Masters Bridge Program **Summer 2021**

Session 2: Random variables

Instructor: Bryan Liu *Contact: runjing_liu@berkeley.edu*

A random variable X is a variable that represents a randomly produced number. The set of possible values the variable X can take on is called the *range* of X.

Typically, we will use capital letters towards the end of the alphabet to denote a random variable, and lower case letters to denote constants. For a random variable X , we can ask, what is the probability that $X = x$? Or $X \leq x$? Or $a \leq X \leq b$? And so forth. We return to the same examples that opened the previous session for concreteness.

Example 2.1 (Rolling two dice)**.** Two dice are rolled, and we record the numbers showing on the top faces. Let S be the sum of the two numbers.

If the die are fair, then the probabilities of the values in the range of S are shown in Table 2.1

k	2	3	4	5	6	7	8	9	10	11	12
$\mathbb{P}(S = k)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Table 2.1: The probability distribution of S.

□

Example 2.2 (Particle decay)**.** Let T be the time it takes for a radioactive particle to decay. The probability that it takes at least t hours to decay is given by

$$\mathbb{P}(T \geq t) = \int_t^\infty \lambda \exp^{-\lambda s} \; ds,$$

where λ is a physical constant unique to the particle in study. □

Example 2.3. (Stock prices) Let $\Delta \log S$ be the change in log price of a stock after one unit time. The probability that the change lands between two numbers l and u is modeled as

$$\mathbb{P}(\Delta \log S \in [l, u]) = \int_l^u \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{1}{2\sigma^2}(x-\mu)^2} \; dx \tag{2.1}$$

where, μ is the *exponential growth rate of the stock* and σ^2 is the *volitility* of the stock. □

In comparing these examples with the motivating examples from yesterday, you should see that we have not really defined anything new. Yesterday, we introducted the notion of *events* in an outcome space; here, random variables determine events. Every event can be expressed as $X \in B$ for some subset B, where B is a subset of the range of the random variable X. The probability of this event is written $\mathbb{P}(X \in B)$; yesterday, we simply wrote $\mathbb{P}(B)$. By including the variable X, we are more explicitly showing where the randomness is coming from (randomness is coming from X).

2.1 Distributions

In the above examples, each random variable was assigned probabilities for the values it could take on. These probabilities are called the *the distribution* of the random variable.

Often these distributions have a name. The next exercise constructs our first named distribution, the *Binomial distribution*:

Exercise 1 (The binomial distribution)**.** Consider a coin which comes up heads with probabiilty p. Consider flipping this coin n times. Let X be the number of heads in the n trials. Show that

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{2.2}$$

When a random variable X has assignemnt probabilities given by (2.2), we say that X has a Binomial distribution with parameters (n, p). The two constants n and p specify this distribution.

Here are a few more examples of named distributions.

Example 2.4. (Poisson) A random variable X which takes values on the non-negative integers is *Poisson distributed* with parameter λ if

$$\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}. \tag{2.3}$$

The Poisson distribution is often used to model the number of occurences of a certain event in a window of time. We will see its appearance in the next session. □

Example 2.5. (Exponential) A random variable X which takes values on the positive real line is *exponentially* distributed with parameter λ if

$$\mathbb{P}(X \in [l, u]) = \int_t^u \lambda \exp^{-\lambda s} \; ds \tag{2.4}$$

for all real numbers $0 \leq l \leq u$.

The time until radioactive decay in Example 2.2 followed an exponential distribution. □

Example 2.6. (Guassian) A random variable X which takes values on the real line follows a *Gaussian* or *normal* distribution with mean μ and variance σ^2 if

$$\mathbb{P}(X \in [l, u]) = \int_l^u \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{1}{2\sigma^2}(x-\mu)^2} \; dx \tag{2.5}$$

for all real numbers $\leq l \leq u$.

Many natural phenomenon such as the distribution of human heights, IQ scores, and measurement error in physical experiments are well-approximated by the familiar bell-shaped curve that characterizes the normal distribution.

The normal distribution is ubiquitous in statistical inference as a result of the celebrated central limit theorem, which says, loosely, that averages of random variables will converge to a normal distribution. We will explore this further in the next session. □

Notation

When a random variable follows some distribution, we will use the $\sim$ symbol. For example, if X is normally distributed with parameters (μ, σ^2), we will write

$$X \sim \text{Normal}(\mu, \sigma^2).$$

We will often use F to denote a generic distribution, and will write $X \sim F$, read as "X follows distribution F", or "X is sampled from distribution F."

Discrete vs continuous random variables

For a *discrete* random variable X, the distribution of X is specified by the probabilities $\mathbb{P}(X = x)$, for all values x in the range of X can take. The function $f(x) := \mathbb{P}(X = x)$ is called the *probability mass function*, or *p.m.f.* The sum of two dice rolls, Binomial random variables, and Poisson random variables are discrete, with probability mass functions displayed in Table 2.1, Equation (2.2), and Equation (2.3), respectively.

For a *continuous* random variable X, such as exponential or normal random variables, the probability that $\mathbb{P}(X = x) = 0$. Instead, probabilities were defined on sets using integrals (2.4,2.5). The integrand is called a *probability density function* (or simply a "density", also abbreviated p.m.f). Loosely, we can interpret a density $f(x)$ as specifying the probability that X lies an infintessimal interval around x, that is

$$\mathbb{P}(X \in dx) = f(x)\ dx.$$

For generic sets $[a, b]$ in $\mathbb{R}$, we sum the infintessimal probabilities to obtain

$$\mathbb{P}(X \in [a, b]) = \int_a^b f(x)\ dx$$

In later sessions, we will have more exercises on manipulating probability density functions. Our current definition will suffice for the concepts introduced below.

2.1.1 Conditional distributions

Last session, we defined conditional probabilities and independence for events. These definitions are similarly defined in the context of random variables.

- **Conditional distributions:** [1] Let X and Y be two random variables. Suppose we know that $Y \in B$, B being a subset of the range of Y. Then the *conditional distribution* of X given

[1]We will come back to this formula specifically for continuous random variables ... in particular we need to be careful if we want to condition on the event $Y = y$, which has probabiilty 0 if Y is continuous.

$Y \in B$ is

$$\mathbb{P}(X \in A | Y \in B) = \frac{\mathbb{P}(X \in A \text{ and } Y \in B)}{\mathbb{P}(Y \in B)}.$$

- **Independence:** Random variables X and Y are independent if

$$\mathbb{P}(X \in A | Y \in B) = \mathbb{P}(X \in A)$$

for all sets A and B.

Intuitively, X and Y are independent if the values of X are unaffected by the values of Y.

Example 2.7 (Conditional distributions and independence)**.** Suppose I roll two die. Let X_1 and X_2 be the result of first and second dice, respectively. X_1 and X_2 are independent random variables, because conceivably, the value of one dice roll does not affect the probability distribution of the other dice roll: each dice has probability $1/6$ on numbers $\{1, ..., 6\}$ regardless of the outcome of the other dice.

Let $S = X_1 + X_2$ be the sum of two dice rolls. X_1 and S are not independent. For example, given that $X_1 = 3$, then S has the conditional distribution displayed in Table 2.2 (compare with Table 2.1).

k	2	3	4	5	6	7	8	9	10	11	12
$\mathbb{P}(S = k)$	0	0	1/6	1/6	1/6	1/6	1/6	1/6	0	0	0

Table 2.2: Conditional distribution of S given $X_1 = 3$.

2.2 Expectation and variance

We now turn to computing scalar quantities which summarize random variables and their distributions. We start with

The *expectation* (or *expected value*) of a random variable X, denoted $\mathbb{E}(X)$, is defined as

$$\mathbb{E}(X) = \sum_x x\mathbb{P}(X = x)$$

if X is a discrete random variable.
If X is a continuous random variable with probability density function f, then

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) \, dx$$

Exercise 2.

(a) Let S be the sum of two dice rolls (see Example 2.1). Compute the expectation of S.

(b) Show that if X is Poisson distributed with parameter λ, then X has expectation λ.

(c) Confirm that if X is normally distributed with parameters (μ, σ^2), then the expectation of X is μ.

Next, we record some properties of expectations:

- **Linearity:** For scalars $a, b \in R$,
$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b.$$
- **Addition rule:** For two random variables X and Y,
$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$
- **Multiplication rule:** For two *independent* random variables X and Y,
$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

Notice that we do not require independence for addition, but we do require independence for multiplication.

Exercise 3. Let S_n be the sum of numbers obtained from n dice. Find $\mathbb{E}(S_n)$.

Exercise 4. Let X be Binomially distributed with parameters (n, p). Show that $\mathbb{E}(X) = np$.

Exercise 5 (Markov's inequality)**.** Let X be a non-negative random variable with expectation μ. Show that for any number $a \geq 0$

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a} \tag{2.6}$$

We use the definition of an expectation to broaden our scope of distributional summaries. Next, we consider the *variance* of a random variable.

Let $\mu = \mathbb{E}(X)$. The *variance* of a random variable X, abbreviated Var(X), is defined as the expected squared deviation of X from μ,

$$\text{Var}(X) = \mathbb{E}\Big((X - \mu)^2\Big).$$

Exercise 6 (Alternative formula for variance)**.** Show that Var$(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$.

Exercise 7.

(a) Let X be the number returned by a single dice roll. Compute the variance of X.

(b) Show that if X is Poisson distributed with parameter λ, then Var$(X) = \lambda$.

(c) Confirm that if X is normally distributed with parameters (μ, σ^2), then Var$(X) = \sigma^2$.

Next, we now record some properties of variances:

- **Scaling and shifting:** For scalars $a, b \in R$,

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

- **Addition rule:** For two *independent* random variables X and Y,

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

Exercise 8 (iid random variables)**.** Suppose $X_1, ..., X_n$ are random variables drawn independently from a common distribution F. Let μ and σ^2 be the mean and variance, respectively, of the distribution F. Let

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

be the average of the n random variables. Show that

(a) $\mathbb{E}(\bar{X}) = \mu$.

(b) $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$.

Exercise 9 (Chebychev's inequality)**.** Use Markov's inequality (2.6) to show that

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \frac{\text{Var}(X)}{a}.$$

Chebychev's inequality puts an upper bound on the *tail probability*, the probability that a random variable X is far from its mean.

Tail probabilities are of particular interest in hypotheis testing. Here, the random variable X would represent a test statistic; if the the deviation of the observed statistic from its mean occurs with small probability under the null hypothesis, then we would consider this evidence in favor or *rejecting* the null. Chebychev inequality is one way to show that the probability of deviations are small. This inequality is useful because it does not depend on distributional assumptions of X (such as normality); we only need to know (or be able to approximate) its mean and variance.

Exercise 10 (Empirical distribution)**.** Suppose I have collected n data points $x_1, ..., x_n$. The *empirical distrbution* is defined as

$$\mathbb{P}(X = x) = \frac{\#\{x_i = x\}}{n}.$$

Show that if X is drawn from the empirical distribution, then

(a) $\mathbb{E}(X) = \frac{1}{n} \sum_{i=1}^{n} x_i$, the mean of the n-data points.

(b) What is $\text{Var}(X)$?

In statistics, we usually view the data points $x_1, ..., x_n$ as coming from some data generating process which is unknown to us. Let F represent the unknown, true data generating distribution. We want to know things about F, for example its mean and variance. Ideally, we would like to compute $\mathbb{E}(X)$ and Var(X) under the assumption that X has distribution F. But, F is unknown. So what do statisticians do?

The *plug-in principle* says that we can try to replace the unknown distribution F with the empirical distribution $\hat{F}$ – "empirical" because $\hat{F}$ is constructed from observed data – and then compute quantities of interest, such as the mean and variance, using $\hat{F}$ in place of F. Theory[2] says that for large n, $\hat{F}$ should well-approximate F. The plug-in principle appears in many forms in statistical practice; the exercise with the mean and variance above is a simple example.

2.2.1 Predictions

We conclude this session with a discussion of risk-minimization for prediction problems. Suppose we have a random variable X coming from some distribution F. In machine learning applications, we would like to predict the value of this random variable. How should we make this prediction?

One possible framework is to define a *loss function* $L(x, b)$, which is the penalty I pay for guessing b if the true value of X is x. The goal is to choose b to minimize the expected loss, or *risk*, defined as

$$r(b) = \mathbb{E}(L(X, b)). \tag{2.7}$$

Exercise 11.

(a) Suppose our penalty is the squared error, $L(x, b) = (x - b)^2$. Show that $b = \mathbb{E}(X)$ minimizes (2.7).

(b) Suppose our penalty is the absolute error, $L(x, b) = |x - b|$. Show that the median of the distribution of X minimizes (2.7).

Exercise 12 (The newsvendor problem)**.** I am operating a newspaper stand, and I need to decide the number of newspapers I should stock every day. Let X be the number of newspapers I sell each day, which I model as a random quantity with distribution F. I buy newspapers at C-dollars per paper, and sell to customers for a price $P > C$. Overstocking results in wasted inventory, while understocking results in lost sales. By choosing to stock b items, my profit when x items are sold is

$$\text{Profit}(x, b) = P \min(b, x) - Cx. \tag{2.8}$$

Define the loss to be $L(x, b) = -\text{Profit}(x, b)$, so that minimizing L is equivalent to maximizing profit.

Show that for this profit model, the optimal b is given by the $(\frac{P-C}{P})$-th quantile of the distribution F.

[2] specifically, the Glivenko–Cantelli theorem

To minimize the risk in (2.7), the data generating distribution F is typically not known. The plug-in principle is in play once again, resulting in technique of *empirical risk minimization* in machine learning. Here, the expectation over F in (2.7) is replaced with the empirical distribution $\hat{F}$.

In the case of squared error loss, we saw that the optimal predictor is the expectation of F. The optimal predictor is unknown if X comes from an unknown distribution F. The plug in principle says that we could try to replace F with $\hat{F}$, the empirical distribuion. The resulting predictor is the expectation of $\hat{F}$, which by Exercise 10 we showed is the equal to the sample mean.