CS70 - Lecture 10 Notes

Name: Felix Su SID: 25794773

Spring 2016 GSI: Gerald Zhang

Polynomials

- **Fact**: Any $d+1$ points specifies a distinct degree d polynomial
- **Modular Fact**: Any $d+1$ points specifies a distinct degree d polynomial in mod p space when p is prime
- **Uniqueness Fact**: At most one degree d polynomial hits $d+1$ points

Uniqueness.

Uniqueness Fact. At most one degree d polynomial hits $d+1$ points.

Proof:

Roots fact: Any degree d polynomial has at most d roots.

Assume two different polynomials $Q(x)$ and $P(x)$ hit the points.

$R(x) = Q(x) - P(x)$ has $d+1$ roots and is degree d.
Contradiction.

- **Roots Fact**: Any degree d polynomial has at most d roots

Only d roots.

Lemma 1: $P(x)$ has root a iff $P(x)/(x-a)$ has remainder 0:
$P(x) = (x-a)Q(x)$.

Proof: $P(x) = (x-a)Q(x)+r$.
Plugin a: $P(a) = r$.
It is a root if and only if $r = 0$. □

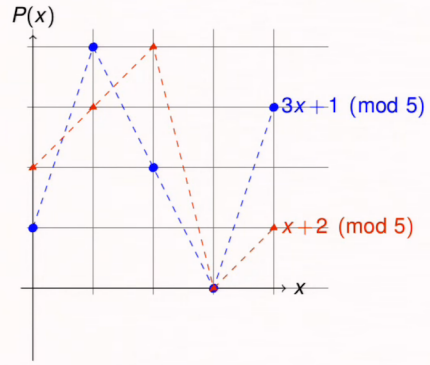
Lemma 2: $P(x)$ has d roots; $r_1, \ldots, r_d$ then
$P(x) = c(x-r_1)(x-r_2)\cdots(x-r_d)$.
Proof Sketch: By induction.

Induction Step: $P(x) = (x-r_1)Q(x)$ by Lemma 1. $Q(x)$ has smaller degree so use the induction hypothesis. □

$d+1$ roots implies degree is at least $d+1$.

Roots fact: Any degree d polynomial has at most d roots.

Polynomial: $P(x) = a_d x^4 + \cdots + a_0 \pmod p$



Finding an intersection.
$x+2 \equiv 3x+1 \pmod 5$
$\implies 2x \equiv 1 \pmod 5 \implies x \equiv 3 \pmod 5$
3 is multiplicative inverse of 2 modulo 5.
Good when modulus is prime!!

- Polynomials only map to f(x) at integer values of x
- f(x) is contained in the mod space
- Use delta functions to create meaningful polynomials in mod space

Shamir's k out of n scheme:
Secret $s \in \{0, ..., p-1\}$
Set $a_0 = s$, randomly assign $a_1, ..., a_{k-1}$
Let $P(x) = a_{k-1}x^{k-1} + a_{k-2}x^{k-2} + ... + a0$ with $P(0) = a_0 = s$
Share $(i, P(i) \bmod p)$ with i-th person
k shares gives secret (degree $= d = k-1$, Modular fact, $d+1 = k$ shares gives the polynomial which reveals $P(0) = s$
Solve for polynomial given $d+1$ coordinates

In general..

Given points: $(x_1, y_1); (x_2, y_2) \cdots (x_k, y_k)$.

Solve...

$$\begin{aligned} a_{k-1}x_1^{k-1} + \cdots + a_0 &\equiv y_1 \pmod p \\ a_{k-1}x_2^{k-1} + \cdots + a_0 &\equiv y_2 \pmod p \\ &\vdots \\ a_{k-1}x_k^{k-1} + \cdots + a_0 &\equiv y_k \pmod p \end{aligned}$$

Will this always work?

As long as solution **exists** and it is **unique!** And...

Modular Arithmetic Fact: Exactly 1 degree $\leq d$ polynomial with arithmetic modulo prime p contains $d+1$ pts.

- $d = k-1, d+1 = k$
- Solve system of linear equations to get a_0

Lagrange Interpolation

Delta Function

$$\Delta_i(x) = \begin{cases} 1, & x = x_i \\ 0, & x = x_j \text{ for } j \neq i \\ \text{doesn't matter}, & x = \text{anything else} \end{cases}$$

- 1 at one point (x-value), 0 everywhere else
- valid for a set of x values $x_1, ..., x_{d+1}$
- $y_i\Delta_i(x) = y_i$ because $\Delta_i(x)$ is 1 at x_i and 0 everywhere else
  - ⋆ $P(x) = y_1\Delta_1(x) + y_2\Delta_2(x) + ... + y_{d+1}\Delta_{d+1}(x)$ because at x_i you only get y_i (Δx_i is 0 at anything except x_i)

Formation of Delta Function:
Given points: $(x_1, y_1); (x_2, y_2); ...(x_{d+1}, y_{d+1})$

$$\Delta_i(x) = \frac{\prod_{j \neq i}(x - x_j)}{\prod_{j \neq i}(x_i - x_j)} \tag{1}$$

CS70 - Lecture 11 Notes

Name: Felix Su SID: 25794773

Spring 2016 GSI: Gerald Zhang

Secret Sharing

Minimality

- Use mod p space where p is prime
- $p > n$ where n is the amount of shares you want to hand out
- $p > 2^b$ where b is the number of bits you want in your secret
- Uses **Theorem**(There is always a prime between n and $2n$). This strategy chooses a p that is within 1 bit of secret size (minimality).

Runtime

- Polynomial in terms of k, n, and $\log p$
- Evaluate $k-1$ degree polynomials n times as a system of linear equations, using $\log p$-bit numbers
- Reconstruct secret by solving system of k equations using $\log p$-bit arithmetic.

Counting

- m^{d+1}: $d+1$ coefficients must be $\in \{0, ..., m-1\}$
- m^{d+1}: $d+1$ points with y-values that must be $\in \{0, ..., m-1\}$

Erasure Codes

Solution

- n packet message, loses k packets in channel
- must send $n+k$ packets
- Use n point values to construct an $n-1$ degree polynomial

Erasure Coding Scheme:

1. n packet message: $m_0, m_1, ..., m_{n-1}$
2. Choose prime $p \approx 2^b$ for mod space where each packet has b bits
3. $p > n + k$
4. $P(x) = m_{n-1}x^{n-1} + ... + m_0 \pmod{p}$
5. Send, $P(1), ..., P(n+k)$

Any n of the $n+k$ packets gives polynomial and the entire message (all coefficients or y-values)

Erasure Coding Example:
Sending
Send message 1, 4, 4 (3 packets, 2 bits)
Make $P(x)$: $P(1) = 1, P(2) = 4, P(3) = 4$
Try mod5 because 5 is the closest prime to $2^b = 4$, but only gives 5 possible shares, so work mod7
Use Lagrange Interpolation
$P(x) = 2x^2 + 4x + 2 \bmod 7$
Send $(0, P(0))(1, P(1))...(6, P(7))$: 6 points
Receieving
Retrieve $P(x)$ using Lagrange or system of linear equations
Need to know which x-value the correct packets correspond to

Error Correction

- Need to recover information sent AND which packets are corrupted
- Send $n + 2k$ packets because if k errors exist, multiple original messages are possible if $< n + 2k$ packets sent.

Reed-Solomon Code:

1. Encoding polynomial $P(x)$ of degree n-1
   - $P(1) = m_1, ..., P(n) = m_n$
   - Can encode with packets as coefficients (check HW6)
2. Use **Lagrange Interpolation** to get $P(x)$
3. Send $(P1), ..., P(n + 2k)$
4. After noisy channel, receive $R(1), ..., R(n + 2k)$
5. $P(i) = R(i)$ for at least $n + k$ points i; $P(i) \neq R(i)$ for k points
6. Do not know where errors occurred
7. $P(x) =$ unique degree $n - 1$ polynomial

Error Locator Polynomial: $\boldsymbol{E(x) = (x - e_1)(x - e_2) \cdots (x - e_k)}$

- Errors at points $e_1, ...e_k$; E(i) = 0 iff $e_j = i$ for some j; $E(x)$ has degree k
- Idea: Multiply equation i by $E(x) = (x - i)$ iff $P(i) \neq R(i)$, but this creates $n + 2k$ **non-linear** equations with n_k unknowns.
- **Solution:** Let $Q(x) = E(x)P(x) = a_{n+k-1}x^{n+k-1} + \cdots + a_0$
  - Now you have $n + 2k$ linear equations $Q(i) = R(i)E(i)$
  - **Find $E(x)$ and $Q(x)$**
    - $E(x) = x^k + b_{k-1}x^{k-1} \vdots b_0$ w/ k unknown coefficients
    - $Q(x) = a_{n+k-1}x^{n+k-1} + \cdots + a_0$ w/ $n + k$ unknown coefficients
    - Solve for coefficients of $Q(x)$ and $E(x)$; Total Unknowns: $n + 2k$
  - $\boldsymbol{P(x) = Q(x)/E(x)}$

Brute force: BAD

- Remove every possible combination of k received packets one at a time and form a degree $n + k - 1$ polynomial with remaining $n + k$ points. First consistent solution gives the corrupted packet.
- Runtime: $(n/k)^k$: exponential in k with $\binom{n+2k}{k}$ possibilities

RS Code Example:
Problem:

- Message 3,0,6 : tolerate $k = 1$ errors (send $n + 2k = 5$ packets)
- Lagrange Encoding $P(x) = x^2 + x + 1 \pmod 7$
- Send: $P(1) = 3, P(2) = 0, P(3) = 6, P(4) = 0, P(5) = 3$
- Receive: $R(1) = 3, R(2) = 1, R(3) = 6, R(4) = 0, R(5) = 3$

Solution: Berklekamp-Welsh Algorithm

- $Q(x) = E(x)P(x) = a_3x^3 + a_2x^2 + a_1x + a_0$
- $E(x) = x - b_0$
- $Q(i) = R(i)E(i)$

$$\begin{aligned}
a_3 + a_2 + a_1 + a_0 &\equiv 3(1 - b_0) \pmod 7 \\
a_3 + 4a_2 + 2a_1 + a_0 &\equiv 1(2 - b_0) \pmod 7 \\
6a_3 + 2a_2 + 3a_1 + a_0 &\equiv 6(3 - b_0) \pmod 7 \\
a_3 + 2a_2 + 4a_1 + a_0 &\equiv 0(4 - b_0) \pmod 7 \\
6a_3 + 4a_2 + 5a_1 + a_0 &\equiv 3(1 - b_0) \pmod 7
\end{aligned}$$

- Gaussian Elimnation: $a_3 = 1, a_2 = 6, a_1 = 6, a_0 = 5$; $b_0 = 2$
- $Q(x) = x^3 + 6x^2 + 6x + 5$
- $E(x) = x - 2$
- Polynomial Long Division: $P(x) = Q(x)/E(x) = x^2 + x + 1 \pmod 7$

$$\begin{array}{r|l}
x - 2 & \begin{array}{r} x^2 + 8x + 22 \\ \hline x^3 + 6x^2 + 6x + 5 \\ -x^3 + 2x^2 \\ \hline 8x^2 + 6x \\ -8x^2 + 16x \\ \hline 22x + 5 \\ -22x + 44 \\ \hline 49 \end{array}
\end{array}$$

- **Message = 3,0,6**
- RS Code: $P(x) = x^2 + x + 1 \pmod 7$ where $P(1) = 3, P(2) = 0, P(3) = 6$

CS70 - Lecture 12 Notes

Name: Felix Su SID: 25794773

Spring 2016 GSI: Gerald Zhang

Berklekamp-Welsh Algorithm

Existence:

- Exists because packets constructed using $P(x)$

Unique:

- Proved assuming $\frac{Q'(x)}{E(x)} = \frac{Q(x)}{E'(x)} = P(x$
- $E(x)$ and $E'(x)$ have at most k roots each
- Cross multiply assumption at n valid points, so claim is true for n points, which make $P(x)$ a unique $< n$ degree polynomial
- **Fact:** $Q'(x)E(x) = Q(x)E'(x)$ on $n + 2k$ values of x
  - Holds when $E(x)$ or $E'(x)$ are 0
  - Use above method of cross multiplication when not zero

Encoding/Decoding Polynomial Summary:

Summary: polynomials.

Set of $d+1$ points determines degree d polynomial.

Encode secret using degree $k-1$ polynomial:
Can share with n people. Any k can recover!

Encode message using degree $n-1$ polynomail:
n packets of information.

Send $n+k$ packets (point values).
Can recover from k losses: Still have n points!

Send $n+2k$ packets (point values).
Can recover from k corruptionss.
Only one polynomial contains $n+k$
Efficiency.
Magic!!!!
Error Locator Polynomial.

Relations:
Linear code.
Almost any coding matrix works.
Vandermonde matrix (the one for Reed-Solomon)..
allows for efficiency. Magic of polynomials.

Counting

- Counting Numbers: 0,1,2... all Natural Numbers $\mathbb{N}$
- **Countable if there is a bijection between S and some subset of $\mathbb{N}$**
- if subset of $\mathbb{N}$ = finite, S has finite **cardinality**
- if subset of $\mathbb{N}$ = infinite, S is **countably infinite**
  - Evens are countably infinite
  - Integers are countably infinite
  - Pairs of Natural Numbers are countably infinite
  - Rationals are countably infinite (subset of pairs of natural numbers with gcd of 1)
  - Reals are uncountable
- All countably infinite sets have the same cardinality

Isomorphism Principle:

- If there is $f : D \to R$ that is one to one and onto, (bijective) then $|D| = |R|$

Listing:

- A bijection with a subset of natural numbers
- The nth item in the list is a mapping $n \in \mathbb{N} \to f(n)$
- If you can list a set you can show a bijection
- Finite List: Bijection with subset of $\mathbb{N}\{0, ..., |S| - 1\}$
- Infinite List: Bijection with $\mathbb{N}$

Enumerating $\equiv$ Countability = Listing:

- Enumerating a set $\implies$ countability
- Corollary: Any subset T of a countable set S is also countable
- Each element of $x \in S$ has a specific, finite position in a list (ex. $\mathbb{Z} = \{0, 1, -1, 2, -2, ...\}$
- Fails for integers if you list positive integers before negative integers
  - $\mathbb{Z} = \{\{0, 1, 2, ..., \}$ and then $\{-1, -2, ...\}\}$
  - -1s position is not finite, because there are ∞ positive integers
  - So.. you must **interweave**

Diagonalization:

- **Diagonal Number** Number that is not in the list of $f(n)$
- Ex. Method to create diagonal number for Reals: Digit i is 7 if number i's ith digit is not 7, 6 otherwise. For every nth position on the list, at least the nth digit will be different than the diagonal number's nth digit. Contradiction because the diagonal number is real.
- Check if this creates contradicton. If diagonal number is in the questionable set, the list could not have existed and thes et is not countable.

CS70 - Lecture 13 Notes

Name: Felix Su SID: 25794773

Spring 2016 GSI: Gerald Zhang

Countability Summary:

- S is countable if there is a bijection between S and some subset of $\mathbb{N}$.
- If the subset of $\mathbb{N}$ is finite, S has finite cardinality.
- If the subset of $\mathbb{N}$ is infinite, S is countably infinite.
- Bijection with natural numbers $\implies$ countably infinite.
- Enumerable (listing) $\equiv$ countable.
- Subset of countable set is countable.
- All countably infinite sets have the same cardinality
- Natural numbers have finite number of digits

Digonalization:

- The set of all subsets S_i of $\mathbb{N}$ (powerset of $\mathbb{N}$ is not countable
  - Arbitrary Listing: L
  - Diagonal set D: For each index i of L, if $i \notin S_i$, put i in D, otherwise omit i
  - D is not in L by construction: D is different from each ith set S_i in L, for every i
  - D is a subset of N: every element in D is a natural number
  - L does not contain all subsets of N: Contradiction

Diagonalization Algorithm:

1. Assume that set S can be enumerated.
2. Consider an arbitrary list of all the elements of S.
3. Use the diagonal from the list to construct a new element t.
4. Show that t is different from all elements in the list $\implies$ t is not in the list.
5. Show that t is in S.
6. Contradiction.

Cardinalities:

Continuum Hypothesis:

- Goedel proved this hypothesis cannot be proven with math we currently know
- There is no infinite set whose cardinality is between the cardinality of an infinite set and its power set.

Uncountable Sets:

- Prove equivalence between cardinalities
- Show bijection exists between two sets: uncountable sets cannot be enumerated
- Create function $f : B \to A$ (can include multiple casesfor certain domains)
- Prove mapping is one to one by testing on arbitrary values: x, y (Need to validate for multiple cases)
  - Example: $|[0, 1]| \equiv |\mathbb{R}|$
  - $f : \mathbb{R}^+ \to [0, 1]$

Undecidability:

Russell's Paradox:

- Naive Set Theory: Any definable collection is a set.

$$\exists y \forall x (x \in y \iff P(x)) \tag{1}$$

- NST : $y =$ the set of elements that satifies $P(x)$
- Make statement: $P(x) = x \notin x$
- By NST: There exists a y that satisfies above statement for $P(x)$
- Plug in $x = y$ to NST

$$y \in y \iff y \notin y \tag{2}$$

- Methematicl system is broken, because conditions and statements are false and contradictions

HALT: DNE

- $HALT(P, I)$: P = program, I = input to program
  - Theoretically determines if $P(I)$ halts or loops forever

Halt Turing Proof:

- Assume $HALT(P, I)$ exists
- Set $P = Turing(P)$
- Use Diagonalization

```
def Turing(P):
    if(HALT(P,P)): #halts
        go into infinite loop
    else
        halt immediately
```

- Assume $Turing(Turing)$ halts
- Run $HALTS(Turing, Turing)$
  - if 'halts', Turing(Turing) 'goes into infinite loop'
  - if loops forever, Turing(Turing) 'halts immediately'
- Contradiction, so $HALT(P, P)$ does not exist

Halt Diagonlization Proof:

- Program and input are both enumerable (fixed length strings)
- Program either halts or loops on any input
- Create list: $P_i \to P_j(P)$ where $i, j \in \mathbb{N}$
- Each entry of list is arbitrarily $HALT$ or $LOOP$
- Diagonal exists, so create $Turing()$ s.t. it returns opposite values along the diagonal
- This means $Turing()$ is not in the list $\implies$ $Turing()$ is not a prgoram
  - $Turing$ is a simple function constructed from $HALT$
  - $\therefore$ $Turing()$ DNE $\implies$ $HALT()$ DNE

Undecidable Problems

-

CS70 - Lecture 14 Notes

Name: Felix Su SID: 25794773

Spring 2016 GSI: Gerald Zhang

Counting

Tree Counting: Slow

- Build up string by bits, total amount of leaves is total possibilities

First Rule of Counting: Product Rule:

- If objects constructed from a sequence of choices $n_1, n_2, ..., n_k$
- Total number of objects $= n_1 \times n_2 \times \cdots \times n_k$

Counting Functions/Polynomials

- There are $|T|^{|s|}$ functions $f : S \to T$
  - $|T|$ choices for mapping of $f(s_i)$ (Use product rule)
- p^{d+1} polynomials of degree d mod p
  - p choices for each of the $d + 1$ coefficients

Permutations

- Derived from the first rule of counting (product rule)
- Choose from less items each step
- Permutations of n objects: number of orderings of n objects (no replacements)
  - $n \times (n-1) \times (n-2) \times \cdots \times 1 = n!$
- Number of one to one functions $|S| \to |S|$
  - Decreasing choices every step: $|S| \times |S| - 1 \times \cdots \times 1 = |S|!$

Permutation Formula

- Number of different samples of saize k from n numbers **without replacement**

$$nPk = n \times (n-1) \times (n-2) \times \cdots \times (n-(k-1)) = \frac{n!}{(n-k)!} \tag{1}$$

Counting Sets: When order doesn't matter

Second Rule of Counting: Order Doesn't Matter (Combination):

- If order doesn't matter, count the number of ordered objects (permutations) and divide by number of orderings
- Choose k out of n possibilities

$$\binom{n}{k} = nCk = \frac{n!}{k!(n-k)!} \tag{2}$$

Sampling:

- Sample k items out of n
- Without replacement:
  - If order matters (first rule): $\frac{n!}{(n-k!)}$
  - If order does not matter (second rule): $\frac{n!}{k!(n-k!)}$
- With replacement:
  - If order matters (first rule): n^k
  - **see Stars and Bars formula (3)**

Anagrams:

- First rule on total number of letters N: $N!$ total permutations
- Divide by the number of duplicate permutations generated due to D duplicate letters: First rule: $D!$
- total distinct permutations = $\frac{N!}{A!B!\cdots D!}$ (can have multiple duplicate sets of letters)

Stars and Bars:

- Ways k people split n things
- Ways to add up k numbers to sum to n
- k undordered choices from set of n possibilities
- $\binom{\text{total} + (\text{sections - 1})}{\text{sections - 1}}$

$$\binom{n+k-1}{k-1} \tag{3}$$

Summary

First Rule (Product)

- k samples
- With replacement: n^k
- Without replacement: $\frac{n!}{(n-k)!}$

Second Rule (Division)

- When order doesn't matter (sometimes): can divide
- Without replacement (order doesn't matter): $\binom{n}{k} = \frac{n!}{(n-k)!k!}$ n choose k
  - You pick a different object every time. The total amount of orderings for your k objects is $k!$, so divide sample without replacement by $k!$ because order doesn't matter

One-to-one Rule

- Equal in number if one-to-one (Bijection)
- With replacement (order doesn't matter): $\binom{k+n-1}{n-1}$

CS70 - Lecture 15 Notes

Name: Felix Su SID: 25794773

Spring 2016 GSI: Gerald Zhang

Review Turing and Halt

- Turing(P) includes Halt(P,P)
- If halts (infinite loop), if doesn't halt, halt (diagonalization)
- Halt program exists ⇒ Turing program exists
- Turing("Turing") interpret program 'Turing' as text
  - Neither halts nor loops (diagonlized statement on itself)
  - Therefore, Turing program DNE
  - No Turing program ⇒ No Halt program (Contrapositive)

Review Stars and Bars

Wikipedia: Stars and Bars

- Choose n from k with replacement, order doesn't matter
- Stars and Bars Bijection (Counting rule)
- Place n (total) objects in k bins
- k bins are distinguishable, objects are not
- $k-1$ bars represent k bins
- Thm 1 (positive nums): Each bin must contain an object, there can only be ≤ 1 bar between each star. You must choose $k-1$ bars from the $n-1$ available positions.
- Thm 2 (non-negative nums): Bins can contain any no. of objects, there can be $\geq k-1$ bars between each star. You must choose $k-1$ bars from the $n+(k-1)$ available positions.

With Replacement/Order Doesn't Matter:

$$\text{Positive Groups} = \binom{n-1}{k-1} \tag{1}$$

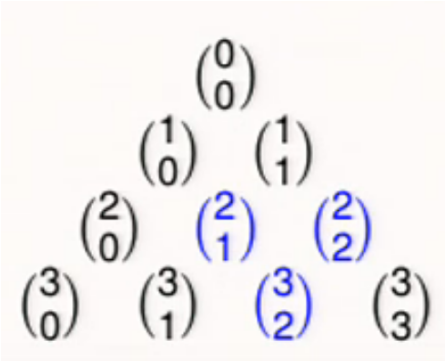
$$\text{Non-Negative Groups} = \binom{n+(k-1)}{k-1} \tag{2}$$

Combinatorial Proofs

- Define what the Left side counts and what the Right side counts, then equate them
- Ask same question for both sides and answer them using the correct approach corresponding to the side

- $\binom{n}{k} = \binom{n}{n-k}$
  - Left: Ways to choose k out of n items
  - Right: Ways to (not) choose $n-k$ out of n items, which is the same as choose k out of n items
- $\binom{n}{k} = \binom{n-1}{k-1} + \cdots + \binom{k-1}{k-1}$
  - Left: Ways to choose k out of n items
  - Right: Sum number of subsets that include the first i items and the number of subsets that do not include the first i items
- $2^n = \binom{n}{kn} + \binom{n}{n-1} + \cdots + \binom{n}{0}$: **Binomial Thm**
  - Sum of coefficients of an $(1+x)^n$ binomial (nth row in Pascal's Triangle)
  - Left: No. of subsets of n choices (element i is either in or out of the subset, 2 poss.)
  - Right: Sum of $\binom{n}{i}$ from i to n
    - $\binom{n}{i}$ ways to choose i elements of n choices

Pascal's Triangle



- Row n = coefficients of $(1+x)^n$
- Choose 2^n terms: 1 or x from $(1+x)$
  - Combine all terms corresponding to x^k
  - Coefficient of x^k is $\binom{n}{k}$: you choose k factors (products) that include x and there are n x's to choose from

Pascal's Rule:

- Left: No. of k subsets from $n+1$ choices
- Right: No. of subsets that choose the first item + No. of subsets that do not choose the first item = Left
  - $\binom{n}{k-1}$: No. of subsets of n that contain the first item. (Take away first item, left with n items and $k-1$ remaining choices)
  - $\binom{n}{k}$: No. of k subsets of n that do not contain the first item. (Take away first item, left with n items, but still need to make k choices.

$$\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1} \tag{3}$$

Simple Inclusion/Exclusion

- **Sum Rule: For disjoint sets** S and T : $|S \cup T| = |S| + |T|$
- **Inclusion/Exclusion Rule: For any** S and T : $|S \cup T = |S| + |T| - |S \cap T|$
  - Ex. No. of 10 digit phone numbers that have 7 as the first or second digit
  - S = first digit 7. $|S| = 10^9$
  - T = second digit 7. $|T| = 10^9$
  - $S \cap T$ = first and second digit. $|S \cap T| = 10^8$
  - $|S| + |T| - |S \cap T| = 10^9 + 10^9 - 10^8$

Probability Space

- Random Experiment: Define possible outcomes and likelihoods (percentages) have Statistical regularity
- Set of Ω outcomes: (Ex. $\Omega = \{H, T\}$)
  - Probabilities assigned to each outcome: $\Pr[H] = 0.5, \Pr[T] = 0.5$
  - Elements of Ω describes one outcome of the complete experiment
- Assign probability to each outcome $\Pr[A]$
  - Probabilities assigned to each outcome: Ex. $\Pr[H] = 0.5, \Pr[T] = 0.5$

- Ω = sample space (can be countable or uncountable)
- $\omega \in \Omega$ = sample point
- probability $\Pr[\omega]$ s.t. $0 \leq \Pr[\omega] \leq 1$ and $\sum_{\omega \in \Omega} \Pr[\omega] = 1$

Uniform Probability Space

- Each outcome ω is equally probable: $\Pr[\omega] = \frac{1}{\Omega}$ for all ω
- Ω must be finite

Non-Uniform Probability Space

- Each outcome ω is any $\Pr[\omega]$ s.t. $0 \leq \Pr[\omega] \leq 1$ and $\sum_{\omega \in \Omega} \Pr[\omega] = 1$

CS70 - Lecture 16 Notes

Name: Felix Su SID: 25794773

Spring 2016 GSI: Gerald Zhang

Set Notation Review

- Set A, Complement $\bar{A}$
- Union (In either: or): $A \cup B$
- Intersection (In both: and): $A \cap B$
- Difference (In A, not B) $A \setminus B$
- Symmetric Difference (In only one: xor) $A \Delta B$

Probability

- event E = subset of outcome: $E \subset \Omega$
- Any Sample Space: $\Pr[E] = \sum_{\omega \in E} \Pr[\omega]$
- Uniform Space: $\Pr[E] = \frac{|E|}{|\Omega|}$
- $p_n := \Pr[E_n] = \frac{|E_n|}{|\Omega|}$
  - $p_n := \frac{\binom{n}{k}}{|\omega|^n}$ if $E = n$ coin tosses with exactly k heads

Stirling Formula: (for large n)

- $n! \approx \sqrt{2\pi n}(\frac{n}{e})^n$
- $\Pr[E] = \frac{|E|}{|\Omega|}$
  - Can apply Stirling Formula because $|E|$ and $|\Omega|$ are defined by combinations (factorials)

Probability is Additive

- If events A and B are disjoint, then sum probabilities
- Non-disjoint sets, use Inclusion/exclusion property: $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$
- **Union bound**: $\Pr[A_1 \cup \cdots \cup A_n] \leq \Pr[A_1] + \cdots + \Pr[A_n]$
- If $A_1, ..., A_N$ are a pairwise disjoint partition of Ω and $\cup_{m=1}^{N} A_m = \Omega$, then $\Pr[B] = \Pr[B \cap A_1] + \cdots + \Pr[B \cap A_N]$

Inclusion/Exclusion Property:
$\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$
Union Bound:
$\Pr[A_1 \cup \cdots \cup A_n] \leq \Pr[A_1] + \cdots + \Pr[A_n]$
Law of Total Probability:
If $A_1, ..., A_N$ are a pairwise disjoint partition of Ω and $\cup_{m=1}^{N} A_m = \Omega$ then,
$\Pr[B] = \Pr[B \cap A_1] + \cdots + \Pr[B \cap A_N]$

Conditional Probability

- Probability of A given B
- $\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]}$

Product Rule

$$\Pr[A_1 \cap \cdots \cap A_n] = \Pr[A_1]\Pr[A_2|A_1] \cdots \Pr[A_n|A_1 \cap \cdots \cap A_{n-1}] \tag{1}$$

Total Probabability $\times$ Product Rule

$$\Pr[B] = \Pr[A_1]\Pr[B|A_1] + \cdots + \Pr[A_N]\Pr[B|A_N] \tag{2}$$

Causality vs. Correlation

- Events A and B are **positively correlated** if $\Pr[A \cap B] > \Pr[A]\Pr[B]$, but this does not imply causation
- Eliminate external/common causes to test causality

Bayes Rule

- Let m = number of situations where A and B occurred, and n = number of situations where $\bar{A}$ and B occurred.
- Therefore: $\Pr[A|B] = \frac{m}{m+n}$

Bayes Rule (Simplified using Law of Total Probability)

$$\Pr[A_n|B] = \frac{\Pr[A_n]\Pr[B|A_n]}{\sum_m \Pr[A_m]\Pr[B|A_m]} = \frac{\Pr[A_n]\Pr[B|A_n]}{\Pr[B]} \tag{3}$$

Independence

- Two events A and B are independent if $\Pr[A \cap B] = \Pr[A]\Pr[B]$
- Two events A and B are independent if and only if $\Pr[A|B] = \Pr[A]$
- $\Pr[A]$ decreases/increases given B

If A and B are **independent** sets:

$$\Pr[\bar{A} \cap \bar{B}] = 1 - A \cup B \tag{4}$$

$$\Pr[A \cap B] = \Pr[A] + \Pr[B] - \Pr[A \cup B] \tag{5}$$

CS70 - Lecture 17 Notes

Name: Felix Su SID: 25794773

Spring 2016 GSI: Gerald Zhang

Review

- Example: $B \subset A \Rightarrow A$ and B are positively correlated
  - $\Pr[A|B] = 1 > \Pr[A]$ and $\Pr[A \cap B] = \Pr[B] > \Pr[A]\Pr[B]$
- Example: $A \subset B = \emptyset \Rightarrow A$ and B are negatively correlated
  - $\Pr[A|B] = 0 < \Pr[A]$ and $\Pr[A \cap B] = 0 < \Pr[A]\Pr[B]$
- For uniformly distributed probability space Ω, $\Pr[A] = \frac{|A|}{|\Omega|}$

Probability of A given B:

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]} \tag{1}$$

Probability of A and B (intersection):

$$\Pr[A \cap B] = \Pr[B]\Pr[A|B] = \Pr[A]\Pr[B|A] \tag{2}$$

A and B are positively correlated if:

$$\Pr[A|B] > \Pr[A] \quad , \quad \Pr[A \cap B] > \Pr[A]\Pr[B] \tag{3}$$

A and B are negatively correlated if:

$$\Pr[A|B] < \Pr[A] \quad , \quad \Pr[A \cap B] < \Pr[A]\Pr[B] \tag{4}$$

A and B are independent iff:

$$\Pr[A|B] = \Pr[A] \quad , \quad \Pr[A \cap B] = \Pr[A]\Pr[B] \tag{5}$$

Find prior probability given some observation B (A given B)

1. Total probability of B given prior probabilities
   - **Law of Total probability**
   - $\Pr[B] = \Pr[A_1]\Pr[B|A_1] + \cdots + \Pr[A_n]\Pr[B|A_n]$
2. Find $\Pr[A|B]$
   - **Bayes Rule**
   - $\Pr[A|B] = \frac{\Pr[A]\Pr[B|A]}{\Pr[B]}$

Terms

- **Most likely A Posteriori (MAP) of** B: The A_m that gives the highest $\Pr[A_m]\Pr[B|A_m]$
- **Maximum Likelihood Estimate (MLE) of** B: The A_m that gives the highest $\Pr[B|A_m]$

Mutual Independence

- A subset of events $A_1, ..., A_k$ where $A_k, k \in J$ are **mutually independent** if the probability that they all occur is equal to the product of their individual probabilities

Mutual Independence
Definition

$$\Pr[\cap_{k\in K} A_k] = \prod_{k\in K} \Pr[A_k], \text{ for all finite } K \subseteq J \tag{6}$$

Theorem

- If the events $\{A_j, j \in J\}$ are mutually independent, and if K_n are pairwise disjoint finite subsets of J, then all the events $\cap_{k\in K_n} A_k$ are independent (same is true if we replace some of the A_k by $\bar{A}_k$

Collision Calculation
Let m = no. of elements, n= no. of bins, C = collision

$$\Pr[\bar{C}] \approx e^{(-\frac{m^2}{2n})} \tag{7}$$

When $m = 1.2\sqrt{n}$

$$\Pr[C] \approx \frac{1}{2} \tag{8}$$

Collision Derivation
If A_i = no collision when the ith ball is placed in a bin

$$\Pr[A_i | A_{i-1} \cap \cdots \cap A_1] = 1 - \frac{i-1}{n} \tag{9}$$

No collisions = $A_1 \cap \cdots \cap A_m$
Product Rule:

$$\Pr[A_1 \cap \cdots \cap A_m] = \Pr[A_1]\Pr[A_2|A_1]\cdots\Pr[A_m|A_1 \cap \cdots \cap A_{m-1}] \tag{10}$$

Apply to $\Pr[\bar{C}]$:

$$\Pr[\bar{C}] = (1 - \frac{1}{n})\cdots(1 - \frac{m-1}{n}) \tag{11}$$

Natural log of both sides:

$$\ln(\Pr[\bar{C}]) = \sum_{k=1}^{m-1} \ln(1 - \frac{k}{n}) \approx \sum_{k=1}^{m-1} \ln(-\frac{k}{n})^{*} = (-\frac{1}{n})(\frac{m(m-1)}{2}) \approx -\frac{m^2}{2n} \tag{12}$$

* Use property that $\ln(1-\varepsilon) \approx -\varepsilon$ for $|\varepsilon| << 1$
Gauss Summation: $1 + 2 + \cdots + m - 1 = \frac{m(m-1)}{2}$

Example: Checksums

- m = no. of files, b = no. of bits in the checksum, C = files share a checksum
- Find b s.t. $\Pr[C] \leq 10^{-3}$
  - * $\Pr[C] \approx 1 - e^{(-\frac{m^2}{2(2^b)})}$
  - * $b = \frac{\ln(-\frac{m^2}{2\ln(1-10^{-3})})}{\ln(2)} = 2.9\ln(m) + 9$
- $\therefore b \geq 2.9\ln(m) + 9$

Probability of Getting n_i out of n with m picks

- Define event of failure A_m (not success)
- Determine probability of failing on each iteration of m
  - $\Pr[A_i|A_{i-1} \cap \cdots \cap A_1] = 1 - \Pr[\bar{A}_i]$ for $i = \{1, ..., m\}$)
  - If not intuitive, try brute force and find a pattern for each $\Pr[A_i]$
- Use Product Rule to get $\Pr[A_m]$
  - $\Pr[A_1 \cap \cdots \cap A_m] = \Pr[A_1]\Pr[A_2|A_1] \cdots \Pr[A_m|A_1 \cap \cdots \cap A_{m-1}]$
  - If events are **independent** $\Pr[A_1 \cap \cdots \cap A_m] = \Pr[A_1]\Pr[A_2] \cdots \Pr[A_m|A_{m-1}]$
- Take natural log of both sides and simplify using the property that $\ln(1 - \varepsilon) \approx -\varepsilon$ for $|\varepsilon| << 1$
- Raise e to the power of both sides (e^n) to derive approximate solution for $\Pr[A_m]$
  - $\Pr[A_m] \approx e^{expression}$

Probability of Complete Collection

- Define event of failure of one iteration E_k
  - E_k for $k = \{1, ..., n\}$
  - Derive $\Pr[E_k]$ using method above: **Probability of Getting n_i out of n with m picks**
- find probability of failing any iteration (or/union)
  - $p := \Pr[E_1 \cup E_2 \cup \cdots \cup E_n]$
- Estimate p using Union Bound
  - $p := \Pr[E_1 \cup E_2 \cup \cdots \cup E_n] \leq \Pr[E_1] + \Pr[E_2] + \cdots + \Pr[E_n]$
- Plug in $\Pr[E_k]$ expression derived above to find Pr[failure of at least one iteration] $\leq expression$
- Use expression to derive minimum value of m to get a certeain $\Pr[miss]$ s.t. $\Pr[miss] \leq p$

CS70 - Lecture 18 Notes

Name: Felix Su SID: 25794773

Spring 2016 GSI: Gerald Zhang

Random Variables

- Random variable is a known, deterministic, function that maps outcome to a number (onto, but not necessarily one-to-one)
- Random Variable X for an experiment with sample space Ω is a function $X : \Omega \to \mathbb{R}$
  - X assigns $X(\omega) \in \mathbb{R}$ to each $\omega \in \Omega$
  - X is not random, nor a variable, it is called a random variable, because its outcome depends on the intial probability/experiment (varies from experiment to experiment)
  - After deriving X and knowing the initial probabilities, can determine the likelihood of each outcome of X
- $X^{-1}(A) = \{\omega | X(\omega) = A\}$: Inverse image of value A
  - Set of outcomes that map to A

Distribution

- The probability of X taking on a value A.
- Definition: The distribution of a random variable X is $\{(a, \Pr[X = a]) : a \in \mathbb{A}\}$ where $\mathbb{A}$ is the range of X
- $\Pr[X = A] = \Pr[X^{-1}(A)]$

Random Variable Method

- Examine the elements of the Sample Space (ex. $\{HHH, THH, HTH, TTH, HHT, THT, HTT, TTT\}$
- Define the Random Variables by given params ex. +1 on heads, -1 on tails $\Rightarrow \{3, 1, 1, -1, 1, -1, -1, -3\}$
- Determine the probability of any generalized event (ex. getting i heads and $n - i$ tails if the prob. of getting heads is p: $p^i(1-p)^{n-i}$)
- Multiply the above probability by the amount of times it occurs (summation) (ex. All ways to get i heads out of n flips: $\binom{n}{i}$)
- This determines the probability of each element in the Random Variable: the distribution of outcomes. (ex. Binomial Distribution)

Binomial Distribution

$$B(n, p) : \Pr[X = i] = \binom{n}{i} p^i (1-p)^{n-1}, i \in \{0, \cdots, n\} \tag{1}$$

- Flip n coins with probability p to get heads, Random var: No. of heads
- Ways to choose i heads out of n flips: $\binom{n}{i}$
- Determine probability of $\omega = i$ heads (probability of heads in any position is p): Pr of i heads and $n-1$ tails is p^i and $(1-p)^{n-i}$ respectively $\therefore \Pr[\omega] = p^i(1-p)^{n-i}$
- Probability of $X = i$ is the sum of all $\Pr[\omega]$ where ω contains i heads: $\binom{n}{i}$

Error Channel

- Apply Binomial Distribution
- Packet is corrupted with probability p
- Send $n + 2k$ packets
- Find probability of at most k corruptions
- $\sum_{i \leq k} \binom{n+2k}{i} p^i (1-p)^{n+2k-i}$: Sum gets total probability of all i corruptions s.t. $i \leq k$
- For RS Code, choose k s.t. the above probability is large

Combining Random Variables

- Let X and Y be two Random Vars in the same Probability space
- Then, $X + Y$ is an RV that assigns value $X(\omega) + Y(\omega)$ to ω
- General Case: $g(X, Y, Z)$ assigns value $g(X(\omega), Y(\omega), Z(\omega))$ to ω

Expectation

- $E[X] = \sum_a a \times \Pr[X = A] \approx \frac{X_1 + \cdots + X_n}{N}$
- Random variable X, X has a possible values (gains). Multiply each possible gain a by the probability of RV $X = a$ and sum over all RVs to get expected value.
- Average $= E(X)$ holds for uniform probability
- Expectation is linear

Expectation Theorem

$$E[X] = \sum_{\omega} X(\omega) \times \Pr[\omega] \tag{2}$$

- Sum of all products of X_i and the probability of getting that X_i (also called the mean by frequentist interpretation): **Law of Large Numbers**

Expectation Derivation Methods

- Two ways to compute the mean value
  - Given: Distribution of X (set of values a and their probabilities)
    - $E[X] = \sum_a a \times \Pr[X = A] \approx \frac{X_1 + \cdots + X_n}{N}$
  - Given: Probability Space
    - Sum over all ω's in probability space
    - $E[X] = \sum_{\omega} X(\omega) \times \Pr[\omega]$

Example of Both Derivation Methods

- Flip fair coin 3 times
- $\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$
- $X =$ number of H's $: \{3, 2, 2, 2, 1, 1, 1, 0\}$
- Method 1: $E[X] = \sum_a a \times \Pr[X = A] = 3(\frac{1}{8}) + 2(\frac{3}{8}) + 1(\frac{3}{8}) + 0(\frac{1}{8}) = \frac{3}{2}$
- Method 2: $E[X] = \sum_{\omega} X(\omega) \times \Pr[\omega] = (3 + 2 + 2 + 2 + 1 + 1 + 1 + 0)(\frac{1}{8})$

CS70 - Lecture 19 Notes

Name: Felix Su SID: 25794773

Spring 2016 GSI: Gerald Zhang

Random Variables

Real number values assigned to each outcome

- **Definition:** A function X that assigns a real number $X(\omega)$ to each $\omega \in \Omega$
- Set of outcomes s.t. the RV assigned to that outcome $X(\omega)$ is some value $a \in \mathbb{R}$
  - Defined as the inverse image of number (RV) a under the function X.
  - Ex: Two dice roll, RV = total of both dice = 4 (3 possible outcomes) $X^{-1}(4) = \{(1,3),(2,2),(3,1)\}$
- Set of outcomes s.t. the RV to that outcome $X(\omega)$ is some value in $A \in \mathbb{R}$
- Probability that RV $X = a$ is the same as the probability of getting an outcome that maps to a
- The **distribution** of RV X is the set of possible RV values paired with their respective probabilities.

Distribution

All coordinate pairs of X $(\text{RV}, \Pr[\text{RV}])$

Combining Random Variables

- Let X, Y, Z be RVs on Ω and function $g : \mathbb{R}^3 \mapsto \mathbb{R}$
- $g(X, Y, Z)$ = RV that assigns value $g(X(\omega), Y(\omega), X(\omega))$ to ω
- Ex: Three dice roll; X, Y, Z = values of each die; $g(X, Y, Z)$ = max value of X, Y, Z

Set of outcomes s.t. the RV assigned to that outcome $X(\omega)$ is some value $a \in \mathbb{R}$:

$$X^{-1}(a) := \{\omega \in \Omega | X(\omega) = a\} \tag{1}$$

Set of outcomes s.t. the RV to that outcome $X(\omega)$ is some value in $A \in \mathbb{R}$:

$$X^{-1}(A) := \{\omega \in \Omega | X(\omega) = A, A \in \mathbb{R}\} \tag{2}$$

Probability that RV $X = a$ is the same as the probability of getting an outcome that maps to a:

$$\Pr[X = a] = \Pr[X^{-1}(a)] \text{ and } \Pr[X = A] = \Pr[X^{-1}(A)] \tag{3}$$

Distribution

$$\{(a, \Pr[X = a]) : a \in \mathscr{A}\} \text{ where } \mathscr{A} = \{X(\omega), \omega \in \Omega\} \tag{4}$$

Expectation

> 1. Multiply each RV in the distribution of X with its respective probability
> 2. Sum all products

- **Definition:** the **expected value** (meanor expectation) of a random variable X
- Not a common value: Expected value may not be a possible value of X

Law of Large Numbers

> Expectation = average value per experiment if it is performed many times

Expected Value:

$$E[X] = \sum_a a \times \Pr[X = a] \tag{5}$$

Thm: Can sum over outcomes instead of RVs

$$E[X] = \sum_\omega X(\omega) \times \Pr[\omega] \tag{6}$$

Law of Large Numbers: When $n >> 1$

$$E[X] = \frac{X_1 + \cdots + X_n}{n} \tag{7}$$

Indicators

> Random variable that is 1 when ω is in desired event A and 0 otherwise

- **Definition:** Let A = event; **Indicator** of event A = RV X:

$$X = \begin{cases} 1, & \text{if } \omega \in A \\ 0, & \text{if } \omega \notin A \end{cases}$$

Expectation of Indicator

$$E[X] = 1 \times \Pr[X = 1] + 0 \times \Pr[X = 0] = \Pr[A] \tag{8}$$

Alternative form of indicator

$$X(\omega) = 1\{\omega \in A\} \text{ or } 1_A(\omega) \tag{9}$$

$$X = 1_A \tag{10}$$

Linearity of Expectation

- Expectation is linear

Examples

> **Roll dice n times**
>
> - X_m = number of dots on roll m; $X = X_1 + \cdots + X_n$ = total number of dots after n rolls
> - $E[X] = E[X_1 + \cdots + X_n]$
> - $= E[X_1] + \cdots + E[X_n]$ (by linearity)
> - $= nE[X_1]$ because all X_m have the same distribution
> - $E[X_1] = 1 \times \frac{1}{6} + \cdots + 6 \times \frac{1}{6} = \frac{6 \times 7}{2} \times \frac{1}{6} = \frac{7}{2}$

> **Flip n coins with heads prob. $= p$ and RV $X =$ no. of heads**
>
> - Hard method:
> - $\Pr[X = i] = \binom{n}{i} p^i (1-p)^{n-i}$
> - $E[X] = \sum_i i \times \Pr[X = i] = \sum_i i \times \binom{n}{i} p^i (1-p)^{n-i}$
> - Linearity Method:
> - Used X_i as an indicator: 1 if ith flip is heads, 0 otherwise
> - $E[X_i] = 1 \times \Pr[H] + 0 \times \Pr[T] = p$
> - $X = X_1 + \cdots + X_n$
> - $E[X] = E[X_1] + E[X_2] + \cdots + E[X_n] = n \times E[X_i] = np$

Linear Expectation

$$E[a_1 X_1 + \cdots + a_n X_n] = a_1 E[X_1] + \cdots + a_n E[X_n] \tag{11}$$

Union of Indicators

$$1_{A \cup B}(\omega) = 1_A(\omega) + 1_B(\omega) - 1_{A \cap B}(\omega) \tag{12}$$

Probability of Event = Expected value of Indicator RV

$$\Pr[A] = E[1_A] \tag{13}$$

Calculating E[g(x)]

- Let $Y = g(X)$. Assume we know the distribution of X
- Method 1 **(bad)**: Calculate distribution of Y
  - $\Pr[Y = y] = \Pr[X \in g^{-1}(y)]$ where $g^{-1}(x) = \{x \in \mathbb{R} : g(x) = y\}$
- Method 2 **(good)**: Use following Theorem
  - $E[g(X)] = \sum_x g(x) \Pr[X = x]$

Method 2 Example

> Let X be uniform in $\{-2, -1, 0, 1, 2, 3\}$; $g(X) = X^2$
> $E[g(X)] = \sum_{x=-2}^{3} x^2 \frac{1}{6} = \frac{19}{6}$

Calculate E[g(X)]

$$E[g(X)] = \sum_x g(x)\Pr[X = x] \tag{14}$$

Calculate E[g(X,Y,Z)]

$$E[g(X,Y,Z)] = \sum_{x,y,z} g(X,Y,Z)\Pr[X = x, Y = y, Z = z] \tag{15}$$

Least Squares

1. Least Squares: $(X - a)^2$ is used to denote the amount of error
2. $a = E[X]$ minimizes $E[(X - a)^2]$, so it is a god guess for X

- **Thm:** The value of a that minimizes $E[(X - a)^2]$ is $a = E[X]$ $\therefore$ if you only know the distribution of X, $E[X]$ is a good guess for X

Least Absolute Deviation

1. Least Absolute Deviation: $|X - a|$ is used to denote the amount of error
2. $a =$ median of X minimizes $E[|X - a|]$, so it is a god guess for X

- **Thm:** The value of a that minimizes $E[|X - a|]$ is $a =$ median of X $\therefore$ if you only know the distribution of X, the median of X is a good guess for X

Monotonicity

Mean value of a bigger RV is bigger than the mean value of a smaller RV

- Let X, Y be 2 RVs on Ω
- $X \leq Y$ is $X(\omega)$ is always less than $Y(\omega)$ for all $\omega \in \Omega$, vice verse for $X \geq Y$
- $X \geq a$ for some constant a if $X(\omega)$ is always greater than a
- **Facts**
  - $X \geq 0 \Rightarrow E[X] \geq 0$
  - $X \leq Y \Rightarrow E[X] \leq Y$

Uniform Distribution

RV X is equally likely to take on any of its values

- X is uniformly distributed in $\{1, 2, \cdots, n\}$ if $\Pr[X = m] = \frac{1}{n}$ for $m = 1, 2, 3 \cdots, n$
- $E[X] = \sum_{m=1}^{n} m\Pr[X = m] = \sum_{m=1}^{n} m \times \frac{1}{n} = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2}$

Geometric Distribution

1. Flip a coin with $\Pr[H] = p$ until you get H
2. Geom. Dist. w/ Parameter p: $\Pr[X = n] = (1-p)^{n-1}p, n \geq 1$
3. Mean value $E[X]$ will increase as p become **smaller** and vice versa

- Flip a coin with $\Pr[H] = p$ until you get H
- Let $X =$ no. of flips until first H
- $X(\omega_n) = n$
- $\Pr[X = n] = (1-p)^{n-1}p, n \geq 1$

Geometric Distribution w/ Parameter p:

$$\Pr[X = n] = (1-p)^{n-1}p, n \geq 1 \tag{16}$$

Sum of Geometric Series:

$$\text{if } |a| < 1, S := \sum_{n=0}^{\infty} a^n = \frac{1}{1-a} \tag{17}$$

CS70 - Lecture 20 Notes

Name: Felix Su SID: 25794773

Spring 2016 GSI: Gerald Zhang

Expectation

Expected Value Defintion:

$$E[X] := \sum_x x\Pr[X = x] = \sum_\omega X(\omega)\Pr[\omega] \tag{1}$$

Expectation of Function of RVs

$$E[g(X,Y)] = \sum_{x,y} g(x,y)\Pr[X = x, Y = y] = \sum_\omega g(X(\omega), Y(\omega))\Pr[\omega] \tag{2}$$

Linearity of Expectation

$$E[aX + bY + c] = aE[X] + bE[Y] + c \tag{3}$$

Uniform Distribution

1. RV X equally likely to take on any of its values
2. $E[X]$ of uniformly distributed X is avg of all outcomes

Uniform Distribution Example

$$\Pr[X = m] = \frac{1}{n} \text{ for } m = 1, 2, \cdots, n \tag{4}$$

Expectation Uniform Distribution

$$E[X] = \sum_{m=1}^{n} m\Pr[X = m] = \sum_{m=1}^{n} m\frac{1}{n} = \frac{n+1}{2} \tag{5}$$

Geometric Distribution

1. Let X = Number of flips of coin with $\Pr[H] = p$ until we get H, so $X(\omega_n) = n$
2. Higher $p \Rightarrow$ smaller expected X

Geometric Distribution

$$\Pr[X = n] = (1-p)^{n-1}p, n \geq 1 \tag{6}$$

Use Geometric Sum

$$\sum_{n=1}^{\infty} \Pr[X_n] = \sum_{n=1}^{\infty} (1-p)^{n-1}p = p\sum_{n=0}^{\infty}(1-p)^n = \frac{p}{1-(1-p)} = 1 \tag{7}$$

Derive Geometric Sum

If $|a| < 1$ and $S := \sum_{n=0}^{\infty} a^n = \frac{1}{1-a}$:

$$S = 1 + a + a^2 + a^3 + \cdots \tag{8}$$
$$aS = \quad a + a^2 + a^3 + \cdots \text{ (shifted right 1)} \tag{9}$$
$$(1-a)S = 1 + a - a + a^2 - a^2 + \cdots = 1 \text{ (subtract above two terms)} \tag{10}$$

Geometric Distribution Expectation

1. $X =_D G(p)$ where ${}_DG(p) \equiv \Pr[X = n] = (1-p)^{n-1}p, n \geq 1$

- $E[X] = \sum_{n=1}^{\infty} n\Pr[X = n] = \sum_{n=1}^{\infty} n(1-p)^{n-1}p$

Expectation of Geometric Distribution

$$E[X] = \sum_{n=1}^{\infty} n\Pr[X = n] = \sum_{n=1}^{\infty} n(1-p)^{n-1}p = \frac{1}{p} \tag{11}$$

Derive Expectation of Geometric Distribution

$$E[X] = p + 2(1-p)p + 3(1-p)^2p + 4(1-p)^3p + \cdots \tag{12}$$
$$(1-p)E[X] = \quad (1-p)p + 2(1-p)^2p + 3(1-p)^3p + \cdots \text{ (shifted right 1)} \tag{13}$$
$$pE[X] = p + (1-p)p + (1-p)^2p + (1-p)^3p + \cdots = \sum_{n=1}^{\infty} \Pr[X = n] = 1 \text{ (subtract above two terms)} \tag{14}$$

Time to Collect Coupons

- Note: $H(n) = 1 + \frac{1}{2} + \cdots + \frac{1}{n}$ (Harmonic Number)
- $H(n) = 1 + \frac{1}{2} + \cdots + \frac{1}{n} \approx \int_1^n \frac{1}{x}dx = \ln(n)$
- $H(n) \approx \ln(n) + \gamma$ where $\gamma \approx 0.58$ (Euler-Mascheroni constant) to account for bars sticking above graph

X - time to get n coupons
$\Pr[\text{get ith coupon}|\text{got i-1 coupons}] = \frac{n-(i-1)}{n} = \frac{n-i+1}{n}$
$E[X_i] = \frac{1}{p} = \frac{n}{n-i+1}, i = 1, 2, \cdots, n$
$E[X] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} \frac{n}{n-i+1} = n(1 + \frac{1}{2} + \cdots + \frac{1}{n}) =: nH(n) \approx n(\ln n + \gamma)$

Stacking

- Cards have width 2
- Keep putting cards 1/2 to the right
- At center of mass, all mass on left = all mass on right

- So, if $x =$ dist. between C.O.M and right side of base card and n cards weigh n, $nx = 1 - x$ and $x = \frac{1}{(n+1)}$
- Induction shows that the C.O.M after n cards is $H(n)$ away from the rightmost edge of the bas card

Geometric Distribution: Memoryless

1. If $X = G(p)$, probability of any X occurring is not dependent on previous events

- Let $X = G(p)$. For $n \geq 0$: $\Pr[X \geq n] = \Pr[\text{first n flips are T}] = (1-p)^n$
- **Thm**: $\Pr[X > n + m | X > n] = \Pr[X > m], m, n \geq 0$
- $\Pr[X > n + m | X > n] = \dfrac{\Pr[X > n + m \cap X > n]}{\Pr[X > n]} = \dfrac{\Pr[X > n + m]}{\Pr[X > n]} = \dfrac{(1-p)^{n+m}}{(1-p)^n} = (1-p)^m = \Pr[X > m]$

Memoryless Geometric Distribution Theorem

$$\Pr[X > n + m | X > n] = \Pr[X > m], m, n \geq 0 \tag{15}$$

Expectation of Natural Numbers (works for Geometric Distribution)

- **Thm:** For RV X that takes values $\{0, 1, 2, \cdots\}$: $E[X] = \sum_{i=1}^{\infty} \Pr[X \geq i]$

For Natural Number RVs X

$$E[X] = \sum_{i=1}^{\infty} \Pr[X \geq i] \tag{16}$$

Poisson

1. For binomial distribution where $\Pr[H] = \lambda/n$
2. $X = P(\lambda) \iff \Pr[X = m] \approx \frac{\lambda^m}{m!} e^{-\lambda}$
3. Expecation of Poisson Distribution $= \lambda$

- Flip coin n times where $\Pr[H] = \lambda/n$
- RV $X =$ no. of heads (Binomial), thus $X = B(n, \lambda/n)$
- **Poisson Distribution** is the distribution of X "for large n" and λ is constant
- Binomial Representation: $\Pr[X = m] = \binom{n}{m} p^m (1-p)^{n-m}$, with $p = \lambda n$

Poisson Proof

- $\Pr[X = m] = \frac{n(n-1)\cdots)(n-m+1)}{m!} (\frac{\lambda}{n})^m (1 - \frac{\lambda}{n})^{n-m} = \frac{n(n-1)\cdots)(n-m+1)}{n^m} (\frac{\lambda^m}{m!}) (1 - \frac{\lambda}{n})^{n-m}$
- $\approx (1)(\frac{\lambda^m}{m!})(1 - \frac{\lambda}{n})^{n-m} \approx (\frac{\lambda^m}{m!})(1 - \frac{\lambda}{n})^n$ (Because $n >> m$)
- $\approx \frac{\lambda^m}{m!} e^{-\lambda}$ (Because $(1 - \frac{a}{n})^n \approx e^{-a}$)

Poisson Expectation Proof

- $E[X] = \sum_{m=1}^{\infty} m \times \frac{\lambda^m}{m!} e^{-\lambda} = e^{-\lambda} \sum_{m=1}^{\infty} \frac{\lambda^m}{(m-1)!} = e^{-\lambda} \sum_{m=0}^{\infty} \frac{\lambda^{m+1}}{m!} = e^{-\lambda}\lambda \sum_{m=0}^{\infty} \frac{\lambda^m}{m!} = e^{-\lambda}\lambda e^{\lambda} = \lambda$

Poisson Distribution

$$\Pr[X = m] \approx \frac{\lambda^m}{m!} e^{-\lambda}, m \geq 0 \tag{17}$$

Expectation of Poisson

$$E[X] = \lambda \tag{18}$$

Distributions Summary

Uniform Distribution

$$U[1, \cdots, n] : \Pr[X = m] = \frac{1}{n}, m = 1, \cdots, n; E[X] = \frac{n+1}{2} \tag{19}$$

Binomial Distribution

$$B(n, p) : \Pr[X = m] = \binom{n}{m} p^m (1-p)^{n-m}, m = 0, \cdots, n; E[X] = np \tag{20}$$

Geometric Distribution

$$G(p) : \Pr[X = n] = (1-p)^{n-1} p, n = 1, 2, \cdots; E[x] = \frac{1}{p} \tag{21}$$

Poisson Distribution

$$P(\lambda) : \Pr[X = m] = \frac{\lambda^m}{m!} e^{-\lambda}, m \geq 0; E[X] = \lambda \tag{22}$$

Independent Random Variables

1. Two RVs X and Y are **independent** if and only if, all the corresponding events are independent.
2. Same as independence of events.
3. RVs are **mutually independent** if the product of their combined intersection (and) is the same as the product of their individual probabilities.
4. Events A, B, C are pairwise (resp. mutually) independent iff their Indicator RVs $1_A, 1_B, 1_C$ are also pairwise independent.

- Independence Thm Proof:
- 'if' left direction: if you choose $A = \{a\}, B = \{b\}$, then the thm eq. is the same as:
- $\Pr[X = a \cap Y = b] = \Pr[X = a]\Pr[Y = b], \forall a, b$
- 'only if' right direction: sum over all possible pairs in A and B of the probability that X is in A and Y is in B
- $\sum_{a \in A}^{\Pr[X \in A \cap Y \in B]} \sum_{b \in B} \Pr[X = a \cap Y = b] = \sum_{a \in A} \sum_{b \in B} \Pr[X = a]\Pr[Y = b]$
- $= \sum_{a \in A} \Pr[X = a] \sum_{b \in B} \Pr[Y = b] = \Pr[X \in A]\Pr[Y \in B]$

Functions of Independent RVs

1. Functions of independent RVs are independent
2. If X, Y, Z are pairwise independent, but not mutually independent, it may indicate that $f(x)$ and $g(Y, Z)$ are not independent
3. Functions of disjoint collections of mutually independent RVs are also mutually independent

- Independent functions of inpdendent RVs Proof:
  - Definition of Inverse Image: $h(z) \in C \iff z \in h^{-1}(C) := \{z | h(z) \in C\}$
  - $\Pr[f(X) \in A, g(Y) \in B] = \Pr[X \in f^{-1}(A) \cap Y \in g^{-1}(B)]$, by def. of Inv. Im.
  - $= \Pr[X \in f^{-1}(A)]\Pr[Y \in g^{-1}(B)]$, because X, Y ind.
  - $= \Pr[f(x) \in A]\Pr[g(Y) \in B]$
- Functions of disjoint collections of mutually independent RVs are also mutually independent
  - Let $\{X_n, n \geq 1\}$ be mutually independent. And $Y_1 := f(X_1, X_2), Y_2 := f(X_3, X_4), Y_3 := f(X_5, X_6)$
  - Then, Y_1, Y_2, Y_3 are mutually independent

Mean($E[X]$) of product of Ind. RVs

1. Expectation of XY is equal to exp. of X times exp. of Y

- Proof:
- $E[g(x, y, z)] = \sum_{x,y} g(x, y)\Pr[X = x \cap Y = y]$, so, $E[XY] = \sum_{x,y} xy\Pr[X = x \cap Y = y] = \sum_{x,y} xy\Pr[X = x]\Pr[Y = y]$
- $= \sum_{x} x\Pr[X = x] \sum_{y} y\Pr[Y = y] = E[X]E[Y]$

Independence of X and Y

$$\Pr[Y = b | X = a] = \Pr[Y = b], \forall a, b \tag{23}$$

$$\Pr[X = a \cap Y = b] = \Pr[X = a]\Pr[Y = b], \forall a, b \tag{24}$$

Thm: X and Y are independent iff

$$\Pr[X \in A \cap Y \in B] = \Pr[X \in A]\Pr[Y \in B], \forall A, B \subset \mathbb{R} \tag{25}$$

Independent Functions

$$\text{If } X, Y \text{ are independent RVs } \implies f(X), g(Y) \text{ are independent } \forall f(), g() \tag{26}$$

Mean of product of Independent RV (Only for Independent RVs)

$$\text{If } X, Y \text{ are ind. RVs } \implies E[XY] = E[X]E[Y] \tag{27}$$

X, Y, Z are Mutually Independent If

$$\Pr[X = x, Y = y, Z = z] = \Pr[X = x]\Pr[Y = y]\Pr[Z = z], \forall x, y, z \tag{28}$$

Independent RV Examples
X, Y, Z are pairwise independent and $U[1, 2, \cdots, n]$
Let $E[X] = E[Y] = E[Z] = 0, E[X^2] = E[Y^2] = E[Z^2] = 1$

- $E[(X+2Y+3Z)^2] = E[X^2+4Y^2+9Z^2+4XY+12YZ+6XZ]$
- $= 1+4+9+(4)(0)(0)+12(0)(0)+6(0)(0) = 14$ (Because Independent RV product and Lin. of Exp.)

- $E[(X-Y)^2] = E[X^2+Y^2-2XY] = 2E[X^2]-2E[X]^2$
- $= \frac{1+3n+2n^2}{3} - \frac{(n+1)^2}{2}$

CS70 - Lecture 21 Notes

Name: Felix Su SID: 25794773

Spring 2016 GSI: Gerald Zhang

Distributions Review

1. **Uniform** $(U[1, \cdots, n]); m = 1, \ldots, n$
   - $\Pr[X = m] = \frac{1}{n}$
   - $E[X] = \frac{n+1}{2}$
2. **Binomial or Bernoulli** $(B(n, p)); m = 0, \ldots, n$
   - $\Pr[X = m] = \binom{n}{m} p^m (1-p)^{n-m}$
   - $E[X] = np$
3. **Geometric** $(G(p)); n = 1, 2, \ldots$
   - $\Pr[X = n] = (1-p)^{n-1} p$
   - $E[X] = \frac{1}{p}$
4. **Poisson** $(P(\lambda)); n \geq 0$
   - $\Pr[X = n] = \frac{\lambda^n}{n!} e^{-\lambda}$
   - $E[X] = \lambda$
5. **Indicator** $X = 1$ **or** 0
   - $\Pr[X = 1] = p$, $\Pr[X = 0] = (1-p)$
   - $E[X] = p$

Poisson and Queuing

(Derivation in previous notes)

1. Flip coin n times, $\Pr[H] = \frac{\lambda}{n}$
2. RV X = no. of heads (Bernoulli indicator -when 1)
3. $X = B(n, \frac{\lambda}{n})$
4. Distribution of X "for large n"

- Distribution of the number of events in an interval
- The average value comes out to λ
- Cut up situation into $n \to \infty$ intervals described by Beronoulli indicators
- This means you can assume no two events occur in the same interval and there is a $\frac{\lambda}{n}$ chance the indicator is 1 in any interval

Independence Review

1. X, Y are independent if and only if:
   - $\Pr[X = x, Y = y] = \Pr[X = x]\Pr[Y = y], \forall x, y$
   - $\Pr[X \in A, Y \in B] = \Pr[X \in A]\Pr[Y \in B], \forall A, B$
2. $X, Y, Z, \ldots$ are mutually indpendent if and only if:
   - $\Pr[X = x, Y = y, Z = z, \ldots] = \Pr[X = x]\Pr[Y = y]\Pr[Z = z] \cdots, \forall x, y, z, \ldots$
   - $\Pr[X \in A, Y \in B, Z \in C, \ldots] = \Pr[X \in A]\Pr[Y \in B]\Pr[Z \in C] \cdots, \forall A, B, C, \ldots$
3. If $U, V, W, X, Y, Z, \ldots$ are all mutually independent then:
   - $f(U, V), g(W, X, Y), h(Z, \ldots), \ldots$ are mutually independent

Variance

1. Measures deviation from the mean value (Standard deviation ($\sigma(X)$) squared)
2. Use squared distance as a continuous function that you can do derivatives and other operations on.
3. To calculate intermediate value $E[X^2]$ of expectations with infinite series:
   - Calculate $E[X^2] - (1 - p)E[X^2] = pE[X^2]$
   - This gives you $pE[X^2]$ in terms of $E[X]$ and a known distribution (total dist.=1)
   - Divide both sides by p
4. Variande of the constant is c that constant squared $\text{Var}[c] = c^2$

- $\text{Var}[X] = E[(X - E[X])^2]$
- $= E[X^2 - 2XE[X] + E[X]^2]$
- $= E[X^2] - 2E[X]E[X] + E[X]^2$
- $= E[X^2] - E[X]^2$

Variance of X

$$\sigma^2(X) := \text{Var}[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2 \tag{1}$$

Potential Final Question:
Give an example where this ratio $\to \infty$

$$\frac{\sigma(X)}{E[|X - E[X]|]} \tag{2}$$

Uniform Variance
Assume that $\Pr[X = i] = \frac{1}{n}$ for $i \in \{1, \ldots, n\}$

$$E[X] = \frac{1}{n}\sum_{i=1}^{n} i = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2} \tag{3}$$

$$E[X^2] = \sum_{i=1}^{n} i^2 \Pr[X = i] = \frac{1}{n}\sum_{i=1}^{n} i^2 = \frac{1 + 3n + 2n^2}{6} \tag{4}$$

$$\text{Var}[X] = \frac{1 + 3n + 2n^2}{6} - \frac{(n+1)^2}{4} = \frac{n^2 - 1}{12} \tag{5}$$

Geometric Distribution Variance
$X = G(p)$, $\Pr[X = n] = (1-p)^{n-1}p$ for $n \geq 1$

$$E[X] = \frac{1}{p} \tag{6}$$

$$\begin{aligned}
E[X^2] &= p + 4p(1-p) + 9p(1-p)^2 + \dots \\
(1-p)E[X^2] &= p(1-p) + 4p(1-p)^2 + 9p(1-p)^3 + \dots \\
pE[X^2] &= 2(p + 2p(1-p) + 3p(1-p)^2 + \dots) \\
&\quad - (p + p(1-p) + p(1-p)^2 + \dots) \\
&= 2E[X] - \text{Distribution} \\
pE[X^2] &= (2E[X] - 1) = (2(\frac{1}{p}) - 1) = \frac{2-p}{p}
\end{aligned}$$

$$E[X^2] = \frac{2-p}{p^2} \tag{7}$$

$$\text{Var}[X] = E[X^2] - E[X]^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2} \tag{8}$$

$$\sigma(X) = \frac{\sqrt{1-p}}{p} \tag{9}$$

Fixed Points Variance
Number of fixed points in a random permutation of n items
i.e. Number of students that get hw back with RV $X = X_1 + X_2 + \cdots + X_n$
X_i = indicator for ith student getting hw back

$$E[X] = 1 \tag{10}$$

$$\begin{aligned}
E[X^2] &= \sum_i E[X_i^2] + \sum_{i \neq j} E[X_i X_j] \\
&= n \times (1 \times \Pr[X_i = 1] + 0 \times \Pr[X_i = 0] = \frac{1}{n}) \\
&\quad + n(n-1) \times (1 \times \Pr[X_i = 1 \cap X_j = 1] + 0 \times \Pr[\text{“anything else”}] = \frac{1 \times 1 \times (n-2)!}{n!} = \frac{1}{n(n-1)})
\end{aligned}$$

$$E[X^2] = 2 \tag{11}$$

$$\text{Var}[X] = E[X^2] - E[X]^2 = 2 - 1 = 1 \tag{12}$$

Binomial
Flip coin with $\Pr[H] = p$
$X = X_1 + X_2 + \cdots + X_n$ no. of heads (X_i, X_j are independent)
$X_i = 1$ if ith flip is heads, $X_i = 0$ otherwise (indicator RV)

$$E[X] = np \tag{13}$$

$$E[X_i^2] = 1^2 \times p + 0^2 \times (1-p) = p$$

$$\text{Var}[X_i] = p - (E[X_i])^2 = p - p^2 = p(1-p)$$

$$\text{Var}[X] = \text{Var}[X_1 + X_2 + \cdots + X_n] = n\text{Var}[X_i] = np(1-p) \tag{14}$$

Properties of Variance

1. $\text{Var}[cX] = c^2\text{Var}[X]$ where c is a constant
2. $\text{Var}[X + c] = \text{Var}[X]$ where c is a constant (shifts center)

Variance of the Sum of 2 Independent RVs

- **Thm:** If X and Y are independent, then $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$
  - Assume $E[X] = E[Y] = 0$ (You can just shift this to accountn for any other RVs)
  - $E[XY] = E[X]E[Y] = 0$ (By Independence)

$$\begin{aligned}\text{Var}[X + Y] &= E[(X + Y)^2] \\ &= E[X^2 + 2XY + Y^2] \\ &= E[X^2] + 2E[XY] + E[Y^2] \\ &= E[X^2] + EY^2 \\ &= \text{Var}[X] + \text{Var}[Y]\end{aligned}$$

Variance of the Sum of Multiple Independent RVs

- **Thm:** If $X, Y, Z, \ldots$ are independent, then $\text{Var}[X + Y + Z + \cdots] = \text{Var}[X] + \text{Var}[Y] + \text{Var}[Z] + \cdots$
  - Assume $E[X] = E[Y] = E[Z] = \cdots = 0$ (You can just shift this to accountn for any other RVs)
  - $E[XY] = E[XZ] = E[YZ] = \cdots = 0$ (By Independence)

$$\begin{aligned}\text{Var}[X + Y + Z + \cdots] &= E[(X + Y + Z + \cdots)^2] \\ &= E[X^2 + Y^2 + Z^2 + \cdots + 2XY + 2XZ + 2YZ + \cdots] \\ &= E[X^2] + E[Y^2] + E[Z^2] + 0 + \cdots + 0 \\ &= \text{Var}[X] + \text{Var}[Y] + \text{Var}[Z] + \cdots\end{aligned}$$

If X and Y are Independent

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] \tag{15}$$

If $X, Y, Z, \ldots$ are Independent

$$\text{Var}[X + Y + Z + \cdots] = \text{Var}[X] + \text{Var}[Y] + \text{Var}[Z] + \cdots \tag{16}$$

Inequalities: Overview

1. Markov: $\Pr[X \geq a] \leq \frac{E[f(X)]}{f(a)}$, for all a such that $f(a) > 0$ and $f : \mathbb{R} \to [0, \infty)$ is non-decreasing
   - Bound the probability of being at least a away from the mean $E[X]$
2. Chabyshev: $\Pr[|X - E[X]| > a] \leq \frac{\text{Var}[X]}{a^2}$
   - Bound probability of getting at least a away from the mean $E[X]$

Markov's Inequality Proof

- Assume $f : \mathbb{R} \to [0, \infty)$ is non-decreasing.
- $\Pr[X \geq a] \leq \frac{E[f(X)]}{f(a)}$, for all a such that $f(a) > 0$

- Observe that: $1\{X \geq a\} \leq \frac{f(X)}{f(a)}$ because:
  - When $X > a$:
    - Left side = 1
    - Right side ¿ 1 because $f : \mathbb{R} \to [0, \infty)$ is non-decreasing
  - When $X = a$:
    - Left side = 1
    - Right side = 1
  - When $X < a$:
    - Left side = 0
    - Right side ≥ 0 because $f \geq 0$.
- Take expectation of both sides (because expectation is monotone):
  - $E[1\{X \geq a\}] \leq E[\frac{f(X)}{f(a)}] \implies \Pr[X \geq a] \leq \frac{E[f(X)]}{f(a)}$

Markov's Inequality
$f : \mathbb{R} \to [0, \infty)$ is non-decreasing and for all a such that $f(a) > 0$

$$\Pr[X \geq a] \leq \frac{E[f(X)]}{f(a)} \tag{17}$$

Markov's Inequality Example $G(p)$:

- Let $X = G(p), E[X] = \frac{1}{p}, E[X^2] = \frac{2-p}{p^2}$
- Using $f(x) = x$: $\Pr[X \geq a] \leq \frac{E[X]}{a} = \frac{1}{ap}$
- Using $f(x) = x^2$: $\Pr[X \geq a] \leq \frac{E[X^2]}{a^2} = \frac{2-p}{a^2p^2}$

Markov's Inequality Example $P(\lambda)$:

- Let $X = P(\lambda), E[X] = \lambda, E[X^2] = \lambda + \lambda^2$
- Using $f(x) = x$: $\Pr[X \geq a] \leq \frac{E[X]}{a} = \frac{\lambda}{a}$
- Using $f(x) = x^2$: $\Pr[X \geq a] \leq \frac{E[X^2]}{a^2} = \frac{\lambda+\lambda^2}{a^2}$

Chebyshev's Inequality Proof

- Let $Y = |X - E[X]|, f(y) = y^2$
- Use Markov's Inequality: $\Pr[Y \geq a] \leq \frac{E[f(Y)]}{f(a)} = \frac{\text{Var}[X]}{a^2}$
- Confirms that the variance measures the "deviations from the mean"

Chebyshev's Inequality
For all $a > 0$

$$\Pr[|X - E[X]| > a] \leq \frac{\text{Var}[X]}{a^2} \tag{18}$$

Chebyshev's Inequality Example $P(\lambda)$:

- Let $X = P(\lambda), E[X] = \lambda, \text{Var}[X] = \lambda$
- $\Pr[|X - \lambda| > n] \leq \frac{\lambda}{n^2}$

Use Markov to get Chebyshev Bounds

- Let $X = P(\lambda), E[X] = \lambda, E[X^2] = \lambda + \lambda^2, \text{Var}[X] = \lambda$

- Using Markov's with $f(x) = x^2$: $\Pr[X \geq a] \leq \frac{E[X^2]}{a^2} = \frac{\lambda+\lambda^2}{a^2}$
- If $a > \lambda$, then $X \geq a \implies X - \lambda \geq a - \lambda > 0 \implies |X - \lambda| \geq a - \lambda$
- So, for $a > \lambda$, $\Pr[X \geq a] \leq \Pr[|X - \lambda| \geq a - \lambda] \leq \frac{\lambda}{(a-\lambda)^2}$

Fraction of H's

- How likely is it that the fraction of H's differs from 50%?
- Let $X_m = 1$ if the mth flip of a fair coin is H and $X_m = 0$ otherwise
  - Any type of polling system mimics this scenario
  - $\Pr[H] = p$ Yes to poll question
  - $\Pr[T] = 1 - p$ No to poll question
- Define $Y_n = \frac{X_1+\cdots+X_n}{n}$, for $n \geq 1$
  - Ratio of people who said yes to question
- Estimate $\Pr[|Y_n - 0.5| \geq 0.1] = \Pr[Y_n \leq 0.4 \text{ or } Y_n \geq 0.6]$
  - If poll result is $\geq 10\%$ away from the mean $\implies$ mistake
- By Chebyshev: $\Pr[|Y_n - 0.5| \geq 0.1] \leq \frac{\text{Var}[Y]_n}{0.1^2} = 100\text{Var}[Y_n] = \frac{25}{n}$
  - $\text{Var}[Y_n] = \frac{1}{n^2}(\text{Var}[X_1] + \text{Var}[X_2] + \cdots + \text{Var}[X_n]) = \frac{n\text{Var}[X_i]}{n^2} = \frac{\text{Var}[X_i]}{n} \leq \frac{1}{4n}$
  - Because all X_i's are independent and the variance of a constant is its square.
  - As $n \to \infty$, $Y_n \to 0$
  - $\text{Var}[X_i] = p(1-p) \leq (.5)(.5) = \frac{1}{4}$
- For $n = 1000$, $\Pr[|Y_n - 0.5| \geq 0.1] \leq 2.5\%$

Law of Large Numbers
As $n \to \infty$, $\Pr[|X - 0.5| \geq \varepsilon] \to 0$

$$\Pr[|T_n - 0.5| \leq \varepsilon] \to 1 \tag{19}$$

Weak Law of Large Numbers

$$\Pr[|\frac{X_1 + \cdots + X_n}{n} - \mu| \geq \varepsilon] \to 0, \text{ as } n \to \infty \tag{20}$$

Weak Law of Large Numbers

Proof

- Let $Y_n = \frac{X_1+\cdots+X_n}{n}$
- $\Pr[|Y_n - \mu| \geq \varepsilon] \leq \frac{\text{Var}[Y_n]}{\mu^2} = \frac{\text{Var}[X_1+\cdots+X_n]}{n^2\varepsilon^2} = \frac{\text{Var}[X_i]}{n\varepsilon^2} \to 0, \text{ as } n \to \infty$

Summary

Variance

$$\mathrm{Var}[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2 \tag{21}$$

Fact

$$\mathrm{Var}[aX + b] = a^2\mathrm{Var}[X] \tag{22}$$

Sum

$$X, Y, Z, \ldots \text{ mutually independent} \implies \mathrm{Var}[X + Y + Z + \cdots] = \mathrm{Var}[X] + \mathrm{Var}[Y] + \mathrm{Var}[Z] + \cdots \tag{23}$$

Markov
$f : \mathbb{R} \to [0, \infty)$ is non-decreasing and for all a such that $f(a) > 0$

$$\Pr[X \geq a] \leq \frac{E[f(X)]}{f(a)} \tag{24}$$

Chebyshev
For all $a > 0$

$$\Pr[|X - E[X]| \geq a] \leq \frac{\mathrm{Var}[X]}{a^2} \tag{25}$$

Weak Law of Large Numbers

$$X_m \text{ i.i.d} \implies \frac{X_1 + \cdots + X_n}{n} \approx E[X] \tag{26}$$

CS70 - Lecture 22 Notes

Name: Felix Su SID: 25794773

Spring 2016 GSI: Gerald Zhang

Confidence Intervals

1. Use Chebyshev: $\Pr[|X - E[X]| \geq a] \leq \frac{\text{Var}[X]}{a^2}$ where $\frac{\text{Var}[X]}{a^2} \leq 5\%$ to find a
2. Define $[A_n - a, A_n + a]$ as a 95% **confidence interval**

Example

- Fip coin n times. Let A_n be the fraction fo Hs
- Know $o = \Pr[[]H] \approx A_n$ for n large (WLLN)
- Use Chebyshev to find a such that $\Pr[p \in [A_n - a, A_n + a]] \geq 95\%$
  - $[A_n - a, A_n + a]$ is a 95% **confidence interval** for p
  - $[A_n - \frac{2.25}{\sqrt{n}}, A_n + \frac{2.25}{\sqrt{n}}]$ is a 95%-CI for p
  - $a = \frac{1}{\sqrt{n}}$ works

Confidence Intervals: Result

- Let X_n be i.i.d. with mean μ and variance σ^2, $A_n = \frac{X_1 + \cdots + X_n}{n}$
- Then: $\Pr[\mu \in [A_n - 4.5\frac{\sigma}{\sqrt{n}}, A_n + 4.5\frac{\sigma}{\sqrt{n}}]] \geq 95\%$
- Thus, $[A_n - 4.5\frac{\sigma}{\sqrt{n}}, A_n + 4.5\frac{\sigma}{\sqrt{n}}]$ is a 95%-CI for μ
  - Let $X_n = 1\{\text{coin } n \text{ yields H}\}$, Then:
  - $\mu = E[X_n] = p = \Pr[H]$ and $\sigma^2 = \text{Var}[X_n] = p(1-p) \leq \frac{1}{4}$
  - Hence: $[A_n - 4.5\frac{0.5}{\sqrt{n}}, A_n + 4.5\frac{0.5}{\sqrt{n}}] = [A_n - \frac{2.25}{\sqrt{n}}, A_n + \frac{2.25}{\sqrt{n}}]$ is a 95%-CI for μ

Confidence Intervals: Result

- $A_n \pm 4.5\frac{\sigma}{\sqrt{n}}$ is a 95%-CI for μ
- Use Chebyshev:
  - $\Pr[|A_n - \mu| \geq 4.5\frac{\sigma}{\sqrt{n}}] \leq \frac{\text{Var}[A_n]}{(4.5\frac{\sigma}{\sqrt{n}})^2} \leq \frac{\frac{\sigma^2}{n}}{20\frac{\sigma^2}{n}} = 5\%$
  - Thus, $\Pr[|A_n - \mu| \leq 4.5\frac{\sigma}{\sqrt{n}}] \geq 95\%$

Confidence Interval

$$A_n \pm 4.5\frac{\sigma}{\sqrt{n}} \text{ is a 95\%-CI for } \mu \tag{1}$$

Chebyshev's Intequality

$$\Pr[|X - E[X]| \geq a] \leq \frac{\text{Var}[X]}{a^2} \tag{2}$$

Linear Regression: Preamble

1. The best guess about Y if we know only the distribution of Y is $E[Y]$.
2. Use observation of some RV X related to Y to improve the guess about Y by constructing a function $g(X)$ to guess Y.

- **Proof**
  - Let $\hat{Y} = Y - E[Y]$. Then, $E[\hat{Y}] = 0$. So, $E[\hat{Y}c] = 0, \forall c$.
  - Now: $E[(Y-a)^2] = E[(Y - E[Y] + E[Y] - a)^2]$
  - $= E[(\hat{Y} + c)^2]$ with $c = E[Y]a$
  - $= E[\hat{Y}^2 + 2\hat{Y}c + c^2] = E[\hat{Y}^2] + 2E[\hat{Y}c] + c^2$
  - $= E[\hat{Y}^2] + 0 + c^2 \geq E[\hat{Y}^2]$.
  - Hence, $E[(Ya)^2] \geq E[(Y - E[Y])^2], \forall a$.

Linear Regression: Motivation

1. Best linear fit: **Linear Regression**

- Minimizes Least Squared distance from the line and the points.

Covariance

1. Defintiion: $\text{cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$
2. $E[X] = 0 \cap E[Y] = 0 \implies \text{cov}(X,Y) = E[XY]$
3. $\text{cov}(X,Y) > 0 \implies X$ and Y are positively correlated
4. $\text{cov}(X,Y) < 0 \implies X$ and Y are negatively correlated
5. $\text{cov}(X,Y) = 0 \implies X$ and Y are uncorrelated
6. **Properties**
   - $\text{Var}[X] = \text{cov}(X,X)$
   - X, Y independent $\implies \text{cov}(X,Y) = 0$
   - $\text{cov}(a+X, b+Y) = \text{cov}(X,Y)$
   - $\text{cov}(aX+bY, cU+dV) = ac.\text{cov}(X,U) + ad.\text{cov}(X,V) + bc.\text{cov}(Y,U) + bd.\text{cov}(Y,V)$

- Covariance Fact Proof:
  - $E[(XE[X])(YE[Y])] = E[XYE[X]YXE[Y] + E[X]E[Y]]$
  - $= E[XY]E[X]E[Y]E[X]E[Y] + E[X]E[Y]$
  - $= E[XY]E[X]E[Y]$.
- $E[X] = 0 \cap E[Y] = 0 \implies \text{cov}(X,Y) = E[XY]$
- $\text{cov}(X,Y) > 0 \implies$ RVs X and Y are large or small together. So, X and Y are said to be positively correlated.
- $\text{cov}(X,Y) < 0 \implies$ when X is larger, Y tends to be smaller. So, X and Y are said to be negatively correlated.
- $\text{cov}(X,Y) = 0 \implies X$ and Y are uncorrelated.

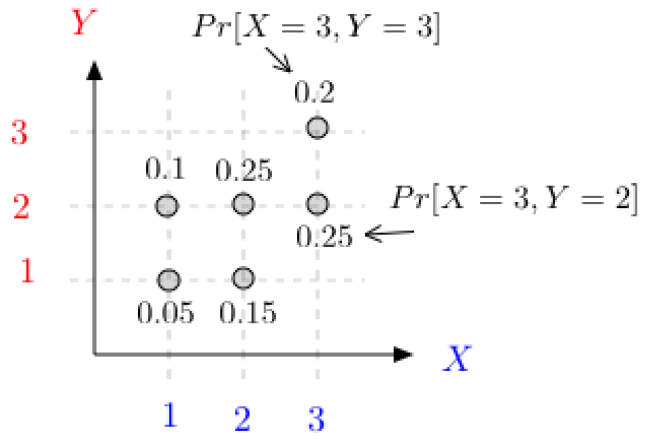
Covariance Definition

$$\text{cov}(X,Y) = E[(X - E[X])(Y - E[Y])] \tag{3}$$

Covariance Fact

$$\text{cov}(X,Y) = E[XY] - E[X]E[Y] \tag{4}$$

Covariance Calculation Example



$E[X] = 1 \times 0.15 + 2 \times 0.4 + 3 \times 0.45 = 1.9$
$E[X^2] = 1^2 \times 0.15 + 2^2 \times 0.4 + 3^2 \times 0.45 = 5.8$
$E[Y] = 1 \times 0.2 + 2 \times 0.6 + 3 \times 0.2 = 2$
$E[XY] = 1 \times 0.05 + 1 \times 2 \times 0.1 + \cdot + 3 \times 3 \times 0.2 = 4.85$
$\text{cov}(X,Y) = E[XY] - E[X]E[Y] = 1.05$
$\text{Var}[X] = E[X^2] - E[X]^2 = 2.19$

LLSE

Covariance

1. LLSE slope correponds with covariance sign (logical)
2. Tip: Use trick of $-\hat{Y} + \hat{Y}$, then expand and use linearity of $E[]$

Proof 1:

- $Y - \hat{Y} = (Y - E[Y]) - \frac{cov(X,Y)}{\text{Var}[X]}(X - E[X])$
  - $E[(Y - E[Y]]$ and $E[(X - E[X])] \implies E[Y - \hat{Y}] = 0$
- $E[(Y - \hat{Y})X] = 0$
  - $E[(Y - \hat{Y})X] = E[(Y - \hat{Y})(X - E[X])]$ because $E[(Y - \hat{Y})E[X]] = 0$
  - $E[(Y - \hat{Y})(X - E[X])] = E[(Y - E[Y])(X - E[X])] - \frac{\text{cov}(X,Y)}{\text{Var}[X]}E[(X - E[X])(X - E[X])] =$ (*) $\text{cov}(X,Y) - \frac{\text{cov}(X,Y)}{\text{Var}[X]}\text{Var}[X] = 0$
  - (*) Recall that $\text{cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$ and $\text{Var}[X] = E[(X - E[X])^2]$
- $E[(Y - \hat{Y})(c + dX)] = 0$
  - Expand to $E[c(Y - \hat{Y})] + dE[(Y - \hat{Y})X] = 0$
  - This means any linear function applies, so, $E[(Y - \hat{Y})(\hat{Y} - a - bX)] = 0, \forall a, b$

- Any mean squared error between Y and $a+bX$ will be greater than or equal to the the mean squared error between $\hat{Y}$ and $a+bX$
  - $E[(Y-a-bX)^2] = E[(Y-\hat{Y}+\hat{Y}-a-bX)^2] = E[(Y-\hat{Y})^2] + E[(\hat{Y}-a-bX)^2] + 0 \geq E[(Y-\hat{Y})^2]$
  - This shows that $E[(Y-\hat{Y})^2] \leq E[(Y-a-bX)^2]$, for all (a,b)
  - Thus $\hat{Y}$ is the LLSE.

Proof 2:

- $Y-a-bX$
  - $= Y-EY-(a-EY)-b(X-EX)+bEX$
  - $= Y-EY-(a-EY+bEX)-b(X-EX)$
  - $= Y-EY-c-b(X-EX)$
- with $c = a-E[Y]+bE[X]$.
- From the first part, we know that the best values of c and b are:
  - $c=0$
  - $b = \text{cov}(X-E[X], Y-E[Y])/var(X-E[X]) = \frac{\text{cov}(X,Y)}{\text{Var}[X]}$
- Thus, $0 = c = a-E[Y]+bE[X]$, so that $a = E[Y]-bE[X]$.
- Hence, $a+bX = E[Y]-bE[X]+bX = E[Y]+b(X-E[X]) = E[Y]+\frac{\text{cov}(X,Y)}{\text{Var}[X]}(X-E[X])$.

Estimation Error

- $E[|Y-L[Y|X]|^2] = E[(Y-E[Y]-(\text{cov}(X,Y)/\text{Var}[X])(X-E[X]))^2] = E[(Y-E[Y])^2]-2(\text{cov}(X,Y)/\text{Var}[X])E[(Y-E[Y])(X-E[X])] + (\text{cov}(X,Y)/\text{Var}[X])^2E[(X-E[X])^2] = \text{Var}[Y] - \frac{\text{cov}(X,Y)^2}{\text{Var}[X]}$
- Without observations, the estimate is $E[Y]$, the error is $\text{Var}[Y]$
- Observing X reduces the error. If X and Y are highly correlated, the mean squared error is reduced (intuitive)

LLSE where X and Y have a given dist. $\Pr[X=x, Y=y]$

$$L[Y|X] = E[Y] + \frac{\text{cov}(X,Y)}{\text{Var}[X]}(X-E[X]) \tag{5}$$

CS70 - Lecture 23 Notes

Name: Felix Su SID: 25794773

Spring 2016 GSI: Gerald Zhang

Review: LLSE and LR

Let X and Y be RVs on Ω

Covariance

$$\text{cov}(X,Y) = E[XY] - E[X]E[Y] \tag{1}$$

LLSE (Given Distribution of X)

$$L[Y|X] = E[Y] + \frac{cov(X,Y)}{\text{Var}[X]}(X - E[X]) = a + bX \text{ where } a, b \text{ minimizes } E[(Y-a-bX)^2] \tag{2}$$

Mean Squared Error of LLSE (Given Distribution of X)

$$E[(Y - L[Y|X])^2] = \text{Var}[Y] - \frac{\text{cov}(X,Y)^2}{\text{Var}[X]} \tag{3}$$

Non-Bayesian (LR)

Calculate each component of $L[Y|X]$

Given samples $(X_1,Y_1),\ldots,(X_K,Y_K)$, no distribution

Define RVs (X,Y) s.t. $\Pr[(X,Y) = (X_k,Y_k)] = \frac{1}{K}, k = 1,\ldots,K$

$$E[X] = \frac{1}{K}\sum_{k=1}^{K} X_k,\ E[Y] = \frac{1}{K}\sum_{k=1}^{K} Y_k,\ E[X^2] = \frac{1}{K}\sum_{k=1}^{K} X_k^2,\ E[XY] = \frac{1}{K}\sum_{k=1}^{K} X_k Y_k \tag{4}$$

$$\text{cov}(X,Y) = E[XY] - E[X]E[Y] \tag{5}$$

$$\text{Var}[X] = E[X^2] - E[X]^2 \tag{6}$$

1. LR line goes through $(E[X], E[Y])$
2. Slope is $\frac{cov(X,Y)}{\text{Var}[X]}$

Conditional Expectation

1. $E[Y|X]$ is the best guess about Y given X

- Let X and Y be RVs on Ω. The conditional expectation of Y given X is defined as $E[Y|X] = g(X)$
  - where $g(x) = E[Y|X=x] = \sum_y y\Pr[Y=y|X=x]$,
  - with $\Pr[Y=y|X=x] = \frac{\Pr[X=x,Y=y]}{\Pr[X=x]}$
  - Theorem: $E[Y|X]$ is the best guess about Y given X
  - That is, for any function $h()$, one has $E[(Y-h(X))^2] \geq E[(Y-E[Y|X])^2]$

Conditional Expectation: $E[Y|X]$ is the best guess about Y given X
Given $\Pr[Y = y|X = x] = \frac{\Pr[X=x, Y=y]}{\Pr[X=x]}$

$$E[Y|X] = E[Y|X = x] = \sum_y y \Pr[Y = y|X = x] \tag{7}$$

Calculating E[Y—X]

- Linearity also applies to E[Y—X]
- Calculate $E[2 + 5X + 7XY + 11X^2 + 13X^3Z^2|X]$ assuming X, Y, Z are i.i.d with mean 0 and variance 1.
- Solve for $E[Y|X] = g(X)$
  - $= 2 + 5X + 7XE[Y|X] + 11X^2 + 13X^3E[Z^2|X]$
  - $= 2 + 5X + 7XE[Y] + 11X^2 + 13X^3E[Z^2]$
  - $= 2 + 5X + 11X^2 + 13X^3(\text{Var}[Z] + E[Z]^2)$
  - $= 2 + 5X + 11X^2 + 13X^3$

Projection Property

- **Claim:** $E[(Y - E[Y|X])f(X)] = 0, \forall f()$ in other words, $E[Yf(X)] = E[E[Y|X]f(X)]$
- In particular, choosing f(x) = 1, we get $E[Y] = E[E[Y|X]]$
- **Proof:** $E[E[Y|X]f(X)] =$
  - $\sum_x E[Y|X = x]f(x)PrX = x$
  - $= \sum_x \sum_y yf(x)\Pr[Y = y|X = x]PrX = x$
  - $= \sum_x \sum_y yf(x)\Pr[X = x, Y = y] = E[Yf(X)]$

Projection Property

$$E[Yf(X)] = E[E[Y|X]f(X)] \tag{8}$$

Smoothing Property

$$E[Y] = E[E[Y|X]] \tag{9}$$

Conditional Probability Application

1. **Find $E[X]$:**
2. Use formula: $E[Y|X] = \sum_y y\Pr[Y = y|X = x]$ and given values for $Y = y$ and $X = x$ to solve for $E[Y|X]$ in terms of X
3. Use fact that $E[Y] = E[E[Y|X]]$ and previous value to solve for $E[Y]$ in terms of $E[X]$
4. Use algebra to solve for $E[X]$

- Given, $X_n = m$, $X_{n+1} = m - 1$ w.p. $\frac{m}{N}$ (picks red ball) and $X_{n+1} = m$ otherwise
  - $E[X_{n+1}|X_n = m] = m - \frac{m}{N} = m\frac{N-1}{N} = X_n\rho$ w. $\rho = \frac{N-1}{N}$
  - So, $E[X_{n+1}] = E[E[X_{n+1}|X_n]] = \rho E[X_n], n \geq 1 \implies E[X_n] = \rho^{n-1}E[X_1] = N(\frac{N-1}{N})^{n-1}, n \geq 1$

Application: Going Viral

- **Fact:** Let $X = \sum_{n=1}^{\infty} X_n$. Then, $E[X] < \infty$ iff $pd < 1$
- **Proof:**
- Given $X_n = k, X_{n+1} = B(kd, p)$. Hence, $E[X_{n+1}|X_n = k] = kpd$
- Thus, $E[X_{n+1}|X_n] = pdX_n$. Consequently, $E[X_n] = (pd)^{n-1}, n \geq 1$.
- If $pd < 1$, then $E[X_1 + \cdots + X_n] \leq (1 - pd) - 1 \implies E[X] \leq (1 - pd) - 1$
- If $pd \geq 1$, then for all C one can find n s.t. $E[X] \geq E[X_1 + \cdots + X_n] \geq C$.
- In fact, one can show that $pd \geq 1 \implies \Pr[X = \infty] > 0$

Wald's Identity

- Assume that $X_1, X_2, \ldots$ and Z are independent, where Z takes values in $\{0, 1, 2, \ldots\}$
- and $E[X_n] = \mu$ for all $n \geq 1$
- Then, $E[X_1 + \cdots + X_Z] = \mu E[Z]$
- **Proof:**
  - $E[X_1 + \cdots + X_Z|Z = k] = \mu k$
  - Thus, $E[X_1 + \cdots + X_Z|Z] = \mu Z$
  - Hence, $E[X_1 + \cdots + X_Z] = E[\mu Z] = \mu E[Z]$

Wald's Identity
If $X_1, X_2, \ldots$ and Z are independent, where $Z \in \{0, 1, 2, \ldots\}$ and $E[X_n] = \mu, \forall n \geq 1$

$$E[X_1 + \cdots + X_Z] = \mu E[Z] \tag{10}$$

Conditional Expectation Best Guess (CE=MMSE)

> 1. $E[Y|X]$ is the best guess about Y based on X.
> 2. Specifically, it is the function $g(X)$ that minimizes $E[(Y - g(X))^2]$

- **Thm:** $CE = MMSE$
  - $g(X) = E[Y|X]$ is the function of X that minimizes $E[(Y - g(X))^2]$
- **Proof:**
  - Let $h(X)$ be any function of X.
  - Then $E[(Y - h(X))^2] = E[(Y - g(X) + g(X) - h(X))^2]$
  - $= E[(Y - g(X))^2] + E[(g(X) - h(X))^2] + 2E[(Y - g(X))(g(X) - h(X))]$
  - But, $E[(Y - g(X))(g(X) - h(X))] = 0$ by the projection property.
  - Thus, $E[(Y - h(X))^2] \geq E[(Y - g(X))^2]$

Summary

1. $E[Y]$ Best single guess for Y
2. $L[Y|X]$ Linear Least Squares Guess for Y given X
3. $E[Y|X]$ Best Estimate of Y given X

CS70 - Lecture 24 Notes

Name: Felix Su SID: 25794773

Spring 2016 GSI: Gerald Zhang

Two State Markov Chain

1. Describes a random motion in $\{0, 1\}$

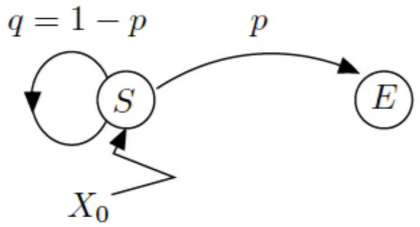
Finite Markov Chain

1. What happens in the future only depends on the current state (amnesic, but successive states are dependent on previous value)
2. A finite set of states: $\mathscr{X} = \{1, 2, \ldots, K\}$
3. A probability distribution π_0 on $\mathscr{X}$: $\pi_0(i) \geq 0, \sum_i \pi_0(i) = 1$
4. Transition probabilities: $P(i, j)$ for $i, j \in \mathscr{X}$
   - $P(i, j) \geq 0, \forall j; jP(i, j) = 1, \forall i$
5. $\{X_n, n \geq 0\}$ is defined so that
   - X_n = state at time n from time $0, 1, \ldots$
   - Define how you start: $\Pr[X_0 = i] = \pi_0(i), i \in \mathscr{X}$ (initial distribution)
   - Define how you move: $\Pr[X_{n+1} = j | X_0, \ldots, X_n = i] = P(i, j), i, j \in \mathscr{X}$
     - $P(i, j)$ does not depend on what happened in the past or time.

Markov Chain Calculations

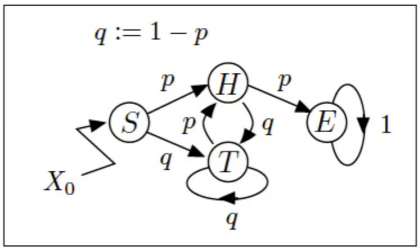
First Passage Time: Example 1

- Flip a coin with $\Pr[H] = p$ until we get H (Use Markov Chain to determine why it will take $\frac{1}{p}$ flips on average $(G(p))$
- Define a Markov Chain:
  - $X_0 = S$ (start)
  - $X_n = S$ for $n \geq 1$ if the last flip was T w. no H yet
  - $X_n = E$ for $n \geq 1$, if we already got H (end)



- Let $\beta(S)$= the avg. time until we reach E, starting from S, then...
- **Claim**: $\beta(S) = 1 + q\beta(S) + p0$ decomposes into:
  - First step (1)
  - Returns to S: still need $\beta(S)$ steps to get to E w. prob. q ($q\beta(S)$)
  - Got to E (Found heads, needs 0 steps to get to E w. prob p ($p0$)
- Subtract $q\beta(S)$ from both sides to get $\beta(S) = \frac{1}{p}$
- Time until E is $G(p)$, so the mean of $G(p)$ is $\frac{1}{p}$

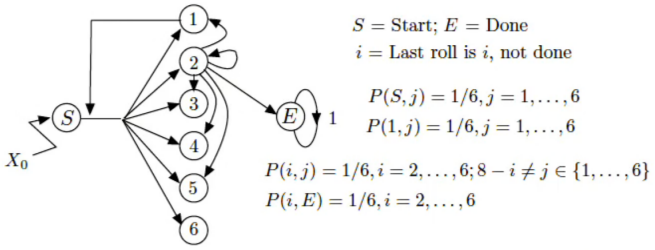
First Passage Time: Example 2



S: Start
H: Last flip $= H$
T: Last flip $= T$
E: Done

- Let $\beta(i)$= avg. time from state i until E (end)
- **First Step Equations**
  - $\beta(S) = 1 + p\beta(H) + q\beta(T)$
  - $\beta(H) = 1 + p0 + q\beta(T)$
  - $\beta(T) = 1 + p\beta(H) + q\beta(T)$
- **Solve:** $\beta(S) = 2 + 3qp^{-1} + q^2p^{-2}$ (E.g., $\beta(S) = 6$ if $p = 1/2$)

First Passage Time: Example 3



S = Start; E = Done
i = Last roll is i, not done

$P(S, j) = 1/6, j = 1, \ldots, 6$
$P(1, j) = 1/6, j = 1, \ldots, 6$
$P(i, j) = 1/6, i = 2, \ldots, 6; 8 - i \neq j \in \{1, \ldots, 6\}$
$P(i, E) = 1/6, i = 2, \ldots, 6$

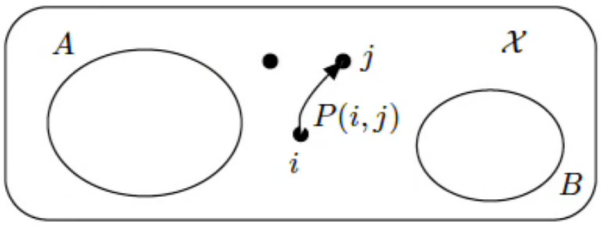
The arrows out of $3, \ldots, 6$ (not shown) are similar to those out of 2.

- $\beta(S) = 1 + \frac{1}{6}\sum_{j=1}^{6} \beta(j)$
- $\beta(1) = 1 + \frac{1}{6}\sum_{j=1}^{6} \beta(j)$
- $\beta(i) = 1 + \frac{1}{6}\sum_{j=1,\ldots,6; j\neq 8i} \beta(j), i = 2, \ldots, 6$
- Symmetry: $\beta(2) = \cdots = \beta(6) = \gamma$. Also, $\beta(1) = \beta(S)$.
  - Thus, $\beta(S) = 1 + (5/6)\gamma + \beta(S)/6; \gamma = 1 + (4/6)\gamma + (1/6)\beta(S)$
  - $\implies \cdots \beta(S) = 8.4$

Summary:

First Step Equations

- Given X_n is a Markov Chain on $\mathscr{X}$ and $A, B \subset \mathscr{X}$ with $A \cap B = \emptyset$



- Define $T_A = min\{n \geq 0 | X_n \in A\}$ and $T_B = min\{n \geq 0 | X_n \in B\}$
- Let $\beta(i) = E[T_A | X_0 = i]$ and $\alpha(i) = \Pr[T_A < T_B | X_0 = i], i \in X$
- $\beta(i)$ denotes a timestep so it adds 1
  - $\beta(i) = 0, i \in A$
  - $\beta(i) = 1 + \sum_j P(i,j)\beta(j), i \notin A$
- $\alpha(i)$ denotes probabilities, so there is no 1
  - $\alpha(i) = 1, i \in A$
  - $\alpha(i) = 0, i \in B$
  - $\alpha(i) = \sum_j P(i,j)\alpha(j), i \notin A \cup B$

CS70 - Lecture 25 Notes

Name: Felix Su SID: 25794773

Spring 2016 GSI: Gerald Zhang

Markov Chain Review

1. Markov Chain:
   - Finite MC set $\mathscr{X}$
   - Initial Distribution π_0
   - Transition Probabilities $P = \{P(i,j), i,j \in \mathscr{X}\}$
   - $\Pr[X_0 = i] = \pi_0(i), i \in \mathscr{X}$
   - $\Pr[X_{n+1} = j | X_0, \ldots, X_n = i] = P(i,j), i,j \in \mathscr{X}, n \geq 0$
   - Note: $\Pr[X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n] = \pi_0(i_0)P(i_0,i_1)P(i_{n-1},i_n)$.
2. First Passage Time:
   - $A \cap B = \emptyset; \beta(i) = E[T_A | X_0 = i]; \alpha(i) = \Pr[T_A < T_B | X_0 = i]$
   - $\beta(i) = 1 + \sum_j P(i,j)\beta(j); \alpha(i) = \sum_j P(i,j)\alpha(j)$

First Passage Time:
Given disjoint sets of states $A \cap B = \emptyset$
Expected Timesteps to get to state in A

$$\beta(i) = E[T_A | X_0 = i] = 1 + \sum_j P(i,j)\beta(j) \tag{1}$$

Probability of reaching A before B

$$\Pr[T_A < T_B | X_0 = i] \tag{2}$$

Distribution of X_n

1. Use $\pi_n = \pi_0 P^n$ function to check if it converges to a vector that does depend on π_0 or not

- Let $\pi_m(i) = \Pr[X_m = i], i \in X$. Note that
- $\Pr[X_{m+1} = j] = \sum_i \Pr[X_{m+1} = j, X_m = i]$
  - $= \sum_i \Pr[X_m = i]\Pr[X_{m+1} = j | X_m = i]$
  - $= \sum_i \pi_m(i)P(i,j)$
  - Hence, $\pi_{m+1}(j) = \sum_i \pi_m(i)P(i,j), \forall j \in X$.
- With π_m, π_{m+1} as a row vectors, these identities are written as $\pi_{m+1} = \pi_m P$.

– Thus, $\pi_1 = \pi_0 P, \pi_2 = \pi_1 P = \pi_0 PP = \pi_0 P^2, \ldots$
– Hence, $\pi_n = \pi_0 P^n, n \geq 0$

Distribution of X_n
Given that $\pi_m(i) = \Pr[X_m = i], i \in \mathscr{X}$

$$\pi_{m+1}(j) = \sum_i \pi_m(i) P(i,j), \forall j \in X \tag{3}$$

With π_m, π_{m+1} as row vectors

$$\pi_{m+1} = \pi_m P \tag{4}$$

General case of π_n

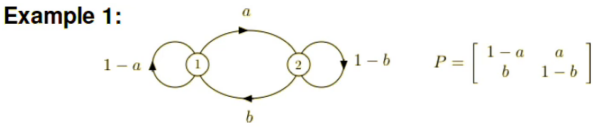
$$\pi_n = \pi_0 P^n, n \geq 0 \tag{5}$$

Balance Equations

1. A distribution π_0 such that $\pi_m = \pi_0, \forall m$ is said to be an invariant distribution ($\pi_0 P = \pi_0$).

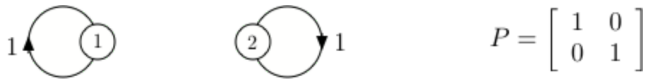
- **Theorem** A distribution π_0 is **invariant** iff $\pi_0 P = \pi_0$. These equations are called the balance equations.
- **Proof:** $\pi_n = \pi_0 P^n$, so that $\pi_n = \pi_0, \forall n$ iff $\pi_0 P = \pi_0$
  - π_0 is invariant $\implies$ the distribution of X_n is always equal to X_0.
  - This does not mean that X_n does not move. It means that the probability that it leaves a state i is equal to the probability that it enters state i.
  - The balance equations say that $\sum_j \pi_(j) P(j,i) = \pi_(i)$
  - That is, $\sum_{j \neq i} \pi_(j) P(j,i) = \pi_(i)(1 - P(i,i)) = \pi_(i) \sum_{j \neq i} P(i,j)$.
  - Thus, $\Pr[\text{enter } i] = \Pr[\text{leave } i]$.

Example 1:



$\pi P = \pi \iff \begin{bmatrix}\pi(1) & \pi(2)\end{bmatrix} \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix} = \begin{bmatrix}\pi(1) & \pi(2)\end{bmatrix}$
$\iff \pi(1)(1-a) + \pi(2)b = \pi(1)$ and $\pi(1)a + \pi(2)(1-b) = \pi(2)$
$\iff \pi(1)a = \pi(2)b$
Equations are redundant, so add an equation: $\pi(1) + \pi(2) = 1$. Then we find
$\pi = \begin{bmatrix}\frac{b}{a+b} & \frac{a}{a+b}\end{bmatrix}$

Example 2:



$\pi P = \pi \iff \begin{bmatrix}\pi(1), \pi(2)\end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
$= \begin{bmatrix}\pi(1), \pi(2)\end{bmatrix} \iff \pi(1) = \pi(1)$ and $\pi(2) = \pi(2)$
Every distribution is invariant for this Markov chain. This is obvious, since $X_n = X_0$ for all n. Hence, $\Pr[X_n = i] = \Pr[X_0 = i], \forall (i,n)$

Irreducibility

1. MC is **irreducible** if it can go from every state i to every state j in any amount of steps

Existence and Uniqueness of Invariant Distribution

- **Theorem:** A **finite irreducible** Markov chain has **one and only one invariant distribution**.
  - There is a unique positive vector $\pi = \begin{bmatrix}\pi(1) & \dots & \pi(K)\end{bmatrix}$ such that $\pi P = \pi$ and $\sum_k \pi(k) = 1$
- **Fact:** If a Markov chain has **two different invariant distributions** π and ν, then it has **infinitely many invariant distributions**.
  * $p\pi + (1-p)\nu$ is then invariant since $[p\pi + (1-p)\nu]P = p\pi P + (1-p)\nu P = p\pi + (1-p)\nu$

Finite irreducible Markov chain has one and only one invariant distribution
There is a unique positive vector $\pi = \begin{bmatrix}\pi(1) & \dots & \pi(K)\end{bmatrix}$ such that

$$\pi P = \pi \text{ and } \sum_k \pi(k) = 1 \tag{6}$$

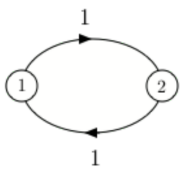
Long Term Fraction of Time in States

- **Theorem** Let X_n be an irreducible Markov chain with invariant distribution π.
- Then, for all i, $\frac{1}{n}\sum_{m-0}^{n-1} 1\{X_m = i\} \to \pi(i)$, as $n \to \infty$.
- The left-hand side is the fraction of time that $X_m = i$ during steps $0, 1, \dots, n-1$. Thus, this fraction of time approaches $\pi(i)$.

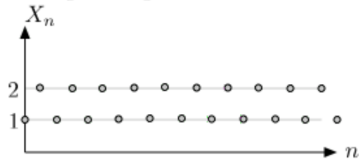
Long Term Fraction of Time in States

$$\text{for all } i, \frac{1}{n}\sum_{m-0}^{n-1} 1\{X_m = i\} \to \pi(i) \text{ as } n \to \infty \tag{7}$$

Example 1

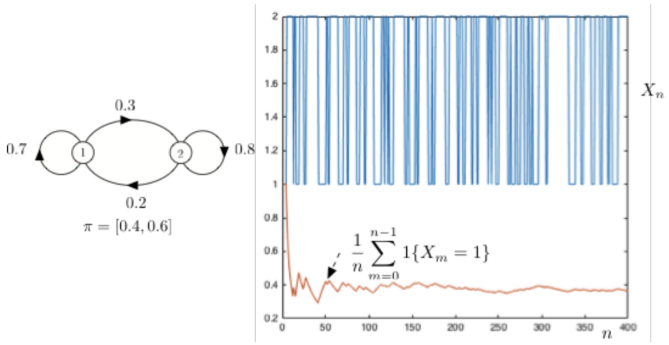


$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad \pi P = \pi \Rightarrow \pi = [1/2, 1/2]$$



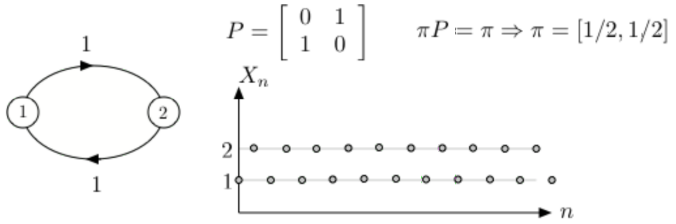
- The fraction of time in state 1 converges to 1/2, which is $\pi(1)$.

Example 2



Convergence to Invariant Distribution

- Assuming the MC is irreducible π_n does not necessarily approach a unique invariant distribution π
- **Example:**



> Assume $X_0 = 1$. Then $X_1 = 2, X_2 = 1, X3_{=}2, \ldots$
> Thus, if $\pi 0 = [1, 0], \pi 1 = [0, 1], \pi 2 = [1, 0], \pi 3 = [0, 1]$, etc.
> Hence, π_n does not converge to $\pi = [1/2, 1/2]$.

Periodicity

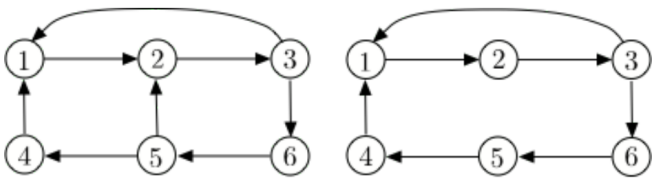
> 1. If the Markov chain is irreducible, $d(i)$ is the same for all i. Check for 1 state.

- **Definition** If $d(i) = 1$, the Markov chain is said to be **aperiodic**.
  - Otherwise, it is periodic with period $d(i)$.
- **Theorem** (see below)
  - Gcd of the set of all numbers of steps it takes to go from from state i, back to state i where the probability of that path is greater that 0
  - Proof: See Lecture notes 24.

Theorem: Periodicity
Assume that the MC is irreducible

$$d(i) := g.c.d.\{n > 0 | \Pr[X_n = i | X_0 = i] > 0\} \text{ has the same value for all states } i \tag{8}$$

Example



$[A] : \{n > 0 | \Pr[X_n = 1 | X_0 = 1] > 0\} = \{3, 6, 7, 9, 11, ...\} \implies d(1) = 1.$
$\{n > 0 | \Pr[X_n = 2 | X_0 = 2] > 0\} = \{3, 4, ...\} \implies d(2) = 1.$
$[B] : \{n > 0 | \Pr[X_n = 1 | X_0 = 1] > 0\} = \{3, 6, 9, ...\} \implies d(i) = 3.$
$\{n > 0 | \Pr[X_n = 5 | X_0 = 5] > 0\} = \{6, 9, ...\} \implies d(5) = 3.$

Convergence of π_n

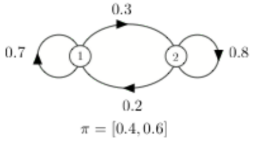
- Irreducible MC $\implies$ fraction of time spent in state i is equal to the invariant probability of that state
- Irreducible + Aperiodic MC $\implies$ fraction of time spent in state i is equal to and converges to the invariant probability of that state
- **Proof:** See EE126, or Lecture notes 24.

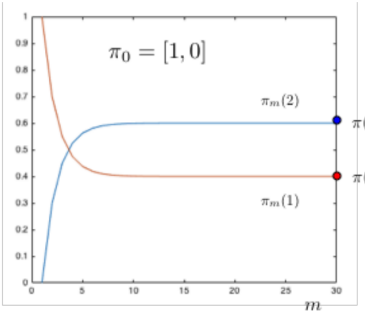
Theorem: Convergence of π_n
Let X_n ba an irreducible and aperiodic MC with invariant distribution π

$$\text{For all } i \in X, \pi_n(i) \to \pi(i), \text{ as } n \to \infty \tag{9}$$

Example 1

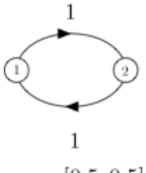
- Irreducible + Aperiodic MC $\implies$ fraction of time spent in state i is equal to and converges to the invariant probability of that state

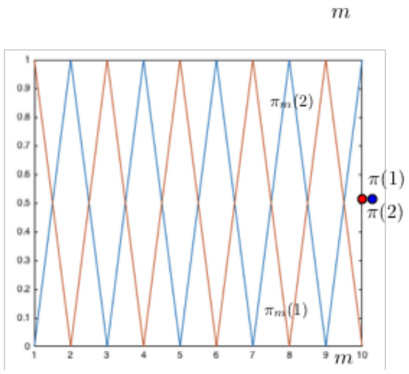




Example 2

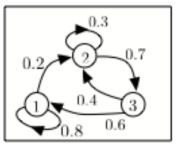
- Irreducible MC $\implies$ fraction of time spent in state i is equal to the invariant probability of that state

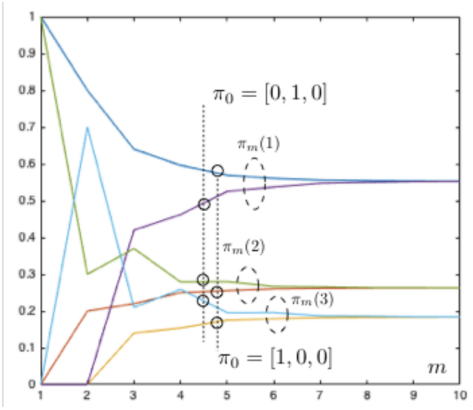




Example 3

- Loop implies aperiodicity





Calculating π

Method

1. Let P be irreducible
2. $\pi P = \pi \implies \pi[P - I] = 0$
3. Replace the last equation with ones $\pi 1 = 1$ to get $\pi P_1 = [0, 0, 1]$
   - Observe the sum of the columns of $P - I = 0$, which shows the equations are redundant, which means the equations are redundant
4. Solve $\pi = [0, 0, 1]P_1^{-1}$

Example:

Let P be irreducible. Find π where $P = \begin{bmatrix} 0.8 & 0.2 & 0 \\ 0 & 0.3 & 0.7 \\ 0.6 & 0.4 & 0 \end{bmatrix}$

One has $\pi P = \pi$, i.e., $\pi[P - I] = 0$ where I is the identity matrix:

$$\pi \begin{bmatrix} 0.8-1 & 0.2 & 0 \\ 0 & 0.3-1 & 0.7 \\ 0.6 & 0.4 & 0-1 \end{bmatrix} = [0, 0, 0].$$

However, the sum of the columns of $P - I$ is 0. This shows that these equations are redundant: If all but the last one hold, so does the last one. Let us replace the last equation by $\pi 1 = 1$, i.e., $\sum_j \pi(j) = 1$:

$$\pi \begin{bmatrix} 0.8-1 & 0.2 & 1 \\ 0 & 0.3-1 & 1 \\ 0.6 & 0.4 & 1 \end{bmatrix} = [0, 0, 1].$$

Hence, $\pi = [0, 0, 1] \begin{bmatrix} 0.8-1 & 0.2 & 1 \\ 0 & 0.3-1 & 1 \\ 0.6 & 0.4 & 1 \end{bmatrix}^{-1} \approx [0.55, 0.26, 0.19]$

Summary: Markov Chains

1. Markov Chain: $\Pr[X_n + 1 = j | X_0, \ldots, X_n = i] = P(i, j)$
2. FSE: $\beta(i) = 1 + \sum_j P(i, j)\beta(j); \alpha(i) = \sum_j P(i, j)\alpha(j)$.
3. $\pi_n = \pi_0 P^n$
4. π is invariant iff $\pi P = \pi$

5. Irreducible $\implies$ one and only one invariant distribution π
6. Irreducible $\implies$ fraction of time in state i approaches $\pi(i)$
7. Irreducible + Aperiodic $\implies$ $\pi_n \to \pi$.
8. Calculating π: One finds $\pi = [0, 0 \ldots, 1]Q - 1$ where $Q = \cdots$.

CS70 - Lecture 26 Notes

Name: Felix Su SID: 25794773

Spring 2016 GSI: Gerald Zhang

Continuous Probability

1. Use **unions of intervals** to describe events

- Choose a real number X, uniformly at random in [0,L].
- Let [a,b] denote the event that the point X is in the interval [a,b].
  - $\Pr[[a,b]] = \frac{\text{length of } [a,b]}{\text{length of } [0,L]} = \frac{b-a}{L} = \frac{b-a}{1000}$
- Events in this space are unions of intervals.
- **Example:** the event A - within 50 of 0 or 1000 is
  - $A = [0,50] \cup [950,1000]$. Thus, $\Pr[A] = \Pr[[0,50]] + \Pr[[950,1000]] = \frac{1}{10}$

Finite vs. Continuous Probability Spaces

1. Start with probability of events (unions of intervals): $\Pr[A]$ for some events A
2. Probability is then a function from events to [0,1]
3. Function must be additive

- **Finite probability space:** $\Omega = \{1, 2, \ldots, N\}$
  - Started with probabilities of each outcome $\Pr[\omega] = p_\omega$
  - Defined probability of event is sum of probability of outcomes in the event: $\Pr[A] = \sum_{\omega \in A} p\omega$ for $A \subset \Omega$.
  - We used the same approach for countable Ω.
- **Continuous space:** $\Omega = [0, L]$,
  - Cannot start with $\Pr[\Omega]$, because this will typically be 0.
  - Start with probability of events (unions of intervals): $\Pr[A]$ for some events A. Here, we started with A = interval, or union of intervals.
  - Probability is then a function from events to [0,1]
  - Function must be additive. In our example, $\Pr[[0,50] \cup [950,1000]] = \Pr[[0,50]] + \Pr[[950,1000]]$

Example:
James Bond Shooting

Chance of landing in a one foot radius circle that is inside a 4×5 rectangle.
$\Omega = \{(x,y) : x \in [0,4], y \in [0,5]\}$.
The size of the event is $\pi(1)^2 = \pi$.
The "size" of the sample space which is 4×5.
Since uniform, probability of event is $\frac{\pi}{20}$.

Continuous Random Variables: CDF

1. Define $\Pr[a < X \leq b] = \Pr[X \leq b] - \Pr[X \leq a] = F_X(b) - F_X(a)$

- Find function to define all intervals between a and b: $\Pr[a < X \leq b]$
- Cumulative probability Distribution Function of X (CDF of X) is
  - $F_X(x) = \Pr[X \leq x]$
- So, $\Pr[a < X \leq b] = \Pr[X \leq b] - \Pr[X \leq a] = F_X(b) - F_X(a)$.
  - Idea: two events $X \leq b$ and $X \leq a$.
  - Difference is the event $a < X \leq b$.
  - Indeed: $\{X \leq b\} - \{X \leq a\} = \{X \leq b\} \cap \{X > a\} = \{a < X \leq b\}$.

Cumulative Probability Distribution Function of X: CDF

$$F_X(x) = \Pr[X \leq x] \tag{1}$$

Define Probability of all Intervals

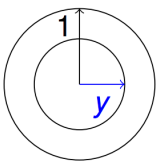
$$\Pr[a < X \leq b] = \Pr[X \leq b] - \Pr[X \leq a] = F_X(b) - F_X(a) \tag{2}$$

Example:
CDF: Value of X in [0,L] with L = 1000.

$$F_X(x) = \Pr[X \leq x] = \begin{cases} 0 & \text{for } x < 0 \\ \frac{x}{1000} & \text{for } 0 \leq x \leq 1000 \\ 1 & \text{for } x > 1000 \end{cases} \tag{3}$$

Probability that X is within 50 of center:
$\Pr[450 < X \leq 550] = \Pr[X \leq 550] - \Pr[X \leq 450] = \frac{550}{1000} - \frac{450}{1000} = \frac{100}{1000} = \frac{1}{10}$

Example:
CDF: Hitting random location on a unit circle.



Random Variable: Y distance from center.

Probability within y of center:

$$\Pr[Y \leq y] = \frac{\text{area of small circle}}{\text{area of dartboard}} = \frac{\pi y^2}{\pi} = y^2 \tag{4}$$

Hence,

$$F_Y(y) = \Pr[Y \leq y] = \begin{cases} 0 & \text{for } y < 0 \\ y^2 & \text{for } 0 \leq y \leq 1 \\ 1 & \text{for } y > 1 \end{cases} \tag{5}$$

Calculation Probability between .5 and .6 of center
$\Pr[0.5 < Y \leq 0.6] = \Pr[Y \leq 0.6] - \Pr[Y \leq 0.5] = F_Y(0.6) - F_Y(0.5) = .36 - .25 = .11$

Density function

1. Find probability of a certain value (within δ): $\lim_{\delta \to 0} \frac{\Pr[x < X \leq x+\delta]}{\delta} = \frac{d(F_X(x))}{dx}$

- Is the dart more like to be (near) .5 or .1?
- Probability within δ of x is $\Pr[x < X \leq x + \delta]$.
- Goes to 0 as δ goes to zero.
- Find the limit as δ goes to zero. $\lim_{\delta \to 0} \frac{\Pr[x < X \leq x+\delta]}{\delta}$
  - $= \lim_{\delta \to 0} \frac{\Pr[X \leq x+\delta] - \Pr[X \leq x]}{\delta}$
  - $= \lim_{\delta \to 0} \frac{F_X(x+\delta) - F_X(x)}{\delta}$
  - $= \frac{d(F_X(x))}{dx}$

Density

1. A **probability density function** for RV X with cdf $F_X(x) = \Pr[X \leq x]$ is the derivate of the cdf: $f_X(x) = \frac{d(F_X(x))}{dx}$
2. Probability that X is within δ of x, is $f_X(x)\delta$

- Definition: (Density) A probability density function for a random variable X with cdf $F_X(x) = \Pr[X \leq x]$ is the function $f_X(x)$ where:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du \tag{6}$$

- Thus, $\Pr[X \in (x, x+\delta]] = F_X(x+\delta) - F_X(x) \approx f_X(x)\delta$

Probability Density Function
For random variable X with cdf $F_X(x) = \Pr[X \leq x]$ is the function $f_X(x)$ where:

$$\Pr[X \in (x, x+\delta]] = F_X(x+\delta) - F_X(x) \approx f_X(x)\delta \tag{7}$$

Example:
Uniform over interval [0,1000]

$$f_X(x) = F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{1}{1000} & \text{for } 0 \leq x \leq 1000 \\ 0 & \text{for } x > 1000 \end{cases} \tag{8}$$

Example:
Uniform over interval [0,L]

$$f_X(x) = F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{1}{L} & \text{for } 0 \leq x \leq L \\ 0 & \text{for } x > L \end{cases} \tag{9}$$

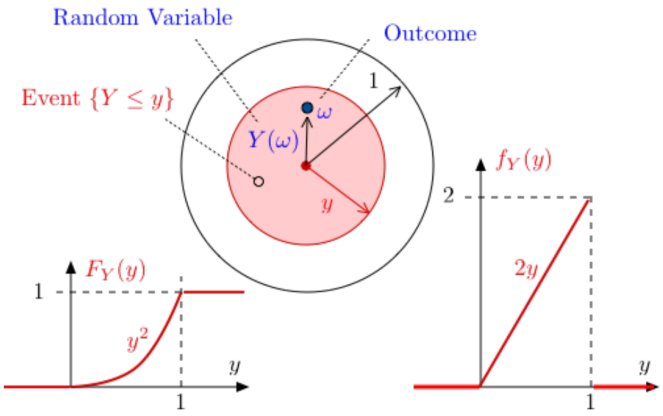
Example:
"Dart" board

- The cumulative distribution function (cdf) and probability distribution function (pdf) give full information.
- Use whichever is convenient.

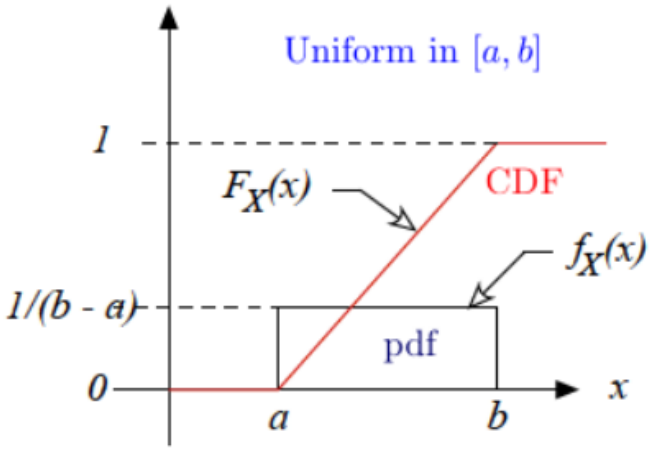
$$F_Y(y) = \Pr[Y \le y] = \begin{cases} 0 & \text{for } y < 0 \\ y^2 & \text{for } 0 \le y \le 1 \\ 1 & \text{for } y > 1 \end{cases} \tag{10}$$

$$f_Y(y) = F'_Y(y) = \begin{cases} 0 & \text{for } y < 0 \\ 2y & \text{for } 0 \le y \le 1 \\ 0 & \text{for } y > 1 \end{cases} \tag{11}$$

Target

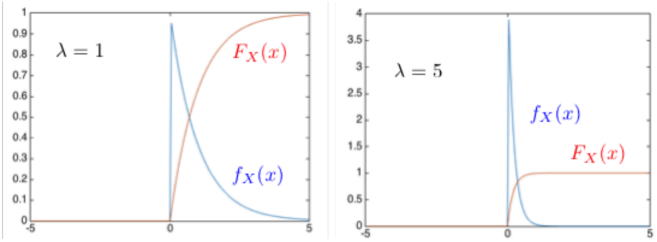


Uniform Distribution: U[a,b]



Expo(λ)

- Note that $\Pr[X > t] = e^{-\lambda t}$ for $t > 0$

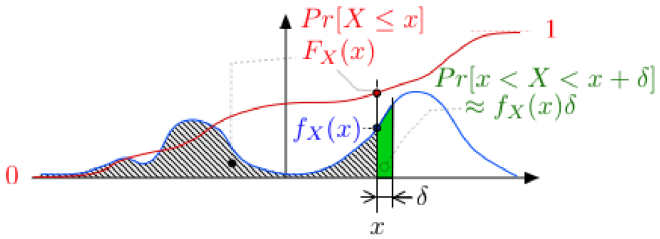


The exponential distribution with parameter $\lambda > 0$ is defined by

$$f_X(x) = \lambda e^{-\lambda x} 1\{x \geq 0\} \tag{12}$$

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-\lambda x} & \text{if } x \geq 0 \end{cases} \tag{13}$$

Random Variables



Continuous random variable X, specified by

1. $F_X(x) = \Pr[X \leq x], \forall x$
   - **Cumulative Distribution Function (cdf)**: $\Pr[a < X \leq b] = F_X(b) - F_X(a)$
   - Non-decreasing between 0 and 1
     - $0 \leq F_X(x) \leq 1 \forall x \in \mathbb{R}$.
     - $F_X(x) \leq F_X(y)$ if $x \leq y$.
2. Or $f_X(x)$, where $F_X(x) = \int_{-\infty}^{x} f_X(u)du$ or $f_X(x) = \frac{d(F_X(x))}{dx}$
   - **Probability Density Function (pdf)**: $\Pr[a < X \leq b] = \int_a^b f_X(x)dx = F_X(b) - F_X(a)$
   - Non-negative and integrates to 1.
     - $f_X(x) \geq 0 \forall x \in \mathbb{R}$.
     - $\int_{-\infty}^{\infty} f_X(x)dx = 1$
3. Recall that $\Pr[X \in (x, x + \delta)] \approx f_X(x)\delta$.
   - Probability that you are δ away from x is $\approx f_X(x)\delta$
   - If density ($f_X(x)$) is large, more likely to be at x.
4. Think of X taking discrete values $n\delta$ for $n = \ldots, -2, -1, 0, 1, 2, \ldots$ with $\Pr[X = n\delta] = f_X(n\delta)\delta$

Some Examples

a. **Expo is memoryless**. Let X =Expo(λ). Then, for $s, t > 0$
   - $\Pr[X > t + s | X > s] = \frac{\Pr[X > t+s]}{\Pr[X > s]} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t} = \Pr[X > t]$.
   - 'Used is a good as new.'

b. **Scaling Expo**. Let X =Expo(λ) and $Y = aX$ for some $a > 0$. Then

- $\Pr[Y > t] = \Pr[aX > t] = \Pr[X > t/a] = e^{-\lambda(t/a)} = e^{-(\lambda/a)t} = \Pr[Z > t]$ for Z =Expo(λ/a)
- Thus, $a\times$ Expo(λ) = Expo(λ/a).

c. **Scaling Uniform** Let $X = U[0,1]$ and $Y = a + bX$ where $b > 0$. Then

- $\Pr[Y \in (y, y+\delta)] = \Pr[a + bX \in (y, y+\delta)] = \Pr[X \in (\frac{y-a}{b}, \frac{y+\delta-a}{b})]$
  - $= \Pr[X \in (\frac{y-a}{b}, \frac{y-a}{b} + \frac{\delta}{b})] = \frac{1}{b}\delta$, for $0 < \frac{y-a}{b} < 1$
  - $= \frac{1}{b}\delta$, for $a < y < a + b$.
- Thus, $f_Y(y) = \frac{1}{b}$ for $a < y < a + b$. Hence, $Y = U[a, a+b]$.

d. **Scaling pdf**. Let $f_X(x)$ be the pdf of X and $Y = a + bX$ where $b > 0$. Then

- $\Pr[Y \in (y, y+\delta)] = \Pr[a + bX \in (y, y+\delta)]$
  - $= \Pr[X \in (\frac{y-a}{b}, \frac{y+\delta-a}{b}]$
  - $= \Pr[X \in (\frac{y-a}{b}, \frac{y-a}{b} + \frac{\delta}{b}]$
  - $= f_X(\frac{y-a}{b})\frac{\delta}{b}$
- Now, the left-hand side is $f_Y(y)\delta$. Hence, $f_Y(y) = \frac{1}{b} f_X(\frac{y-a}{b})$.

Expectation

- **Definition** The expectation of a random variable X with pdf $f_X(x)$ is defined as $E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$.
- Justification: Say $X = n\delta$ w.p. $f_X(n\delta)\delta$. Then,
  - $E[X] = \sum_n (n\delta)\Pr[X = n\delta] = \sum_n (n\delta) f_X(n\delta)\delta = \int_{-\infty}^{\infty} x f_X(x) dx$.
- Indeed, for any g, one has $\int g(x)dx \approx \sum_n g(n\delta)\delta$. Choose $g(x) = x f_X(x)$.

Expectation of function of RV

- **Definition** The expectation of a function of a random variable is defined as $E[h(X)] = \int_{-\infty}^{\infty} h(x) f_X(x) dx$.
- Justification: Say $X = n\delta$ w.p. $f_X(n\delta)\delta$. Then
  - $E[h(X)] = \sum_n h(n\delta)\Pr[X = n\delta]$
  - $= \sum_n h(n\delta) f_X(n\delta)\delta$
  - $= \int_{-\infty}^{\infty} h(x) f_X(x) dx$.
- Indeed, for any g, one has $\int g(x)dx \approx \sum_n g(n\delta)\delta$. Choose $g(x) = h(x) f_X(x)$.
- **Fact** Expectation is linear.
- **Proof** As in the discrete case.

$E[X]$ **with pdf** $f_X(x)$

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx \tag{14}$$

Exepection of a function of RV X**:** $E[h(X)]$ **with pdf** $f_X(x)$

$$E[h(X)] = \int_{-\infty}^{\infty} h(x) f_X(x) dx \tag{15}$$

Variance

- **Definition:** The variance of a continuous random variable X is defined as $\text{Var}[X] = E[(X - E[X])^2] = E[X^2] - (E[X])^2 = \int_{-\infty}^{\infty} x^2 f(x)dx - (\int_{-\infty}^{\infty} xf(x)dx)^2$

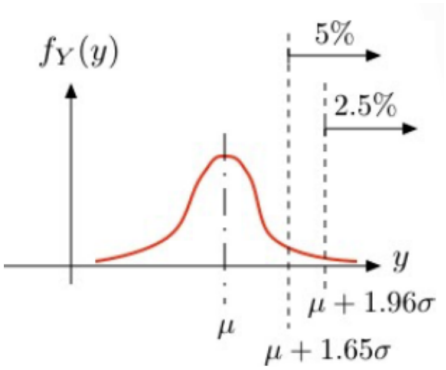
Variance

$$\text{Var}[X] = \int_{-\infty}^{\infty} x^2 f(x)dx - (\int_{-\infty}^{\infty} xf(x)dx)^2 \tag{16}$$

Motivation for Gaussian Distribution

- **Key fact:** The sum of many small independent RVs has a Gaussian distribution.
- This is the Central Limit Theorem. (See later.)
- Examples: Binomial and Poisson suitably scaled.
- This explains why the Gaussian distribution (the bell curve) shows up everywhere.

Normal Distribution



- For any mean: μ and std. dev: σ, a **normal** (aka **Gaussian**) random variable Y, which we write as $Y = \mathcal{N}(\mu, \sigma^2)$, has pdf $f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2}$
- Standard normal has $\mu = 0$ and $\sigma = 1$.
- Note: $\Pr[|Y - \mu| > 1.65\sigma] = 10\%$;$\Pr[|Y - \mu| > 2\sigma] = 5\%$.
- Guassian RV is within 2σ of the mean with 95%

Normal Distribution
For any μ and σ, a Gaussian RV, $Y = \mathcal{N}(\mu, \sigma^2)$ has pdf:

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2} \tag{17}$$

Summary: Continuous Probability

1. **pdf**: $\Pr[X \in (x, x + \delta]] = f_X(x)\delta$.
2. **CDF**: $\Pr[X \leq x] = F_X(x) = \int_{-\infty}^{x} f_X(y)dy$.
3. **U[a,b], Expo(λ), target**
4. **Expectation**: $E[X] = \int_{-\infty}^{\infty} xf_X(x)dx$.

5. **Expectation of function**: $E[h(X)] = \int_{-\infty}^{\infty} h(x) f_X(x) dx$.
6. **Variance**: $\text{Var}[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$.
7. **Gaussian**: $N(\mu, \sigma^2) : f_X(x) = \ldots$ "bell curve"

CS70 - Lecture 27 Notes

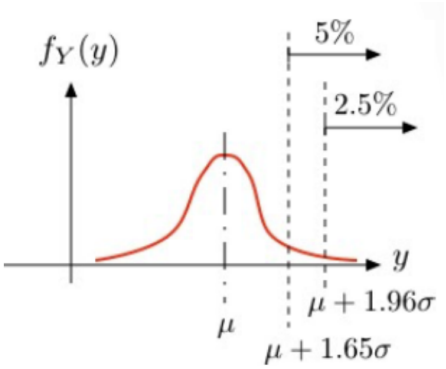
Name: Felix Su SID: 25794773

Spring 2016 GSI: Gerald Zhang

Review: Continuous Probability

1. **pdf**: $\Pr[X \in (x, x+\delta]] = f_X(x)\delta$.
   - Probability of point in $\Omega = 0$, so define prob. of events as intervals = pdf(δ)
   - Pdf is non-negative and integrates to 1
2. **CDF**: $\Pr[X \leq x] = F_X(x) = \int_{-\infty}^{x} f_X(y)dy$.
   - $\Pr[a < x \leq b] = \Pr[X \leq b] - \Pr[X \leq a]$
3. **U[a,b], Expo(λ), target**
4. **Expectation**: $E[X] = \int_{-\infty}^{\infty} x f_X(x)dx$.
   - $x*$probability of $X = x$ in that interval
5. **Expectation of function**: $E[h(X)] = \int_{-\infty}^{\infty} h(x) f_X(x)dx$.
6. **Variance**: $\text{Var}[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$.
7. **Gaussian**: $N(\mu, \sigma^2) : f_X(x) = \ldots$ "bell curve"
   - When you add up many small RVs, the CDF comes out as a bell shape

Normal Distribution



- For any mean: μ and std. dev: σ, a **normal** (aka **Gaussian**) random variable Y, which we write as $Y = \mathcal{N}(\mu, \sigma^2)$, has pdf $f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2}$
- Standard normal has $\mu = 0$ and $\sigma = 1$.
- Note: $\Pr[|Y - \mu| > 1.65\sigma] = 10\%$;$\Pr[|Y - \mu| > 2\sigma] = 5\%$.
- Guassian RV is within 2σ of the mean with 95%

Normal Distribution
For any μ and σ, a Gaussian RV, $Y = \mathcal{N}(\mu, \sigma^2)$ has pdf:

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2} \tag{1}$$

Scaling and Shifting

1. If you scale a Gaussian RV (by $Y = \mu + \sigma X$), you get another Gaussian RV
2. When you multiply at RV by a constant (σ), you multiply its variance by the square of that constant (σ^2)

- **Theorem** Let $X = \mathcal{N}(0, 1)$ and $Y = \mu + \sigma X$. Then
  - $Y = \mathcal{N}(\mu, \sigma^2)$.
- **Proof:** $f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{x^2}{2}\}$.
  - Now, $f_Y(y) = \frac{1}{\sigma} f_X(\frac{y-\mu}{\sigma})$ (See Lec. 26, slide 19.)
  - $= \frac{1}{\sqrt{2\pi\sigma 2}} \exp\{-\frac{(y-\mu)^2}{2\sigma^2}\}$

Expectation, Variance

- **Theorem** If $Y = \mathcal{N}(\mu, \sigma^2)$, then
  - $E[Y] = \mu$ and $\text{Var}[Y] = \sigma^2$
- **Proof:** It suffices to show the result for $X = \mathcal{N}(0, 1)$ since $Y = \mu + \sigma X, \ldots$
- Thus, $f_X(x) = \frac{1}{2\pi} \exp\{-\frac{x^2}{2}\}$.
  - First note that $E[X] = 0$, by symmetry.
  - $\text{Var}[X] = E[X^2] = \int x^2 \frac{1}{\sqrt{2\pi}} \exp\{-\frac{x^2}{2}\} dx$
  - $= -\frac{1}{\sqrt{2\pi}} \int x d\exp\{-\frac{x^2}{2}\}$
  - $= \frac{1}{\sqrt{2\pi}} \int \exp\{-\frac{x^2}{2}\} dx$ (Integration by Parts: $\int_a^b f dg = [fg]_a^b - \int_a^b g df$)
  - $= \int f_X(x) dx = 1$

Central Limit Theorem

1. Tells us how many samples to take in order for the airthmetic mean to tend to the expected value of an RV
2. Derive the Normalized Sample mean $S_n = \frac{A_n - \mu}{\sigma/\sqrt{n}} = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$
3. **CLT:** As $n \to \infty$, the distribution of $S_n \to$ the standard normal distribution $\mathcal{N}(0, 1)$
   - Expectation of S_n is always 0
   - Variance of S_n is always 1

- **Law of Large Numbers:** For any set of independent identically distributed random variables, X_i, $A_n = \frac{1}{n} \sum X_i$ "tends to the mean."
  - Say X_i have expectation $\mu = E[X_i]$ and variance σ^2.
  - Mean of A_n is μ, and variance is σ^2/n.

- Let $S_n = \frac{A_n - \mu}{\sigma/\sqrt{n}} = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$
- Then, $E[S_n] = \frac{1}{\sigma/\sqrt{n}}(E[An] - \mu) = 0$
- $\text{Var}[S_n] = \frac{1}{\sigma^2/n}\text{Var}[A_n] = 1$.

- **Central limit theorem:** As n goes to infinity the distribution of S_n approaches the standard normal distribution.
  - Expectation of S_n is always 0
  - Variance of S_n is always 1

Central Limit Theorem

Normalized Sample Mean

$$S_n = \frac{A_n - \mu}{\sigma/\sqrt{n}} = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \tag{2}$$

Expected Value of Normalized Sample Mean

$$E[S_n] = \frac{1}{\sigma/\sqrt{n}}(E[An] - \mu) = 0 \tag{3}$$

Variance of Normalized Sample Mean

$$\text{Var}[S_n] = \frac{1}{\sigma^2/n}\text{Var}[A_n] = 1 \tag{4}$$

Central Limit Theorem

- Let $X_1, X_2, \ldots$ be i.i.d. with $E[X_1] = \mu$ and $\text{Var}[X_1] = \sigma^2$.
- Define $S_n := \frac{A_n - \mu}{\sigma/\sqrt{n}} = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$.
  - Then, $S_n \to \mathcal{N}(0,1)$,as $n \to \infty$.
  - That is, $\Pr[S_n \leq \alpha] \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha} e^{-x^2/2} dx$.
  - **Proof:** See EE126.
- THe CDF of the RV S_n approches the CDF of $\mathcal{N}(0,1)$
  - PDF begins to look like a bell shape
  - CDF looks like the integral of a bell shape

Central Limit Theorem:

$$\Pr[S_n \leq \alpha] \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha} e^{-x^2/2} dx \tag{5}$$

Confidence Interval (CI) for Mean

- Let $X_1, X_2, \ldots$ be i.i.d. with mean μ and variance σ^2 and $A_n = \frac{X_1 + \cdots + X_n}{n}$.
- The CLT states that $\frac{A_n - \mu}{\sigma/\sqrt{n}} = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \to \mathcal{N}(0,1)$ as $n \to \infty$.
  - Thus, for $n >> 1$, one has $\Pr[|\frac{A_n - \mu}{\sigma/\sqrt{n}}| \leq 2] \approx 95\%$.
  - Equivalently, $\Pr[\mu \in [A_n - 2\frac{\sigma}{\sqrt{n}}, A_n + 2\frac{\sigma}{\sqrt{n}}]] \approx 95\%$.
  - That is, $[A_n - 2\frac{\sigma}{\sqrt{n}}, A_n + 2\frac{\sigma}{\sqrt{n}}]$ is a 95% -CI for μ.

Confidence Interval (CI) for Mean

$$\Pr[|\frac{A_n - \mu}{\sigma/\sqrt{n}}| \leq 2] \approx 95\% \tag{6}$$

$$[A_n - 2\frac{\sigma}{\sqrt{n}}, A_n + 2\frac{\sigma}{\sqrt{n}}] \text{ is a } 95\% \text{ -CI for } \mu \tag{7}$$

- Let $X_1, X_2, \ldots$ be i.i.d. with mean μ and variance σ^2 and $A_n = \frac{X_1+\cdots+X_n}{n}$.
- The CLT states that $\frac{X_1+\cdots+X_n-n\mu}{\sigma\sqrt{n}} \to \mathcal{N}(0,1)$ as $n \to \infty$.
  - Also, $[A_n - 2\frac{\sigma}{\sqrt{n}}, A_n + 2\frac{\sigma}{\sqrt{n}}]$ is a 95% -CI for μ.
  - Recall: Using Chebyshev, we found that (see Lec. 22, slide 6)
  - $[A_n - 4.5\frac{\sigma}{\sqrt{n}}, A_n + 4.5\frac{\sigma}{\sqrt{n}}]$ is a 95% -CI for μ.
- Thus, the CLT provides a smaller confidence interval.
  - Chebyshev works for all values of n.
  - The CLT assumes n is large enough.

Example Question: Question like this will be on the FINAL

Example:
Coins and normal

> Let $X_1, X_2, \ldots$ be i.i.d. $B(p)$. Thus, $X_1 + \cdots + X_n = B(n,p)$.
> Here, $\mu = p$ and $\sigma = \sqrt{p(1-p)}$.
> CLT states that $\frac{X_1+\cdots+X_n-np}{\sqrt{p(1-p)n}} \to \mathcal{N}(0,1)$ and $[A_n - 2\frac{\sigma}{\sqrt{n}}, A_n + 2\frac{\sigma}{\sqrt{n}}]$ is a 95% -CI for μ with $A_n = (X_1+\cdots+X_n)/n$.
> Hence, $[A_n - 2\frac{\sigma}{\sqrt{n}}, A_n + 2\frac{\sigma}{\sqrt{n}}]$ is a 95% -CI for p.
> Solve $\frac{d\text{Var}[X]}{dp} = \frac{d\sigma^2}{dp}$, test each result to find the p that returns the max of the Variance
> Substitute the returned p back in to solve for the max value of σ
> Since $\sigma \leq 0.5$, Substitute σ with the upper bound: $[A_n - 2\frac{0.5}{\sqrt{n}}, A_n + 2\frac{0.5}{\sqrt{n}}]$ is a 95% -CI for p.
> Thus, $[A_n - \frac{1}{\sqrt{n}}, A_n + \frac{1}{\sqrt{n}}]$ is a 95% -CI for p.

Summary: Gaussian and CLT

> 1. **Gaussian:** $\mathcal{N}(\mu, \sigma^2) : f_X(x) = \ldots$ "bell curve"
> 2. **CLT:** X_n i.i.d. $\implies \frac{A_n-\mu}{\sigma/\sqrt{n}} \to \mathcal{N}(0,1)$
> 3. **CI:** $[A_n - 2\frac{\sigma}{\sqrt{n}}, A_n + 2\frac{\sigma}{\sqrt{n}}]$= 95%-CI for μ.