



Probability in Computing

CS
237

Reminders

- HW 9 due Thursday

Reading

- LLM 19.4.6, P 3.1.5

Lecture slides are in Piazza Resources
and will be on class web page later

LECTURE 19

Last time

- Variance and its properties
- Discrete Distributions: Bernoulli, Uniform, and Binomial

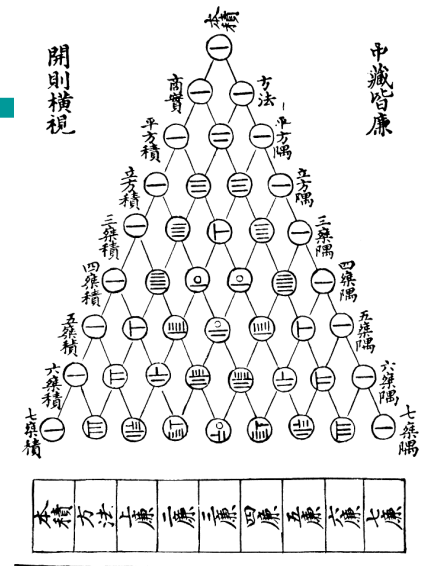
Today

- Discrete Distributions: Binomial concluded, Geometric, Negative Binomial

$X \sim \text{Binomial}(N, p)$:

- N **independent** Bernoulli(p) trials
- X = number of successes
- Canonical experiment: Take a coin with probability of heads p , flip it N times and count the number of heads.
- $\text{Range}(X) = \{0, \dots, N\}$
- $\Pr(X = k) =$

$$\binom{N}{k} \cdot p^k \cdot (1 - p)^{N-k}$$



Yang Hui's triangle
(Pascal's triangle)

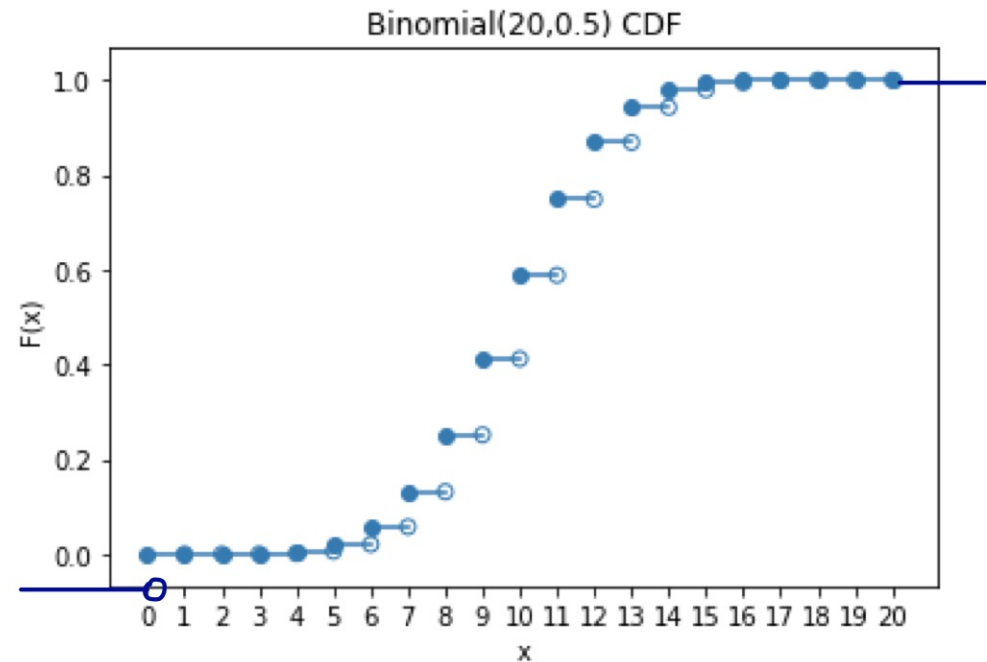
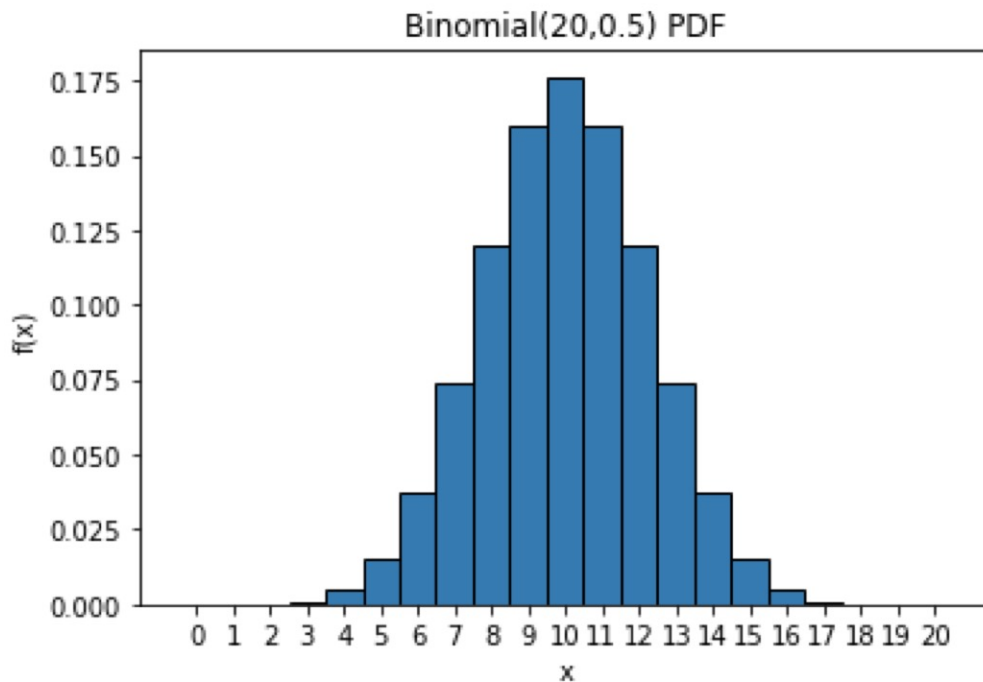
NOTE: Binomial is the **sum** of independent Bernoulli RVs:

$$X = X_1 + X_2 + \dots + X_n$$

Tophat Question One

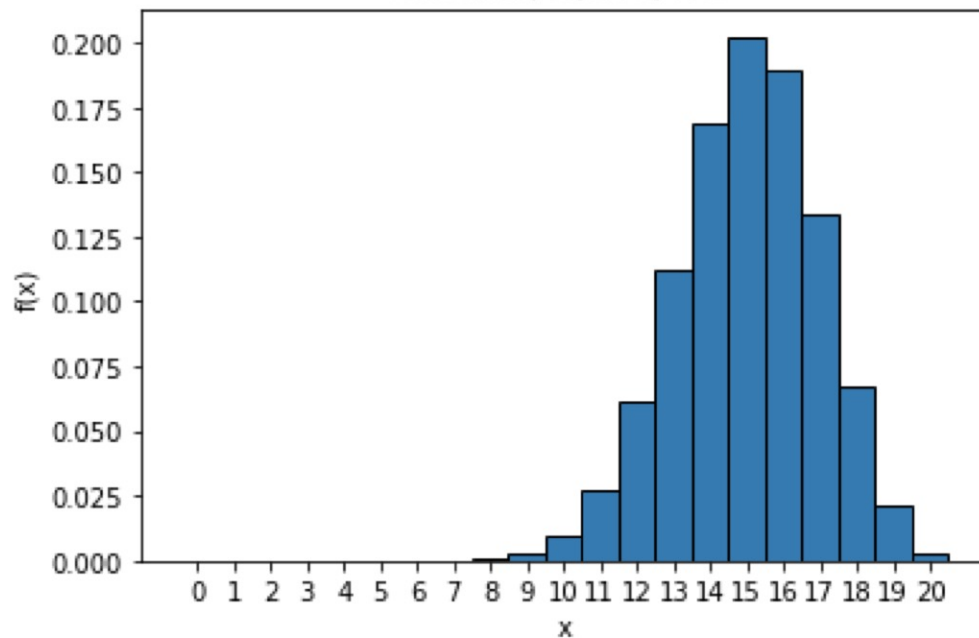
Binomial Distribution

- $X \sim \text{Binomial}(N, p)$ example

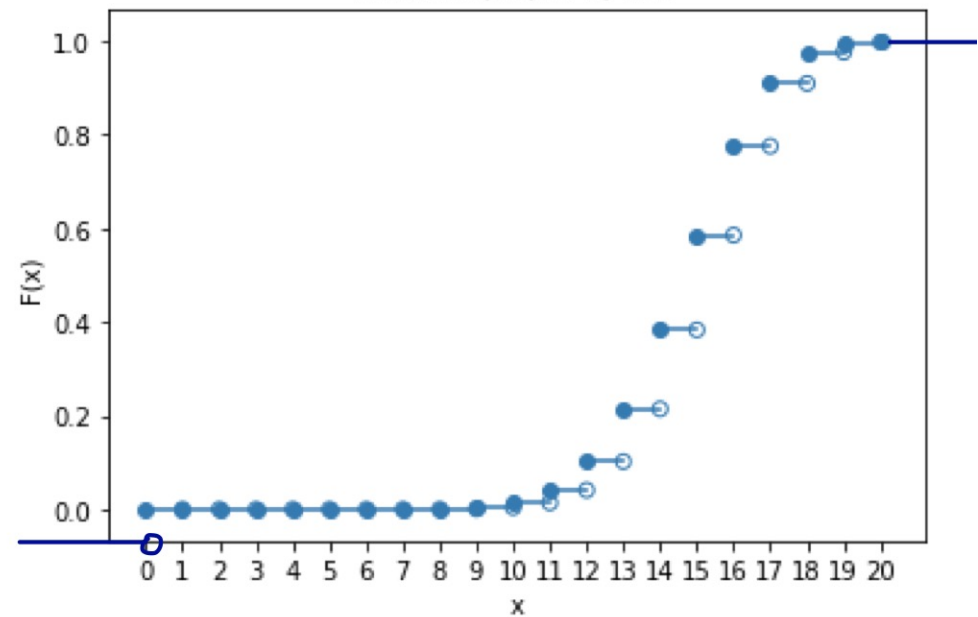


- $X \sim \text{Binomial}(N, p)$ example

Binomial(20,0.75) PDF



Binomial(20,0.75) CDF

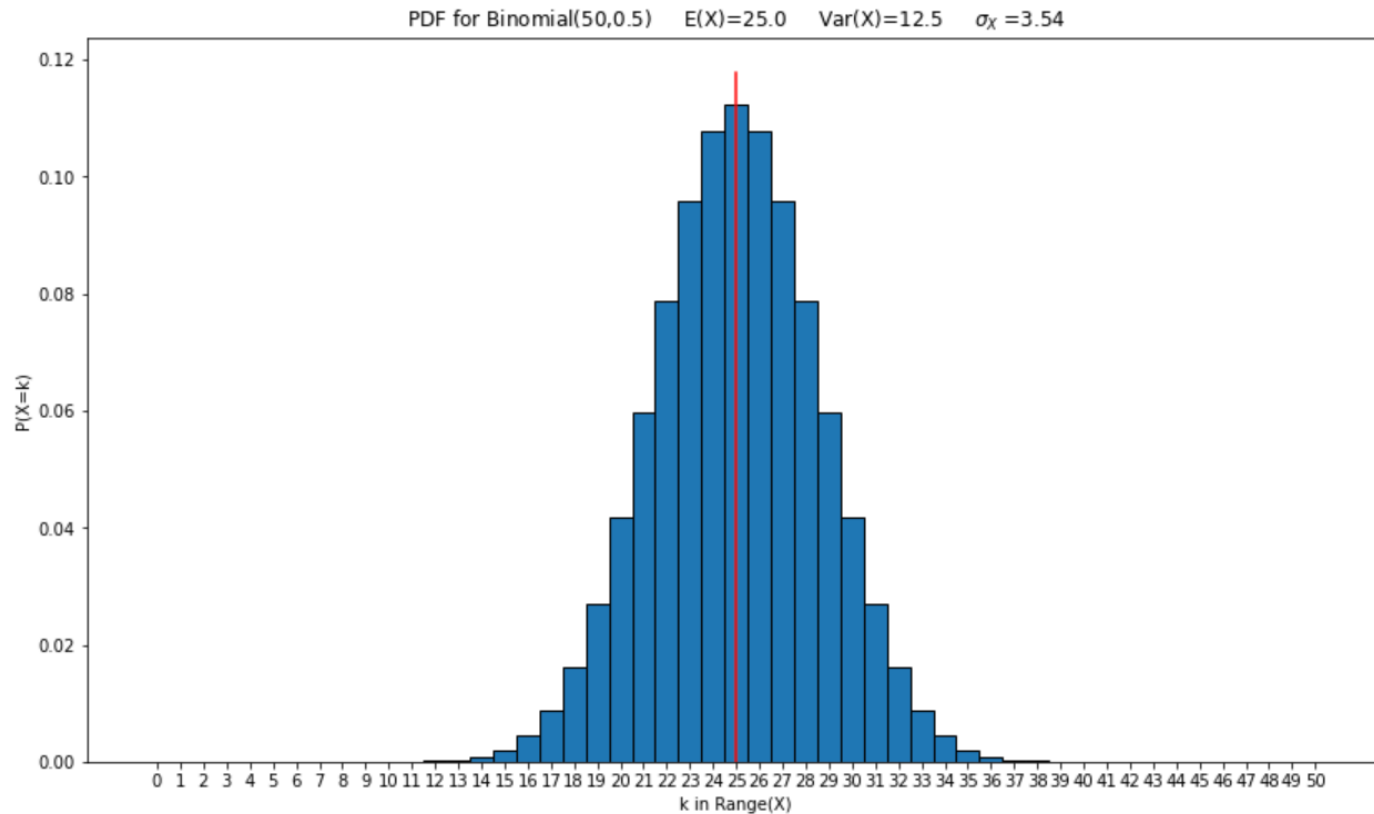


- Let $X \sim \text{Binomial}(N, p)$
- $E(X) = ?$

NOTE: Binomial is the **sum** of independent Bernoulli RVs:

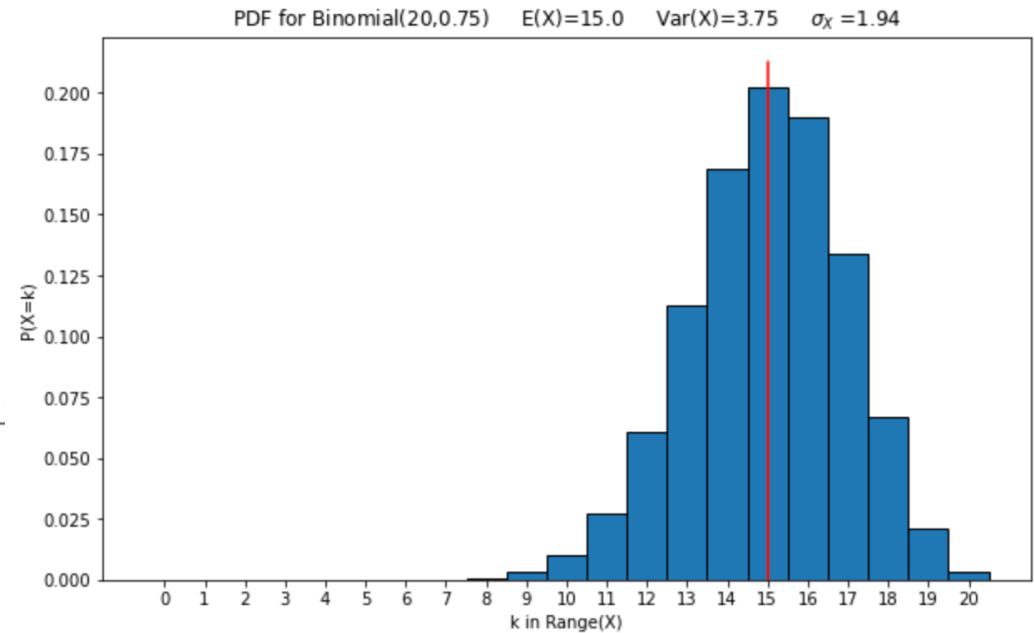
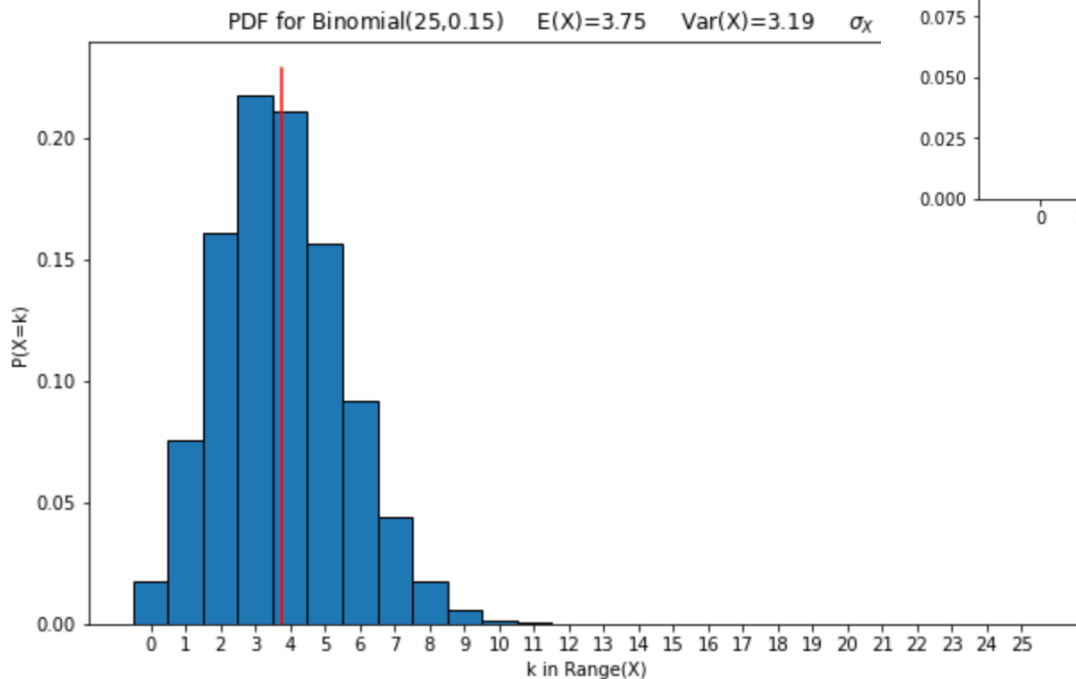
$$X = X_1 + X_2 + \dots + X_n$$

- Let $X \sim \text{Binomial}(N, p)$
- $E(X) = n * p$



Binomial Distribution

- Let $X \sim \text{Binomial}(N, p)$
- $E(X) = n * p$



- Let $X \sim \text{Binomial}(N, p)$
- $\text{Var}(X) = ?$

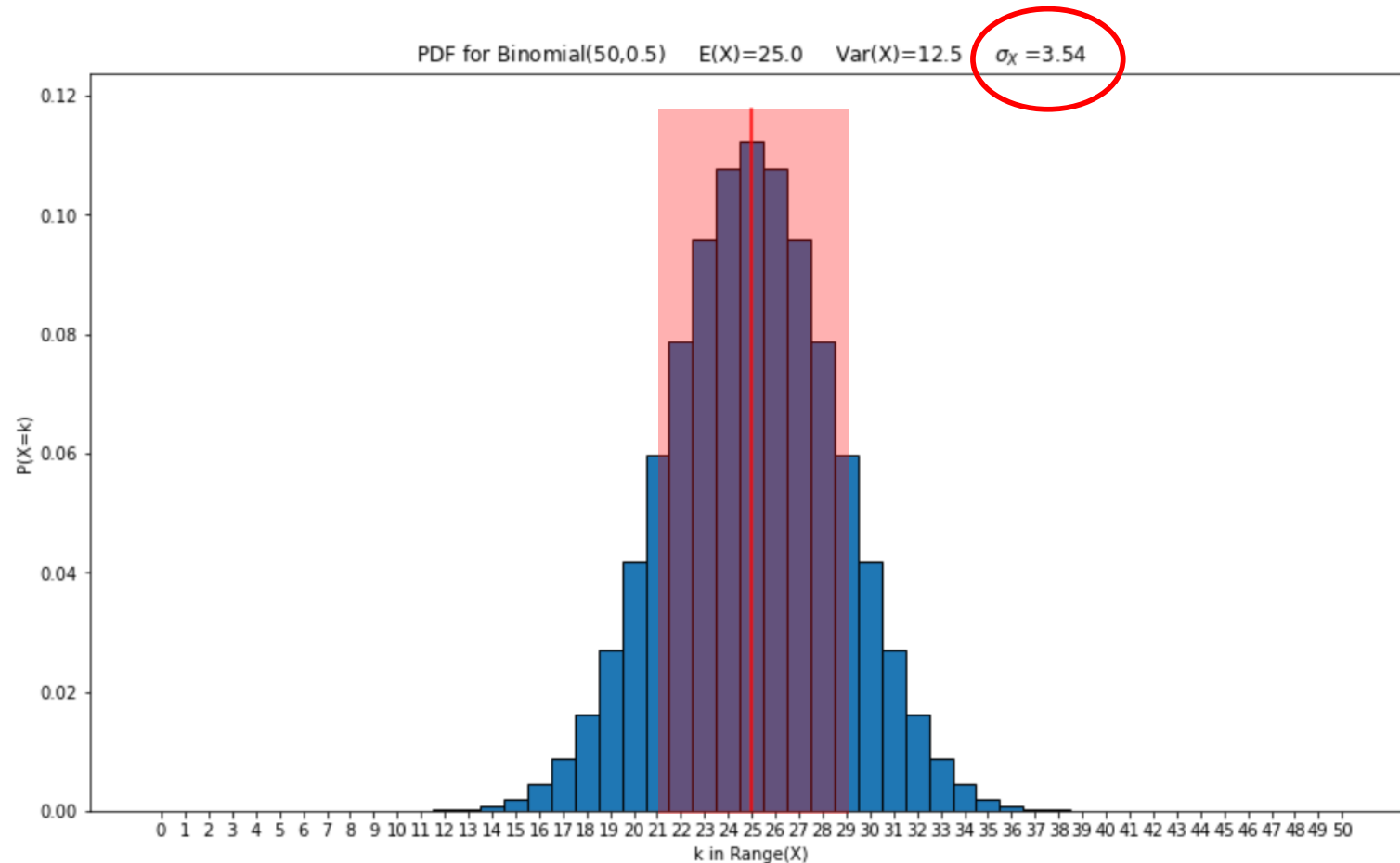
Tophat Question Two

NOTE: Binomial is the **sum** of independent Bernoulli RVs:

$$X = X_1 + X_2 + \dots + X_n$$

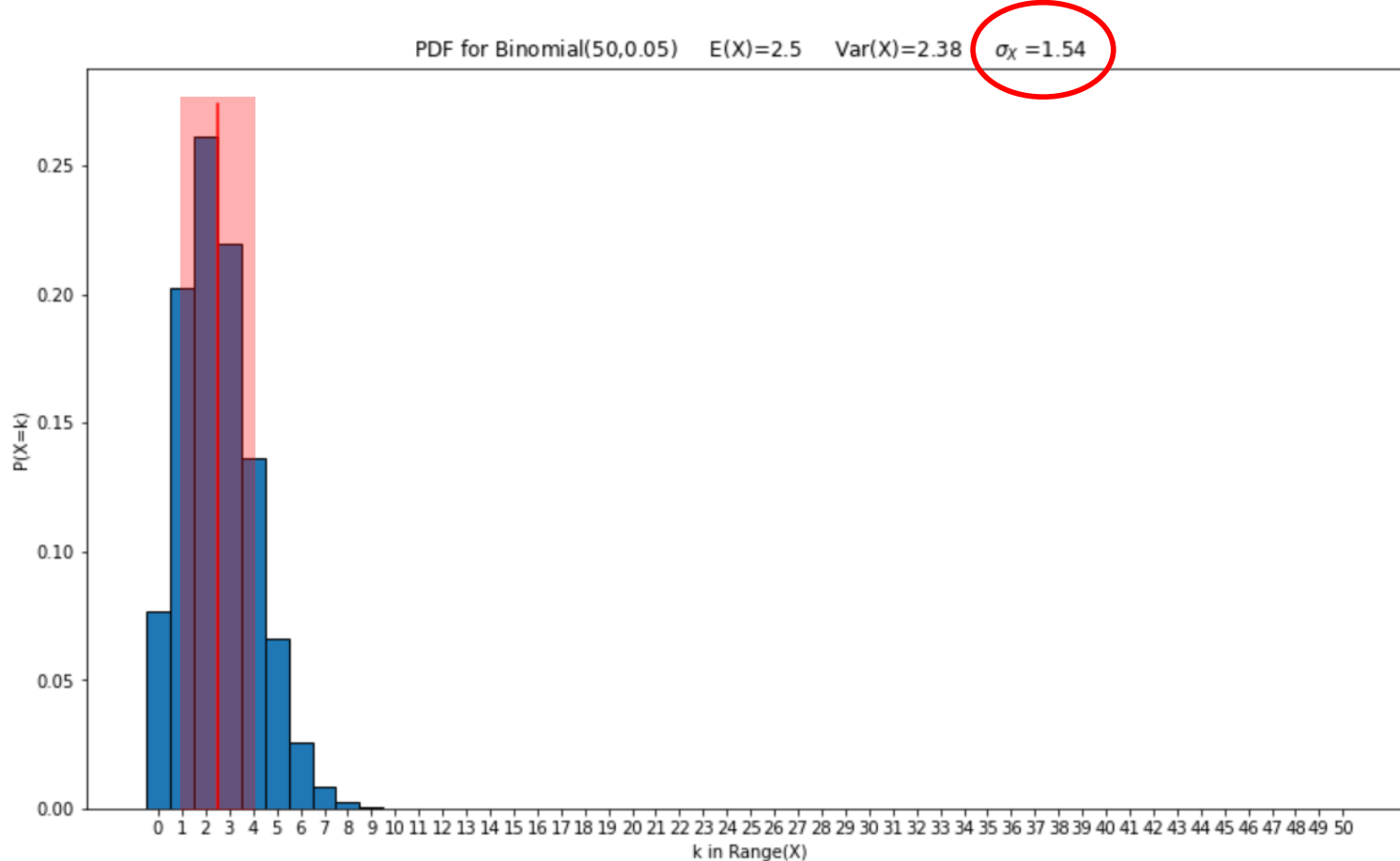
Binomial Distribution

- Let $X \sim \text{Binomial}(N, p)$
- $\text{Var}(X) = N * p * (1 - p)$



Binomial Distribution

- Let $X \sim \text{Binomial}(N, p)$
- $\text{Var}(X) = N * p * (1 - p)$



The motivation for this distribution comes from the fact that many complex phenomena are composed of the additive effect of many small binary choices or events (Bernoulli Trials!); a vivid illustration of this can be seen in the Galton Board or Quincunx:

<https://www.mathsisfun.com/data/quincunx.html>

<https://www.youtube.com/watch?v=J7AGOptcR1E>

The problem with the binomial is there is no simple way to calculate the PDF for large N:

$$\Pr(X = k) = \binom{N}{k} \cdot p^k \cdot (1 - p)^{N-k} \quad \Pr(X \leq k) = \sum_{i=0}^k \binom{N}{i} p^i (1 - p)^{N-i}$$

Example: There are about 20K genes in the human body. Supposing (very naively) that there is a 0.45 probability that a gene is dominant, what is the probability that 9000 are dominant?

$$\binom{20,000}{9,000} (0.45)^{9,000} (0.55)^{11,000}$$

The problem is calculating the binomial coefficient....

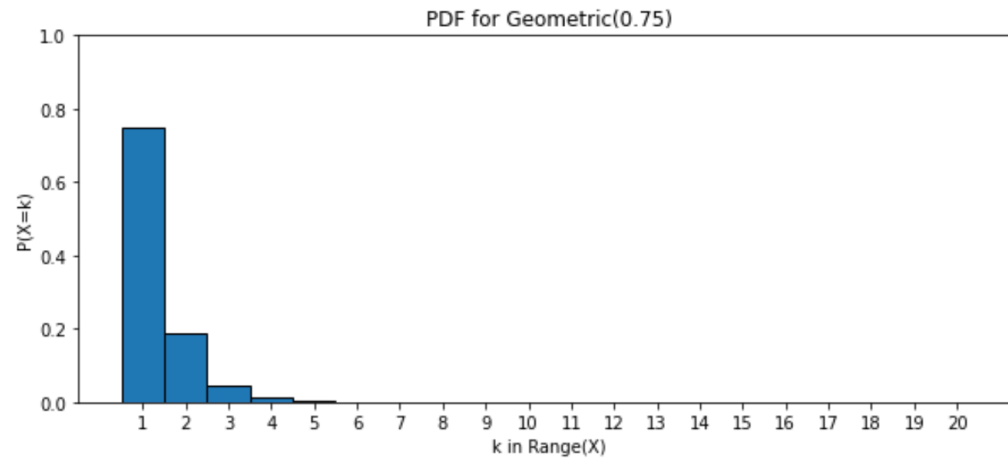
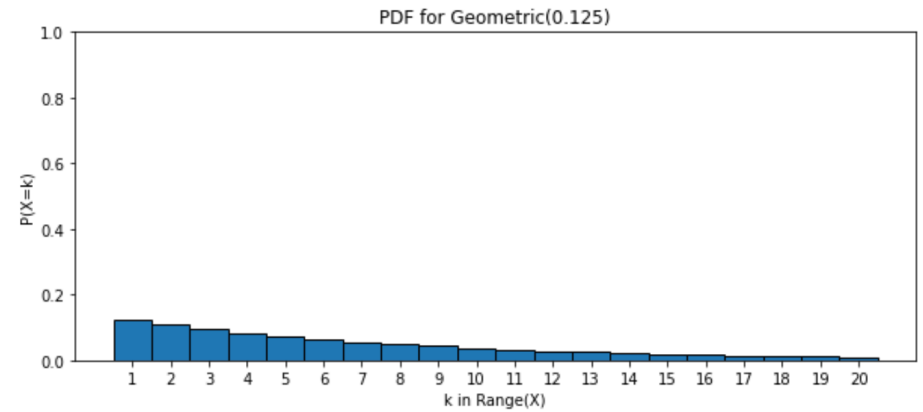
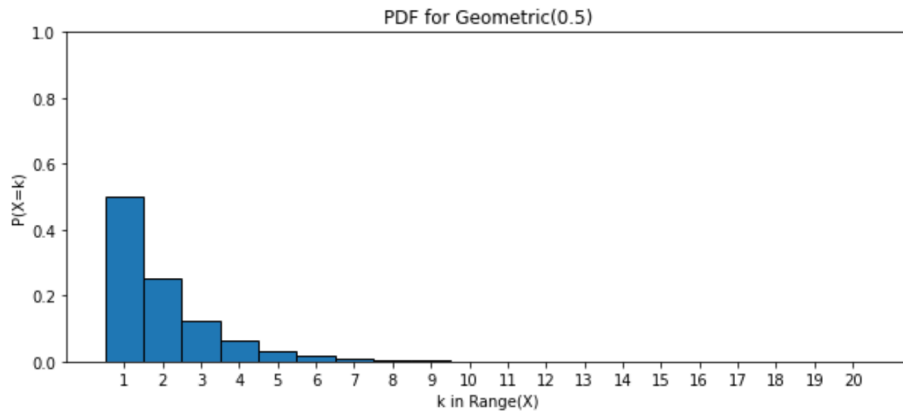


017102160589794829012301007746171778567521588915058841117534090731622704232496308933461758317509424634977207693680961
64766235300691487892523901216163145802534189784427278102661693600072968022831792925342059988541879034812145726594477
24885783970380045361522609589454333575435229584312027097371957397496610391726926221946516763024890340650869934368996
001000486815989173337517058173795352029896149251362864744109963897455046533627237113721335583463188230168234107247
663627766711821480637048749699256481576396797558440296666745019108377561061085772202450227198908472754777154009764488
810339034741385958282080195431904026139595088871662534736561526990080685169980064637294534376537356333444709630
60871268601241105782472108442808621372619249626839947348152989506436281038709158878184111979639577191586304952439741
031014553967451800744589559960974218592205263982442520110590118643104348689635668752622961565577993408914590451947558
542926588869881713267139595303963417232270047656583645874843658687455775369770338699734649542866125552493304647405989
2215639190024820371682446977595249332666959790037042670113750442825052882858884244799158937422640123646473433730
580005535127464003521417001999774603812012067323839873978754498842164928005720899501464714716378173641888697968990329
041286803965598164981377999797198389139537177371387529233184535026059113362810457964663220425972047991701586094805453
58197455731465281260726175115488698886751834169724647163198881595904870581095061450296552248294064273841844365831781
456721733016696264362322423002631281457187956938768520929852230752322863789400985542060938324941418339428406853
569099659769431429822418279556313974779501068070049381038369248739212943955202357039920458646213740364136423266895
578539155542750121511684830603728256206420154567968275835427314507743903732610202153783543908445304344532033265947180
51691323693064446164621132974268718056676613083421614403936271959657132428421231799245092767186813571440980086467891
5758828583214317565167414640988147840448606989822901782791606027024859794738191546144444620631985540011470078400061
57424754910116039582646936760645019739076959088569364726490427211973984780895548007246457176572349491202242994276331
340616425257317827862328626641724704365305387157851445506722439381936525099799870807647427003844677660134390209433607
71131720112510339293223338851902488380178016365544181035354657766146929231025306020324377416019094135451898632992007
7356467571047029034712824053300224308584815302646450712314854705492334414492278801651536799646349027992108398311492
276822736698108539941299836981590018109317359302947285116563887907757366888796072853883339504621074549928190144434838
933800378320281332888756432886021327794304244236918162844916157593538521725670311551463762302271907970929770355671946
70716570268850587779501022571680864930688200270247680069302876142897473367028677938804028547661732956255265347129450
19367875662933679345814982974551695394137493756604558527808589353418486204200203645512862230823119909290185642443850
554808131662581261288401823962588824482896479035078825937187484223754123622308031268101563660714857751512787430276309
015936362962817431846154734230934955240275340447647981037666401551239233971591624509753913765588455578015887925467346
1415319733789389375294197188764957796898518272837904602757928943824195608388942159591199235842868304996583510456153
5315013194074249057546695455720396504748826951913302430089281791710891150560050747131085612380052984641976681005337150
677126311937236261699009823882325728322304115033263476607390165046894862552605369499122284702031800642946301780609574
3027927564272748332014405918231167121142421714282675992009402035388207136729096754340816183198732823285502136148033862
833322388188272794687439244059391467771164109051439225888304452625838929861858304911855824840274350780046031290507063
857181048613930870651541701703650653968750658011101649740052251613539158444396393925867502709142783668944879961358164
404585760159874984505215215974449784401093015525401080302362716208419434338396624734

[illegible]

- The Geometric Distribution models the following scenario
 - There are **independent** Bernoulli trials (possibly infinitely many)
 - Each trial is a success with probability p
 - The process stops at the first success
- X = number of trials until the first success (including the successful trial)
- Notation: $X \sim \text{Geometric}(p)$
- Canonical Problem: Take a coin whose probability of heads is p and count the number of flips until the first head.

Geometric Distribution



- $X \sim \text{Geometric}(p)$
- PDF is simple to compute:

$$\Pr(X = k) = (1 - p)^{k-1} p$$

F F F F F S



k-1 failures before first success

- So is CDF:

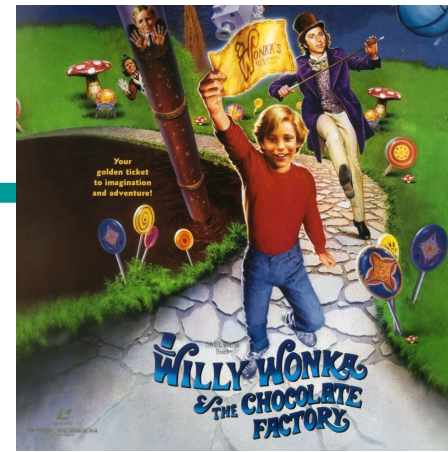
$$\Pr(X > k) = (1 - p)^k$$

F F F F F



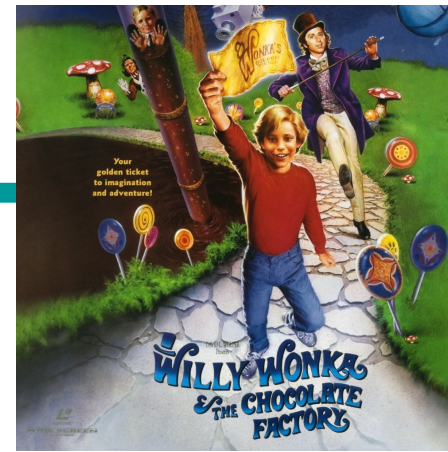
k failures so far....

$$\Pr(X \leq k) = 1 - (1 - p)^k$$



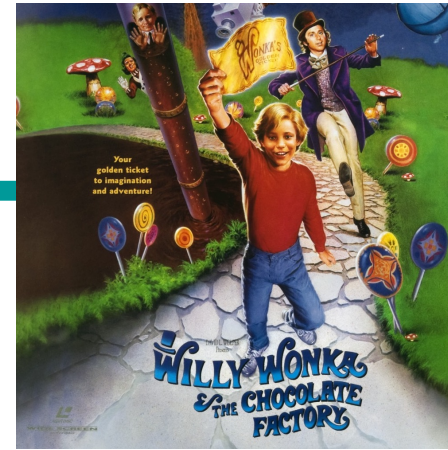
- Willy Wonka puts a golden ticket in each chocolate bar with probability 0.01, independently
- Charlie buys chocolate bars until he gets a golden ticket
- What is the Pr that Charlie buys **exactly** 5 chocolates?

(A) 0.99^4 (B) $0.99^4 \cdot 0.01$ (C) 0.99^5 (D) none of the above



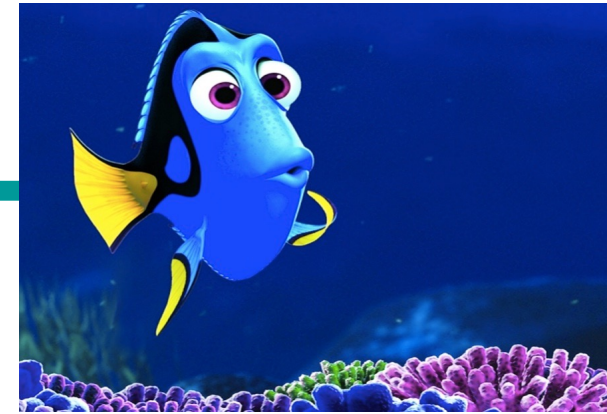
- Willy Wonka puts a golden ticket in each chocolate bar with probability 0.01, independently
- Charlie buys chocolate bars until he gets a golden ticket
- What is the Pr that Charlie buys **at least** 5 chocolates?

(A) 0.99^4 (B) $0.99^4 \cdot 0.01$ (C) 0.99^5 (D) none of the above



- Suppose Charlie already bought 5 chocolates, none of which had a golden ticket
- What is the Pr that he buys at least 5 **more** chocolates?

(A) 0.99^4 (B) $0.99^4 \cdot 0.01$ (C) 0.99^{10} (D) none of the above



- **Theorem** Let $X \sim \text{Geometric}(p)$.

$$P(X > n + m \mid X > m) = P(X > n)$$

It will take $> n+m$
total flips.

You've flipped
it m times.

It will take $> n$
more flips.

- Version for a 10-year old:

You flip a coin until heads. The coin doesn't remember how many tails have occurred, so you start over exactly the same with every flip.

- Proof for a 10-year:

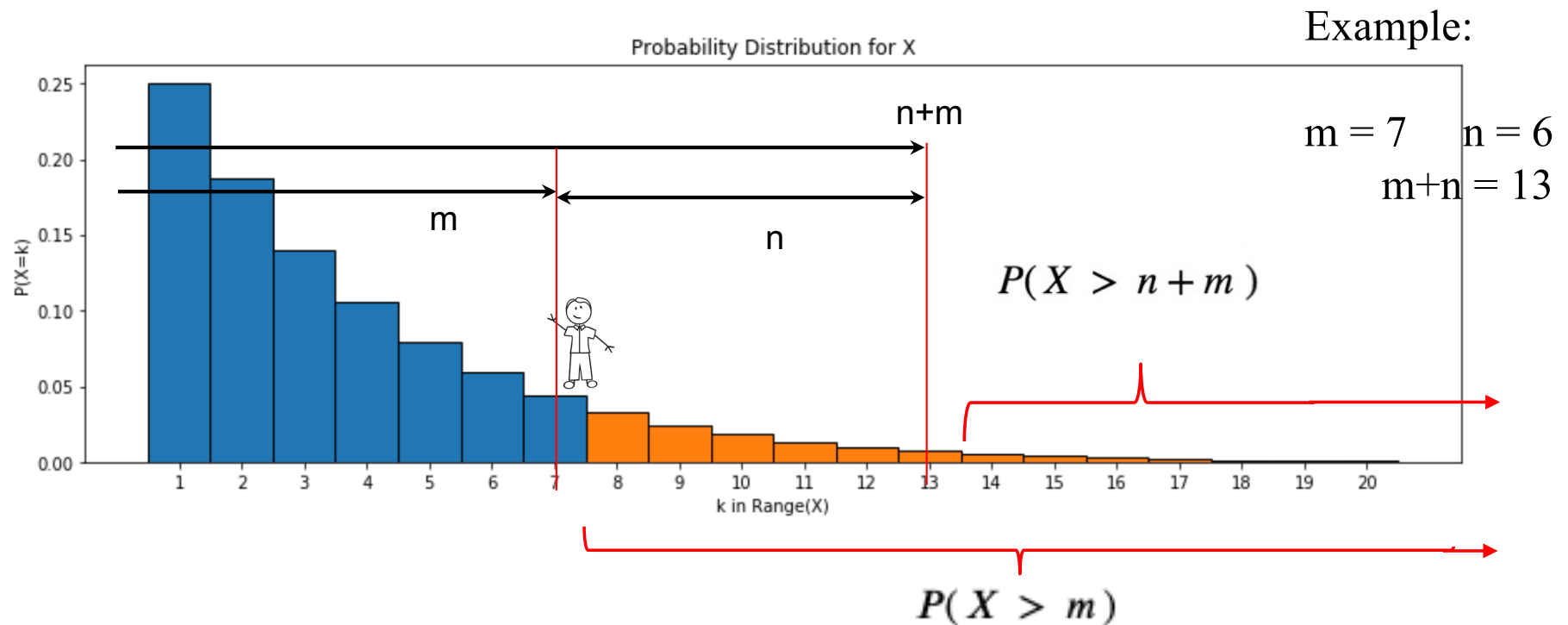
Suppose I'm flipping a coin to get heads and while I'm doing it, someone walks in the room, not knowing how many tails I've gotten already. Why would the coin behave differently with this person watching?

Memoryless Property: Version for CS 237

Theorem A random variable X is called **memoryless** if, for any $n, m \geq 0$,

$$P(X > n + m \mid X > m) = P(X > n)$$

Fact: For any probability p , $X \sim \text{Geometric}(p)$ has the memoryless property.

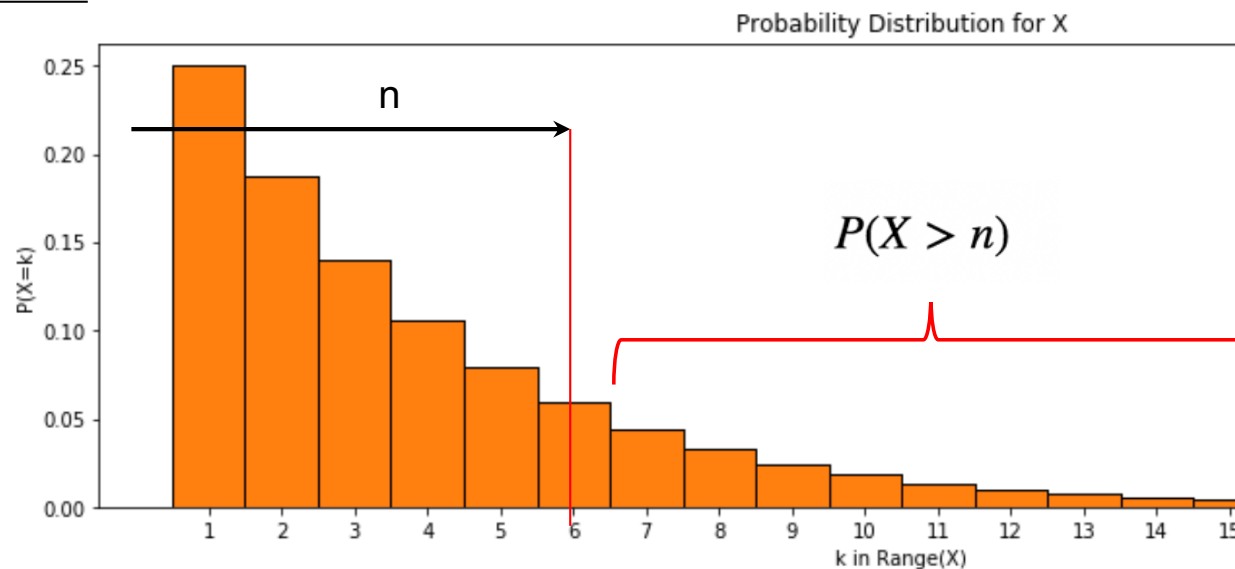
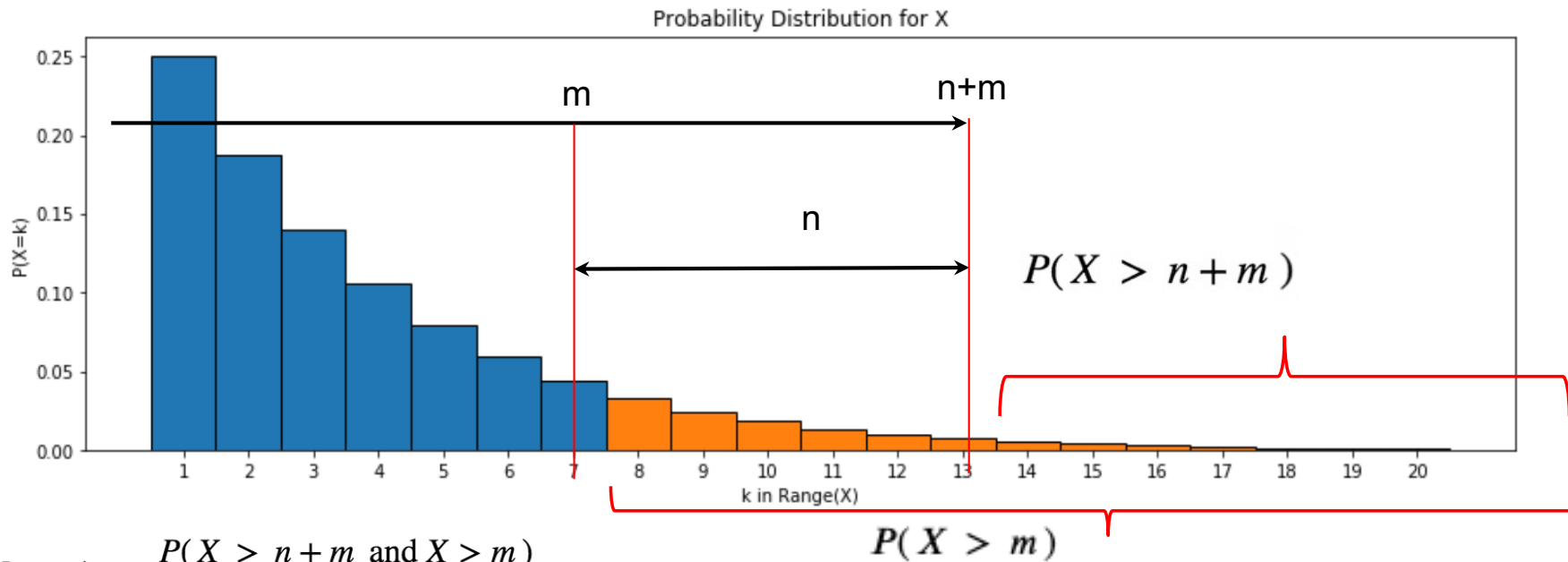


(In fact, the Geometric is the only discrete distribution with this property; a continuous version of the Geometric, called the Exponential, is the other one.)

Memoryless Property: Version for CS 237

Proof:

$$\begin{aligned}
 P(X > n+m | X > m) &= \frac{P(X > n+m \text{ and } X > m)}{P(X > m)} \\
 &= \frac{P(X > n+m)}{P(X > m)} \\
 &= \frac{(1-p)^{(n+m)}}{(1-p)^m} \\
 &= (1-p)^n \\
 &= P(X > n)
 \end{aligned}$$



Suppose in Park Street a train arrives every 15 minutes, and the probability that an arriving train is for the Green line is $1/5$. Suppose you have just arrived at Park Street and are waiting for a Green Line train.

What is the probability that you will have to wait at least 45 minutes (3 train arrivals) for a Green Line train?

Consider these three situations:

- (1) When you arrive, you just miss a Green Line train. How long to wait?
- (2) When you arrive, you see a Red Line train leaving. How long to wait?
- (3) Five trains have come and gone, one of them a Green Line train, but you missed it because you were reading your phone. How long to wait?

How are these different?

BACK TO TOPHAT!

- The memoryless property also simplifies the analysis of the Geometric
- $X \sim \text{Geometric}(p)$
- $E(X) = ?$
- $\text{Var}(X) = ?$

Derivation using law of total expectation (aka case analysis)

On the board:

$$E(X) = E(X|X = 1) \cdot \Pr(X = 1) + E(X|X > 1) \cdot \Pr(X > 1)$$

$$= 1 \cdot p + E(X|X > 1) \cdot (1 - p)$$

$$= p + (E(X) + 1) \cdot (1 - p)$$

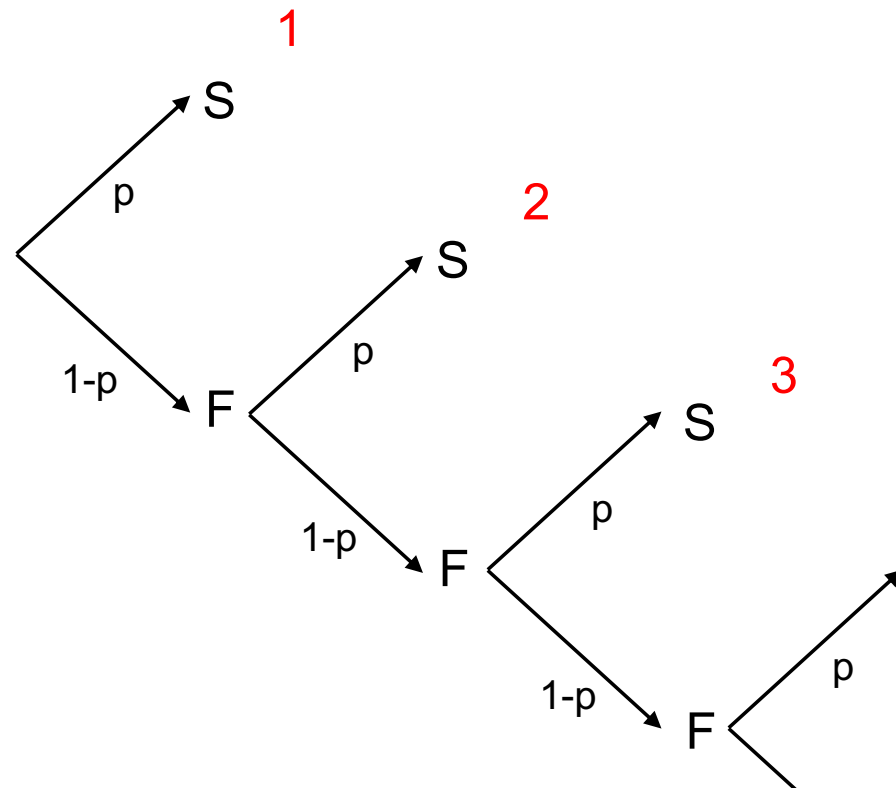
$$= p + E(X) - E(X) \cdot p + 1 - p$$

$$E(X) = E(X) - E(X) \cdot p + 1$$

$$0 = -E(X) \cdot p + 1$$

$$E(X) \cdot p = 1$$

$$E(X) = \frac{1}{p}$$

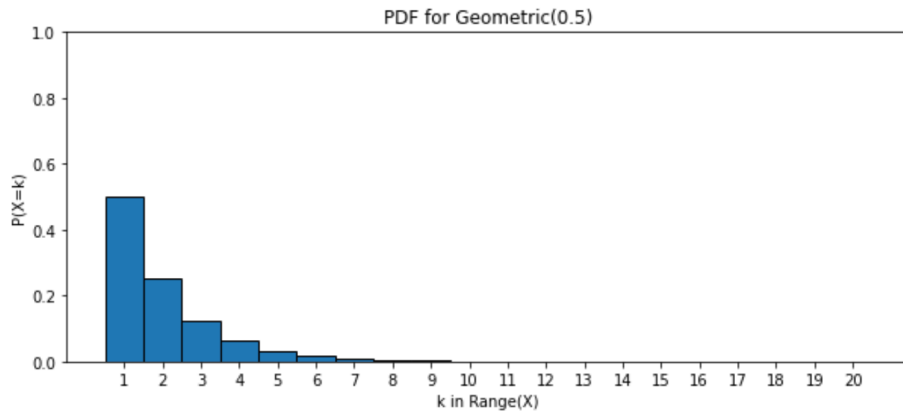


- $X \sim \text{Geometric}(p)$

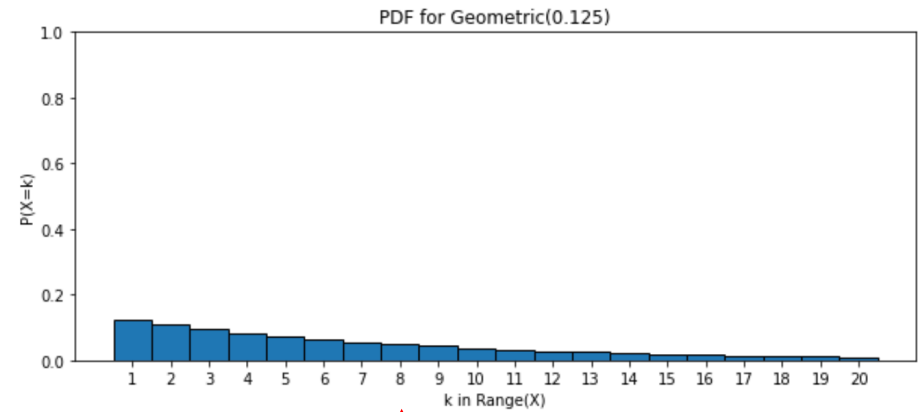
$$E(X) = \frac{1}{p}$$

$$\text{Var}(X) = \frac{1-p}{p^2}$$

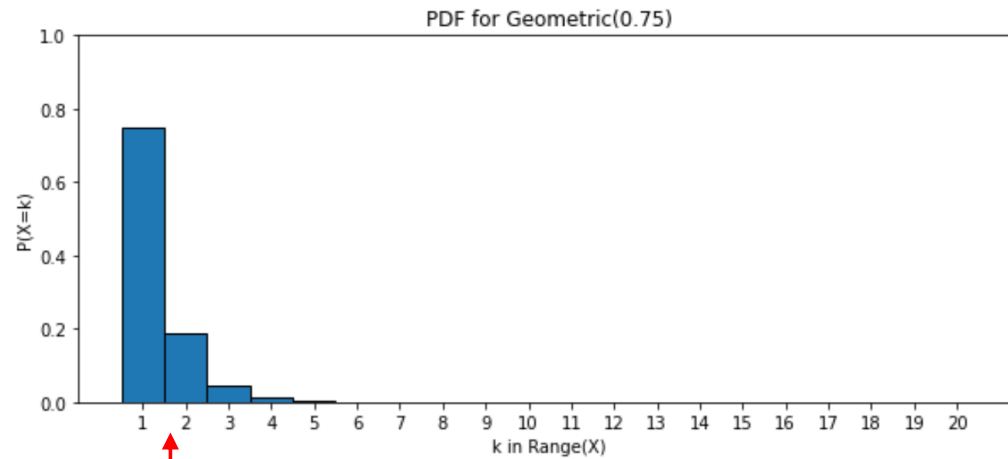
Geometric Distribution



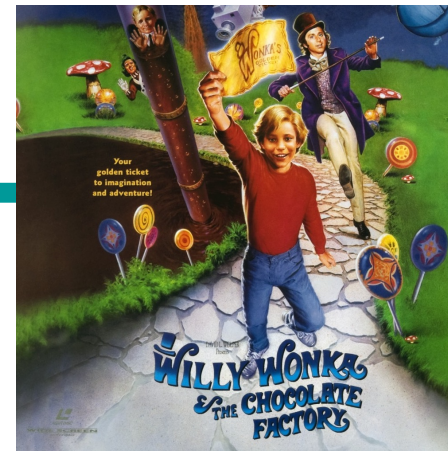
$\mu = 2$



$\mu = 8$



$\mu = 4/3$



- Willy Wonka puts a golden ticket in each chocolate bar with probability 0.01, independently
- Charlie buys chocolate bars until he gets a golden ticket
- Let C = number of chocolates that Charlie buys

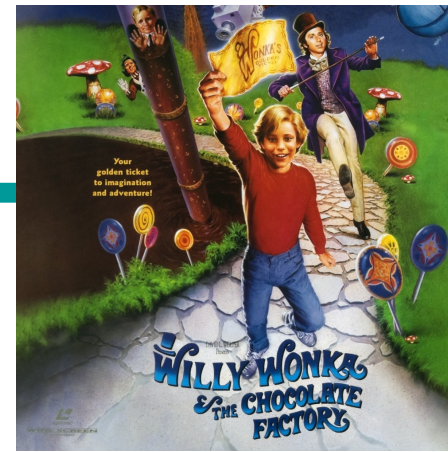
What is $E(C)$?

(A) 2

(B) 3

(C) 100

(D) none of the above



- Willy Wonka puts a golden ticket in each chocolate bar with probability 0.01, independently
- Charlie buys chocolate bars until he gets a golden ticket
- He's already bought 5 chocolates and no golden ticket.
- How many **more** bars does he need to buy in expectation?

(A) 2

(B) 3

(C) 100

(D) none of the above

A natural generalization of the Geometric is to ask how long before $r \geq 1$ successes...

Formally, if $Y \sim \text{Bernoulli}(p)$, and

$X =$ “The number of trials of Y until the first r successes”

then we say that X is distributed according to the Negative Binomial Distribution with parameters r and p , and write this as:

$$X \sim \text{NegativeBinomial}(r, p)$$

Example: An oil company conducts a geological study that indicates that an exploratory oil well should have a 20% chance of striking oil. In order to be profitable, they must strike oil 5 times a year. What is the probability that the 5th strike comes on the 30th well drilled?

Negative Binomial Distribution

$$X \sim \text{NegativeBinomial}(r, p)$$

$$\Pr(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$$

$$E(X) = \frac{r}{p}$$

$$\text{Var}(X) = \frac{r(1-p)}{p^2}$$

NOTE: Negative Binomial is the **sum** of independent Geometric RVs:

$$X = X_1 + X_2 + \dots + X_r$$

If it takes exactly k trials to get the r th success:

then the probability of any specific sequences of k trials, with r successes and k-r failures, is

$$p^r (1 - p)^{k-r}$$

Since the last trial is S, then we count the number of such sequences by choosing $r-1$ successes out of $k-1$ trials, yielding:

$$\Pr(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$$

Single

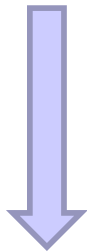
Sum Of

Counting:

Bernoulli(p)



Binomial(N, p)



Waiting:

Geometric(p)



NegativeBinomial(r, p)