

Session 2: Random variables

Instructor: Bryan Liu

We introduce the notion of a random variable. We saw several examples already in the previous session: the sum of two dice roles, the time until radioactive decay, the change in the stock price.

Informally, a random variable X is a variable that can take a set of possible values, called the *range* of X , and the possible values in its range are assigned a probability.

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.

We computed yesterday that if the die are fair, the probability of rolling a 5 is

$$\mathbb{P}(S = 5) = \frac{4}{36} = \frac{1}{9}.$$

□

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 log change between two numbers l and u is modeled as

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

where, μ is the *exponential growth rate of the stock* and σ^2 is the *volatility* 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 introduced the notion of *events* in an outcome space; here, random variables X determine events. Every event can be written as “ $X \in B$ ” where B is a subset of the range of 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: we call these assignment probabilities *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 probability p . Consider flipping this coin n times. Let X be the number of heads in the n trials. Show that

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

When a random variable X has assignment 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

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

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

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

$$\mathbb{P}(X \in [l, u]) = \int_l^u \lambda \exp^{-\lambda s} ds \quad (2.3)$$

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

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

Example 2.6. (Gaussian) 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

$$\mathbb{P}(X \in [l, u]) = \int_l^u \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx \quad (2.4)$$

for all real numbers $l \leq u$.

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. \square

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 possible values x that the random variable X can take. The sum of two dice rolls, Binomial random variables, and Poisson random variables are discrete.

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 (2.3, 2.4). The integrand is called a *probability density function* (or simply a “density”). Loosely, we can interpret a density $f(x)$ as specifying the probability that X is in an infinitesimal interval around x , that is

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

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

Definitions for conditional probabilities and independence are similarly defined as before 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 $Y \in B$ 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 .

¹We 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 probability 0 if Y is continuous.

2.2 Expectation and variance

We now turn 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

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

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 $a \geq 0$, that

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a} \tag{2.5}$$

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

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

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

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

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

Next, we now record some properties of variances:

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

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

- **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, \dots, X_n are random variables drawn independently from a common distribution. Let μ and σ^2 be the mean and variance, respectively, of this distribution. 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 (2.5) to show that

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \frac{\text{Var}(X)}{a^2}.$$

In statistics, we are often interested in the *tail probability*, the probability that a random variable X is far from its mean. Chebychev's inequality puts an upper bound on how large this probability can be.

In hypothesis testing, this random variable X would represent a test statistic; if 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 of *rejecting* the null. Chebychev inequality is one way to show that the probability of deviations are small, which may allow us to reject a null hypothesis. Notice that this inequality does not depend on distributional assumptions of X (such as normality); we only need to know (or be able to estimate) its mean and variance.

Exercise 10 (Empirical distribution). Suppose I have collected n data points x_1, \dots, x_n . The *empirical distribution* is defined as the distribution that places mass $1/n$ on each data point. Let \hat{F} be the empirical distribution, so we say that a random variable X has distribution \hat{F} if

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

Show that if X has distribution \hat{F} , 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, \dots, 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 $\text{Var}(X)$ 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 computed from data – and then compute quantities of interest, such as the mean and variance, using \hat{F} in place of F . Theory² 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 loss-minimization for prediction problems. Suppose we have a random variable X coming from some distribution, call it F . In machine learning applications, we would like to predict the value of this random variable. What value should we pick?

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

$$r(b) = \mathbb{E}(L(X, b)). \quad (2.6)$$

Exercise 11.

- (a) Suppose our penalty is the squared error, $L(x, b) = (x - b)^2$. Show that $b = \mathbb{E}(X)$ minimizes (2.6).
- (b) Suppose our penalty is the absolute error, $L(x, b) = |x - b|$. Show that the median of the distribution of X minimizes (2.6).

Exercise 12 (The newsvendor problem). I am operating a newspaper stand, and I need to decide how many number of newspapers I should stock every day. Let X be the number of newspapers I sell each day, which is a random quantity with distribution F . I buy newspapers at C per item, 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

$$\text{Profit}(x, b) = P \min(b, x) - Cx. \quad (2.7)$$

Define my 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 $((P - C)/P)$ -th quantile of the distribution F .

To minimize the risk in (2.6), the data generating distribution F is typically not known. The plug-in principle is in play once again, resulting in the idea of *empirical risk minimization* in machine learning. Here, the expectation over F in (2.6) is replaced with the empirical distribution \hat{F} .

In the case of squared error loss, we saw that the optimal predictor was $\mathbb{E}(X)$. This 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 resulting predictor is $\mathbb{E}(X)$ under \hat{F} , which by Exercise 10 we showed is the equal to the sample mean.