Stat 230 Introductory Statistics Course Introduction & Descriptive Statistics

Instructor: Dr. Yas Yamin

University of British Columbia Okanagan
yas.yamin@ubc.ca

Examples —- Unequally likely Outcomes Random Process

- The numbers of children born to a pregnant woman.
- The numbers of people got lightening hit.

There are many processes that is unequally likely outcomes.
We will introduce more probability distribution.

Random Variable

Before continuing with probability distributions, we need the idea of a random variable.

- Often the outcomes of a sample space could be numbers on the real line or not. e.x. rolling a die or tossing a quarter.
- The real number is needed in mathematical calculation

Definition 3.1 (random variable)

A random variable (RV) X is a function that maps the sample space Ω to the real line; that is, for each element $\omega \in \Omega$, $X(\omega) \in \mathbb{R}$.

- We denote random variables by capital letters (typically from the end of the alphabet)
- The real number associated with outcome ω is a realization (or value) of $X(\omega)$ and is denoted by the corresponding lower case letter x.
- Herein we will use X in place of $X(\omega)$

Random Variables

Example 3.1 *If the experiment is tossing a coin 3 times then we could have a random variable X representing the number of heads obtained. Then we ask questions like: what is $P(X = 2)$?*
$\Omega = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}$

Hence, we have defined a *function* over the sample space Ω. This is an example of what is called a **discrete random variable**.

Random Variables

There two kinds of random variables

1. Discrete randome variable
2. Continuous random variable

We are going to learn discrete random variable first in this chapter then to learn continuous random variable in the next charpter.

Discrete Random Variable

Definition 3.2

A **discrete random variable** is an rv whose possible values either constitute a finite set or else can be listed in an infinite sequence in which there is a first element, a second element, and so on.

Definition 3.3

If a discrete random variable X can take values $x_1, x_2, \ldots, x_n$ with probabilities $p_1, p_2, \ldots, p_n$ such that

1. $p_1 + p_2 + \cdots + p_n = 1$, and
2. $p_i \geq 0$ for all i,

then this defines a **discrete probability distribution** for X.

Discrete Probability Distribution

Example 3.2 *Toss two fair coins. Let X represent the number of tails obtained. What is the discrete probability distribution for X.*

Example 3.3 *Suppose we roll a pair of fair dice, one green and one red. If we are interested in the number of dots facing up on two dice, we can define the sample space and present it by an array;*

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

Let X be a random variable to represent the sum of two dies.

The realized values of X for the entire sample space would be:

$$x = \left\{\begin{matrix} 2 & 3 & 4 & 5 & 6 & 7 \\ 3 & 4 & 5 & 6 & 7 & 8 \\ 4 & 5 & 6 & 7 & 8 & 9 \\ 5 & 6 & 7 & 8 & 9 & 10 \\ 6 & 7 & 8 & 9 & 10 & 11 \\ 7 & 8 & 9 & 10 & 11 & 12 \end{matrix}\right\}$$

Hence the distribution can be written

x	2	3	4	5	6	7	8	9	10	11	12
$P(X = x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

We can often represent the probability distribution function (pdf) as a function in the more classical sense

Returning to Ex 3.3 we could represent the $P(X)$ by the function:

Definition 3.4 (probability distribution function (pdf))

If X is a discrete random variable, the function given by $f(x) = P(X = x)$ for each x within the range (or support) of X is called the probability distribution of X.

The probability distribution or pmf also can be written

$$p(x)=P(X=x)=\begin{cases}\frac{1}{36}, & x=2\\ \frac{2}{36}, & x=3\\ \frac{3}{36}, & x=4\\ \frac{4}{36}, & x=5\\ \frac{5}{36}, & x=6\\ \frac{6}{36}, & x=7\\ \frac{5}{36}, & x=8\\ \frac{4}{36}, & x=9\\ \frac{3}{36}, & x=10\\ \frac{2}{36}, & x=11\\ \frac{1}{36}, & x=12\end{cases}$$

$$p(x)=\begin{cases}\frac{6-|x-7|}{36}, & \text{if} x = 2, 3, \ldots, 12\\ 0, & \text{otherwise}\end{cases}$$

Cumulative distribution function (cdf)

Definition 3.5

The **cumulative distribution function** (cdf) that is denoted as F(x) of a discrete rv X with pmf p(x) is defined for every number x by

$$F(x) = P(X \leq x) = \sum_{y:y \leq x} f(y), \quad \text{for} -\infty < x < \infty$$

F(x) is the probability that the observed value of X will be at most x. In the other words $F(x)$ accumulates the probabilities of $X \leq x$

Returning to Example 3.3

$F(4) =$

Cumulative distribution function (cdf)

Theorem 3.6

The values $F(x)$ of a RV X satisfy the conditions

(1) $\lim_{x \to -\infty} F(x) = 0$,

(2) $\lim_{x \to \infty} F(x) = 1$.

(3) *If $a < b$, then $F(a) \leq F(b)$ for any real numbers a and b.*

Returning to Example 3.3

$F(12) =$

$3 < 4 \to F(3) \leq F(4)$

Cumulative distribution function (cdf)

$$F(x) = F(X \leq x) = \begin{cases} 0, & x < 2 \\ \frac{1}{36}, & 2 \leq x < 3 \\ \frac{3}{36}, & 3 \leq x < 4 \\ \frac{6}{36}, & 4 \leq x < 5 \\ \frac{10}{36}, & 5 \leq x < 6 \\ \frac{15}{36}, & 6 \leq x < 7 \\ \frac{21}{36}, & 7 \leq x < 8 \\ \frac{26}{36}, & 8 \leq x < 9 \\ \frac{30}{36}, & 9 \leq x < 10 \\ \frac{33}{36}, & 10 \leq x < 11 \\ \frac{35}{36}, & 11 \leq x < 12 \\ \frac{36}{36}, & \mathrm{x} \geq 12 \end{cases}$$

Cumulative distribution function (cdf)

Theorem 3.7

For any two numbers a and b with $a \leq b$,

$$P(a \leq X \leq b) = F(b) - F(a^-)$$

where $F(a^-)$ represents the maximum of $F(x)$ values to the left of a. Equivalently, if a is the limit of values of x approaching from the left, then $F(a^-)$ is the limiting value of F(x). In particular, if the only possible values are integers and if a and b are integers, then

$$\begin{aligned} P(a \leq X \leq b) &= P(X = a \text{ or } a+1 \text{ or } \cdots \text{ or } b) \\ &= F(b) - F(a-1) \end{aligned}$$

Taking $a = b$ yields $P(X = a) = F(a) - F(a-1)$ in this case.

Cumulative distribution function (cdf)

Example 3.4 *Back to Example 3.3*

Find out F(3), F(8).

Find out the probability of getting sum result between 4 and 8, that is $4 \leq x \leq 8$

Special Discrete Probability Distributions

Definition 3.8

A **Bernoulli Trial** is an experiment with only two possible outcomes: success and failure. If $X \sim \text{Bernoulli}(p)$, then

- $P(\text{success}) = P(X = 1) = p$, and
- $P(\text{failure}) = P(X = 0) = 1 - p$.

- For example, tossing a coin and looking for a head is a Bernoulli trial — getting a head is success and getting a tail is failure.
- Rolling a die and looking for a 6 is also a Bernoulli trial.

Special Discrete Probability Distributions

Definition 3.9 (Bernoulli distribution)

A random variable X has a *Bernoulli distribution*, denoted by $X \sim \text{Bernoulli}(p)$, if and only if it has probability distribution function as

$$f(x) = p^x(1-p)^{1-x} \qquad \text{for } x = 0, 1$$

where $p = P(X = 1)$.

Example 3.5 *If $X = 1$ if a coin lands heads, and $X = 0$ if the coin lands tails, $X \sim \text{Bernoulli}(p = 0.5)$. What is the probability that the coin lands tails?*

Special Discrete Probability Distributions

Definition 3.10 (Binomial Distribution)

The **Binomial Probability Distribution** comes about as the result of n independent Bernoulli trials, each trial having success probability p. The probability of obtaining x successes in these n trials is given by

$$P(X = x) = \binom{n}{x} p^x (1-p)^{(n-x)}$$

for $x = 0, 1, 2, \dots, n$

Special Discrete Probability Distributions

- Looking at this formula from an intuitive viewpoint, it makes sense
  1. there are $\binom{n}{x}$ ways of getting x successes from n trials,
  2. $\binom{n}{x} = \frac{n!}{x!(n-x)!}$, where $n! = n(n-1)(n-2)\cdots 3(2)(1)$,
  3. p^x is the probability of success, x times and
  4. $(1-p)^{n-x}$ is the probability of failure, $n-x$ times.

Example 3.6 *A fair coin is tossed six times. What is the probability of obtaining exactly four heads?*

Looking at this formula from an intuitive viewpoint, it makes sense

1. there are $\binom{6}{4} = \frac{6!}{4!(6-4)!} = 15$ ways of getting 4 successes in 6 trials,
2. $\left(\frac{1}{2}\right)^4$ is the probability of success, 4 times and
3. $\left(\frac{1}{2}\right)^2$ is the probability of failure, 2 times.

Exercises I

Exercise 3.1 A fair die is rolled 10 times.

3.1.1 What is the probability that it shows the number 6 exactly 5 times?

3.1.2 What is the probability that it does not show the number 1 at all?

3.1.3 What is the probability that it shows the number 4 less than 2 times?

Exercise 3.2 A couple decide to keep having children until they have a boy, then they stop.

3.2.1 What is the probability that the couple have 5 children? (Geometric Distribution)

3.2.2 Eight couples take this approach. What is the probability that more than 2 of these couples have 5 children?

Sampling With Replacement

- Sampling such that each unit is replaced before the next sample is drawn is called **sampling with replacement**.
- The binomial distribution will be applicable in cases where we sample with replacement.
- We have already seen some examples, such as tossing a coin or rolling dice.
- What if we don't sample with replacement, such as in a lottery or a raffle?

Sampling Without Replacement

- Sampling such that each unit is <u>not</u> replaced before the next sample is drawn is called **sampling without replacement**.
- The binomial distribution will not be applicable.
- Consider a situation where we are looking for defects in units.

Special Discrete Distribution: The Hypergeometric Distribution

- Suppose that there are N items in total (N is finite) and we are interested in M of these (M 'successes').
- If we sample n of these items then the probability that x are 'successes' is given by

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \qquad \text{for } x = 0, 1, 2, \ldots, n$$

Example 3.7 *A lottery involves drawing 6 numbers from 1–49, without replacement. Each player chooses 6 numbers and prizes are awarded for matching 4, 5 or 6 of the numbers that are drawn. What is the probability of matching 4 numbers?*

Exercises II

Exercise 3.3 An athlete conceals two performance enhancing tablets in a bottle containing eight vitamin pills that look similar. If a drug surveillance scheme is in operation that involves randomly sampling three of these pills, what is the probability that cheating will be detected? Assume that the analysis of the tablets is not error prone.

Exercise 3.4 Eight cards are chosen at random from a well-shuffled pack. What is the probability of obtaining three spades, two hearts, two diamonds and one club?

The Poisson Distribution

The Poisson distribution is often used to model counts. Usually, the observation process is considered to be taking place over continuous intervals of time or space. For example:

- the number of accidents in an intersection during a given time interval;
- the number of customers that arrive to a service queue during a given time period;
- the number of the incoming calls per hour of a given phone line in day time;
- the number of whales in a circle of radius r.

The Poisson Distribution

X counts the number of events occurring in a fixed interval, given that these events occur randomly and independently with constant rate (λ)

Definition 3.11

Let X be a random variable that follows a **Poisson distribution** with mean λ. Then,

$$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}, \qquad x = 0, 1, 2, \ldots$$

- The Poisson distribution is very powerful, both in its own right and as an approximation to the binomial distribution.

Example 3.8 *Chocolates are manufactured so that the number of surface blemishes on any one chocolate is described by a Poisson distribution with mean* $\lambda = 0.05$. *What is the probability that a box of ten chocolates will contain two or more with surface blemishes?*

Example 3.9 *Customers arrive at an average rate of 3.7/hour. If the store opens at 8:00 what is the probability that there are at least two arrivals by 8:45?*

Exercises III

Exercise 3.5 Coliform bacteria are randomly distributed in river water at an average concentration of 1 per 25cc of water. What is the probability of finding more than two bacteria in a sample of 10cc of river water?

Exercise 3.6 Each morning, after opening her e-mail account, Lucy has to discard, on average, ten spam messages. If the number of spam messages may be described by a Poisson distribution, what is the probability that on any given morning Lucy will receive less than four spam messages?

Connections between the binomial and the Poisson

- The Poisson distribution is the limiting form of the binomial distribution for
  - the total number of trials being very large and
  - the probability of "success" being sufficiently small while the "successes" rate remains constant.
- That is, for $n \to \infty$, $p \to 0$, while $np = \lambda$ for a constant λ. If X is the number of "success", the limiting form of the binomial distribution is the Poisson distribution with $\lambda = np$.
- When can we use this approximation?
  - $n \geq 20$ and $p \leq 0.05$ or
  - $n \geq 100$ and $np \leq 10$

Example 3.10 *A heavy machinery manufacturer has 3849 large generators . If the probability that any one will fail during the given year is 1/1200, find the probability that fewer than 9 generators will fail during a given year.*

The exact answer, *i.e* the one using the binomial distribution would be calculated by . . .

`pbinom(8,3849, 1/1200)` (Calculated in R)

```
[1] 0.9942245
```

Definition 3.12 (expected value)

The *expected value* of a discrete random variable X is given by

$$\mathbb{E}[X] = \sum_x xP(X = x).$$

Expected Value

- The expected value is just the mean, or average of the population.
- The expected value of X is not necessarily
  - a value that you would expect X to take on
  - It may not be the most probable value
  - It may be an unlikely value.
  - It may even be an impossible value
    (e.g. the expected number of children born per woman in Canada is 1.58 children/woman).
- The expected value can be interpreted as a weighted average of all possible values

Expected Value

Theorem 3.13

Let a, b and c be constants, then:

$$\mathbb{E}[a] = a \tag{1}$$

$$\mathbb{E}[aX] = a\mathbb{E}[X] \tag{2}$$

$$\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c \tag{3}$$

Definition 3.14

Note that, if $g(X)$ is some function of X, then

$$\mathbb{E}[g(X)] = \sum_x g(x)P(X = x).$$

Definition 3.15 (variance)

The *variance* of a random variable X is given by

$$\mathbb{V}\mathrm{ar}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right].$$

Variance

Theorem 3.16

Let a, b and c be constants, then:

$$\mathbb{V}ar[a] = 0$$
$$\mathbb{V}ar[aX + b] = a^2\mathbb{V}ar[X]$$

Theorem 3.17

If X and Y are independent then $Cov(X, Y) = 0$ and so

$$\mathbb{V}ar[X \pm Y] = \mathbb{V}ar[X] + \mathbb{V}ar[Y].$$

Theorem 3.18

For any random variable X , the variance of X is the expected value of the squared difference between X and its expected value

$$\mathbb{V}ar[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

Proof:

Variance

Definition 3.19 (standard deviation)

The *standard deviation* of a random variable X can be written as follows:

$$\mathrm{SD}(X) = \sqrt{\mathbb{V}\mathrm{ar}[X]}.$$

- Note that this is not the same as the *sample* standard deviation s because we are using n instead of $n-1$ in the denominator.

Example Expected Value and Variance — Bernoulli Distribution

Find the expectation and variance of X where $X \sim$ Bernoulli with success probability p.

Example Expected Value and Variance—- Binomial Distribution

Find the expectation and variance of X where $X \sim$ Binomial distribution with success probability p and total number trials n.

Example Expected Value —- Poisson Distribution

Find the expectation of X where $X \sim$ Poisson Distribution with rate λ.

Expected Value Table

Distribution	*Mean*	*Variance*
$Bernoulli(p)$	p	$p(1-p)$
$Binomial(n, p)$	np	$np(1-p)$
$Poisson(\lambda)$	λ	λ

Expected Gain / Loss Example

Example 3.1

A player makes a bet such that she wins \$20 with probability 0.25 and loses \$5 with probability 0.75. What is her expected gain / loss?