# Reading for Lesson 1

## Rules of Probability

An event is a capital letter like A

A is event we roll a 4 (on 6 sided dice)

P(A) = 1/6

If we don’t want 6 functions for each result, but 1 function and 6 possible results

Make function be P

P(X=4) = 1/6

All probabilities must have:

1. P < 1 & P > 0
2. For I being every value it could take in this dice fxn

For I in range(6):

sumP(X=i) = 1

3. The complement of an event A\*\*c means the event does not happen

Since probabilities must add to one

P(A\*\*c) = 1 – P(A)

4. If A and B are 2 events, the probability that A OR B happens

(this is an inclusive or, meaning that either A or B or both happen)

Is the probability of the union of the events

Union represents or, intersection represents and

## Odds:

Ex. Rolling 4 on a die is event A

P(A) = 1/6

Odds for event a O(A)

= P(A) / P(A\*\*c)

= P(A) / (1 – P(A))

In event A, 1/6 chance

O(A) = (1/6) / (5/6) = 1/5

Odds can be expressed as 1:5 or 5:1 (odds against)

Ex 2.

Event w P(event B) = 3/10

O(B) = (3/10) / (7/10)

= 30/70 = 3/7 odds

3:7 odds or 7:3 odds against

Ex 3.

P(C) = 4/5

O( c) = 4/5 / 1/5

= 4/1

= 4:1 odds

Can also calculate P(X) from odds

If event D has a:b odds (with both a & b > 0)

P(D) / (1 – P(D)

= a/b

= P(B) \* b

= a – P(b) \* a

= P(B) = a / (a + b)

Ex of odds -> prob:

O(E) = 2:5

P(E) = 2 / (2 + 5)

P(E ) = 2 / 7

## Expectation:

The expected value of random variable X is a weighted avg of values X can take

With weights given by the probabilities of those values

We can calculate the expected value

If X can only be finite num of values:

X1, x2, … xn

Then calculate by

E(X) = sum(xi \* P(X = xi))

Ex of this, find the expected value of a 6-sided die

1 \* 1/6 + 2\*1/6 + … 6\*1/6

= 3.5

This means that if you rolled a ton of times, averaged them out, its likely close or approaches 3.5

# Notes from Lesson 1 videos

Examples of questions we can ask

P(X=4) # prob of rolling 4

P(fair) # prob die is fairly weighted

P(rain) # prob it rains at any point tomorrow

P(drop packet) # prob router does so

P(Y1 > Y2) #prob router from 1 company is more reliable than prob router from 2nd company

P(universe expands forever) # existential

3 frameworks under which we can define probabilities

1) Classical

Outcomes that are equally likely have equal probabilities

In case of fair die, P(X=4) = 6

P(sum of 4 on 2 rolls)

1/6 \* 1/6 = 1/36 equally likely outcomes on the pair of rolls

How many give 4? 3-1, 2-2, 1-3

= 3/36 = 1/12

2) Frequentist

Hypothetical infinite sequence of events – look at the relative frequency

If rolling die infinite number of times

1/6 of time, get a 4 -> P(X=4) = 1/6

Drop packet ex:

Lose 1 / 10,000 packets – define p(drop packet) = 1/10,000

Rain tomorrow – what fraction of infinite tomorrows have rain – strange

Fair – infinite rolls – many rolls wont change it P(fair) = {0, 1}

P(universe expand forever) = {0, 1} – either it does or doesn’t or multiverse – what fraction of them do

Can run into philosophical questions, objectivity is illusory, interpretations that are not intuitive

## Video 1.2

3) Bayesian

Is from a personal perspective – takes into account what you know about problem/situation

If have different info – may be diff

Can work well in math foundation, can be intuitive

P(rains tomorrow) – willing to take bet – win 4 if rains, lose 1 if doesn’t, odds 4:1

If fair, you should be able to take opposite – win 1 if no rain, lose 4 if doesn’t

P(rain) = 1 / (1/4) = 1/5

Expected return

4(1/5) – 1(4/5) = 0

1(4/5) – 4(1/5) = 0

Coherence

Probs must follow rules for probability

If don’t, someone can construct rules where youre guaranteed to lose money – dutch book

# Reading for Lesson 2

When there are 2 discrete possible events



When there are 3



If events A1,…Am are mutually exclusive and only 1 can occur and they sum to 1 (ie if they form a partition of the space)

Then bayes theorem can be written as



For continuous distributions, sum gets replaced w an integral

# Notes from Lesson 2 videos

### Video 2.1 Bayes Theorem

Conditional probability

Considering 2 events that are related to each other

P (A given B happened) = P (both A and B happen) / P(B happens)

Ex. Class of 30 students; 9 females in class; 12 cs majors; 4 female cs majors

P(F in this class) = 9/30 = 3/10

P(CS major in class) = 12/30 = 2/5

P(F and CS major) = 4/30 = 2/15

Now can ask conditional prob questions

P( Female given theyre a cs major) = P(F and CS major) / P(CS major)

= (4/30) / (12/30)

= (1/3)

Can also think of it as subsegment 4 females/12 cs majors -> 1/3

Whats prob of being female given theyre not a cs major

P(F | CS\*\*c)

= P(F and not CS) / P(not CS)

= P(F & !CS) / P(!CS)

= (5/30) / (18/30)

= 5/18

Concept of independence

When event A doesn’t depend on B

Doesn’t matter whether or not B occured

P(A | B) = P(A)

Also

P(A & B) = P(A) \* P(B)

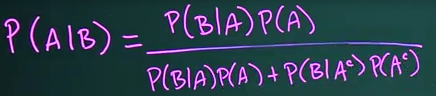
Can see that prob female given cs is not equal to prob female \* prob cs

P(F|CS) != P(F)

Being female and being CS are not independent / are dependent

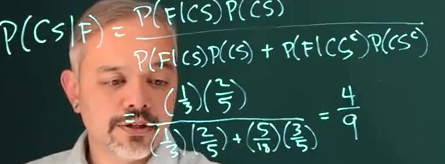
### Video 2.2 Bayes Theorem

P(A|B) = ( P(B|A) \* P(A) ) / (P(B|A) \* P(A) + P(B|A\*\*c) \* P(A\*\*c)



Math works out to be the same as P(A&B) / P(B)

Ex.



Direct def

P(CS|F) = P(CS & F) / P(F)

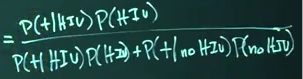
= 4/30 / 9/30 = 4/9

P(+|HIV) = .977

P(-/noHIV) = .926

P(north American HIV) = .0026

Find out P(HIV|test +)



.977 \* .0026

.977 \* .0026 + (1-.926) \* (1 - .0026)

= .033

Num of FPs greatly outnumbers TPs, get more of them than TPs

Test in subpopulation

Start w prior beliefs, collect data, condition on the data, lead to posterior beliefs (guess)

# Reading for Lesson 3

Indicator functions / Heaviside / unit step fxns

Fxn that takes value 1 if argument is true, takes value 0 if argument is false

I(A)(x) or 1(A)(x) or I(A)



Useful for making sure you don’t take log of negative number for example

Ex 2

Function for the exponential distribution (4.1) can be written

F(x) = lambda exp (-lambda x) I(x>=0)

Expected values…

# Notes from for Lesson 3 videos

### Video 3.1 Review of distributions

#### Bernoulli distribution

When 2 possible outcomes

Random variable X is distributed as or follows distribution B(p)

P(X=1) = p

P(X=0) = 1-p

Write as fxn for all possible outcomes

Prob rv X takes the value x given a specific probability p

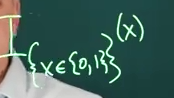
F(X=x | p)

F(x|p)

P\*\*x \* (1-p) \*\* 1-x

#### Indicator fxn

Can write with an indicator fxn



Gets evaluated first. Can avoid squaring neg number

This is the probability mass fxn

Some books make distinction – if discrete variable, call it prob mass fxn

If continuous var, call it prob density fxn

Can call all prob density fxn

Expected value

Theoretical avg/mean

E[X] = xp.sum()

= 1p + 0(1-p) = p

Variance of it – sq. of stddev

= p(1-p)

#### Binomial

Generalization of Bernoulli when there are n repeated trials is a binomial

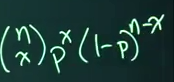
Binomial is the sum of n independent bernoullis

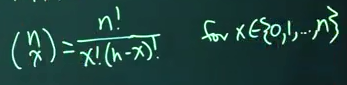
n is number of events

p is the prob of a hit occuring

X follows Bin(n,p)

Probality fxn

P(X=x|p) = f(x|p) = 



##### Example from quiz 3.1 – calc P(X=0) when X ~ Binom(3, .2)

X = 0, p = .2, n = 3

P(0 successes) = .2\*\*0 \* (1- .2) \*\* (3-0)

= 1 \* .8\*\*3

= .512

Cant do

n! / x! \* (n-x)!

= 3! / 0?

Example2 from quiz 3.1 – calc P(X <= 2) when X ~ Binom(3, .2)

X= 0 + X=1 + X=2, p = .2, n = 3

P(1 success) = .2\*\*1 \* (1- .2) \*\* (3-1)

= .2 \* .8\*\*2

= .128

P(2 success) = .2\*\*2 \* (1-.2) \*\* (3-2)

= .04 \* .8 \*\* 1

= .032

= .672

Try again

X = {0, 1, 2}, p = .2, n = 3

X = 2:

(n choose x) = n! / (x! \* (n-x))

3! / (2! \* (3 – 2))

= 6 / 2

= 6

Then

6 \* p\*\*x \* (1-p)\*\*(1-x)

= 6 \* .2\*\*1 \* .8\*\*1

= 6 \* .2 \* .8

= .96??

Expected value

E[X] = np

Var(X) = np \* (1-p)

### Video 3.2 Review of distributions

These are continuous random variables – requires calc

Defined by its Prob density fxn

Which is proportional to the p(x) that a random variable will take a specific value

(can take inf values) – so its like a differential

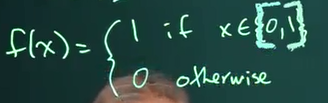
If integrate a pdf over an interval,

Gives the p(x) that the rv will be in that interval

Ex – uniform distribution

X is uniformly distributed b/w 0 and 1

X ~ U[0,1]



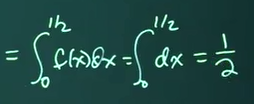
Or as indicator fxn

= I {0 <= x <= 1}(x)

Can then ask questions about probabilities like

P(0 < x < ½)

To get it, integrate the density fxn



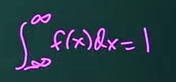
P(0 <= x <= ½)

Integral doesn’t depend on endpoints being in or not, still ½

P(x=1/2) = 0

Because infinite outcomes

#### Prob density fxn rules:

1. 

Has to be probability 1 that something happens

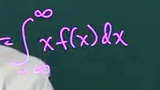
2. 

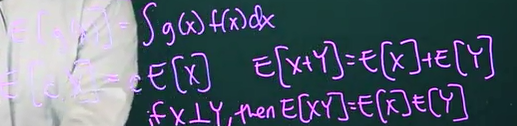
Densities have to be non-negative for all possible values of x

Expected val for continuous random variable

As integral of x \* f(x)dx

Analogous to sum of discrete var

E[X] = 



In general, expected val of f(x)gx is integral of g(x)f(x)dx

Can pull the constant out

If looking at expectation of sum of 2 random variables, just sum them

If x is independent of y, x\*y is the product

### Video 3.3 Review of distributions

#### continous dists

##### Expontential dist

X ~ Exp(lambda)

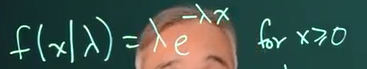
Lambda is events occurring at particular rate

The exponential is the waiting time b/w events

Ex is the waiting time for a bus

Density fxn (for x >= 0)

F(x|lambda) = lambda \* e \*\* (-lambda \* x)



E[X] = 1 / lambda

Ex. Bus every 10 minutes, without knowing schedule, 1/10 chance it comes this minute

Quiz 3.1-3.2 #2

Lambda is 5, so E(X) is 1/5 = .2

Var(X) = 1 / lambda\*\*2

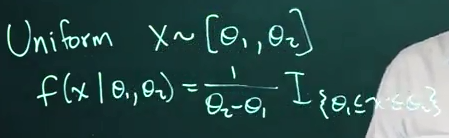
##### Uniform dist

X ~ [thetha1, theta2]

Density

F(x| theta1, theta2) = 1 / (theta2 – theta1) given that or I {theta1 <= x <= theta2}

Uniform dist’s Density fxn is



Q2: If X ~ uniform(0,1) then what is the value of P(-3 < X < .2)

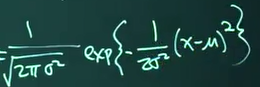
.2 = theta 2 in this case? 0 is theta 1?

= 1 / (.2-0)

= 5? No. somehow is .2

##### Normal fxn

Is X ~ N(mu, sigma\*\*2)

F(x | mu, sigma\*\*2) = 

E[X] = mu

Var(X) = sigma\*\*2

Gamma dist and Beta dist

##### Gamma dist:

If events X1,…Xn are independent and identically distrubted across Exp(lambda) waiting times b/w events

Then the total waiting time for all n events to occur

Y = sum(x)

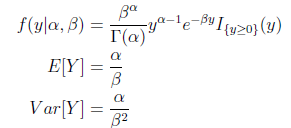
Will follow a gamma distribution w

shape parameter alpha = n

And rate parameter beta = lambda

For a set of Ys across a gamma distrubtion with alpha = num events, beta = time b/w each event

Pdf:



E[Y] is just 100 bus events / 10 minutes = 10 events/minute? no

Var[Y] = 10

Gamma fxn

A generalization of the factorial fxn that can accept non-integer arguments

If n is a + int, then 

Exponential dist is a case where gamma dist has alpha = num events = 1 event

Gamma is used to model +-valued, continuous quantities whose distribution is right-skewed

As alpha / num of events increases, the gamma dist more closely resembles the normal distribution

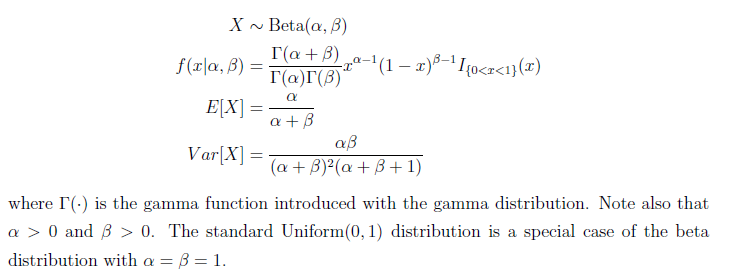
Ex. Bus that comes every 10 minutes, except they are not dependent on each other

##### Beta dist

Random variables which take on values between 0 and 1

Commonly used to model probabilities

Also has alpha = num events? Beta = space b/w numbers?



##### T dist

If we have normal data, can use normal pdf

However we might not know the value of sigma, the sd of the population

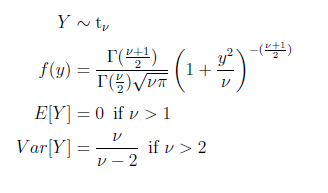
If we estimate it w/ sample data, can replace it with



This causes the normal to be standard t distribution with v = n-1 degrees of freedom

The t dist is symmetric, resembles normal dist but with thicker tails

As df increases, looks more like standard normal dist



#### Discrete Distributions:

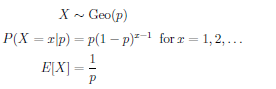
##### Geometric

Number of trials needed to get the first success

The number of Bernoulli events until a success is observed

Ex. How many coinflips until a heads

It takes values on the positive integers starting with one (at least 1 trial is needed to observe a success)



If the p(x) of getting a success is p, then the expected number of trials until the first success is 1/p

Ex. P(4 coinflips with 0 heads)

This is the same as asking what is P(X > 4) where X ~ Geo(1/2)

P(X = 4 | .5)

= .5(.5)\*\*(4-1)

= .5 \* .125

= .0625 = 1/16

##### Multinomial dist

Another generalization of the Bernoulli and binomial dists

A binomial when there are more than 2 possible outcomes

If we have n trials, and k outcomes with occur with prob p1, … pk

Ex: rolling a die that might be loaded

n = total number of rolls

k = 6 (possible outcomes)

p1 = prob of rolling 1 (also call it x1) … p6 = prob of rolling 6 (also call it x6)

sum of rolls = n (6)

sum of probs = 1



N! is n \* n-1 \* … \* 1

4! = 4\*3\*2\*1 = 24

Expected number of observations in category i (not sure what I is)

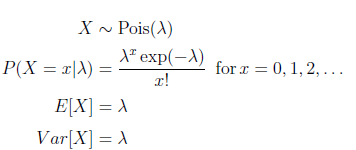
= n \* pi

##### Poisson dist

Used for counts

Poisson process is when events occur on avg at a rate, events occur 1 at a time, and are independent of each other

Lambda > 0 is the rate at which we expect to observe the thing we are counting



###### Ex. Earthquakes in western us follow approx. a Poisson process with rate of 2 earthquakes / week

P(at least 3 earthquakes in the next 2 weeks)?

Answer:

Rate per 2 weeks = 2\*2 = 4, so let X ~ Pois(4)

Want to know P(X >= 3)

= 1 – P(X<= 2)

= 1 – P(X=0) – P(X=1) – P(X=2)

= 

Continuous Dists:

## Lesson 3.1 quiz attempt – got 6 first time but forgot answrs

Q1: When using random var notation, big X denotes

A: A random variable

Q2: When using random var notation, little x denotes

A: a realization of a random var

Q3: When using random var notation, X ~ denotes

A: distributed as

Q4: x=3, what is f(x) = -5 I {X>= 2}(x) + x I {x < -1}(x)

A: no idea, -15

A2: got this one wrong – says only one of the terms will be non-zero

Indicator is always either 0 or 1, so this should be -5? Correct.

Q5: x=0, what is f(x) = -5 I {X>= 2}(x) + x I {x < -1}(x)

A: no idea, 0

Q6: which scenario could be modelled using a Bernoulli random variable?

A: predicting whether hockey team wins next game (tie counts as a loss)

Q7: Expected value of this random var: X takes on values {0,1,2,3} w probs {.5, .2, .2, .1}

A: .9 (added their sumproducts)

Q8: Which scenario could we model using a binomial rand var (with n>1)?

A: predicting number of wins in a series of 3 games against same opponent

Q9: X ~ Binomial(3, .2); calculate P(X=0); round to 2 decimal places

A: got it, see notes

Q10: X ~ Binomial(3, .2); calculate P(X<=2); round to 2 decimal places

A: .672 (adding up 0,1,2)

A2: got this one wrong, is it .96? Wrong

A3:

Want to get the cdf not the prob density fxn

F(x; p, n) = sum(i->x) n! / (i! \* (n – i)! \* p\*\*I \* (1-p)\*\*(n-i)

First do 0:

N=3, p=.2, i=0

= 3! / (0! \* (3-0)!

= 6 / (1 \* 6)

= 1

= p\*\*I \* (1-p)\*\*(n-i)

= .2\*\*0 \* .8\*\*(3 – 0)

= 1 \* .8\*\*3

= .512

Then do 1:

= n! / (i! \* (n – i)!

= 3! / (1! \* (3-1)!

= 6 / (1 \* 2!)

= 3

= .2\*\*(n) \* .8\*\*(n-i)

= .2\*\*(3) \* .8\*\*(3-1)

= .2\*\*3 \* .8\*\*2

= .008 \* .64

= 3 \* .000512

= .01536

Then do 2:

= 3! / (2! \* (3-2)!

= 6 / (2 \* 1)

= 3

= .2\*\*(n) \* .8\*\*(n-i)

= .2\*\*3 \* .8\*\*(3-2)

= .008 \* .8\*\*1

\*3

= .0192

Sum3 = .512 + .01536 + .0192 = .71936 = .72

Still wrong. No idea. No examples online

## Quiz 3.1-3.2 – 11 questions

Q1. If cont rand var X has prob density fxn f(x),

What is interpretation of integral -2->5 f(x)dx

A: P(X>= -2 U X <= 5); union is the same as &, upside-down union is same as |

A2:

Q2: If X ~ uniform(0,1) then what is the value of P(-3 < X < .2)

A: .2

Q3: If X ~ exponential(5) , find the exp val E(X)

A: 1/5 = .2

Q4: Which scenario could we model using an exponentially distributed rand var?

A: events occurring at a rate, exp is waiting time

The lifetime in hours of 1 lightbulb

Q5: If X ~ Uniform(2,6) which is the pdf of x?

A: cant see images, 3rd one

A2:

Q6: If X ~ Uniform(2,6) what is P(2 < X <= 3)?

A: 6-2 = 4.

3-2 = 1

¼

= .25

Q7: If X ~ N(0,1) which is the pdf of x?

A: cant see pictures, 3rd option

A2:

Q8: If X ~ N(2,1) what is expected val of -5X? This is denoted as E(-5X)

A: don’t even understand the question. Asking prob that X is -5?

0?

A2:

Q9: Let X~N(1,1) and Y~N(4, 3\*\*2); what is value of E(X + Y)?

E(X) = 1

E(Y) = 4

= 5?

Q10: The normal dist is also linear in sense that if X ~ N(mu, var), then for any real constants a != 0

and b, the dist of Y = aX + b is distributed N(amu + b, a\*\*2, var)

Using this fact, what is distribution of Z = X – mu / sd

A: N(mu, var) (no idea)

A2:

Q11: which of following random variables would yield highest value of P(-1 < X < 1)?

A: X ~ N(0, .1)

# Reading for Lesson 4

# Notes from for Lesson 4 videos

## 4.1 - Frequentist approach for making prob statements

Flip coin 44H, 56T

View it as a sample of infinite flips

Rand var X ~ Bernoulli(P)

Assuming p is fixed since its 1 coin

Ask best estimate of getting Head or P

Apply the CLT:

the sum of the 100 flips ~ N(100p, 100p\*(1-p))

95% of time, will get result within 1.96 SDs of the mean

100p – 1.96 \* sqrt( 100p \* (1-p))

And

100p + 1.96 \* sqrt (100p \* (1-p))

Observe that 44H or sum(xi) = 44

P-hat or estimate of p = 44/100

Plug this value into theoretical quantity

CI

44 +- 1.96 \* sqrt ( 44\*.56)

= 44 += 9.7

(34.3, 53.7)

Can say 95% confident that p is in (.343, .537)

Reasonable that p is a fair coin

Means that if repeating trial many times,

95% of the intervals we make will contain the true value of p

Rather ask – what’s the probability that this interval contains true value

Either 0 or 1

Bayesian – able to compute interval, and using random interpretation of unknown parameter, actually say p(x) that true value is in our interval

## 4.2 – Frequentist Inference 2: Estimate mortality rate

Ex. Hospital w 400 patients admitted over a month for heartattacks

1 month later – 72 have died, 328 survived

1st – establish our reference population

Heart attack victims in region,

Heart attack patients admitted to this hospital over larger amt of time

But our data is not a random sample of above

All ppl who might have a heart attack, and who might get accepted to this hospital

Patient Yi ~ Bernoulli(unknown theta we want to estimate)

P(Yi = 1) = theta (mortality rate we are trying to guess)

for all the ppl admitted; a success/1 is mortality, living

Pdf in vector form:

P(all Ys = y | theta)

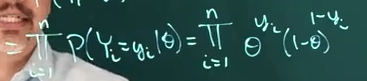
= P(Y1 = y1, … y2 = y2 0/1, …Yn = yn | theta)

Since viewing each person/event as independent, this is same as prob of each event

P(Y1 = y1) … P(Yn = yn)

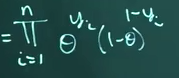
In product notation

Prob of observing actual data | theta



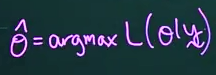
Can think of expression as function of theta

Likelihood fxn is a density fxn as a fxn of thetha

L(theta | Y) = 

One way to estimate theta is by choosing theta that gives us largest value of the likelihood

Maximum Liklihood Estimate – the theta that maximizes the likelihood



Log likelihood 

In practice, often easier to maximize the natural logarithm of the likelihood

If maximize log of a fxn, also max the fxn

Data is independently, identically distributed:

Yi iid B(thetha)

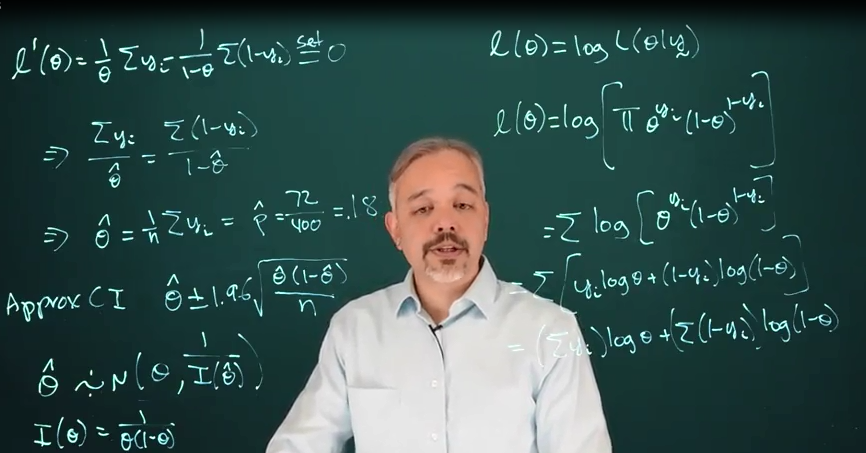
With this independent type of data, likelihood is a product

And so, when we take the log of a product, we get a sum (sums easier to work with)

How to find theta that maxes this fxn?

Take derivative and set it equal to 0

## 4.3 – Frequentist Inference 3 – derivatives



## 4.4 Frequentist Inference 4 – exponential dists

## 4.5 – Intro to R / 4.8 Plotting Likelihood

See notes in bayes.r

## 5.1 Bayesian Inference

Ex: with 2 outcomes – compare it to frequentist

Coin is supposedly loaded, heads 70% of time

Bet that it will be heads

Not sure if its loaded or fair

Flip, get 2 Heads and 3 tails

1st define likelihood

Unknown parameter:

Theta = {fair, loaded}

Data:

What is that probability theta?

X ~ Bin(5, ?)

Likelihood:

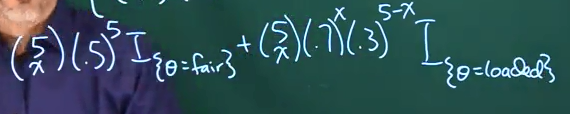
F(x | theta)

{ (5 choose x) (1/2)^5 if theta = fair

(5 choose x) .7^x \* .3^(5-x) if theta = loaded

}

Can re-arrange as indicator fxn



So observed x=2 heads

Whats the likelihood?

F instead of L; in Bayesian, fs everywhere

F(theta(x=2))

= { .3125 if theta = fair,

.1323 if theta = loaded}

MLE theta hat = fair coin

How to answer question how sure are we?

Also, what is p(theta = fair | X = 2)?

= P(theta = fair)

= {0, 1}

## 5.2 Bayesian approach to coin flips

Easily incorporate prior info when you know something before seeing data

You think that 60% chance this coin is loaded

Prior P(loaded) = .6

Updated prior with the data to get posterior beliefs

F(theta | x) = f(x|theta) \* f(theta) / sum f(x|theta) \* f(theta)

