Probability Bootcamp

3. Random Variables

References

1. Chapter 2, Mathematical Statistics and Data Analysis (3rd edition), John A. Rice

Agenda

1. Random Variables
2. Probability Distribution Functions
3. Discrete/Continuous Random Variables
4. Expectation and Variance

Random Variables

Random Variables

1. Frequently, when an experiment is performed, we are interested mainly in some function of the outcome as opposed to the actual outcome itself
    a. Example: When studying the effectiveness of a treatment, it's more meaningful to study the number of patients who recovered, rather than the effect of the treatment on each patient
    b. When determining if a coin is fair, we study the number of heads appearing after repeated tosses, rather than the exact outcome of each individual toss
2. Informally, a random variable refers to the variable of interest we are measuring from the experiment
    a. Formally, a random variable is a function that maps the sample space **Ω** to a real number
3. Because a random variable is random, we may assign probabilities to the possible values of the random variable
    a. These probabilities reflect how we believe the random variable should behave

Introduction

4. Example - consider flipping 3 coins.
    a. **Ω** = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}
    b. Let random variable X denote the number of heads when we flip 3 coins
        i. Hence, X is a function which maps the individual outcomes in **Ω** to a real number (number of heads)
        ii. Example, HHH is mapped to 3, while TTH is mapped to 1
    c. Since each outcome in **Ω** is equally likely,
        i. P(X=0) = P(TTT) = ⅛
        ii. P(X=1) = P(TTH or THT or HTT) = ⅜
        iii. P(X=2) = ⅜
        iv. P(X=3) = ⅛

Probability Distribution Functions

Probability Distribution Functions

1. Recall that in a random variable, we can assign probabilities to all possible outcomes in the support of the random variable
   a. Support of random variable = set of all possible outcomes of random variable
2. A probability distribution function (pdf) is a function that maps all possible outcomes to $[0, 1]$
3. For example, let random variable X = number of heads in 3 coin flips. If f is the pdf, then $f(x) = P(X = x)$
   a. $f(0) = P(X = 0) = P(\text{all tails}) = \frac{1}{8}$
   b. $f(1) = P(X = 1) = P(\text{1 head, 2 tails}) = \frac{3}{8}$
4. In addition to the pdf, it is sometimes convenient to use the cumulative distribution function (cdf). If F is a cdf,
   a. $F(x) = P(X \leq x)$
   b. Using the same X, $F(2) = f(0) + f(1) + f(2) = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} = \frac{7}{8}$

Discrete/Continuous Random Variables

Discrete Random Variables

1. A discrete random variable is a random variable that takes on a countable number of outcomes
   a. Note that countable ≠ infinite
   b. Eg - the set of positive integers {1, 2, 3, …} is countable but is infinite (known as countably infinite)
   c. (from discrete math part) A set is considered countable if there exists a 1-1 mapping from the set to the natural numbers
2. The earlier coin flip example is a discrete random variable. Other examples:
   a. Number of people boarding a bus
   b. Number of coin flips needed to land a first head
3. We will discuss three common discrete random variables here
   a. Bernoulli
   b. Binomial
   c. Geometric

Bernoulli Random Variable

- A Bernoulli random variable takes on a single parameter p, where p $\in$ (0, 1)
- It outputs two values
  - 1 with probability p
  - 0 with probability 1-p
  - May be helpful to think of 1 and 0 as success or failure
- Example. Suppose X denotes the outcome of a coin toss. Assign (1,0) = (H,T)
  - If the coin is fair (meaning it is equally likely for H or T), p = 0.5
  - Suppose the coin is biased and H turns up 60% of the time, then p=0.6
- If X follows a Bernoulli distribution with success probability p, we write as X~Bernoulli(p)
- f(1) = p, f(0) = 1-p

Binomial Random Variable

- Consider the problem of tossing n identical coins independently, each showing heads with probability p. How many heads will we observe?
  - n and p are fixed beforehand
- If X is the random variable for the above, X follows the binomial distribution of an experiment with n trials and probability of success p
  - Written as X~Bin(n, p)
- $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$. Intuition behind pdf is as follows
  - There are $\binom{n}{x}$ ways of observing x heads in a sequence of n coin tosses
  - Given a sequence with x heads and n-x tails, since the individual coin tosses are random, the probability of observing this sequence is $p^x (1-p)^{n-x}$
  - Since there are $\binom{n}{x}$ such sequences, multiplying both terms gives the desired outcome

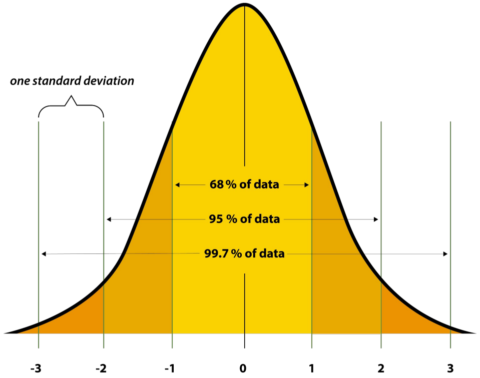
Geometric Random Variable

- Given a coin with probability p of showing heads. Consider the problem of counting the number of tosses it takes to land heads for the first time.
  - If X is the above random variable, it follows the geometric distribution with parameter p
  - Written as X~Geom(p)
- $f(x) = (1-p)^{x-1}p$
  - $f(x) = P(X = x)$ is the probability that it takes x coin flips to observe heads for the first time
  - $(1-p)^{x-1}$ gives the probability of observing tails for the first x-1 trials
  - Multiplying $(1-p)^{x-1}$ with p gives the probability of finally observing a heads in the x-th trial

Continuous Random Variable

- We are often interested in random variables that can take on a continuum of values, rather than just a countable number
  - For example, when studying the waiting time for buses, it is more meaningful to study the number of buses that arrive within 2-3 minutes rather than in exactly 2 minutes
- If X is a continuous random variable with a probability distribution function of f,
  - P(X = x) = 0 for any x and $P(a < X < b) = \int_a^b f(x)\,dx$
  - Using the bus waiting time as an example, the intuition behind this is as follows
    - The probability of a bus arriving in exactly 2 minutes can be regarded as non-existent, since there are an infinite number of possibilities (2.0001 mins, 2.1 mins, 1.999 mins, etc). Hence, the probability of a continuous random variable hitting an exact number can be treated as 0 and P(X = x) = 0
    - Integration can be viewed as an infinite sum across an interval. By integrating the pdf across (a, b), we are summing up all the (infinite number of) probabilities in this region
  - Note that since P(X = x) = 0, $P(a < X < b) = P(a \le X \le b) = P(a < X \le b) = P(a \le X < b)$

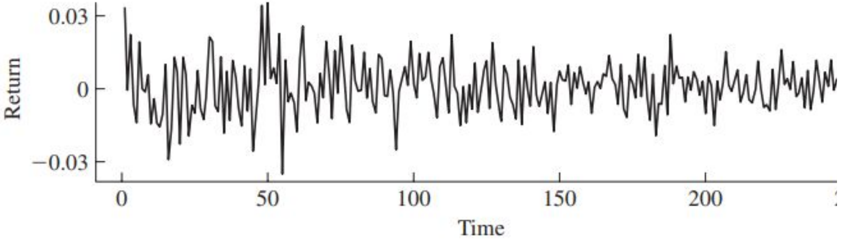
Normal Distribution



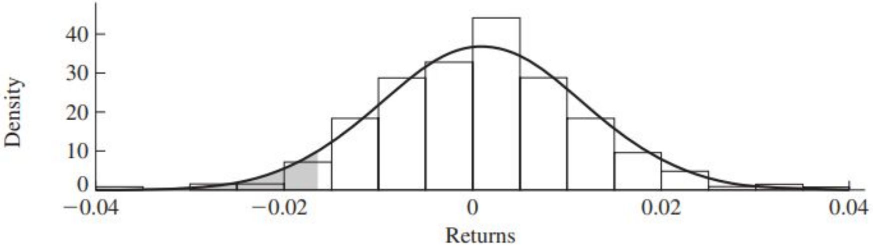
- The normal distribution takes in two parameters
  - μ: The mean (average) of the random variable
  - σ^2: The variance (spread) of the random variable
- It has the following pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$
- Central limit theorem: The sum of many independent and identically distributed random variables follows a normal distribution
- The normal distribution is popular in practice as a lot of real-life data with a unimodal distribution (one peak) can be approximated well by a normal distribution

Normal Distribution

- Average daily returns of S&P 500 in 2003 is given below



- This average daily return is fitted with a normal distribution with μ = 0.001, σ = 0.01



Expectation (Mean) and Variance

Expectation

- From its name, it's what we ‘expect’ to be the value of the random variable
- Mathematically, it’s a weighted average of all possible outcomes of the random variable, weighted by their respective probabilities. Given a random variable X with pdf f, its expectation (written as E(X)) is,
  - If X is discrete with support S, $E(X) = \sum_{x \in S} x f(x)$
  - If X is continuous, $E(X) = \int_{-\infty}^{\infty} x f(x)\, dx$
  - Note that both take on a similar form and notice how they are both weighted averages

Expectation of some common random variables

- For practice, try to derive the expectation for the Bernoulli, binomial and geometric cases.
  - Normal distribution is a bit more tricky and involves more advanced mathematical tricks

Distribution	X	E(X)
Bernoulli	Bernoulli(p)	p
Binomial	Bin(n, p)	np
Geometric	Geom(p)	1/p
Normal	$N(\mu, \sigma^2)$	μ

Variance

- The mean tells us what will the average value will look like in the long run. But it does not tell us the spread of the random variable
  - For example, two funds may have similar 7-day moving average values.
  - But a higher risk fund may invest in more volatile assets, which makes it more likely for the fund to appreciate or depreciate rapidly (higher variance)
  - In contrast, a lower risk fund may invest in less risky assets (eg government bonds), making drastic fluctuations in value less likely (less variance)
- Intuitively, we want to measure: On average, how much does the random variable deviate from its mean?
- Mathematically, for the discrete and continuous cases respectively

$$Var(X) = \sum_{x \in S} (x - E(X))^2 f(x)$$

$$Var(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x)\, dx$$

Variance of some common random variables

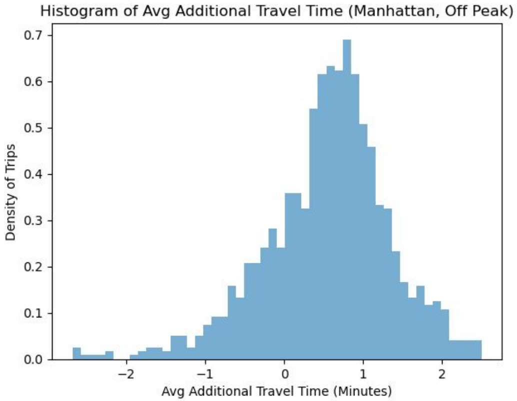
- Variance of the four distributions covered earlier below
  - Computing the variance might be harder as it requires other properties of expectation and integrals
  - If interested, derivation of variance can be found in their respective Wikipedia pages

Distribution	X	E(X)
Bernoulli	Bernoulli(p)	p(1-p)
Binomial	Bin(n, p)	np(1-p)
Geometric	Geom(p)	$(1-p)/p^2$
Normal	$N(\mu, \sigma^2)$	σ^2

- Variance = $(\text{standard deviation})^2$

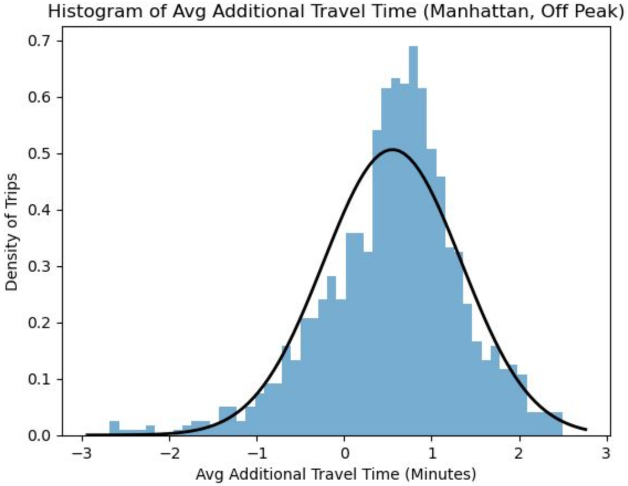
Example 1 - Bus Travelling Time

- We analyze the average additional travel time of Manhattan buses during off-peak hours, taken from the MTA Bus Customer Journey-Focused Metrics: 2017-2019 dataset
  - Average additional travel time = Actual travel time - travel time advertised
  - Positive average travel time = journey took longer than expected
  - Negative average travel time = journey was quicker than expected
- Histogram of average additional travel time given below



Example 1 - Bus Travelling Time

- Looks bell-shaped. Normal approximation may be a good candidate to approximate its distribution
  - Black curve is the normal distribution curve, with the mean and standard deviation parameters derived from the original data
  - $\mu \approx 0.552$, $\sigma^2 \approx 0.621$



Example 1 - Bus Travelling Time

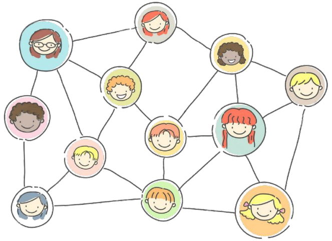
- Let random variable X = commuter’s waiting time. We approximate X with our modelled normal distribution, using μ and σ from previous slide
- Since we have a normal approximation for X, we can (approximately) answer the following

Question	In terms of X	(Approximated) Answer
Proportion of trips on time	$X \le 0$	$P(X \le 0) \approx 0.242$
Proportion of trips more than 3 mins late	$X > 3$	$P(X>3) \approx 0.0009$
Maximum time such that 99% of trips are completed	Minimum value of t such that $P(X<t) = 0.99$	Since $P(X<2.38) = 0.99$, $t = 2.38$ 99% of trips will take at most 2.38 mins more than the scheduled time.

Example 1 - Bus Travelling Time

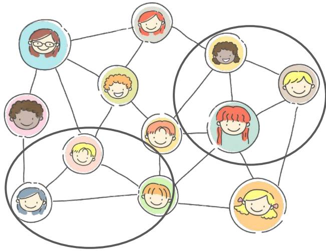
- Suppose now we have a group of 50 commuters in Manhattan travelling during off-peak hours. We are interested in the number of commuters out of the 50 whose journeys are not delayed.
- Let random variable Y = number of commuters (out of 50) with on-time journeys
  - Y ~ Binomial(50, p)
  - What is p? Can take estimate from previous slide, p ≈ 0.242
- Expected number of commuters with on-time trips?
  - E(Y) = 50p = 50*0.242 = 12.1
- Probability that all 50 trips arrive on time?
  - $P(Y=50) = \binom{50}{50} p^{50}(1-p)^{50-50} = p^{50} \approx 10^{-31}$

Example 2 - Subgraph Counts in A Random Graph



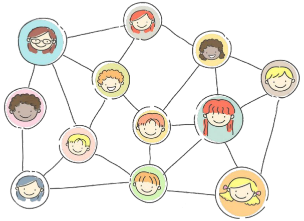
- Consider a community of n people
- If we randomly select two people, both of them are friends with probability p
  - Recall that there are $\binom{n}{2}$ pairs of people.
  - Let Y be the random variable representing the number of pairs of friends. $Y \sim Bin\left(\binom{n}{2}, p\right)$
  - For example, if n = 10 people, and p = 0.1,
    - $Y \sim Bin\left(\binom{10}{2}, 0.1\right) = Bin(45, 0.1)$
    - E(Y) = 45 * 0.1 = 4.5
- In practice, this kind of network is known as an Erdős–Rényi random graph
  - Good for modelling problems that involve relationships among different subjects, like social networks, computer networks, disease spread, etc.

Example 2 - Subgraph Counts in A Random Graph



- Earlier, we modeled the number of pairs of people that are friends. Can we extend this to triads?
  - That is, if we randomly pick 3 people, do all 3 of them know each other?
  - How many such triads are there? Equivalently, how many triangles (like those circled) are there in the friendship graph?
- There are $\binom{n}{3}$ ways of selecting 3 people from a group of n
- Given 3 randomly chosen people, the probability that all of them know each other is p^3
- Therefore, if X is the number of triads that know each other, $X \sim Bin\left(\binom{n}{3}, p^3\right)$
  - If n = 10, and p = 0.1, then $E(X) = \binom{10}{3} \times 0.1^3 = 0.12$
  - If n = 10, and p = 0.5, then $E(X) = \binom{10}{3} \times 0.5^3 = 15$
  - If n = 10, and p = 1, then $E(X) = \binom{10}{3} = 120$

Example 2 - Subgraph Counts in A Random Graph



- More generally, this problem is known as subgraph counts in random graphs
  - A friendship graph with many triangles means that if two people share mutual friends, they are likely to be friends too
  - A network graph with many ‘stars’ would indicate many single points of failures
  - A disease spread graph with many ‘stars’ would also indicate the presence of many super-spreaders
- Active research problem, since counting patterns in large graphs is known as a NP-hard problem (very computationally expensive)
  - But many problems can be modeled as graphs!
- Researchers are turning to probabilistic approximations (like what we did) to statistically model and approximate subgraph counts
  - This has been applied to detect possible financial crimes in financial transaction networks
    - The presence of loops may indicate money laundering
    - The presence of a pyramid/tree like structure may indicate pyramid schemes