Probability Bootcamp

1. Introduction

References

1. Introductory Lecture, Probability: Theory and Examples (3rd Edition), Rick Durrett
2. https://www.investopedia.com/terms/c/central_limit_theorem.asp
3. https://www.investopedia.com/terms/l/lawoflargenumbers.asp

Introduction

- Probability has a left and right hand
  - Right: Probabilistic intuition - thinking of uncertainty in terms of everyday events (coin tosses, dice, etc)
  - Left: Rigorous math - formally expressing this intuition using math
- Probability theorists and statisticians need to use both hands
  - Right: Develop intuition for complex problems & explaining to layperson
  - Left: Mathematically justify intuition from left hand, find exceptions where the intuition fails
- Purpose of this bootcamp
  - To gently (re)-introduce this intuition (right hand), while informally illustrating how we can use mathematical tools to justify this intuition (left hand)
  - The latter won't be emphasized; will be more formally taught in actual modules
- For context, we will give a brief preview of what's to come

Probability Theory

- Two main results in classical probability
  - Law of large numbers
    - If we repeat an experiment repeatedly, would averaging the results give more accurate findings?
    - If we gamble and place the same bet repeatedly, what are our expected earnings (loss)?
  - Central limit theorem
    - If I have a sufficiently large (random) sample from a population, can I approximate the distribution of the true population?
- We motivate this with a gambling example
  - In probability literature, gambling is used as a metaphor for events involving random chance
  - Good understanding of probability should serve as a deterrent against gambling

Probabilities and Expected Value

Roulette Wheel (Scenario 1)

- A roulette wheel has 38 slots
  - 18 black, 18 red, 2 green
- Suppose we bet $1 on red repeatedly
  - If the ball ends up in red, we win additional $1
  - Otherwise, we lose our $1 bet
- Is this a good bet?

Roulette Wheel (Scenario 2)

- A roulette wheel has 38 slots
  - 18 black, 18 red, 2 green
- Suppose we bet $1 on red repeatedly
  - If the ball ends up in red, we win additional $2
  - Otherwise, we lose our $1 bet
- Is this a good bet?

Roulette Wheel (Scenario 3)

- A roulette wheel has 38 slots
  - 18 black, 18 red, 2 green
- Suppose we bet $1 on red repeatedly
  - If the ball ends up in red, we win additional $0.50
  - If ball ends up in green, we win additional $4.50
  - Otherwise, we lose our $1 bet
- Is this a good bet?

Roulette Wheel (Comparing The Three Scenarios)

- A roulette wheel has 38 slots
  - 18 black, 18 red, 2 green
- Bet $1 each round, payoffs summarized below
- Which game would you bet on, and why?

	1	2	3
Payoff	Red: Win $1 Green: Lose $1 Black: Lose $1	Red: Win $2 Green: Lose $1 Black: Lose $1	Red: Win $0.50 Green: Win $4.50 Black: Lose $1

Roulette Wheel (Expected Payout)

- A roulette wheel has 38 slots
  - 18 black, 18 red, 2 green
- Bet $1 each round, payoffs summarized below
- Which game would you bet on, and why?

	1	2	3
Payoff	Red: Win $1 Green: Lose $1 Black: Lose $1	Red: Win $2 Green: Lose $1 Black: Lose $1	Red: Win $0.50 Green: Win $4.50 Black: Lose $1
Expected payout	$1 * (18/38) + (-$1) * 20/38 = -$2/38	$2 * (18/38) + (-$1) * 20/38 = $16/38	$0.50 * (18/38) + $4.50 * (2/38) + (-$1) * 18/38 = $0

Roulette Wheel (Expected Payout)

- How do we interpret expected payout?
  - Intuitively, expected payout is the average payout if I play this game many times
  - Mathematically, it's the average of all possible outcomes, weighted by their likelihood

	1	2	3
Payoff	Red: Win $1 Green: Lose $1 Black: Lose $1	Red: Win $2 Green: Lose $1 Black: Lose $1	Red: Win $0.50 Green: Win $4.50 Black: Lose $1
Expected payout	$1 * (18/38) + (-$1) * 20/38 = -$2/38	$2 * (18/38) + (-$1) * 20/38 = $16/38	$0.50 * (18/38) + $4.50 * (2/38) + (-$1) * 18/38 = $0

Probabilities and Expected Values

- Probability
  - Black/green/red - which is more likely?
  - Used to compare between different events
  - In law, ‘balance of probabilities’ is used in court to determine if its more likely than not an event occurred
- Expected value
  - If I repeat this game repeatedly, what are my average earnings?
  - In data science, expected value tells us the projected value some time in the future
    - If a disease infects the population with a probability of 0.1, the expected number of infections is 0.1 * total population
- In fact, probability is just a special case of expected value (how?)

Law of Large Numbers

Roulette Wheel

- Suppose you played Scenario 2, but you keep losing
- You suspect the casino is cheating, as the ball never lands on green
- Can you prove this?

Law of Large Numbers

- Recall that there are 38 slots
  - 2 green
  - 18 red and 18 black
  - The casino owner claims that the ball falls in each slot with equal probability
- Law of large numbers tells us that if we repeat an experiment many times, the average value is close to the actual value
  - The more we repeat, the closer it gets
- This means that if the game is played repeatedly, expect the proportion of balls that fall in the green slot to be close to 2/38

Law of Large Numbers

- Suppose we played the game 50 times, and green only occurred once
- Is this evidence of the casino cheating?
- No! We cannot conclude that the casino is cheating just yet
- Reason
  - Law of large numbers tells us what happens in the long run.
  - But it does not tell us how many repeats it takes to reach the ‘long run’
  - This gap is filled by the Central Limit Theorem

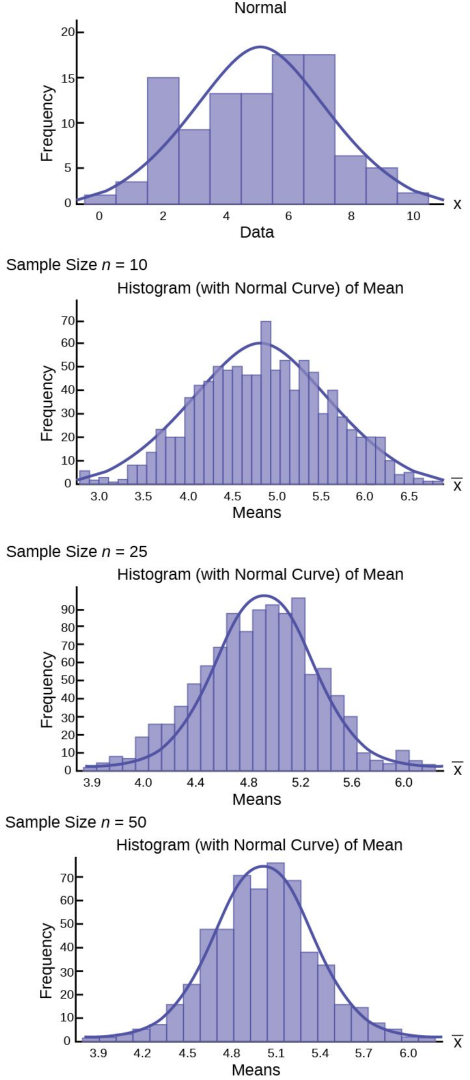
Central Limit Theorem

Central Limit Theorem

- Also known as the ‘bell curve theorem’
- Central limit theorem (CLT) tells us that the distribution of a sample can be approximated by a Normal distribution
  - This approximation improves when the sample size increases
  - In practice, even with moderate sample sizes (eg 20-30), the approximation usually works well
- Mathematical representation of CLT - Let
  - S_n = number of times we observe green after n rounds
  - μ = theoretical probability of observing green (2/38)
  - σ^2 = Variance of a single game (Covered later)
    - For now, take for granted it is (2/38) * (1 - 2/38)
- Then, $(S_n - n\mu) / (n^{1/2} \sigma)$ can be approximated by a normal distribution.
  - This approximation improves as n gets larger

When n gets larger

- Note two things
  - The spread of the normal approx. gets smaller
  - The real data gets closer and closer to the normal approx.



Is the Casino Cheating?

- Recall that $(S_n - n\mu) / (n^{1/2} \sigma)$ can be approximated by a normal distribution.
    - n = number of rounds played (50)
    - S_n = number of times we observe green after n rounds (1)
    - μ = theoretical probability of observing green (2/38)
    - σ^2 = Variance of a single game (Covered later)
        - For now, take for granted it is (2/38) * (1 - 2/38)
- $(S_n - n\mu) / (n^{1/2} \sigma) = -1.03$
    - If Z is the standard normal distribution, $P(Z < -1.03) = 0.15$
- No strong statistical evidence casino is cheating

Is the Casino Cheating?

- Determined to catch the casino cheating, you played a total of 100 rounds
  - After 100 rounds, you only observed 2 greens (S_n)
  - n = number of rounds played (100)
  - Everything else remains the same
- $(S_n - n\mu) / (n^{1/2} \sigma) = -1.46$
  - If Z is the standard normal distribution, $P(Z < -1.46) = 0.07$
- Stronger evidence casino is cheating!
  - In practice, p-values of 0.10 are usually deemed as strong evidence

Central Limit Theorem

- Intuitively, playing more rounds increases your knowledge of how the system behaves
- CLT formalizes this by telling us that the more rounds we play, the better we can approximate and predict how the system will behave in the future
- A useful tool to derive inferences on a large population with a limited sample

Putting Everything Together

Probability for Decision Making

Tool	Expected Value	Law of Large Numbers	Central Limit Theorem
What is it?	Weighted average of all possible outcomes by likelihood of outcome	Taking averages from repeated experiments yields more accurate estimates	Under appropriate (but easily fulfilled) conditions, the statistical distribution of a sample is related to the Normal distribution
When to use?	Comparing between multiple choice of scenarios	Estimating an unknown (but observable) parameter by running an experiment repeatedly	When drawing statistical inferences and predictions based on observed data
How was it used earlier?	Comparing between the three roulette games to see which one is more profitable	Estimating the probability of the ball entering the green slot	Inferring if the casino is cheating when it claims all slots in the roulette wheel are equally likely

Agenda for Today

- Introductory Probability
  - Sample Spaces
  - Probability Measures
  - Counting Methods
  - Conditional Probability
  - Independence
- Introduction to Random Variables
  - Random Variables
  - Probability Distribution Functions
  - Discrete/Continuous Random Variables
  - Expectation and Variance