cdots + \mathsf{Var}(X_k)$$

- The *standard deviation* of X is $\sigma_X \overset{\text{def}}{=} \sqrt{\mathsf{Var}(X)}$

Bias-variance tradeoff

Suppose we wish to estimate a random object Y, and the estimation error is given by Z.

A common way to evaluate the goodness of the estimate is via the mean squared estimation error

$$\mathbb{E}[Z^2] \;=\; (\;\underbrace{\mathbb{E}[Z]}_{\text{error bias}}\;)^2 + \underbrace{\mathsf{Var}(Z)}_{\text{variance}}$$

$$\text{mean squared error} \;=\; \text{bias}^2 + \text{variance}$$

- Achieving optimal tradeoff between bias and variance is a central problem arising in most machine learning / estimation tasks

Mean and variance for common random variables

random variable	mean	variance
Bern(p)	p	$p(1-p)$
Geo(p)	p^{-1}	$\frac{1-p}{p^2}$
Bin(n,p)	np	$np(1-p)$
Poisson(λ)	λ	λ
Unif(a,b)	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exp(λ)	λ^{-1}	λ^{-2}
$\mathcal{N}(\mu,\sigma^2)$	μ	σ^2

Example: coupon collector's problem




- There are n different types of coupons
- Each pack contains one coupon (independently and equally likely)
- How many packs X would you buy to complete the series (i.e. obtain each type of coupon at least once)?

Example: coupon collector's problem

Claim: $\mathbb{E}[X] = n \sum_{i=1}^{n} \frac{1}{i} \approx n \log n.$

Proof: Let X_i be # packs we need to buy in order to obtain the ith new coupon, after $i-1$ different coupons have been collected. Then

$$X = \sum_{i=1}^{n} X_i$$

When exactly $i-1$ coupons have been found, the probability of obtaining a new coupon in a new draw is

$$p_i = \frac{n-(i-1)}{n}$$

This means $X_i \sim$ Geo(p_i), and hence $\mathbb{E}[X_i] = \frac{1}{p_i} = \frac{n}{n-i+1}$. Therefore

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} \frac{n}{n-i+1} = n \underbrace{\sum_{i=1}^{n} \frac{1}{i}}_{\text{harmonic number}}$$

Application: expected runtime of Quicksort

Top 10 algorithms of the 20th century

1. 1946: The Metropolis Algorithm for Monte Carlo
2. 1947: Simplex Method for Linear Programming
3. 1950: Krylov Subspace Iteration Method
4. 1951: The Decompositional Approach to Matrix Computations
5. 1957: The Fortran Optimizing Compiler
6. 1959: QR Algorithm for Computing Eigenvalues.
7. 1962: Quicksort Algorithms for Sorting
8. 1965: Fast Fourier Transform (by James Cooley and $\underbrace{\text{John Tukey}}_{\text{Princeton statistics}}$)
9. 1977: Integer Relation Detection
10. 1987: Fast Multipole Method

https://www.siam.org/pdf/news/637.pdf

Application: expected runtime of Quicksort

Quicksort: a recursive "divide-and-conquer" approach to sorting

- Input: a list $S = \{x_1, \cdots, x_n\}$ of n numbers
- Output: elements of S in sorted order

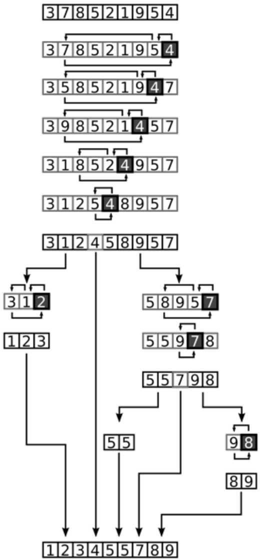
1. choose an element of S as a pivot; call it x
2. compare all other elements of S to x and divide S into 2 sublists

    S_1: all elements of S less than or equal to x

    S_2: all elements of S greater than x

3. Quicksort(S_1) and Quicksort(S_2)
4. return $[S_1, x, S_2]$

Application: expected runtime of Quicksort



Demo:

`http://me.dt.in.th/page/Quicksort/`

Application: expected runtime of Quicksort

Random Quicksort: a pivot is chosen independently and uniformly at random

Theorem 2.1

The expected $\#$ comparisons made by Random Quicksort is

$$(2n+2)\sum_{k=1}^{n}\frac{1}{k} - 4n \;=\; 2n\log n + O(n)$$

- Much better than *worst-case* complexity (i.e. $O(n^2)$)

$f(n) = O(n)$ means there is a constant $c > 0$ s.t. $f(n) \leq cn$ for all n

Application: expected runtime of Quicksort

Proof of Theorem 2.1: Let $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ be the same values as $x_1, \cdots, x_n$ but sorted in increasing order.

- For $i < j$, let Z_{ij} be an indicator s.t.

$$Z_{ij} = \begin{cases} 1, & \text{if } x_{(i)} \text{ and } x_{(j)} \text{ have been compared at any time} \\ 0, & \text{otherwise} \end{cases}$$

- The total number of comparisons $Z = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Z_{ij}$ obeys

$$\mathbb{E}[Z] = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{E}[Z_{ij}] \qquad \text{(linearity of expectation)}$$

Application: expected runtime of Quicksort

Proof of Theorem 2.1 (cont.)

- Observe that $Z_{ij} = 1$ iff either $x_{(i)}$ or $x_{(j)}$ is the 1st pivot selected from the set $\{x_{(i)}, \cdots, x_{(j)}\}$.
  - If neither is the first pivot from this set, then $x_{(i)}$ or $x_{(j)}$ will be separated into distinct sublists and will not be compared
- Since the pivot is chosen uniformly at random, this observation indicates that

$$\mathbb{E}[Z_{ij}] = \frac{2}{j - i + 1}$$

Application: expected runtime of Quicksort

Proof of Theorem 2.1 (cont.) Therefore,

$$
\begin{aligned}
\mathbb{E}[Z] &= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{E}[Z_{ij}] \\
&= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{2}{j-i+1} \\
&= \sum_{i=1}^{n-1} \sum_{k=2}^{n-i+1} \frac{2}{k} && \text{(replace } j-i+1 \text{ by } k\text{)} \\
&= \sum_{k=2}^{n} \sum_{i=1}^{n+1-k} \frac{2}{k} && \text{(switch order of summation)} \\
&= \sum_{k=2}^{n} (n+1-k) \frac{2}{k} \\
&= 2(n+1) \sum_{k=2}^{n} \frac{1}{k} - \sum_{k=2}^{n} 2 \\
&= 2(n+1) \sum_{k=1}^{n} \frac{1}{k} - 4n
\end{aligned}
$$

Expectation involving two random variables

- Let $(X, Y) \sim f_{X,Y}$ and let $g(x, y)$ be a function of x and y. The expectation of $g(X, Y)$ is

$$\mathbb{E}[g(X, Y)] = \int \int g(x, y) f_{X,Y}(x, y) \mathrm{d}x \mathrm{d}y$$

- The *correlation* of X and Y is defined as $\mathbb{E}[XY]$
  - X and Y are said to be *orthogonal* if $\mathbb{E}[XY] = 0$

Expectation involving two random variables

- The *covariance* of X and Y is

  $$\mathsf{Cov}(X,Y) \overset{\text{def}}{=} \mathbb{E}\left[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\right] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$$

  Proof:

  $$\begin{aligned}
  &\mathbb{E}\big[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\big] \\
  &\quad = \mathbb{E}\big[XY - \mathbb{E}[X]Y - X\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y]\big] \\
  &\quad = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \quad \text{(by linearity)} \\
  &\quad = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]
  \end{aligned}$$

- X and Y are said to be *uncorrelated* if $\mathsf{Cov}(X,Y) = 0$, or equivalently, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$

- Note that $\mathsf{Cov}(X,X) = \mathsf{Var}(X)$

Independence implies uncorrelatedness

If X and Y are independent, then they are uncorrelated

Proof:

$$
\begin{aligned}
\mathbb{E}[XY] &= \int\int xy f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y \\
&= \int\int xy f_X(x) f_Y(y)\mathrm{d}x\mathrm{d}y \\
&= \int y f(y)\left(\int x f_X(x)\mathrm{d}x\right)\mathrm{d}y \\
&= \mathbb{E}[X]\int y f(y)\mathrm{d}y \ = \mathbb{E}[X]\,\mathbb{E}[Y]
\end{aligned}
$$

$$
\implies \qquad \mathsf{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = 0
$$

Uncorrelatedness does NOT imply independence

Example: let $X, Y \in \{-2, -1, 1, 2\}$ such that

$$
\begin{aligned}
&p_{X,Y}(1,1) = 2/5, && p_{X,Y}(-1,-1) = 2/5 \\
&p_{X,Y}(-2,2) = 1/10, && p_{X,Y}(2,-2) = 1/10, \\
&p_{X,Y}(x,y) = 0 \text{ otherwise}
\end{aligned}
$$

Are X and Y independent? Are they uncorrelated? (Homework)

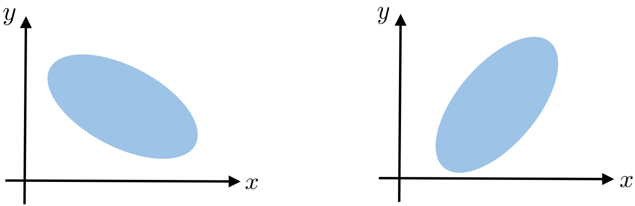
Correlation coefficient

The *correlation coefficient* of X and Y is defined as

$$\rho_{X,Y} \stackrel{\text{def}}{=} \frac{\mathsf{Cov}(X,Y)}{\sqrt{\mathsf{Var}(X)\mathsf{Var}(Y)}}$$

- If $\rho > 0$ (resp. $\rho < 0$), then the values of $X - \mathbb{E}[X]$ and $Y - \mathbb{E}[Y]$ "tend" to have the same (resp. opposite) sign
- The size of $|\rho|$ provides a normalized measure of the extent to which this is true.
- For any constants $a > 0$ and $b \in \mathbb{R}$, one has $\rho_{aX+b,Y} = \rho_{X,Y}$

Correlation coefficient



Examples of negatively (left) and positively (right) correlated r.v.s, if X and Y are uniformly distributed over the ellipses

- **Fact:** $|\rho_{X,Y}| \leq 1$ with equality iff $X - \mathbb{E}[X]$ is a *linear* function of $Y - \mathbb{E}[Y]$ (a corollary of the Cauchy-Schwarz inequality)

Cauchy-Schwarz inequality

For any two random variables X and Y, we have

$$|\,\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]\,\mathbb{E}[Y^2]}$$

(Check why this immediately establishes $|\rho_{X,Y}| \leq 1$)

Proof: For any value c, one has

$$0 \leq \mathbb{E}\left[(X - cY)^2\right] = \mathbb{E}\left[X^2\right] - 2c\mathbb{E}\left[XY\right] + c^2\mathbb{E}\left[Y^2\right].$$

Taking $c = \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}$ and substituting, we have

$$\begin{aligned} 0 &\leq \mathbb{E}\left[X^2\right] - \frac{2\left(\mathbb{E}[XY]\right)^2}{\mathbb{E}[Y^2]} + \frac{\left(\mathbb{E}[XY]\right)^2}{\mathbb{E}[Y^2]} \\ &= \mathbb{E}\left[X^2\right] - \frac{\left(\mathbb{E}[XY]\right)^2}{\mathbb{E}[Y^2]}, \end{aligned}$$

which immediately gives

$$\left(\mathbb{E}[XY]\right)^2 \leq \mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right].$$

Conditional expectation

- We know that $f_{X|Y}(x|y)$ is a pdf for X (function of y), so we can define the expectation of any function $g(X,Y)$ w.r.t. $f_{X|Y}(x|y)$ as

$$\mathbb{E}[g(X,Y) \mid Y=y] = \int g(x,y) f_{X|Y}(x|y) \mathrm{d}x$$

- If $g(X,Y) = X$, then the conditional expectation of X given $Y=y$ is

$$\mathbb{E}(X|Y=y) = \int x f_{X|Y}(x|y) \mathrm{d}x$$

Conditional expectation

- We define the *conditional expectation* of $g(X, Y)$ given Y as the random variable $\mathbb{E}[g[X, Y] \mid Y]$, which is a function of the random variable Y. This r.v. $\mathbb{E}[g[X, Y] \mid Y]$ takes the value of $\mathbb{E}[g[X, Y] \mid Y = y]$ when $Y = y$
- In particular, $\mathbb{E}[X \mid Y]$ is the conditional expectation of X given Y, a r.v. that is a function of Y. As we will see later, this forms an *estimator* of X given Y
- *Example:* consider a biased coin, whose probability of heads, denoted by Y, is itself random. Toss the coin n times and let X be # heads obtained. Then for any $y \in [0, 1]$, we have $\mathbb{E}[X \mid Y = y] = ny$, so $\mathbb{E}[X \mid Y] = nY$

Conditional expectation

Since $\mathbb{E}[g(X,Y) \mid Y]$ is a random variable, it has an expectation $\mathbb{E}\left[\,\mathbb{E}[g(X,Y) \mid Y]\right]$ of its own

- **Law of iterated expectation**

$$\mathbb{E}[g(X,Y)] \;=\; \mathbb{E}\left[\,\mathbb{E}[g(X,Y) \mid Y]\right]$$

- For any function $g(\cdot)$,

$$\mathbb{E}[Xg(Y) \mid Y] = g(Y)\,\mathbb{E}[X \mid Y]$$

  This follows since given the value of Y, $g(Y)$ is a constant and can be pulled outside

Conditional expectation

Proof of law of iterated expectation (for the case when $g(X,Y)=X$)

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[X \mid Y\right]\right] \\
&\quad = \int_{-\infty}^{\infty} \mathbb{E}\left[X \mid Y = y\right] f_Y(y) \mathrm{d}y \\
&\quad = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x f_{X|Y}(x \mid y) \mathrm{d}x \right) f_Y(y) \mathrm{d}y \\
&\quad = \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} f_{X|Y}(x \mid y) f_Y(y) \mathrm{d}y \right) \mathrm{d}x \quad \text{(switch order of integral)} \\
&\quad = \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} f_{X,Y}(x, y) \mathrm{d}y \right) \mathrm{d}x \\
&\quad = \int_{-\infty}^{\infty} x f_X(x) \mathrm{d}x = \mathbb{E}[X]
\end{aligned}
$$

Conditional variance

- Define the *conditional variance* of X given $Y = y$ to be the variance of X using $f_{X|Y}(x \mid y)$, i.e.

$$\begin{aligned}\mathsf{Var}(X \mid Y = y) &= \mathbb{E}\left[(X - \mathbb{E}[X \mid Y = y])^2 \mid Y = y\right] \\ &= \mathbb{E}[X^2 \mid Y = y] - [\mathbb{E}[X \mid Y = y]]^2\end{aligned}$$

- The r.v. $\mathsf{Var}(X \mid Y)$ is simply a function of Y that takes on the value $\mathsf{Var}(X \mid Y = y)$ when $Y = y$.
- *Example:* consider a biased coin, whose probability of heads, Y, is random. Toss the coin n times and let X be # heads obtained. Then for any $y \in [0, 1]$, we have $\mathsf{Var}(X \mid Y = y) = ny(1 - y)$, so $\mathsf{Var}(X \mid Y) = nY(1 - Y)$

Law of conditional variances

$$\mathsf{Var}(X) = \mathbb{E}\left[\mathsf{Var}(X|Y)\right] + \mathsf{Var}\left(\mathbb{E}[X|Y]\right)$$

Proof: The expected value of the r.v. $\mathsf{Var}(X \mid Y)$ is

$$\begin{aligned}
\mathbb{E}\left[\mathsf{Var}(X \mid Y)\right] &= \mathbb{E}\left[\mathbb{E}[X^2 \mid Y] - (\mathbb{E}[X \mid Y])^2\right] \\
&= \mathbb{E}[X^2] - \mathbb{E}\left[(\mathbb{E}[X \mid Y])^2\right] \quad \text{(by law of iterated expectation)}
\end{aligned}$$

Since $\mathbb{E}[X|Y]$ is a r.v., it has a variance

$$\begin{aligned}
\mathsf{Var}(\mathbb{E}[X|Y]) &= \mathbb{E}\left[\ \left(\mathbb{E}[X|Y] - \mathbb{E}[\mathbb{E}[X|Y]]\right)^2\ \right] \\
&= \mathbb{E}\left[(\mathbb{E}[X|Y])^2\right] - \left(\ \mathbb{E}[\mathbb{E}[X|Y]]\ \right)^2 \\
&= \mathbb{E}\left[(\mathbb{E}[X|Y])^2\right] - (\mathbb{E}[X])^2 \quad \text{(by law of iterated expectation)}
\end{aligned}$$

Add the above expressions and use $\mathsf{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ to complete the proof.

Law of conditional variances

Example: Consider n independent tosses of a biased coin whose probability of heads, Y, obeys $Y \sim \mathsf{Unif}(0,1)$. Calculate the variance of X.

Solution: Recall that $\mathbb{E}[X \mid Y] = nY$ and $\mathsf{Var}(X \mid Y) = nY(1-Y)$. Thus,

$$\begin{aligned}
\mathbb{E}\left[\mathsf{Var}(X \mid Y)\right] &= \mathbb{E}\left[nY(1-Y)\right] = n\left\{\mathbb{E}\left[Y\right] - \mathbb{E}\left[Y^2\right]\right\} \\
&= n\left\{\mathbb{E}\left[Y\right] - \left(\mathsf{Var}(Y) + \left(\mathbb{E}\left[Y\right]\right)^2\right)\right\} \\
&= n\left(\frac{1}{2} - \frac{1}{12} - \frac{1}{4}\right) = \frac{n}{6}
\end{aligned}$$

Furthermore,

$$\mathsf{Var}\left(\mathbb{E}\left[X \mid Y\right]\right) = \mathsf{Var}\left(nY\right) = n^2/12$$

By the law of total variance,

$$\mathsf{Var}\left(X\right) = \mathbb{E}\left[\mathsf{Var}(X \mid Y)\right] + \mathsf{Var}\left(\mathbb{E}\left[X \mid Y\right]\right) = \frac{n}{6} + \frac{n^2}{12}$$

Reference

[1] "*Introduction to probability (2nd Edition)*," D. Bertsekas, J. Tsitsiklis, *Athena Scientific*, 2008.

[2] "*Probability and Computing (2nd Edition)*," M. Mitzenmacher, E. Upfal, *Cambridge University Press*, 2017.

[3] "*Lecture notes: Statistical Signal Processing*," A. El Gamal.

[4] "*Statistical Inference (2nd Edition)*," G. Casella, R. Berger, *Cengage*, 2002.

[5] "*Stochastic Processes: Theory for Applications*," R. Gallager, *Cambridge University Press*, 2013.