Probability 101

Cmput466/551

Covering…
HTF Ch2 (kinda)…
+ Review of Probability Theory
+ …

R Greiner
Department of Computing Science
University of Alberta

Thanks to R Parr, C Guesterin

Outline

- Foundations
  - Bayes Theorem
  - (Conditional) Independence
  - Dutch Book Theorem
  - Moments: Mean, Variance
- Estimation
  - MLE (Binomial)
  - Bayesian model
- Gaussian (Normal)

Probability: Who needs it?

- Learning without probabilities is possible, provided…
  - No noise in feature values, no noise in labels
  - Correct "concept" $\in$ Hypothesis class
  - World does not change from train to test

  but rare…
- Learning almost always involves
  - Noise in data (training, testing)
  - Uncertainty about hypothesis class
  - …
- Learning systems
  that don't use probability in some way
  tend to be very, very brittle


Bayesian Training Camp |
oilersnation.com/2014/9/16/bayesian-training-camp
OILERSNATION.COM
W.GRETZKY CAPT.
P.CONACHER L.FOGOLIN G.FUHR
C.HUDDY P.HUGHES D.HUNTER
K.LINSEMAN
HOME
NATIONGEAR
VOICES OF THE NATION
SCHEDULE
MEDIA
STICKERS
CONTACT US
ACCOUNT
BAYESIAN TRAINING CAMP
Jonathan Willis
September 16 2014 03:22PM
The Nation
EDMONTO

Bayesian Training Camp |

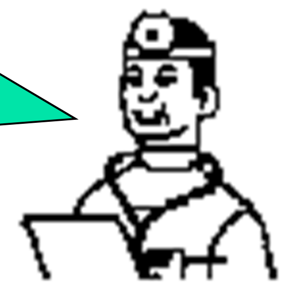
oilersnation.com/2014/9/16/bayesian-training-camp

Let's try it using a player, one in his mid-20's who has played 100-odd NHL games and was an NHL'er for half of last season. We could call him Player X and keep this hypothetical, but to keep this easy to track we'll use a concrete example: Jesse Joensuu. What we're trying to determine is whether Joensuu (or X) is an NHL player; our hypothesis is that he is.

THE PRIOR

The first thing to do is establish what we think about Joensuu *before we see so much as a second of training camp*. This is easiest before camp, when we haven't seen anything and are completely uninfluenced; it's harder to do after the fact. Since I'm writing this, we'll use my estimates – your mileage may vary, and you may have more or less experience watching him play, but it's the process rather than the exact numbers that matter. Here are the main things I know about the player:

- I saw Joensuu play 42 regular season games last year, plus some time in the preseason. He looked good before the year but terrible during it. Based on my observations alone, I don't think much of him.
- Joensuu's numbers from 2013-14 are interesting. His relative Corsi was middling on a lousy team, but he also had flat-out brutal zonestarts. Further he got murdered by PDO (4.4 on-ice



What is P(+hep | -jaun, +blood) ?

Typical Task

- Given observations $\{O_1=v_1, \ldots O_k=v_k\}$
  (J=No, B=Yes [symptoms, history, test results, …])
  what is best DIAGNOSIS Dx_i for patient?
  (Hep=Yes vs Hep=No)
- Compute Probabilities of Dx_i
  given observations $\{O_1=v_1, \ldots O_k=v_k\}$

$$P(Dx = u \mid O_1 = v_1, \ldots, O_k = v_k)$$

Bayes Rule and Its Use

- **Diagnosis** typically involves computing P(Hypothesis | Symptoms)

  What is P(Meningitis | StiffNeck) ?

  ≡ Probability that patient A has meningitis, given that A has stiff neck?

- Typically have . . .
  - Prior probability of meningitis $P(+m) = 1/50{,}000$
  - Prior probability of having a stiff neck $P(\,+sn\,) = 1/20$
  - Probability that meningitis causes a stiff neck $P(\,+sn \mid +m\,) = ½$
- Bayes Rule: $$P(\,\text{hypothesis} \mid \text{symptoms}\,) = \frac{P(\text{symptoms}|\text{hypothesis})\, P(\text{hypothesis})}{P(\,\text{symptoms}\,)}$$
- Eg: $$P(+m \mid +sn) = \frac{P(+sn \mid +m)\, P(+m)}{P(+sn)} = \frac{\frac{1}{2} \times \frac{1}{50000}}{\frac{1}{20}} = \frac{1}{5000}$$
- Only 1 in 5000 stiff necks have meningitis, even though SN is the major symptom of M

Independence of Variables

- Note $P(+m) \neq P(+m \mid +sn)$
  - $P(+m) = 0.00002$
  - $P(+m \mid +sn) = 0.0002$

  $\Rightarrow$ So knowing "stiff neck" changes belief in meningitis
  - M is *dependent* on SN

- But some variables are NOT dependent:
- Coin tosses:
  - H_1: the first toss is a head; T_2: the second toss is a tail
  - $P(T_2 \mid H_1) = P(T_2)$

- α and β ***independent*** iff $P(\beta|\alpha) = P(\beta)$
  - iff $P(\alpha, \beta) = P(\alpha) P(\beta)$
  - In distribution *P*, α independent of β

Independence

Repeat

- Events α and β are independent *iff*
  - $P(\alpha, \beta) = P(\alpha)\,P(\beta)$
  - $P(\alpha \mid \beta) = P(\alpha)$
  - $P(\alpha \vee \beta) = 1 - (1 - P(\alpha))\,(1 - P(\beta))$
- Variables independent

  $\Leftrightarrow$ independent for all values

$\forall a, b \quad P(A = a, B = b) = P(A = a)\;P(B = b)$

i.i.d = "independent and identically distributed"

Probabilities

- Natural way to represent uncertainty
- ∃ **intuitive** notions about probabilities, but …
  - Many notions are wrong or inconsistent
  - Many people don't get what probabilities mean

$\Rightarrow$ Have **FORMAL** description,
that is consistent and useful

  - Overall framework is understood
  - … allows discussion over fine details of "meaning"

- Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations.

- Rank the following by probability
  (1 = most probable; 8 = least probable)
  a. Linda is a teacher in elementary school.
  b. Linda works in a bookstore and takes yoga classes.
  c. Linda is an active feminist.
  d. Linda is psychiatric social worker.
  e. Linda is a member of the League of Women Voters.
  f. Linda is a bank teller.
  g. Linda is an insurance salesperson.
  h. Linda is an active feminist and a bank teller.

Understanding Probabilities

- Two ways to think about Probabilities
  - Relative frequencies: objective

    (≈frequentist view)
  - Degree of belief: subjective

    (≈Bayesian view)

- Neither is entirely satisfying
  - No two events are truly the same (reference class problem)
  - Statements should be grounded in reality in some way

Probability as Relative Frequency?

- What is probability of *event* E ?
- Over long sequence of experiments, ratio of
  - (# of times E occurred)
    number of times E occurs in sequence, to
  - (# of trials)
    total number of experiments
- Estimate: $P(E) \approx \dfrac{\text{(\# of times E occurred)}}{\text{(\# of trials)}}$
- As (# of trials) $\rightarrow \infty$,
  ratio approaches true probability
  - given std assumptions

Examples…

- What is P(S can swim 50′ in ≤15 seconds) ?
  - Swimmer S …
    - tries **100** times to swim 50′ in ≤ 15 secs
    - succeeds **20** occasions
  - Estimate: probability that
    S can swim 50′ in ≤15 seconds is:
    - P(S can swim 50′ in ≤15 seconds) ≈ 20/100 = 0.2
- For probability to be meaningful,
  must clearly defined
  - (repeated) experiments
  - sample space
  - events
- What is the probability of a *nuclear war* ?

Frequencies vs Subjective …

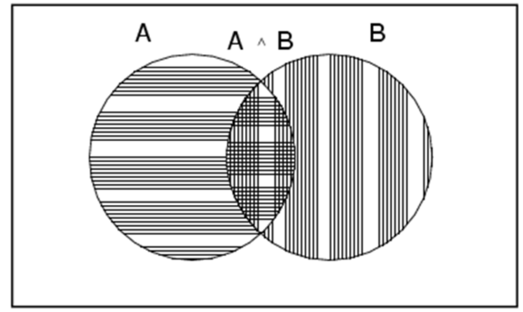
- Frequentists
  - $P(\alpha)$ = the frequency of α in the limit
  - Many arguments against this interpretation
    - What is the frequency of the event "nuclear war tomorrow" ?
- Subjective interpretation
  - $P(\alpha)$ = my degree of belief that α will happen
  - … where "degree of belief" means…

    If I say $P(\alpha)=0.8$, then I am willing to bet…
    at 4-to-1 odds !!

Subjective Beliefs … with Caution

- Subjectivists: *probabilities are degrees of belief*
- AI has used many notions of belief:
  - Certainty Factors
  - Fuzzy Logic
  - …
- Is any *degree of belief* $\equiv$ *probability*?
- NO!!
  - Dutch book argument
  - If you follow rules that do not follow probability theory, you will lose…

Probability Theory



- Axioms:

$0 \leq P(A) \leq 1$

$P(\text{ True }) = 1, \quad P(\text{ False }) = 0$

$P(A \vee B) = P(A) + P(B) - P(A \,\&\, B)$

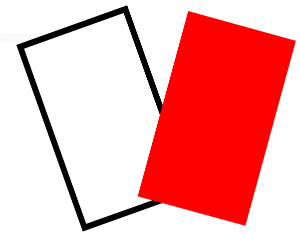
$P(A) + P(\neg A) = 1$

…

- Not arbitrary:
  - If Agent1 use probabilities that violate axioms, then

    ∃ betting strategy s.t.
    Agent1 guaranteed to lose $
  - "Dutch book"

The Three-Card Problem



- Three cards
  - RR = red on both sides
  - WW = white on both sides
  - RW = red on one side, white on the other
- Draw single card randomly and toss it into the air
- What is the probability …

  a. … of drawing red-red? P(D_RR)

  b. … that the drawn cards lands white side up? P(W_up)

  c. … that the red-red card was not drawn,
  assuming that the drawn card lands red side up ?
  P(not-D_RR | R_up)

Fair Bets

B believes

- P(D_RR) = 1/3
- P(W_up) = ½
- P(not-D_RR | R_up) = ~~1/3~~ ½

- A bet is *fair* to an individual B if,
  - according to B's probability assessment,
  - the bet will break even in the long run.
- B thinks these 3 bets are fair :

Bet **(a)** : Win \$4.20 if D_RR;
lose \$2.10 otherwise. [B believes P(D_RR) = 1/3]

Bet **(b)**: Win \$2.00 if W_up;
lose \$2.00 otherwise. [B believes P(W_up) = 1/2]

Bet **(c)**: W/L \$0 if not-R_up;
win \$4.00 if R_up and not D_RR;
lose \$4.00 if R_up and D_RR.
[B believes P(not-D_RR | R_up)=1/2]

Possible Outcomes

(a): Win $4.20 if D_RR;
lose $2.10 otherwise.

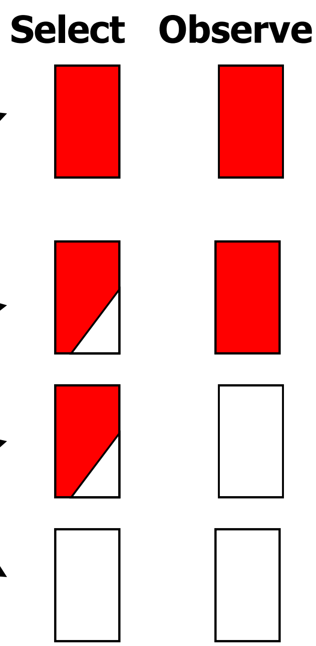
(b): Win $2.00 if W_up;
lose $2.00 otherwise

(c): Win $4.00 if R_up and not D_RR;
lose $4.00 if R_up and D_RR;
win $0 if not-R_up

1. D_RR & R_up:
   Draw RR,
   which lands red side up.

2. not-D_RR & R_up:
   Draw non-RR card,
   which lands *red* side up.

3. not-D_RR & W_up:
   Draw non-RR card,
   which lands *white* side up.



(a)	(b)	(c)	Total
+4.20	-2.00	-4.00	-1.80
-2.10	-2.00	+4.00	-0.10
-2.10	+2.00	±0.00	-0.10
-2.10	+2.00	±0.00	-0.10

Possible Outcomes

(a): Win $4.20 if D_RR;
lose $2.10 otherwise.

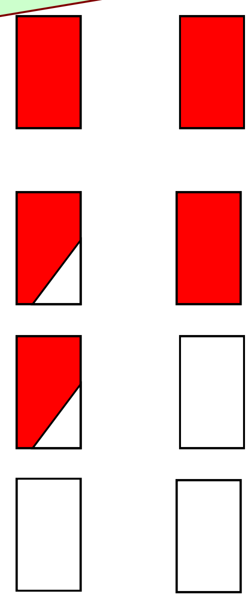
(b): Win $2.00 if W_up;
lose $2.0…

…and not D_RR;
…and D_RR;

B is always guaranteed to lose money…
- whichever card is drawn, &
- however it lands !

1. D_RR
   Draw …
   which lands red side up.
2. not-D_RR & R_up:
   Draw non-RR card,
   which lands *red* side up.
3. not-D_RR & W_up:
   Draw non-RR card,
   which lands *white* side up.

…serve



(a)	(b)	(c)	Total
+4.20	-2.00	-4.00	-1.80
-2.10	-2.00	+4.00	-0.10
-2.10	+2.00	±0.00	-0.10
-2.10	+2.00	±0.00	-0.10

The Dutch Book Theorem

- Spse B accepts any bet it thinks is fair. Then…
- a Dutch book can be made against B (ie, B guaranteed to lose $)

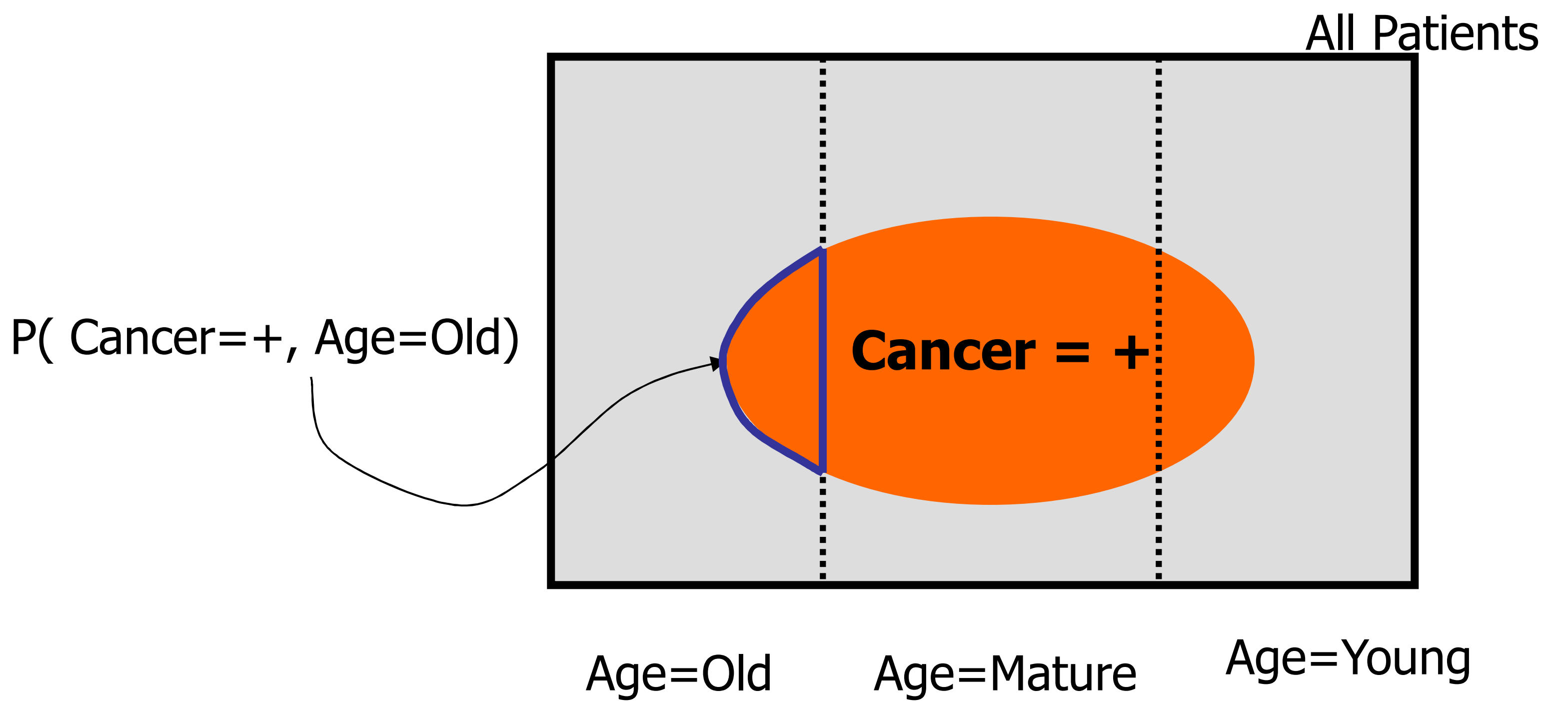
  iff

  B's assessment of probability violates Bayesian axiomatization.

Outline

- Foundations
  - Bayes Theorem
  - (Conditional) Independence
  - Dutch Book Theorem
  - Moments: Mean, Variance
- Estimation
  - MLE (Binomial)
  - Bayesian model
- Gaussian (Normal)

Factoids



$$P(\ +c\ ) = \sum_a P(\ +c, A = a\ )$$

Expected Value

- Discrete
  - $E(X) = \sum_x x\, P(x)$
  - ≈ "average", "mean", arithmetic mean
  - $P(X=1) = \frac{1}{6},\ P(X=2) = \frac{1}{6},\ \ldots,\ P(X=6) = \frac{1}{6}$

$$E[\,X\,] = (1\times\tfrac{1}{6}) + (2\times\tfrac{1}{6}) + \ldots + (6\times\tfrac{1}{6}) = \frac{21}{6} = 3.5$$

- Continuous
  - $E(X) = \int_x x\, P(x)\, dx$

Properties of Expectation

$$\boxed{E(\, f(X)\,) \;=\; \sum_x f(x)\, P(x)}$$

$$E(\, aX\,) \;=\; a\, E(X\,)$$

$$E(aX+b) = a\, E(X) + b$$

$$E(X + Y) = E(X) + E(Y)$$

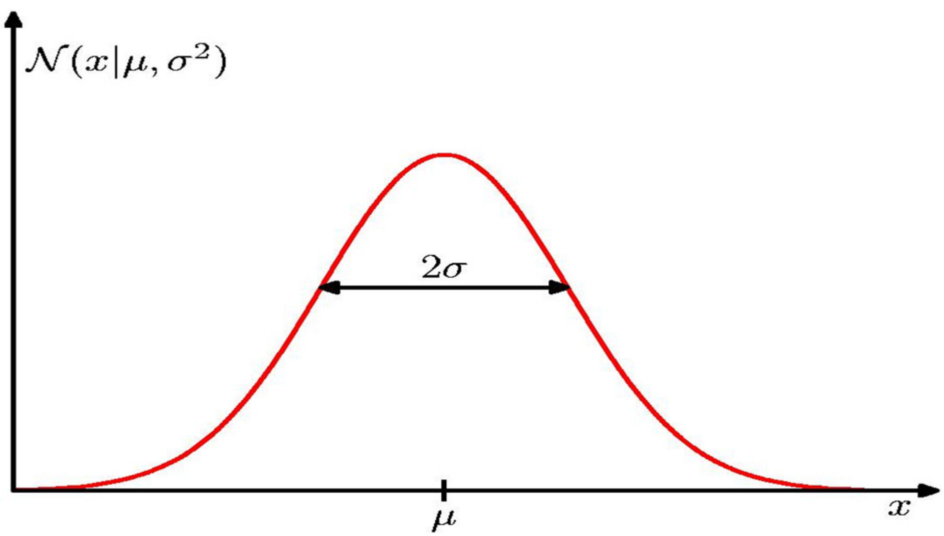
$$E(\, X \times Y) = ???$$

If $X \perp Y$, then $E(X)\, E(Y)$

Variance

- ≈ "How much to *trust* the mean"
  … hard to define in words…

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



Properties of Variance

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

$$\mathrm{Var}(aX) = a^2\, \mathrm{Var}(X)$$

$$\mathrm{Var}(aX+b) = a^2\, \mathrm{Var}(X)$$

$$\mathrm{Var}(X + Y) =$$
$$\mathrm{Var}(X) + \mathrm{Var}(Y) + 2\, E[\,(X-E(X))\,(Y-E(Y)\,]$$

If $X \perp Y$, then ... $= \mathrm{Var}(X) + \mathrm{Var}(Y)$

CoVariance

$$\boxed{Var(X+Y) = Var(X) + Var(Y) + 2\,E[(X-E(X))\,(Y-E(Y)]}$$

- CoVariance captures the "leftover"

$$\boxed{Cov(X,Y) = E[\,(X-E(X))\,(Y-E(Y)\ \,]}$$

$$\boxed{Var(X+Y) = Var(X) + Var(Y) + 2\,Cov(X,Y)}$$

- If $X \perp Y$, then $Cov(X, Y) = 0$

Standard Deviation

$$SD(X) = \sqrt{Var(X)}$$

- Sometimes more natural than variance:
  - SD(a X) = a SD(X)
- Sometimes, not:
  - X ⊥ Y, then

$$SD(X + Y) = \sqrt{SD(X)^2 + SD(Y)^2}$$

Outline

- Foundations
  - Bayes Theorem
  - (Conditional) Independence
  - Dutch Book Theorem
  - Moments: Mean, Variance
- Estimation
  - MLE (Binomial)
  - Bayesian model
- Gaussian (Normal)

Learning involves Estimation

Repeat

- Consider flipping a Thumbtack.
  What is the probability it will land with the nail up?

- Try flipping it a few times…
  observe H,H,T,T,H

- What is your BEST GUESS?

Jump

Repeat

Simple "Learning" Algorithm

$$\hat{\theta} = \arg\max_{\theta} \ \ln P(D \mid \theta)$$
$$= \arg\max_{\theta} \ \ln \theta^{h} (1 - \theta)^{t}$$

- Set derivative to zero: $\boxed{\frac{d}{d\theta} \ \ln P(\mathcal{D} \mid \theta) = 0}$

$$\frac{\partial}{\partial\theta}\ln[\theta^h(1-\theta)^t] = \frac{\partial}{\partial\theta}[h\ln\theta + t\ln(1-\theta)] = \frac{h}{\theta} + \frac{-t}{(1-\theta)}$$

$$\frac{h}{\theta} + \frac{-t}{1-\theta} = 0 \quad \Rightarrow \quad \hat{\theta} = \frac{\text{h}}{\text{t+h}}$$

So just average!!!

Jump

How many flips are "needed"?

$$\hat{\theta}_{MLE} = \frac{\#H}{\#H + \#T}$$

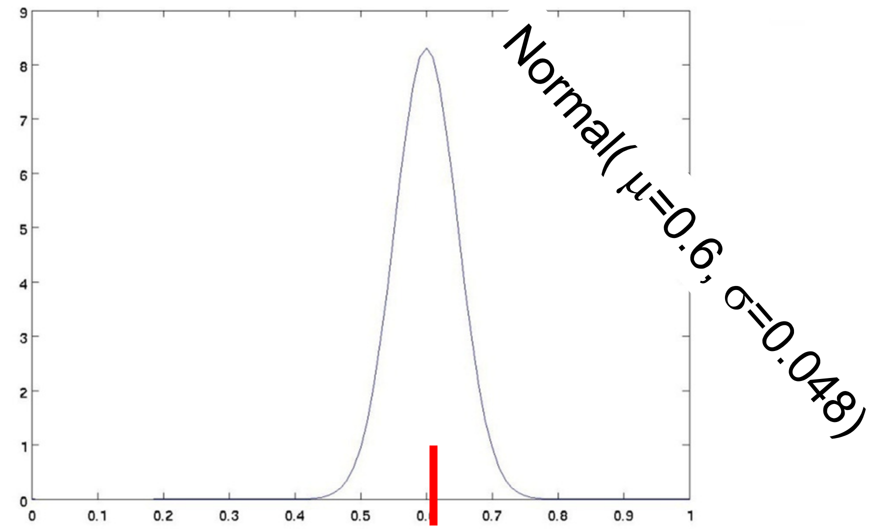
- Given 3 heads and 2 tails, $\theta_{MLE} = \frac{3}{5} = 0.6$
- But…
  Given 30 heads and 20 tails, $\theta_{MLE} = \frac{30}{50} = 0.6$
- SAME!!!
  Which is better? … more precise?

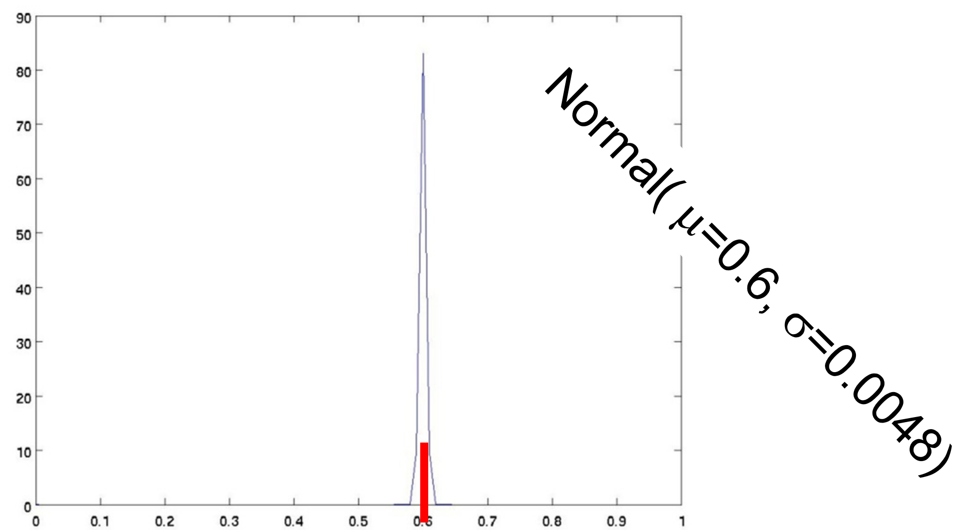
Jump

Using Variance

- Variance measures "spread" around mean
- For Binomial(h, t)
  - Mean: $\mu = \frac{h}{h+t}$
  - Variance: $\sigma = \frac{\mu(1-\mu)}{h+t}$

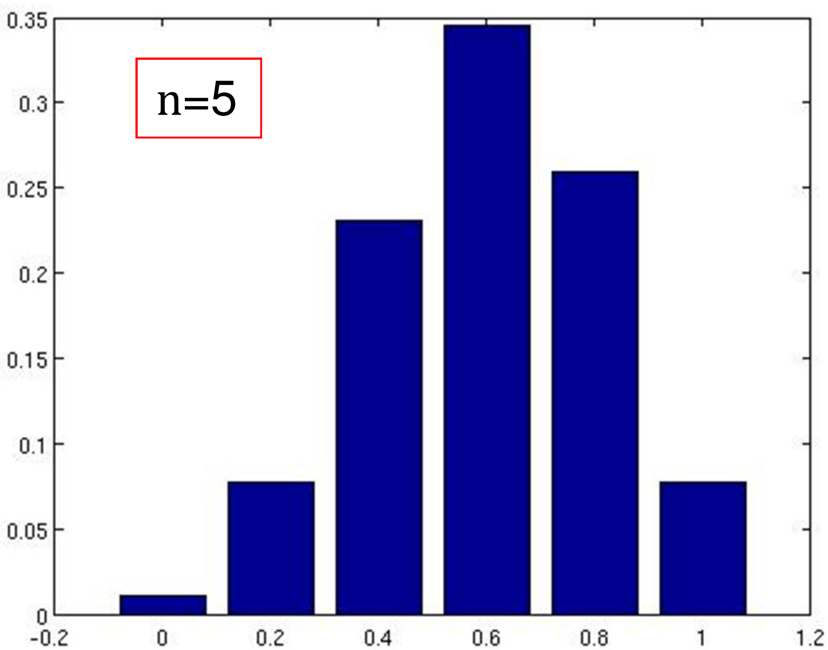
- Binomial(**3**H, **2**T)
  μ=0.6 σ=0.048
- Binomial(**30**H, **20**T)
  μ=0.6 σ=0.0048

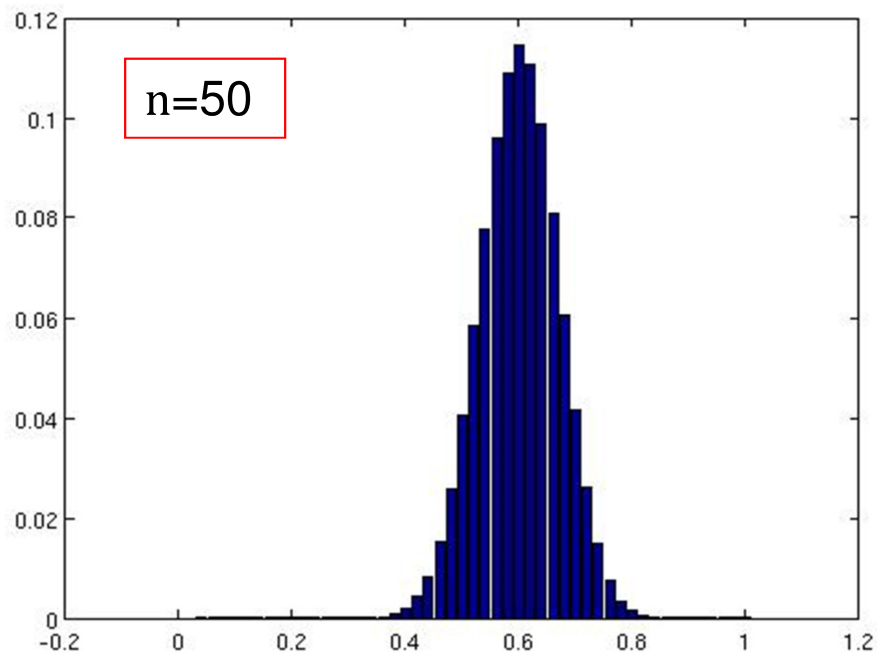




Binomial Distribution

P(D | θ)   for fixed θ=0.6

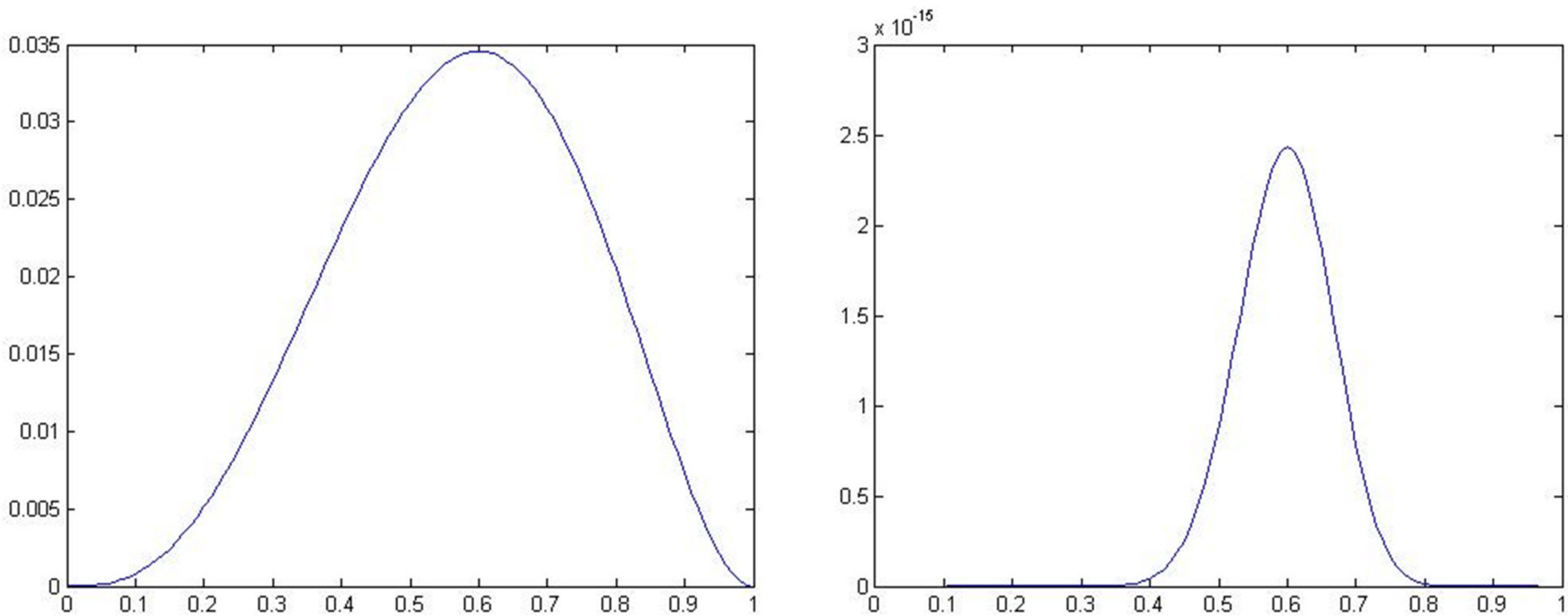




Prob that p=0.6 coin generates $\frac{k}{n}$ heads, in n flips

Probability Functions

$P(\,D \mid \theta\,)$ for fixed D



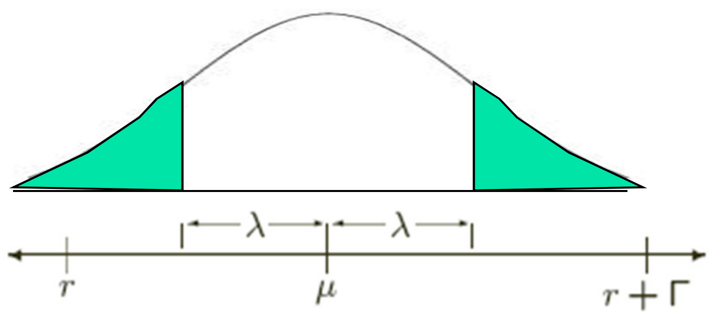
Prob that $p=\theta$ coin generates $\frac{h}{N}$ heads

$(1-\frac{h}{N}$ tails)

Hoeffding's Inequality

Defn: $S_m = \frac{1}{m}\sum_{i=1}^{m} X_i$ observed average over m r.v.s in {0,1}

- $P(\, S_m > \mu + \lambda \,) \; < \; e^{-2\, m\, \lambda^2}$



$$\Pr[\ |S_m - \mu\ | < \lambda\ ] \geq 1 - 2e^{-2m\,(\lambda/\Gamma)^2}$$

- Holds ∀ (bounded) distributions ... not just Bernoulli...
- Sample average likely to be close to true value as #samples (*m*) increases...

Simple bound (using Hoeffding's Inequality)

Here…

- #flips $m = m_H + m_T$
- Sample average = $\hat{\theta}^{(m)} = \frac{m_H}{m_H + m_T}$
- Let θ^* be the true parameter

For any m, $\varepsilon > 0$:

$$\boxed{P(\,|\hat{\theta}^{(m)} - \theta^*| > \epsilon\,) \;<\; 2\,e^{-2\,m\,\epsilon^2}}$$

Using Hoeffding's Inequality

$$P(|\hat{\theta} - \theta^*| > \epsilon) < 2e^{-2m\epsilon^2}$$

- To estimate the thumbtack parameter θ,
  - within $\varepsilon = 0.1$,
  - with probability $\geq 1 - \delta = 0.95$

require #flips $m > \frac{1}{2\epsilon^2} \ln \frac{2}{\delta}$ ≈ 80.1

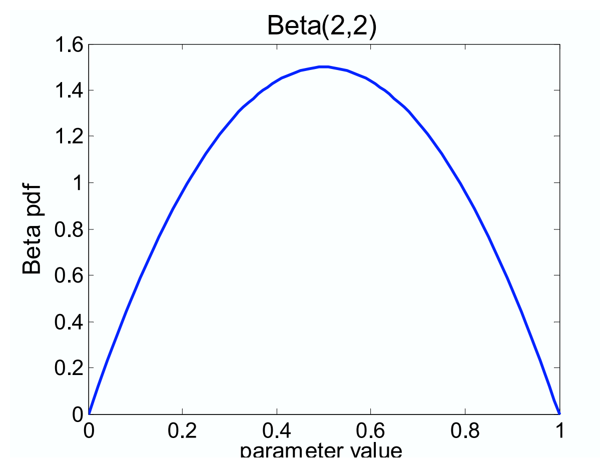
Problems with MLE

- Do you really believe 0% if 0 / 0+2 ?
- 0/0 issues
- Which is better?
  - 3 heads, 2 tails $\theta = \frac{3}{3+2} = 0.6$
  - 30 heads, 20 tails $\theta = \frac{30}{30+20} = 0.6$
  - 3E23 heads, 2E23 tails $\theta = \frac{3E23}{3E3+2E23} = 0.6$
- What if you already know SOMETHING about the variable…

≈ 50/50 …

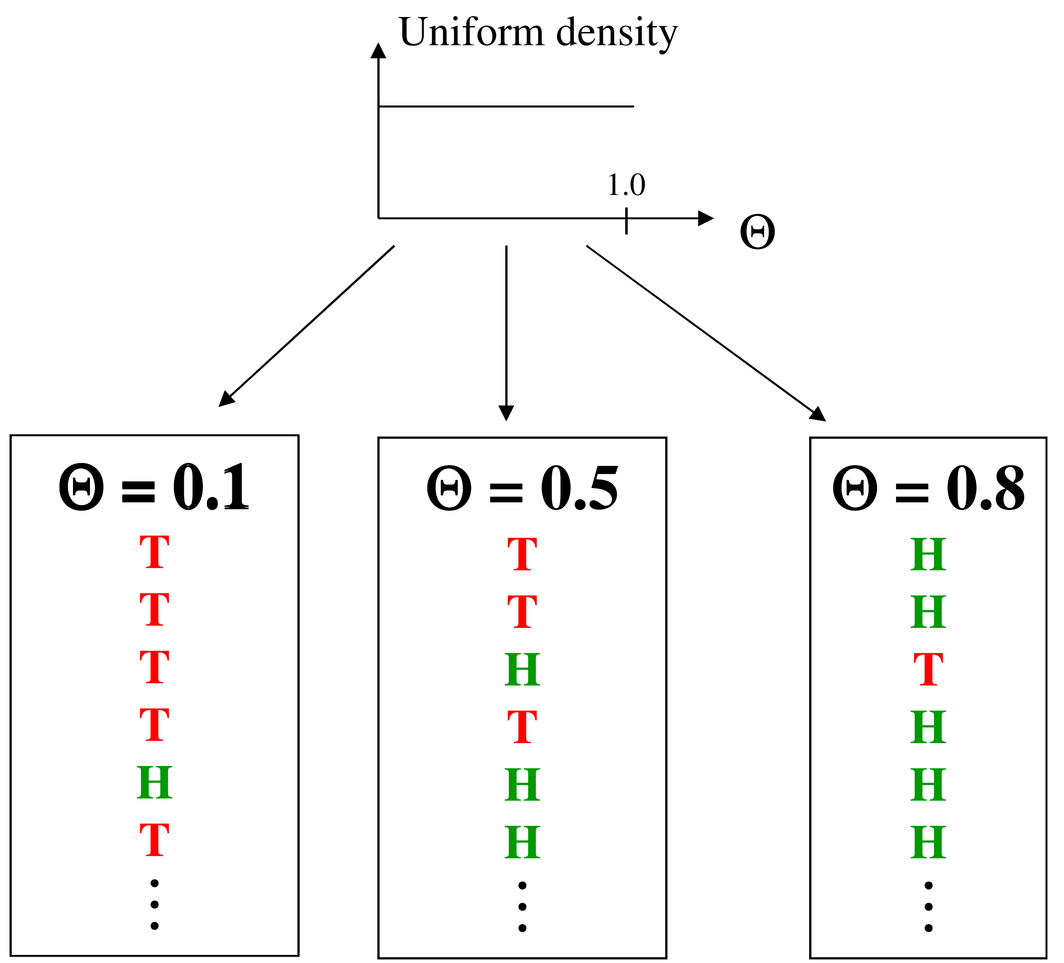
What about prior knowledge?

- Spse you *know* the thumbtack θ is "close" to 50-50
- **You can estimate it the Bayesian way…**
- Rather than estimate a single θ, obtain a *distrib'n* over possible values of θ

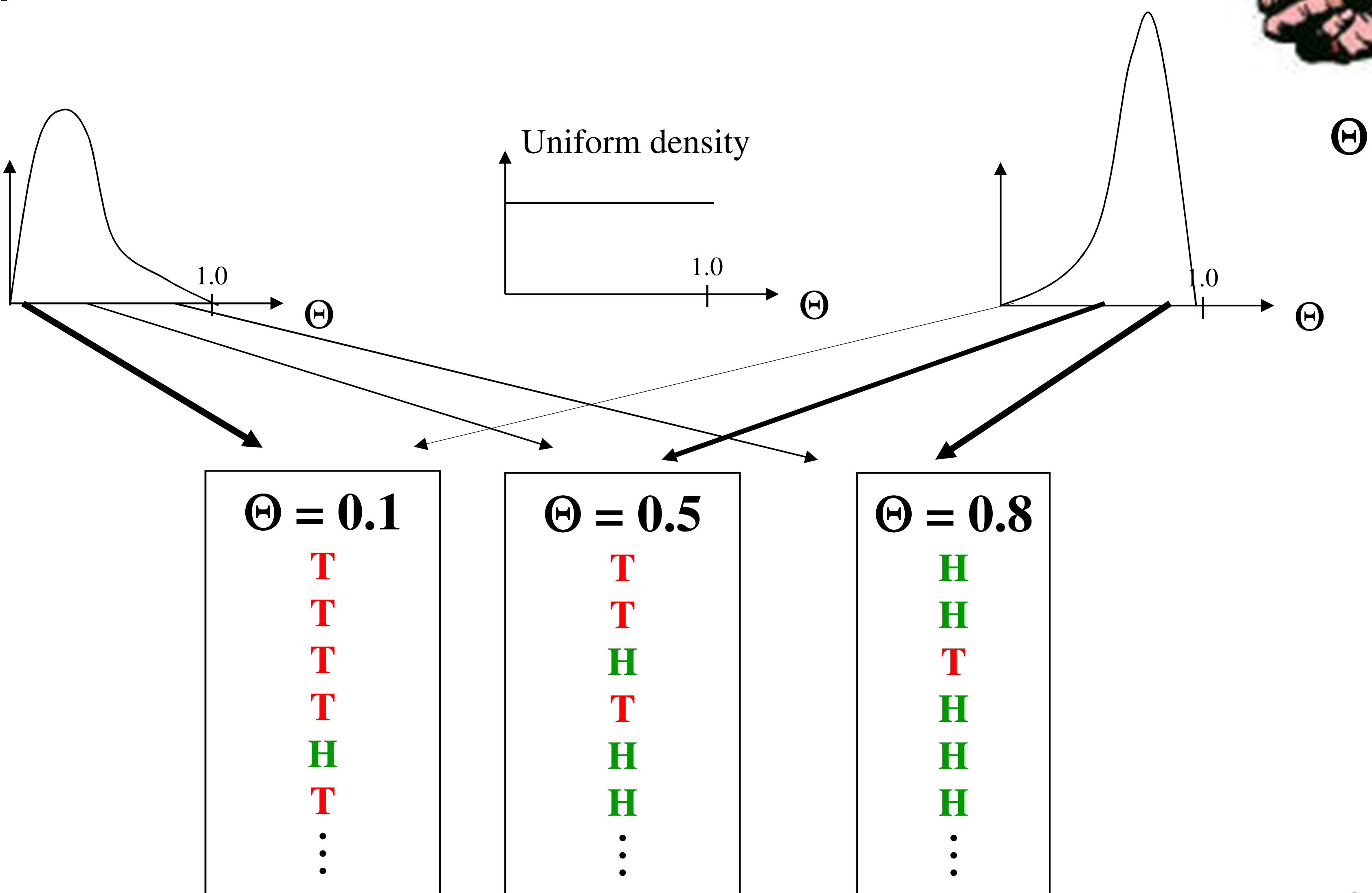


Two (related) Distributions: Parameter, Instances

Θ



Two (related) Distributions: Parameter, Instances



Jump

Bayesian Learning

- Use Bayes rule:

$$\underbrace{P(\theta|D)}_{\text{posterior}} = \frac{\overbrace{P(D|\theta)}^{\text{likelihood}}\overbrace{P(\theta)}^{\text{prior}}}{P(D)}$$

- Or equivalently (wrt $\mathrm{argmax}_\theta\ P(\theta|D)$)

$$\boxed{P(\theta|D) \propto P(D|\theta)P(\theta)}$$

Bayesian Learning for Thumbtack

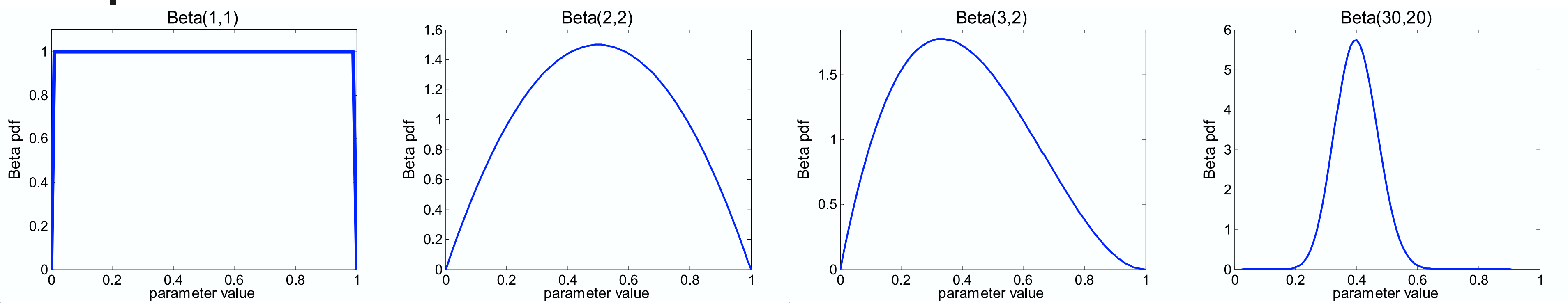
$$\underbrace{\mathrm{P}(\theta \mid D)}_{\textit{posterior}} \propto \underbrace{P(D \mid \theta)}_{\textit{likelihood}} \underbrace{P(\theta)}_{\textit{prior}}$$

- Likelihood function is simply Binomial:

$$P(\mathcal{D} \mid \theta) = \theta^{m_H}(1-\theta)^{m_T}$$

- What about prior, P(θ) ?
  - Represent expert knowledge
  - Simple posterior form

Beta prior distribution – P(θ)

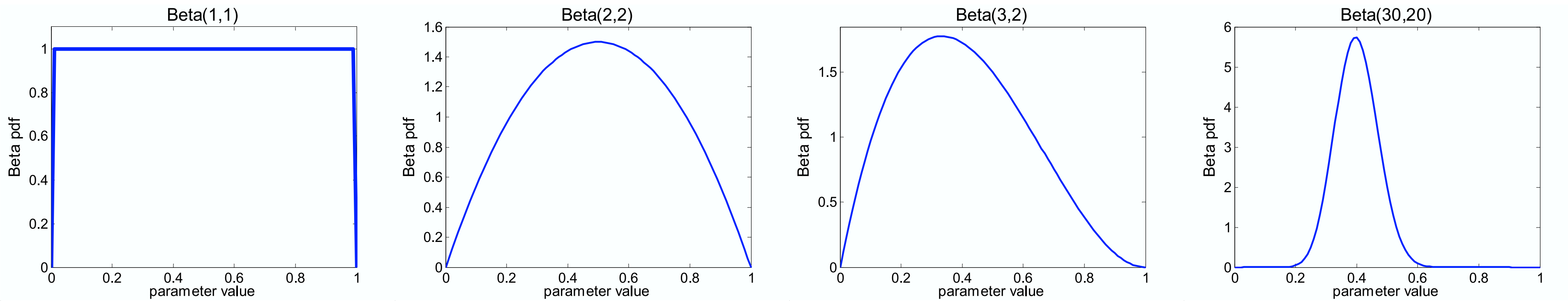


For $\theta \sim \mathrm{Beta}(a, b)$:

- PDF: $P(\theta) = \dfrac{\theta^{a-1}\,(1-\theta)^{b-1}}{B(a,b)}$
- Mean: $E[\theta] = \dfrac{a}{a+b}$
- Variance: $\mathrm{Var}[\theta] = \dfrac{a\,b}{(a+b)^2(a+b+1)} = \dfrac{a}{a+b}\,\dfrac{b}{a+b}\,\dfrac{1}{a+b+1}$
- Unimodal if $a,b>1$
- Likelihood function:

$$P(\,h\text{ “+”s},\, t\text{ “–”s}\ |\theta\,) = \theta^h\,(1-\theta)^t$$

Posterior distribution… from Beta



Prior $P(\theta)$

Likelihood $P(D|\theta)$

$$P(\theta \mid \mathcal{D}) \propto P(\theta)\; P(\mathcal{D} \mid \theta)$$

$$= \boxed{\Theta^{\alpha_H - 1}(1-\Theta)^{\alpha_T - 1}} \times \boxed{\Theta^{m_H}(1-\Theta)^{m_T}}$$

$$= \theta^{\alpha_H + m_H - 1}\,(1-\theta)^{\alpha_T + m_T - 1}$$

$$\sim \text{Beta}(\,\alpha_H + mH, \alpha_T + mT\,)$$

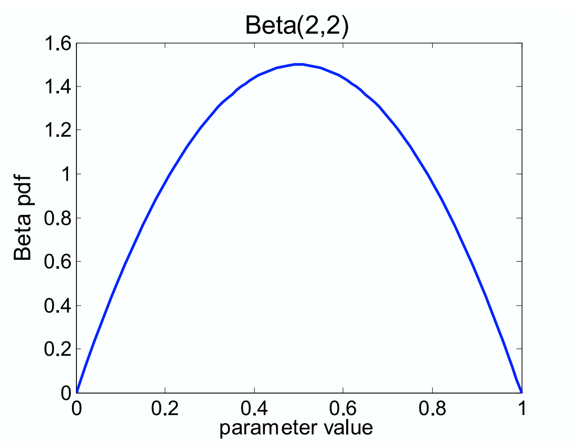
So Posterior is same form as Prior!! Conjugate!

Posterior Distribution

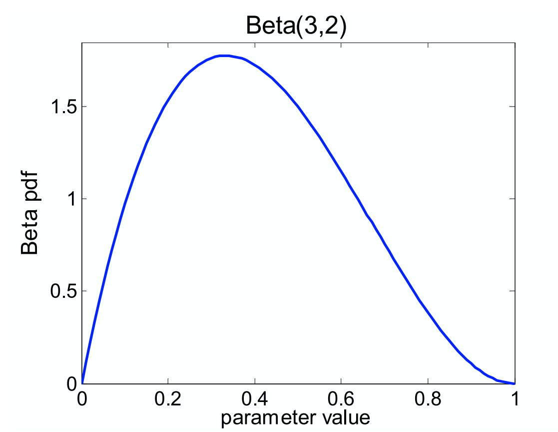
- Prior: $\theta \sim \mathrm{Beta}(\alpha_H, \alpha_T)$
- Data $\mathcal{D}$: m_H heads, m_T tails

$\Rightarrow$ Posterior distribution:

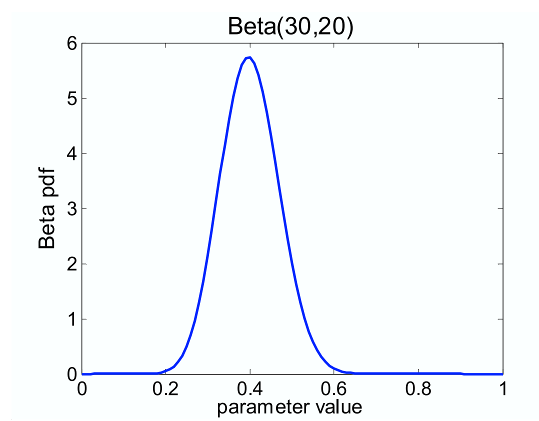
$$\theta \mid \mathcal{D} \sim \mathrm{Beta}(\alpha_H + m_H, \alpha_T + m_T)$$



Prior



+ observe 1 head



+ observe
27 more heads;
18 tails

Conjugate Prior

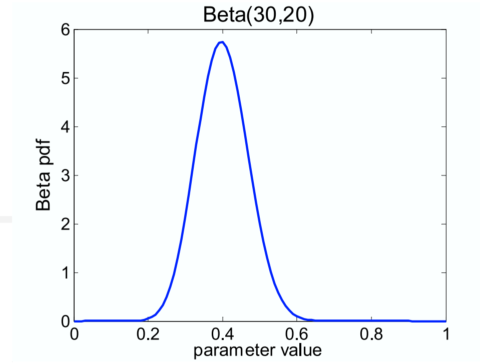
- Given
  - Prior: $\Theta \sim \mathrm{Beta}(\alpha_H, \alpha_T)$
  - Data: $\mathcal{D}$ with m_H heads and m_T tails (binomial likelihood)
- Posterior distribution:

$$\Theta|\mathcal{D} \sim \mathrm{Beta}(\alpha_H + m_H, \alpha_T + m_T)$$

- (Parametric) prior $P(\theta|\alpha)$ is **conjugate** to likelihood function if **posterior is of the same parametric family**, and can be written as:

  $P(\theta|\alpha')$ for some new set of parameters α'

Bayesian Prediction of a New Coin Flip



- Prior: $\Theta \sim Beta(\alpha_H, \alpha_T)$
- Observed m_H heads, m_T tails
- What is probability that next ($m+1^{st}$) flip is heads?

$$P(X_{m+1} = H \mid D) = \int_0^1 P(X_{m+1} = H \mid \Theta, D) \times P(\Theta \mid D)\ d\Theta$$

$$= \int_0^1 \Theta \times \boxed{Beta(\Theta : \alpha_H + m_H, \alpha_T + m_T)}\ d\Theta$$

$$= E_{\theta \sim Beta(\alpha_H + m_H, \alpha_T + m_T)}[\theta] = \frac{\alpha_H + m_H}{\alpha_H + m_H + \alpha_T + m_T}$$

Bayesian learning ≈ Smoothing

- Spse $\theta \sim \text{Beta}(1,4)$
  Then see $\mathcal{D} = \{+-++----+-\}$ = 4 +'s, 6 –'s
- Initially: $E[\theta] = \frac{1}{5}$

  … MLE is $\frac{4}{4+6} = 0.4$
- $\theta | \mathcal{D} \sim \text{Beta}(1+4,\ 4+6) = \text{Beta}(5,\ 10)$
  What is *Mean a posteri*?

$$E[\Theta \mid D] = \frac{\alpha_H + m_H}{\alpha_H + m_H + \alpha_T + m_T} = \frac{\alpha_H}{m+\alpha} + \frac{m_H}{m+\alpha}$$

$$= \left[\frac{\alpha}{m+\alpha}\right] \underbrace{\boxed{\frac{\alpha_H}{\alpha}}}_{\text{prior}} + \left[\frac{m}{m+\alpha}\right] \underbrace{\boxed{\frac{m_H}{m}}}_{\theta_{MLE}}$$

$m = m_H + m_T \qquad \alpha = \alpha_H + \alpha_T$
… equivalent sample size

Bayesian learning ≈ Smoothing

- Spse $\theta \sim \text{Beta}(1,4)$
  Then see $\mathcal{D} = \{+-++----+-\}$ = 4 +'s, 6 –'s
- Initially: $E[\theta] = \frac{1}{5}$

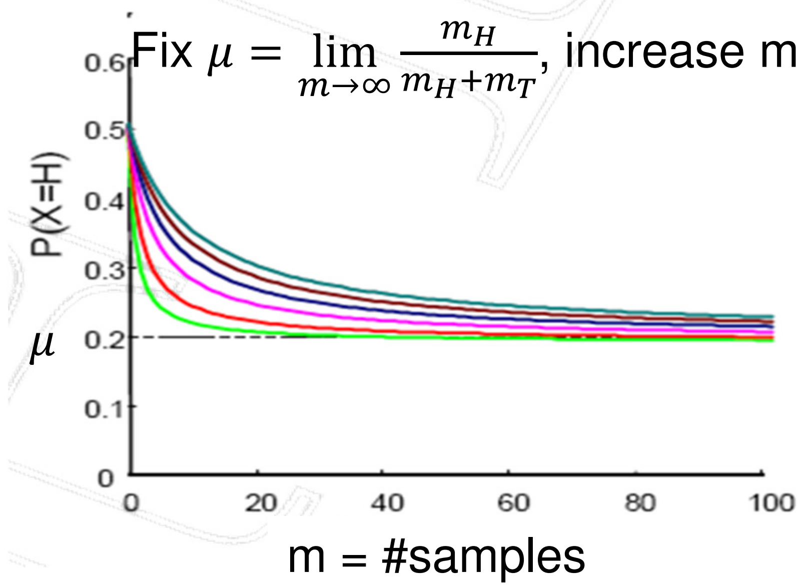
  … MLE is $\frac{4}{4+6} = 0.4$
- $\theta|\mathcal{D} \sim \text{Beta}(1+4,\ 4+6) = \text{Beta}(5,\ 10)$
  What is *Mean a posteri*?

  $$E[\,\theta\ |\mathcal{D}] = \frac{1 \quad + \quad 4}{1+4 \quad + \quad 4+6} = \frac{5}{15}$$
- Note $E[\theta \mid \mathcal{D}]$ is BLUR between $E[\theta]$ and MLE
  - … weighted by $\frac{5}{5+10}$ and $\frac{10}{5+10}$
  - Equivalent sample sizes:

    $\alpha = \alpha_H + \alpha_T = 5 \qquad m = m_H + m_T = 10$

Asymptotic Behavior

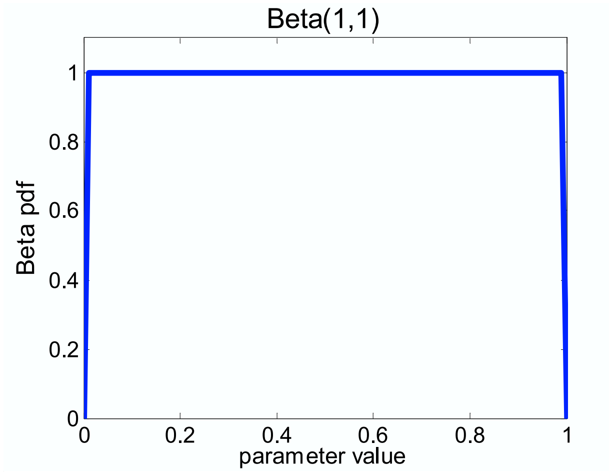
- $E[\theta\,] = \left[\frac{\alpha}{m+\alpha}\right]\frac{\alpha_H}{\alpha} + \left[\frac{m}{m+\alpha}\right]\frac{m_H}{m}$

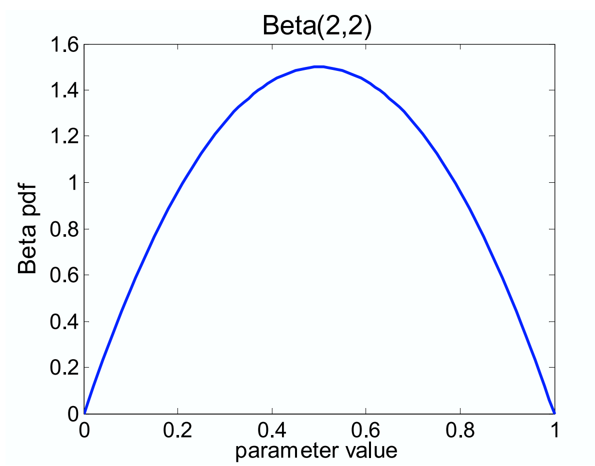


- For small sample size $m \approx 0$, prior $\frac{\alpha_H}{\alpha}$ is important
- As $m = m_T + m_H \to \infty$, prior is “forgotten”…

Alternative “Encoding”

- Beta(a, b) ≡ B′(m, μ) where
  - m = (a+b)
    … effective sample size
  - $\mu = \frac{a}{a+b}$
- Eg
  - Beta(1, 1) = B′(2, 0.5)
  - Beta(10,10) = B′(20, 0.5)
  - Beta(7, 3) = B′(10, 0.7)
  - …





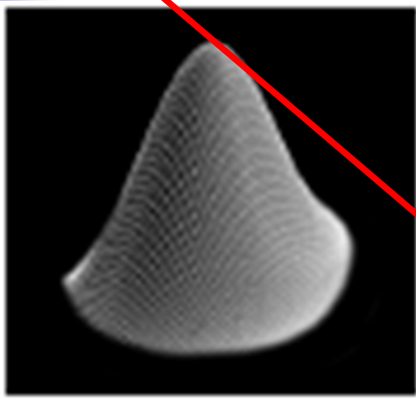
Bayesian learning for *Multi* nomial

- What if you have a k-sided thumbtack???
  - … still just ONE thumbtack (so just one event)
- Likelihood function if **multinomial**:
  - $P(X = i) = \theta_i \quad i = 1..k$
  - $\sum_i \theta_i = 1 \quad \theta_i \geq 0$
- **Conjugate** prior for multinomial is **Dirichlet**:
  - $\theta \sim \mathrm{Dirichlet}(\alpha_1, \ldots, \alpha_k) \sim \prod_i \theta_i^{\alpha_i - 1}$
- **Observe** *m* data points, m_i from assignment i, **posterior**:
  - $\mathrm{Dirichlet}(\alpha_1 + m_i, \ldots, \alpha_k + m_k)$
- **Prediction**: $$P(X_{m+1} = i \mid D) = \frac{\alpha_i + m_i}{\sum_j (\alpha_j + m_j)}$$

Outline

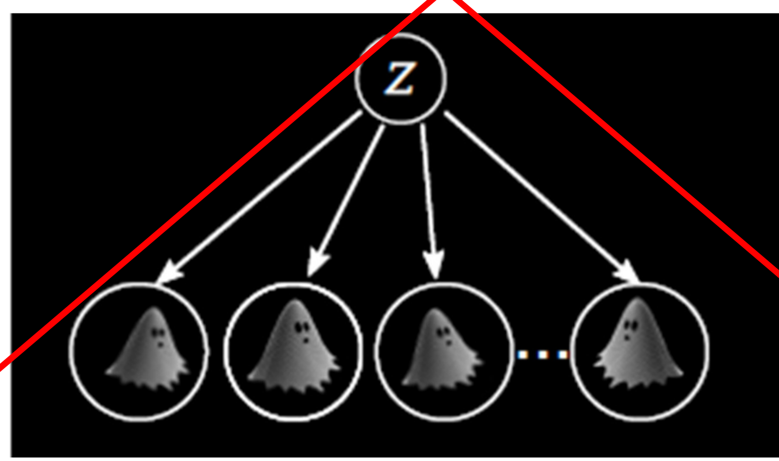
- Foundations
  - Bayes Theorem
  - (Conditional) Independence
  - Dutch Book Theorem
  - Moments: Mean, Variance
- Estimation
  - MLE (Binomial)
  - Bayesian model
- Gaussian (Normal)

Types of Normal Distr’ns



Normal Distribution

ParaNormal Distribution



Mixtures of paranormal distributions with occult variables

D Maturana, A Spectral Approach to Ghost Detection, 2013