Random Variables & Discrete Probability Distributions

Sean Hellingman ©

Introduction to Statistical Data Analysis (ADSC1000)

shellingman@tru.ca

Fall 2024

THOMPSON RIVERS UNIVERSITY

Topics

Introduction

- Many experiments naturally have numerical outcomes.
  - Rolling dice, returns on stocks, or number of goals scored.
- Other experiments have categorical sample spaces.
  - Flipping coins, colours of drawn cards, or type of car driving past.
- A **random variable** is a numerical description of the outcome of an experiment.
  - Function that assigns a real number to each element of a sample space.
- Colours of passing cars example:
  - Red $= 0$, Blue $= 1$, and Green $= 2$.

Probability Distributions

- A **probability distribution** is a characterization of the possible values that a random variable may assume & the corresponding probabilities.
    - Probability distributions can be either discrete or continuous.
- A **discrete random variable** is one for which the number of possible outcomes can be counted.
- A **continuous random variable** has outcomes over one or more continuous intervals of real numbers
- Can use *theoretical*, *relative frequency*, or *subjective* approaches.

Example 1

- Determine if the following random variables are discrete or continuous:
  - Rolling a six sided die.
  - Daily returns on the Apple stock.
  - Olympic sprint times.
  - Drawing cards from a standard deck.
  - Data science test grades.

Example 2

- What is the probability distribution of the sum of two six sided dice?

- Solution:

Outcome	2	3	4	5	6	7	8	9	10	11	12
Probability	1/36	1/18	1/12	1/9	5/36	1/6	5/36	1/9	1/12	1/18	1/36

Probability Mass Function

- The probability distribution of the discrete random variable X is called a **probability mass function**.
- The probability mass function (PMF) is denoted by $p(x)$.
- Where $p(x_i)$ is the probability of the i^{th} value of X.
- From Example 2:
  - $x_1 = 2$, then $p(x_2) = 1/36$.
  - $x_5 = 5$, then $p(x_5) = 1/9$.

Probability Mass Function Assumptions

- Recall from *Probability*:

$$0 \leq p(x_i) \leq 1 \tag{1}$$

$$\sum_i p(x_i) = 1 \tag{2}$$

- Generally, the PMF determines:

$$p(x) = P(X = x) \tag{3}$$

Example 3

- What is the probability mass function of the outcome of flipping two coins (unordered)?

Cumulative Distribution Function

- The **cumulative distribution function** (CDF) $F(x)$, specifies the probability that the random variable X will assume a value less than or equal to a specified value, x.

- *The probability that the random variable X is less than or equal to X.*

$$F(x) = P(X \leq x) \tag{4}$$

Using the Cumulative Distribution Function

- Can use it to find $P(X \leq x)$ or $P(X > x)$
  - Example: $P(X > 5) = 1$ - $P(X \leq 5)$
- *Can also use it to find probabilities over intervals* $P(x_i \leq X \leq x_j)$
  - Example: $P(4 \leq X \leq 8) = P(X \leq 8)$ - $P(X \leq 3)$

Example 4

- What is the cumulative distribution function of the sum of two six sided dice?
  - Use the CDF to determine the probability of rolling a sum between 5 and 7 $P(5 \leq X \leq 7)$.

- Distribution:

Outcome	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Probability	1/36	1/18	1/12	1/9	5/36	1/6	5/36	1/9	1/12	1/18	1/36
Cumulative	1/36	1/12	1/6	5/18	5/12	7/12	13/18	5/6	11/12	35/36	1

Expected Value

- The **expected value** corresponds to the mean or average, for a sample.
- Discrete random variable X the expected value is denoted as $E[x]$
- The expected value is the weighted average of all possible outcomes, where the *weights are the probabilities*.

$$E[X] = \sum x_i \cdot p(x_i) \tag{5}$$

Example 5

- What is the expected value of the sum of rolling two six sided dice?

Outcome	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Probability	1/36	1/18	1/12	1/9	5/36	1/6	5/36	1/9	1/12	1/18	1/36
Cumulative	1/36	1/12	1/6	5/18	5/12	7/12	13/18	5/6	11/12	35/36	1

Variance

- *Recall:* The variance measures the level of uncertainty of the random variable.
- The **variance** of a discrete random variable X as a weighted average of the squared deviations from the expected value:

$$V[X] = \sum (x - E[X])^2 \cdot p(x). \tag{6}$$

- Where the weights are again the probabilities.

Example 6

- Suppose we have the following PMF for a school raffle:

x	$p(x)$
-\$10	0.95
\$500	0.05

- Calculate the expected value $E[X]$ and the variance $V[X]$ of X.

Existing Distributions

- Often easier (theoretically & computationally) to assume our data come from a defined probability distribution.
  - Concerned with estimating the parameters of said distribution.
- Bernoulli Distribution
- Binomial Distribution
- Poisson Distribution

Bernoulli Distribution I

- The **Bernoulli distribution** characterises a random variable having two possible outcomes.
- These outcomes are typically defined as:
  - *success* $(x = 1)$ with probability p
  - *failure* $(x = 0)$ with probability $1 - p$
- Example: Define a success as a flipped coin landing on the tails side.

Bernoulli Distribution II

- The PMF of the Bernoulli distribution is:

$$p(x) = \begin{cases} p, & \text{if } x = 1 \\ 1 - p, & \text{if } x = 0 \end{cases} \tag{7}$$

- Expected value: p
- Variance: $p(1 - p)$

Binomial Distribution I

- The **binomial distribution** models n independent Bernoulli trials each with the probability p of *success*.
- Example: Probability of a coin flipped $n = 10$ times landing on tails 7 times.

Binomial Distribution II

- The PMF of the binomial distribution is:

$$p(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & \text{for } x = 0, 1, ..., n \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

- Expected value: np
- Variance: $np(1-p)$

Binomial Distribution III

- *Hand calculations using the binomial distribution may be a bit complicated*
- R functions:
  - PMF: `dbinom(x, n, p)` calculates $P(X = x)$
  - CDF: `pbinom(x, n, p)` calculates $P(X \leq x)$
  - `qbinom(P, n, p)` used to find the nth quantile.
  - `rbinom(n, N, p)` generates n random variables of a particular probability.

Poisson Distribution I

- The **Poisson distribution** is a discrete distribution used to model the number of occurrences in some unit of measure.
- Examples:
  - Number of customers within an hour.
  - Number of baskets per minute in a basketball game.
  - Number of errors per line of R code.

Poisson Distribution II

- There is no limit on the number of occurrences (X can be any non-negative integer).

- The PMF of the Poisson distribution is:

$$p(x) = \begin{cases} \frac{e^{-\lambda}\lambda^{x}}{x!}, & \text{for } x = 0, 1, \ldots \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

- Expected value: λ

- Variance: λ

Poisson Distribution III

- *Hand calculations using the Poisson distribution may be a bit complicated*

- R functions:
  - PMF: `dpois(x, lambda)` calculates $P(X = x)$
  - CDF: `ppois(x, lambda, lower.tail = TRUE)` calculates $P(X \leq x)$ & $P(X > x)$ if `lower.tail = FALSE`.
  - `qpois(P, lambda)` the number of successes that corresponds to a certain quantile `P`.
  - `rpois(n, lambda)` generates n randomly generated numbers that follow a Poisson distribution with an average number of `lambda` successes.

Example 7

- Suppose we are rolling a six sided die. We define a *success* to be rolling a 3 or a 4.
    - Define the Bernoulli distribution for this experiment.
    - Calculate the expected value and variance (from the theory).

Example 8

- Suppose we are repeating the experiment from Example 7 $n = 10$ times
  - What is the name of the resulting distribution?
  - What is the expected value and variance of this distribution?
  - Using R:
    - Calculate the probability that we get 4 successes.
    - Calculate the probability that we get at most 4 successes.

Example 9

- Suppose we have a Poisson distribution for the number of arrivals at a grocery sore with $\lambda = 10$ per hour.
  - What is the expected value and variance of this distribution?
  - Using R:
    - Calculate the probability that we only have 4 arrivals in an hour.
    - Calculate the probability that we get at **most** 15 arrivals in an hour.
    - Calculate the probability that we get at **least** 11 arrivals in an hour.

Exercise 1

- Identify three continuous and three discrete random variables you might come across in real life.

Exercise 2

- What is the probability mass function (PMF) and the cumulative distribution (CDF) generated by the sum of spinning two 3 numbered spinners of equal probability (The spinners' possible outcomes are 1, 2, or 3 with equal probabilities of 1/3)?

- Calculate the following probabilities:
  - $P(X = 3)$ or $x_3 = 3$
  - $P(X \leq 5)$
  - $P(3 \leq X \leq 6)$

Exercise 3

- What is the expected value and variance of the random variable from Exercise 2?

Exercise 4

- Suppose we have the following PMF:

x	$p(x)$
-10	0.60
20	0.25
15	0.15

- Calculate the expected value $E[X]$ and the variance $V[x]$ of X.

Exercise 5

- Suppose we have $n = 20$ Bernoulli trials with $p = 0.25$.
- From theory, what is he expected value and variance of this distribution?
  - *Note: this is not a Bernoulli distribution*
- Using R:
  - Calculate the probability that we get 4 successes.
  - Calculate the probability that we get at **most** 4 successes.
  - Calculate the probability that we get between 5 and 7 successes.
  - Calculate the probability that we get at **least** 11 successes.

Exercise 6

- Suppose we have a Poisson distribution for the number of goals during a 90 minute soccer game with $\lambda = 3$.
    - What is the expected value and variance of this distribution?
    - Using R:
        - Calculate the probability that we have 4 goals in a 90 minute game.
        - Calculate the probability that we get at **most** 2 goals.
        - Calculate the probability that we get at **least** 4 goals.
        - Calculate the probability that we get between 2 and 4 goals.

References & Resources

1. Evans, J. R., Olson, D. L., & Olson, D. L. (2007). *Statistics, data analysis, and decision modeling.* Upper Saddle River, NJ: Pearson/Prentice Hall.
2. Devore, J. L., Berk, K. N., & Carlton, M. A. (2012). *Modern mathematical statistics with applications (Second Edition).* New York: Springer.

- https://rpubs.com/mpfoley73/458411
- https://en.wikipedia.org/wiki/Binomial_distribution
- https://en.wikipedia.org/wiki/Poisson_distribution