Statistics 213: Random Variables and the "Modeling of Random Events"

When data is collected on a quantitative variable (discrete or continuous) from some population of interest, exploratory data analysis was employed to understand how the variable behaved. How was the variable distributed? What central point did the values on this variable tend to fall about - what is its measure of center? *How* did the values on the variable of interest "fall about" this measure of center - that is, what was the *spread* in the distribution of the variable of interest?

As we have seen, often this preliminary stage of statistical inquiry involves the construction of a density-scale histogram. Recall a density-scale histogram has a total area of one. It is here where we see the motivation behind the construction of a such a histogram.

Some Terminology

Previously a random experiment was defined to be process producing one of many possible and unpredictable outcomes. Some examples of random experiments:

1. tossing a coin three times and observing the outcome.
2. tossing a fair coin in succession until the first head appears.
3. the amount of time it takes you to commute from home to the university.

Each of the random experiments given above have sample spaces, listed below:

1. $\mathcal{S} = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$ $n(\mathcal{S}) = 8$.
2. $\mathcal{S} = \{H, TH, TTH, TTTH, TTTTH, \cdots\}$ $n(\mathcal{S}) =?$
3. $\mathcal{S} =$ set of all real numbers: 0 to ∞, therefore $n(\mathcal{S}) = \infty$.

Examples one and two produced discrete sample spaces: Example 1 had a finite number of outcomes/elements, example 2 was said to be *countably infinite*. The random experiment in Example 3 produces a sample space that has an infinite number of elements - this will be of significance in a few lectures.

Up to now, we have been describing events produced by random experiments in terms of A, or perhaps B, something of that nature. In many instances, we may be interested in investigating the *distribution* of events produced by a random experiment. In order to do so - investigate the shape, measure of center, and measure of spread, we must take these events A, B, etc., and use some measurement scale to quantify these events. Such events are made into numbers through the introduction of a random variable.

A **random variable** is a means to quantify events. A random variable assigned a unique numerical value to an event produced by a random experiment.

Example 1: Consider an experiment that results in three successive tosses of a fair coin. One event that can be produced from this random process is observing two heads, no more nor less, on the three tosses of the coin. Clearly there are three ways in which this can occur: $\{(HHT, HTH, THH)\}$ and its probability is $P(2Heads) = 3/8 = 0.375$. Suppose a random variable[1], X, is introduced to represent the number of heads appearing on the three tosses of this coin. Instead of addressing the event "2 Heads", one could say the value of the random variable X is equal to 2, or $X = 2$. One can then say that the probability that $X = 2$, or $P(X = 2) = 0.375$.

Is this the only value that X can assume? Clearly the experiment of tossing a coin three times can produce values of $X = 0$, 1, 2, or 3. If $X = 0$, that implies no heads are observed after the experiment has finished.

[1]It is customary to reprsent random variables as capital letters.

Notice how the values of X vary - it takes on four different values (0, 1, 2, 3)? Also notice that the value X assumes is "unpredictable" in the sense that one cannot predict *what the value* of X will be? Hence the name "random variable."

Example 2: Three roommates play the following game: they take turns tossing a fair die until one of them tosses a 'six'. The first roommate to toss a six wins the game, with the other two roommates paying the monthly rent. That is, the other two roommates split the winning roommates share of the monthly rent. Letting a random variable Y represent the number of tosses it takes for the first six to be observed. Suppose the sequence of tosses goes like this: Roommate A, Roommate B, then Roommate C tosses the die. What is the probability that Roommate A looses on his/her second toss, or Roommate A looses the game on his/her second try?

Here $Y = 1, 2, 3, \cdots$. $Y = 4$ in simply a way of quantifying the event "the first six appears on the fourth toss", and the probability that Y is equal to 4 is:

$$\begin{aligned} P(Y=4) &= P(6_1^c \cap 6_2^c \cap 6_3^c \cap 6_4) \\ &= P(6_1^c) * P(6_2^c) * P(6_3^c) * P(6_4) \qquad \text{(Independence)} \\ &= \left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{1}{6}\right) \\ &= \left(\frac{125}{1296}\right) \\ &= 0.09645 \end{aligned}$$

Finally, of one were to define a random variable Z as the amount of time, in minutes, it takes you to commute from home to the university, Z could assume an infinite number of possible values. If it took you 20 minutes and 30 seconds to commute to the university today, then Z would assume a value of 20.5, or $Z = 20.5$. Another day it may take you 15 minutes and 10 seconds, or $Z = 15.1667$.

Notationally, the value that the random variable X assumes is x. The probability that the random variable is equal to a certain value x_1 is represented by $P(X = x_1)$.

Probability Distribution of a Random Variable X

Since random variables assume numbers, one can construct a probability distribution of the random variable to understand how the random variable behaves.

The **probability distribution** of a random variable X is a listing, presented in tabular form, which provides each value of the random variable along its probability of occurrance. A probability distribution takes on the form:

X	x_1	x_2	x_3	$\cdots$	x_k
$P(X = x)$	$P(X = x_1)$	$P(X = x_2$	$P(X = x_3)$	$\cdots$	$P(X = x_k)$

There are two conditions in order for the "table" above to be considered a probability distribution:

1. $0 < P(X = x_i) < 1$, for $i = 1, 2, \cdots, k$.
2. $\sum_{i=1}^{k} P(X = x_i) = 1$

Example 1: X was defined to represent the number of heads observed after a fair coin has been tossed three times in succession. The probability distribution of X is given as:

X	0	1	2	3
$P(X = x)$	$1/8 = 0.125$	$3/8 = 0.375$	$3/8 = 0.375$	$1/8 = 0.125$

Notice how

$$\begin{aligned} \text{Condition 2}: &= \sum_{i=1}^{k} P(X = x_i) \\ &= \left(\frac{1}{8}\right) + \left(\frac{3}{8}\right) + \left(\frac{3}{8}\right) + \left(\frac{1}{8}\right) \\ &= 1 \end{aligned}$$

and the probability of each value of X is between 0 and 1?

From the above probability distribution for **Example 1**, one can easily find the probability that one observes at least two heads. This event expressed in terms of the random variable X, is $X \geq 2$. This probability of at most two heads is then:

$$\begin{aligned} P(X \geq 2) &= P(X = 2) + P(X = 3) \\ &= \frac{3}{8} + \frac{1}{8} \\ &= \frac{4}{8} = 0.50 \end{aligned}$$

Example 2: Y was defined to represent the number of times the fair coin was tossed until the first head appears. The probability distribution of Y is given below.

Y	1	2	3	$\cdots$
$P(Y = y)$	$1/6 = 0.1667$	$5/36 = 0.1389$	$25/216 = 0.1157$	$\cdots$

Example 3: A box contains five tickets numbered 1 through 5. Three tickets are selected from the box at random.

(a) A random variable X is defined as the smallest number sampled. Find the probability distribution of the random variable X. If the smallest number chosen is at most 2 ($X \geq 2$), what is the probability that the smallest number is a one ($X = 1$)?

(b) A different random variable is defined to represent the *sum* of the thre numbers selected, or $Y =$ sum of the three tickets selected. Find the probability distribution of the random variable Y. From this probability distribution, find the probability that the sum of the three tickets selected is (i) between 6 and 8 (inclusive) (ii) less than 7.

Measures of Center and Measures of Spread

The random variable X has a distribution of values. All its possible values are distributed around some center of gravity. This "balancing point" is referred to as the mean of the distribution, or its expected value. This measure of central tendency is nothing more than a weighted-mean, calculated by using the expression

$$\mu_X = E(X) = \sum_{all\ x} xP(X = x).$$

Example 4(a): From the random variables given in Example 3, the expected value of the random variable X is given by

$$\begin{aligned} E(X) &= \sum_{all\ x} xP(X = x) \\ &= 1 * P(X = 1) + 2 * P(X = 2) + 3 * P(X = 3) \\ &= 1 * \left(\frac{6}{10}\right) + 2 * \left(\frac{3}{10}\right) + 3 * \left(\frac{1}{10}\right) \\ &= 1.5 \end{aligned}$$

The center of the probability distribution of X is 1.5 - this represent the "center of gravity" around which the values of the random variable X are dispersed.

Example 4(b): The expected value of Y is given by:

$$\begin{aligned} E(Y) &= \sum_{all\ y} yP(Y = y) \\ &= 6 * P(Y = 6) + 7 * P(Y = 7) + 8 * P(Y = 8) + \cdots 11 * P(Y = 11) + 12P(Y = 12)) \\ &= 6 * \left(\frac{1}{10}\right) + 7 * \left(\frac{1}{10}\right) + 8 * \left(\frac{2}{10}\right) + \cdots + 11 * \left(\frac{1}{10}\right) + 12 * \left(\frac{1}{10}\right) \\ &= 9 \end{aligned}$$

One would "expect" the sum of the three tickets randomly chosen to be 9.

In addition, there exists a measure of dispersion which represents how the individual values of the random variable X are distributed around the mean. This measure of spread is called the standard deviation and it measures the "typical distance" each possible value of the random variable lies away from the mean.

The standard deviation of a random variable X, or σ_X, is the square root of the average squared-distance each value lies away from the mean (a quantity deemed the variance). In general,

$$\begin{aligned} \sigma_X^2 = Var(X) &= \sum_{all\ x} (x - \mu_x)^2 P(X = x) \\ &= E(X^2) - \mu_x^2 \\ \sigma_X &= \sqrt{\sigma_X^2} = \sqrt{Var(X)}. \end{aligned}$$

Example 5(a): Find the standard deviation in the probability distribution of the random variable X.
Solution:

$$\begin{aligned} Var(X) &= E(X^2) - [E(X)]^2 \\ &= \sum_{all\ x} x^2 P(X = x) - [E(X)]^2 \\ &= \left[1^2 * P(X = 1) + 2^2 * P(X = 2) + 3^2 * P(X = 3)\right] - [E(X)]^2 \\ &= \left[1^2 * \left(\frac{6}{10}\right) + 2^2 * \left(\frac{3}{10}\right) + 3^2 * \left(\frac{1}{10}\right)\right] - [1.5]^2 \end{aligned}$$

$$
\begin{aligned}
&= \left(\frac{27}{10}\right) - 1.5^2 \\
&= 0.45
\end{aligned}
$$

The standard deviation of X is then:

$$\sigma_X = SD(X) = \sqrt{Var(X)} = \sqrt{0.45} = 0.6708$$

Example 5(b): Find the standard deviation of the random variable Y, or σ_Y.

$$
\begin{aligned}
Var(Y) &= E(Y^2) - [E(Y)]^2 \\
&= \sum_{all\ y} y^2 P(Y = y) - [E(Y)]^2 \\
&= \left[6^2 * P(Y = 6) + 7^2 * P(Y = 7) + 8^2 * P(Y = 8) + \cdots + 12^2 * P(Y = 12)\right] - [E(Y)]^2 \\
&= \left[36 * \left(\frac{1}{10}\right) + 49 * \left(\frac{1}{10}\right) + 64 * \left(\frac{2}{10}\right) + \cdots + 144 * \left(\frac{1}{10}\right)\right] - [9]^2 \\
&= \left(\frac{840}{10}\right) - 9^2 \\
&= 3
\end{aligned}
$$

The standard deviation of Y is then:

$$\sigma_Y = SD(Y) = \sqrt{Var(Y)} = \sqrt{3} = 1.732$$

Example 6: A professor gives his class 10 questions to complete. From this pool of questions, the professor is to randomly select 3 to place on a quiz. Ken, a student in the class, decides to complete only 5 of the 10 questions assigned. Assume that Ken knows how to do each of the five questions he does complete. Defining the random variable X as the number of questions Ken correctly solves on the quiz

(a) find the probability distribution of X.

(b) How many questions can Ken expect to correctly solve on the quiz?

(c) Find the standard deviation in X.

Example 7: Paving Company A submits bids on two separate projects - Project One and Project Two. The manager believes the probability they will win the bid on Project One to be 70%. If they win the bid on the first project, the manager believes they will secure Project Two with probability 0.80. If Project One is not secured, the manager believes the probability Project Two will be secured is 0.4.

If X represents the number of projects Paving Company A secures, how many projects should the manager expect to secure? Provide a standard deviation as well.

The Bernoulli Trial, or a '0 - 1' Trial

Many times in life, you will find yourself in a situation which can be classified as a "0 - 1" trial. Whether it is making an offer to purchase a home, getting married, purchasing a lottery ticket, your grade in a university course, should you take that job in Toronto - all of these scenarios can be categoized into a "0 - 1" trial.

Firstly, a "0 - 1" trial should be well-defined. What is a "0 - 1" trial? Simply stated, a "0 - 1" trial is an experiment which yields one of two possible results, each result occurring with a given probability. Often called a Bernoulli trial, the two possible results are categorized as a "success" or "failure". These outcomes are mutually exclusive and complementary of one another. For example, if you make a offer to purchase a house, you sit down with a realtor and put an offer to purchase on paper. This in essence is an experiment. There are many possible outcomes to this situation. However, the variation in outcomes can be simplified into a choice of two possible outcomes: the vender (or seller) of the home you are offering to purchase may accept your offer, or decline your offer and subsequently counter your offer. After this point, the trial has ended. The home-owner has a choice to accept your offer (which you can classify as a "success") or can reject your offer (or a "failure" has occurred). Each of these two outcomes does have a probability of happening. The probability that the vendor will accept your offer is a percentage between 0 and 1, but is unknown. Alternatively, the probability that the home-owner will not accept your offer (and decide to counter-offer or outright reject the offer) is an unknown percentage between 0 and 1. This mysterious probability is the difference between 1 and the probability that the offer will be accepted. Mathematically, a "0 - 1" trial is expressed as the random variable Y, where

$$Y = \begin{cases} 0 & \text{with probability } (1-p), \quad \text{denoting the outcome of a failure} \\ 1 & \text{with probability } p, \quad \text{denotes a successful outcome.} \end{cases}$$

The expected value and variance of this Bernoulli random variable is given below:

$$\begin{aligned} E(Y) &= \sum_{y=0}^{y=1} yP(Y=y) \\ &= 0 * P(Y=0) + 1 * P(Y=1) \\ &= 0 * (1-p) + 1 * (p) \\ E(Y) &= p \end{aligned}$$

$$\begin{aligned} Var(Y) &= E(Y^2) - [E(Y)]^2 \\ &= \sum_{y=0}^{y=1} y^2 P(Y=y) - [E(Y)]^2 \\ &= [0^2 * P(Y=0) + 1^2 * P(Y=1)] - (p)^2 \\ &= [0^2 * (1-p) + 1^2 * p] - p^2 \\ &= p - p^2 \\ &= p(1-p) \end{aligned}$$

Therefore, the standard deviation of a Bernoulli random variable Y is

$$\sigma_Y = SD(Y) = \sqrt{Var(Y)} = \sqrt{p(1-p)}$$

Example 8: Consider the spin of a roulette wheel. A gambler places a bet that the little white ball will land on a 'black' pocket. Defining a random variable X as the outcome from this wager, one can consider this as a Bernoulli trial: a success will occur if the gambler wins the bet - the white ball lands in the 'black' pocket, a failure as the white ball does not land in a 'black' pocket. In this instance, we can give the probability distribution of the random variable X, where $p = P(X = 1) = \frac{18}{38}$:

X	0	1
$P(X = x)$	0.5263	0.4737

We can see that the expected value of X is $\frac{18}{38} = 0.4737$, $(E(X) = p = \frac{18}{38} = 0.4737)$ and the standard deviation of X is $SD(X) = \sqrt{(1 - \frac{18}{38})(\frac{18}{38})} = 0.4993 \approx 0.50$.