STAT 515 fa 2023 Lec 4

Random variables

Karl Gregory

Random variables

Recall that a *statistical experiment* is a process which generates a single outcome, where:

1. There is more than one possible outcome.
2. It is known in advance what the possible outcomes are.
3. The outcome to be generated cannot be predicted with certainty.

Suppose we define X to be a numeric encoding of the outcome of a statistical experiment. We call X a *random variable.* A *random variable* X is a real number

1. which can take one of many possible values,
2. for which the possible values it can take are known in advance, and
3. whose value cannot be predicted with certainty.

Note that *random variables* and *statistical experiments* go hand-in-hand!

The set $\mathcal{X}$ of all values a random variable X can take is called the *support* of X.

Example. Flip two coins and let X be the number of heads. The sample space $\mathcal{S}$ of the statistical experiment and the support $\mathcal{X}$ of the random variable X are

$$\mathcal{S} = \{HH, HT, TH, TT\} \quad \text{and} \quad \mathcal{X} = \{0, 1, 2\}.$$

Example. Roll a die and let X be the roll. Then we have

$$\mathcal{S} = \{1,2,3,4,5,6\} \quad \text{and} \quad \mathcal{X} = \{1,2,3,4,5,6\}.$$

Example. Roll two dice and record the rolls, letting X be the sum of the rolls. We have

$$\mathcal{S} = \left\{ \begin{array}{cccccc} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{array} \right\}$$

and

$$\mathcal{X} = \{2,3,4,5,6,7,8,9,10,11,12\}.$$

Example. Fill your car with gas and let X be the miles driven until the next fill-up divided by the number of gallons needed to refill the tank. Then

$$\mathcal{S} = [0,\infty) \quad \text{and} \quad \mathcal{X} = [0,\infty).$$

Discrete random variables

A *discrete random variable* is a random variable for which the values in its support can be written down in a list. The list could have a finite or an infinite number of elements.

Example. Roll one die and let X be the roll. Then the support $\mathcal{X}$ of X is

$$\mathcal{X} = \{1,2,3,4,5,6\},$$

so X is a discrete random variable.

Example. Count the number of people of who jaywalk over Broad River Road on a given day and let this be X. Then the support $\mathcal{X}$ of X is

$$\mathcal{X} = \{0,1,2,\dots\}.$$

Continuous random variables

A *continuous random variable* is a random variable for which the support is an interval (or a collection of intervals), and thus the values X can take cannot be written down in a list.

Example. Fill your car with gas and let X be the miles driven until the next fill-up divided by the number of gallons needed to refill the tank. Then the support $\mathcal{X}$ of X is the interval

$$\mathcal{X} = [0, \infty).$$

Example. Let X be the amount of rainfall in the coming month. Then the support $\mathcal{X}$ of X is

$$\mathcal{X} = [0, \infty).$$

Encoding random variables for categorical data

Categorical data take values which have no direct numerical interpretation.

Nominal categorical data are categorical data for which which the values do not admit an ordering.

Example. Choose a USC student at random and write down his or her eye color.

Ordinal categorical data are categorical data for which the values admit an ordering.

Example. Choose a USC student at random and record his or her response to the question, "Would you rate your move-in experience at USC as poor, reasonable, good, or excellent?"

Since *random variables* are technically supposed to be numbers, we sometimes encode the values of categorical data numerically, as in the following example.

Example. [Random variables for encoding categorical data] Choose a USC student at random and write down his or her eye color. Then define three random variables:

$$X_1 = \begin{cases} 1 & \text{if brown} \\ 0 & \text{otherwise} \end{cases} \quad X_2 = \begin{cases} 1 & \text{if blue} \\ 0 & \text{otherwise} \end{cases} \quad X_3 = \begin{cases} 1 & \text{if green} \\ 0 & \text{otherwise} \end{cases}$$

Probability distributions of discrete random variables

A key feature of a random variable X is that its value cannot be predicted with certainty. This is why we call it random! However, if we know the *probability distribution* of the random variable, we can make probabilistic statements about the as-yet-unobserved value of X.

The *probability distribution* of a random variable X tells us which values X can take and assigns probabilities to these values (or to ranges of these values if X is continuous; we will cover this later).

Definition: Probability distribution of a discrete random variable

For a discrete random variable X which can take the values $x_1, x_2, x_3, \ldots$, the probability distribution assigns a probabilities $p_1, p_2, p_3, \ldots$, to the values $x_1, x_2, x_3, \ldots$ such that

1. Each probability must be between zero and one: $p_i \in [0, 1]$, and
2. The probabilities must sum to 1: $\sum_i p_i = 1$.

Discrete random variables with finite support

If the support $\mathcal{X}$ of X is finite (has a finite number of elements), then we can tabulate the probability distribution of X as in the following examples.

Example. Roll one die and let X be the roll. Then the probability distribution of X can be tabulated as

x	1	2	3	4	5	6
$P(X = x)$	1/6	1/6	1/6	1/6	1/6	1/6

.

We see that $p_i = 1/6$ for $i = 1, \ldots, 6$, so that

$$\sum_{i=1}^{6} p_i = P(X = 1) + P(X = 2) + \cdots + P(X = 6) = 1.$$

We often use a lower case x to denote a specific realized value of the random variable X. So the capital X represents the as-yet-unobserved random variable, and x simply represents a number. The table above tells us, for each possible value x of X, the probability that X will assume that value.

Example. Select a USC undergraduate student at random and let X equal 1 or 0 according to whether he or she comes from South Carolina or not. If the proportion of the undergraduate students at USC coming from South Carolina is 0.60, then the probability distribution of X is

x	0 (out-of-state)	1 (in-state)
$P(X = x)$	0.40	0.60

.

So X takes the values 0 and 1 with the probabilities $p_1 = 0.40$ and $p_2 = 0.60$, respectively. Note that $p_1 + p_2 = 1$.

Exercise. Flip a coin twice and let X be the number of heads. Write down the probability distribution of X.

Answer: The possible outcomes of flipping two coins are

$$\mathcal{S} = \{TT, TH, HT, HH\},$$

leading to the possible values

$$\mathcal{X} = \{0, 1, 2\}$$

for the random variable X. We have $X = 0$ for the outcome TT, $X = 1$ for TH, $X = 1$ for HT, and $X = 2$ for HH. Therefore we may write

x	0	1	2
$P(X = x)$	1/4	1/2	1/4

.

Discrete random variables with infinite support

We cannot fully tabulate the probability distribution of a random variable that may take on an infinite number of values, that is if the support $\mathcal{X}$ of X is infinite.

Example. Count the number of people of who jaywalk over Broad River Road on a given day and let this be X. Then we might begin tabulating the probability distribution as

x	0	1	2	3	$\cdots$
$P(X = x)$	0.05	0.07	0.10	0.12	$\cdots$

,

but we cannot finish tabulating the entire probability distribution. We could truncate the table at the value 4 and write

x	0	1	2	3	≥ 4
$P(X = x)$	0.05	0.07	0.10	0.12	.66

.

Note that the probabilities sum to 1.

Nor can we tabulate the probability distribution of a continuous random variable, as the values of continuous random variables cannot be listed. We will discuss probability distributions of continuous random variables later on.

Probabilities of events based on random variables

We can use the probability distribution of a random variable X to compute the probabilities of events based on X.

Exercise. Roll one die and let X be the roll. What is $P(X \geq 4)$?

Answer: Recall that the probability distribution of X can be tabulated as

x	1	2	3	4	5	6
$P(X = x)$	1/6	1/6	1/6	1/6	1/6	1/6

To compute $P(X \geq 4)$, we sum the probabilities corresponding to $x \geq 4$. That is $P(X \geq 4) = P(X = 4) + P(X = 5) + P(X = 6) = 1/2$.

Exercise. Roll two dice and let X be the sum of the rolls. What is $P(X = 4)$?

Answer: We must recall the experiment in which we roll two dice and record the pair of rolls as (roll 1, roll 2). The possible outcomes are

$$\mathcal{S} = \left\{ \begin{array}{cccccc} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{array} \right\},$$

leading to the possible values

$$\mathcal{X} = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

for the random variable X. The outcomes in $\mathcal{S}$ are mutually exclusive and each occurs with probability 1/36. Each of the three outcomes $(1, 3)$, $(2, 2)$, and $(3, 1)$ lead to $X = 4$. So $P(X = 4)$ is the sum

$$P((1, 3)) + P((2, 2)) + P((3, 1)) = 1/36 + 1/36 + 1/36 = 1/12.$$

Expected value of a discrete random variable

To understand what is meant by the *expected value* of a random variable, we need to imagine repeating our statistical experiment over and over again. Imagine repeating a statistical experiment over and over again and then taking the mean of all the X values we have observed. The expected value of X is the value which we believe the mean of all the outcomes will approach as we repeat the experiment more and more times.

Exercise. Suppose we flip a coin and let $X = 1$ if heads comes up and $X = 0$ if tails comes up. What is the expected value of X?

Answer: If we flip the coin many many times, we would expect the sequence of X values to look something like

$$0100110111100100111011010111110010110010\ldots$$

now, if we took the mean of all the X values so far observed, we would get something close to 0.5. The mean of the above sequence is $23/40 = 0.575$. If we were to continue flipping the coin indefinitely, we *expect* that the mean of the ones and zeroes would eventually "converge" to 0.5, that is, it would continue getting closer and closer to 0.5. So the expected value of the random variable X is 0.5.

One may object, saying that the "expected value" of X cannot be equal to 0.5, because X can only be equal to 0 or 1, but when we say "expected value", we are referring to a mean value, which may not be one of the values X can take.

Definition: Expected value of a random variable

The *expected value* of a discrete random variable X which takes the values $x_1, x_2, x_3, \ldots$ with probabilities $p_1, p_2, p_3, \ldots$ is given by

$$\mathbb{E}(X) = x_1p_1 + x_2p_2 + x_3p_3 + \ldots$$

We often use the Greek letter μ to denote the expected value of a random variable. We often refer to the expected value of a random variable as its mean.

Example. Suppose we flip a coin and let $X = 1$ if heads comes up and $X = 0$ if tails comes up. Then

$$\mu_X = \mathbb{E}(X) = 0(1/2) + 1(1/2) = 1/2.$$

Example. Suppose we roll a die and let X be the roll. Then

$$\mu_X = 1(1/6) + 2(1/6) + 3(1/6) + 4(1/6) + 5(1/6) + 6(1/6) = 21/6 = 3.5.$$

We may think of the expected value of a random variable X as the balancing point of all possible values of X when they are weighted by their probabilities of occurrence. If they were sitting on a seesaw, where would the fulcrum need to be placed in order to balance them?

Exercise. Shoot a freethrow and let $X = 1$ if you make a basket and $X = 0$ if you miss. If your probability of making a basket is 0.7, what is the expected value of X?

Answer:

$$\mu_X = 0(0.3) + 1(0.7) = 0.7.$$

Variance of a discrete random variable

The *variance* of a random variable X with mean μ_X is the expected squared distance of X from its mean. We denote the variance of X by $\mathrm{Var}(X)$, and it is defined as

$$\mathrm{Var}(X) = \mathbb{E}(X - \mu_X)^2.$$

We often use the notation σ_X^2 to denote the variance of X.

Definition: Variance of a random variable

The variance $\mathsf{Var}(X)$ of a discrete random variable X having mean μ_X and which takes the values $x_1, x_2, x_3, \ldots$ with probabilities $p_1, p_2, p_3, \ldots$ is given by

$$\mathsf{Var}(X) = p_1(x_1 - \mu_X)^2 + p_2(x_2 - \mu_X)^2 + p_3(x_3 - \mu_X)^2 + \ldots$$

Example. Suppose we flip a coin and let $X = 1$ if heads comes up and $X = 0$ if tails comes up. Then

$$\begin{aligned}\sigma_X^2 &= (1/2)(0 - 1/2)^2 + (1/2)(1 - 1/2)^2 \\ &= (1/2)(1/4) + (1/2)(1/4) \\ &= 1/4.\end{aligned}$$

Example. Suppose we roll a die and let X be the roll. Then

$$\begin{aligned}\sigma_X^2 &= (1/6)(1 - 21/6)^2 + (1/6)(2 - 21/6)^2 + (1/6)(3 - 21/6)^2 \\ &\quad + (1/6)(4 - 21/6)^2 + (1/6)(5 - 21/6)^2 + (1/6)(6 - 21/6)^2 \\ &= 35/12 = 2.916667.\end{aligned}$$

The variance σ_X^2 of X gives us a description of how spread out the values of the random variable X tend to be.

Cumulative probabilities

We are sometimes interested in tabulating the distribution of a random variable X in terms of the values of $P(X \leq x)$ instead of (or in addition to) $P(X = x)$ for all the values x in $\mathcal{X}$. The probabilities $P(X \leq x)$ are called *cumulative probabilities.* Consider the following two examples:

Example. Roll one die and let X be the roll. Then we can tabulate the distribution of X as

x	1	2	3	4	5	6
$P(X = x)$	1/6	1/6	1/6	1/6	1/6	1/6 .
$P(X \leq x)$	1/6	2/6	3/6	4/6	5/6	6/6

The probabilities $P(X \leq x)$ for $x = 1, \ldots, 6$ are the cumulative probabilities. Note that we can compute either row of the table using the other row; we can get the probabilities $P(X = x)$, $x = 1, \ldots, 6$ from the cumulative probabilities by taking differences; for example, we have

$$P(X = 4) = P(X \leq 4) - P(X \leq 3) = 4/6 - 3/6 = 1/6.$$

To get the cumulative probabilities, we just sum up the probabilities $P(X = x)$.

Example. Flip a coin twice and let X be the number of heads. We can write

x	0	1	2
$P(X = x)$	1/4	1/2	1/4
$P(X \leq x)$	1/4	3/4	4/4

We see that the cumulative probabilities are computed by summing up the probabilities of each value of $x = 0, 1, 2$.