10-701

Probability and MLE

http://www.cs.cmu.edu/~pradeepr/701

(brief) intro to probability

Basic notations

- Random variable

  - referring to an element / event whose status is unknown:

  A = "it will rain tomorrow"

- Domain (usually denoted by Ω)

  - The set of values a random variable can take:

  - "A = The stock market will go up this year": Binary

  - "A = Number of Steelers wins in 2015": Discrete

  - "A = % change in Google stock in 2015": Continuous

Axioms of probability (Kolmogorov's axioms)

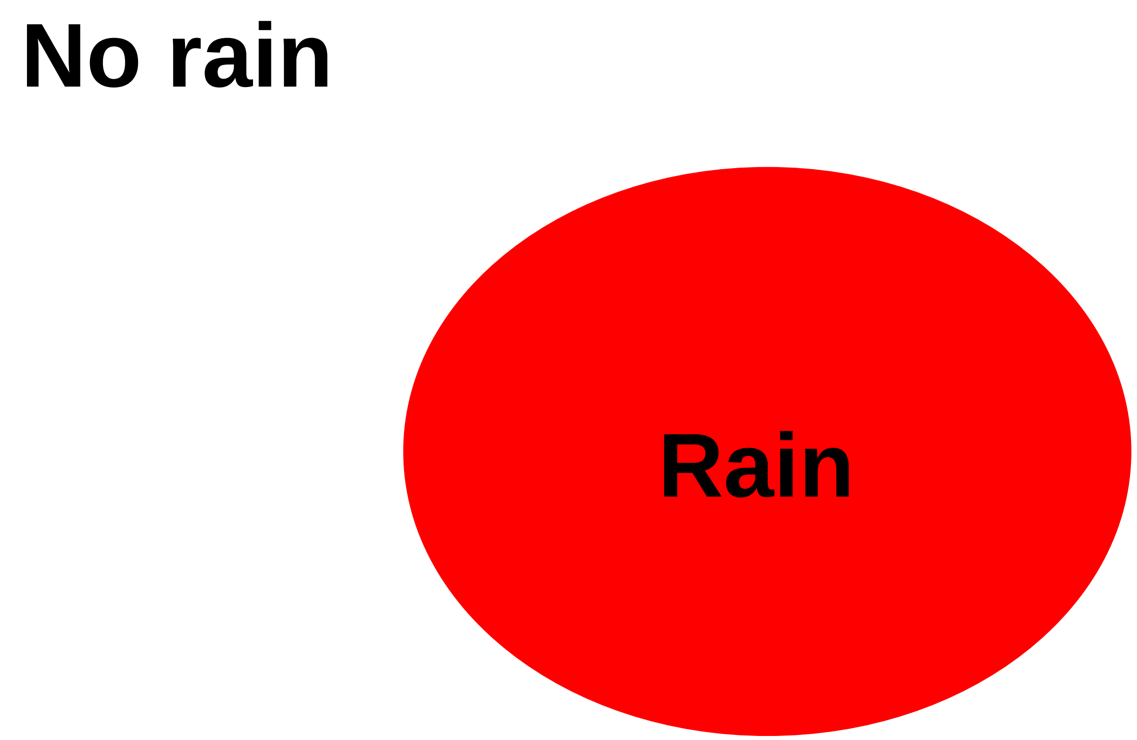
A variety of useful facts can be derived from just three axioms:

1. $0 \leq P(A) \leq 1$
2. $P(\text{true}) = 1$, $P(\text{false}) = 0$
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

There have been several other attempts to provide a foundation for probability theory. Kolmogorov's axioms are the most widely used.

Priors

Degree of belief in an event in the absence of any other information



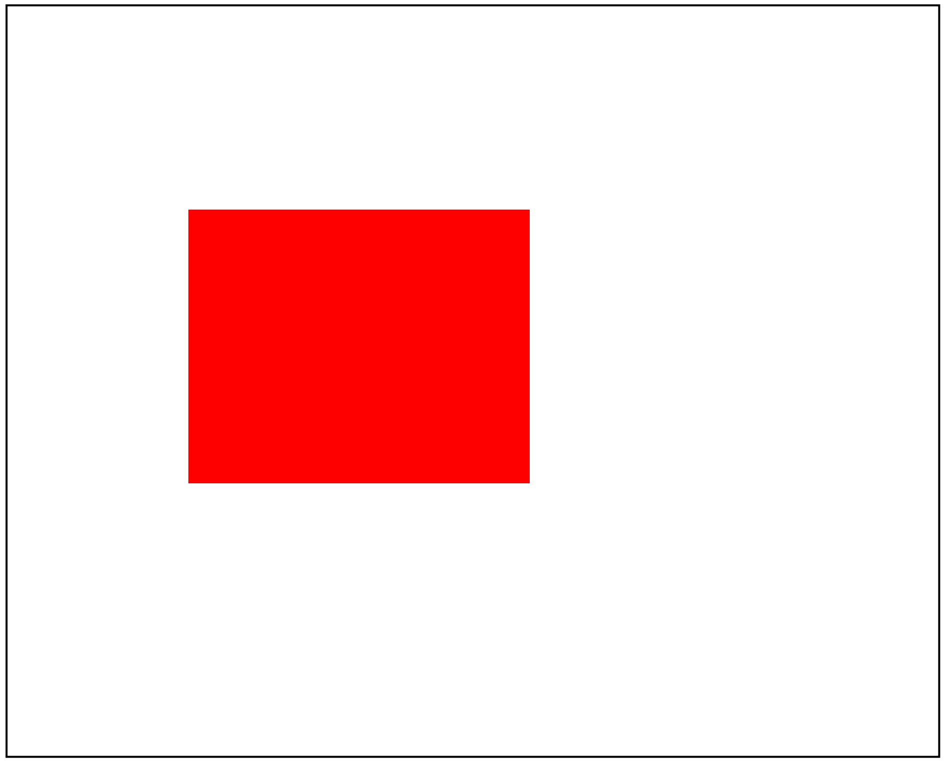
P(rain tomorrow) = 0.2

P(no rain tomorrow) = 0.8

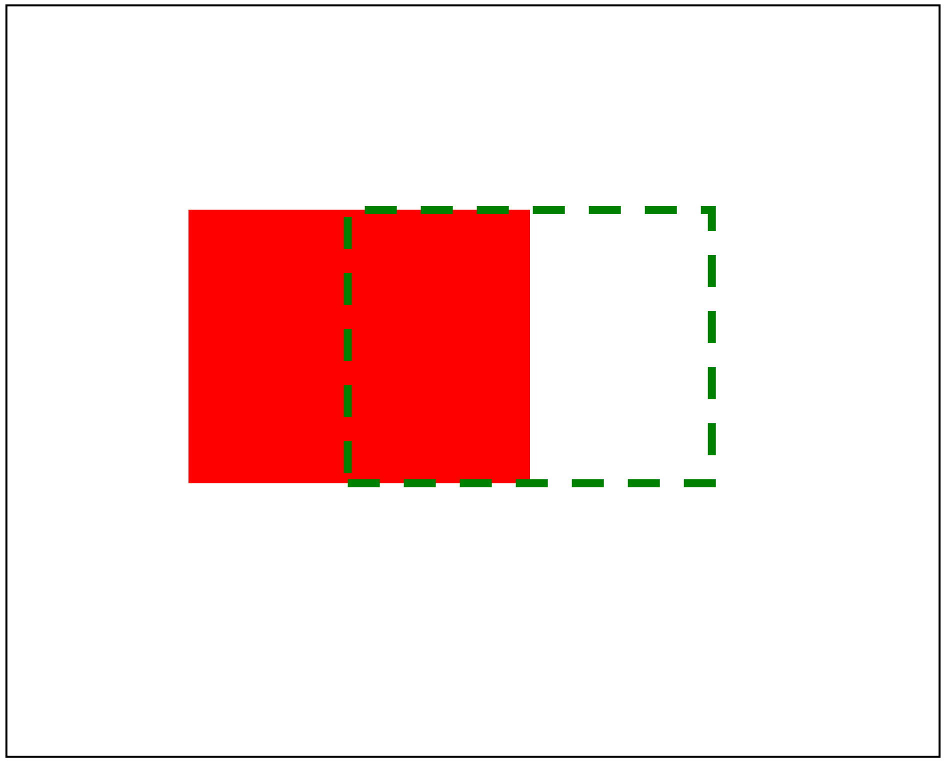
Conditional probability

- P(A = 1 | B = 1): The fraction of cases where A is true if B is true

P(A = 0.2)



P(A|B = 0.5)



Conditional probability

- In some cases, given knowledge of one or more random variables we can improve upon our prior belief of another random variable
- For example:

p(slept in movie) = 0.5

p(slept in movie | liked movie) = 1/4

p(didn’t sleep in movie | liked movie) = 3/4

Slept	Liked
1	0
0	1
1	1
1	0
0	0
1	0
0	1
0	1

Joint distributions

- The probability that a *set* of random variables will take a specific value is their joint distribution.
- Notation: P(A $\wedge$ B) or P(A,B)
- Example: P(liked movie, slept)

If we assume independence then

P(A,B)=P(A)P(B)

However, in many cases such an assumption may be too strong (more later in the class)

Joint distribution (cont)

P(class size > 20) = 0.6

P(summer) = 0.4

P(class size > 20, summer) = ?

Evaluation of classes

Size	Time	Eval
30	R	2
70	R	1
12	S	2
8	S	3
56	R	1
24	S	2
10	S	3
23	R	3
9	R	2
45	R	1

Joint distribution (cont)

P(class size > 20) = 0.6

P(summer) = 0.4

P(class size > 20, summer) = 0.1

Evaluation of classes

Size	Time	Eval
30	R	2
70	R	1
12	S	2
8	S	3
56	R	1
24	S	2
10	S	3
23	R	3
9	R	2
45	R	1

Joint distribution (cont)

P(class size > 20) = 0.6

P(eval = 1) = 0.3

P(class size > 20, eval = 1) = 0.3

Size	Time	Eval
30	R	2
70	R	1
12	S	2
8	S	3
56	R	1
24	S	2
10	S	3
23	R	3
9	R	2
45	R	1

Joint distribution (cont)

P(class size > 20) = 0.6

P(eval = 1) = 0.3

P(class size > 20, eval = 1) = 0.3

Evaluation of classes

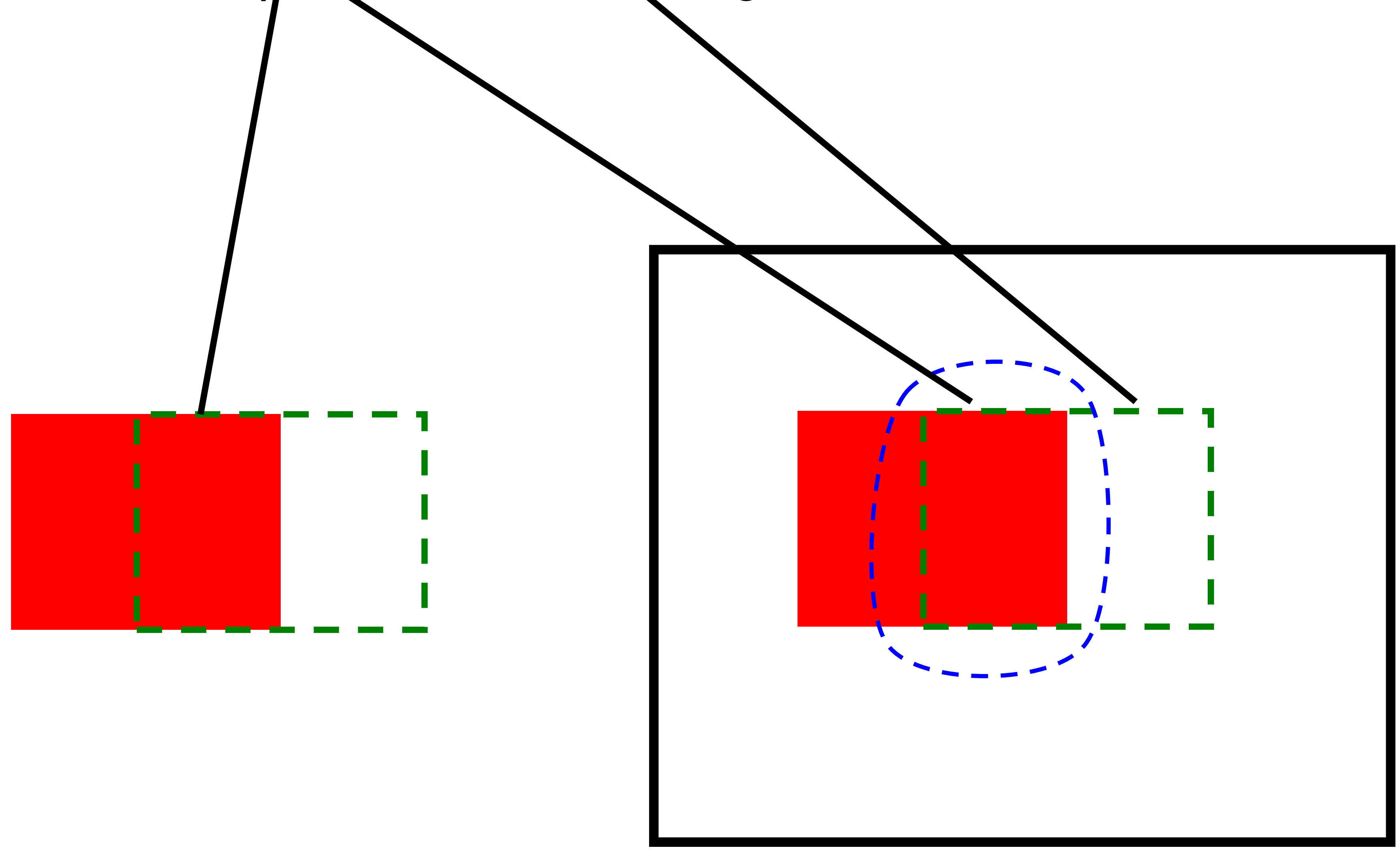
Size	Time	Eval
30	R	2
70	R	1
12	S	2
8	S	3
56	R	1
24	S	2
10	S	3
23	R	3
9	R	2
45	R	1

Chain rule

- The joint distribution can be specified in terms of conditional probability:

  P(A,B) = P(A|B)*P(B)

- Together with Bayes rule (which is actually derived from it) this is one of the most powerful rules in probabilistic reasoning



Bayes rule

- One of the most important rules for this class.
- Derived from the chain rule:

  P(A,B) = P(A | B)P(B) = P(B | A)P(A)

- Thus,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Thomas Bayes was an English clergyman who set out his theory of probability in 1764.

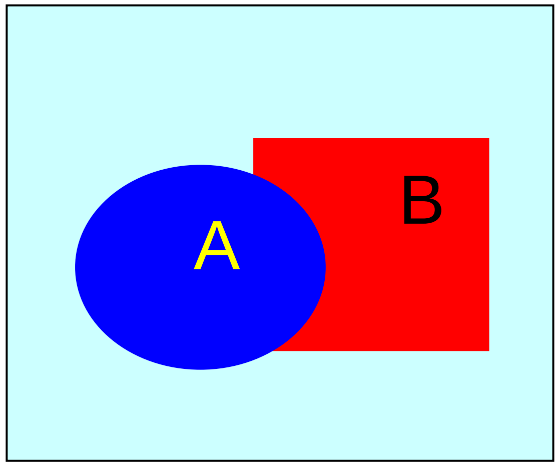
Bayes rule (cont)

Often it would be useful to derive the rule a bit further:

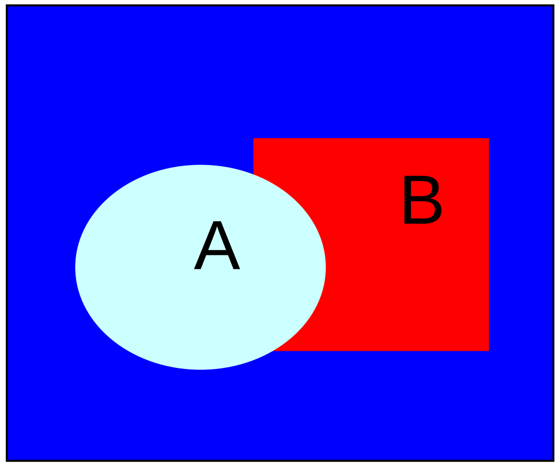
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{\sum_{A} P(B|A)P(A)}$$

This results from:
$P(B) = \sum_{A} P(B,A)$

P(B,A=1)



P(B,A=0)



Recall: Your first consulting job

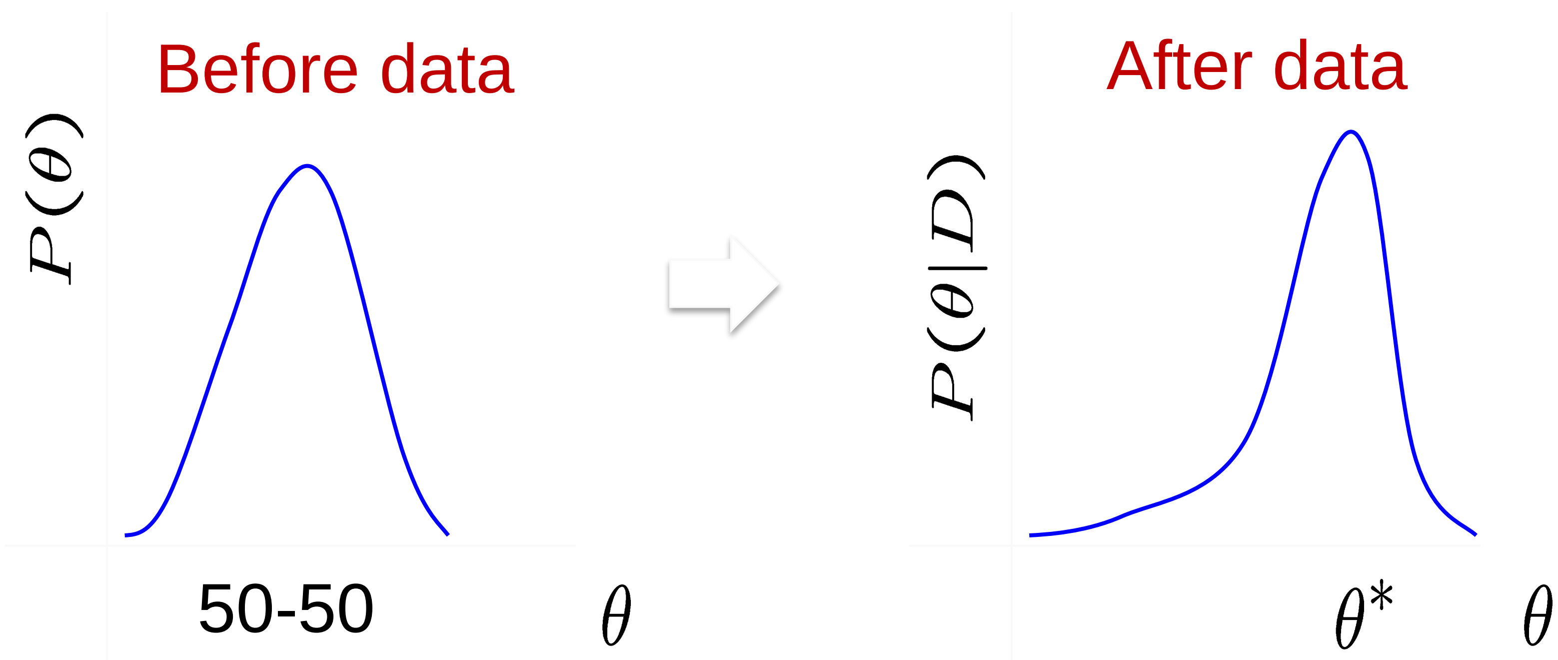
- A billionaire from the suburbs of Seattle asks you a question:
  - He says: I have a coin, if I flip it, what's the probability it will fall with the head up?
  - You say: Please flip it a few times:

  - You say: The probability is: **3/5** because… frequency of heads in all flips
  - **He says: But can I put money on this estimate?**
  - You say: ummm…. Maybe not.
    - Not enough flips (less than sample complexity)

What about prior knowledge?

- Billionaire says: Wait, I know that the coin is “close” to 50-50. What can you do for me now?
- **You say: I can learn it the Bayesian way…**

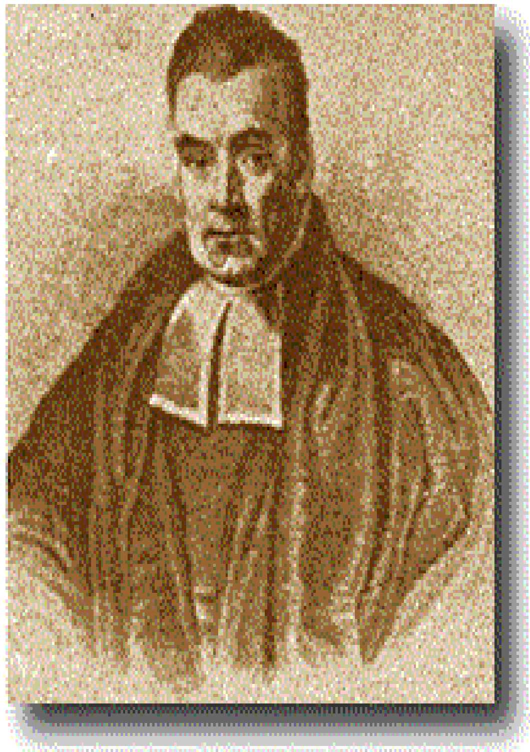
- Rather than estimating a single θ, we obtain a distribution over possible values of θ



Bayesian Learning

- Use Bayes rule:

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta) P(\theta)}{P(\mathcal{D})}$$



- Or equivalently:

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

posterior likelihood prior

AIDS test (Bayes rule)

Data

- Approximately 0.1% are infected
- Test detects all infections
- Test reports positive for 1% healthy people

AIDS test (Bayes rule)

Data

- **Approximately 0.1% are infected**
- **Test detects all infections**
- **Test reports positive for 1% healthy people**

Probability of having AIDS if test is positive:

$$P(a=1|t=1) = \frac{P(t=1|a=1)P(a=1)}{P(t=1)}$$

$$= \frac{P(t=1|a=1)P(a=1)}{P(t=1|a=1)P(a=1)+P(t=1|a=0)P(a=0)}$$

$$= \frac{1 \cdot 0.001}{1 \cdot 0.001 + 0.01 \cdot 0.999} = 0.091$$

Only 9%!...

Prior distribution

- From where do we get the prior?
  - Represents expert knowledge (philosophical approach)
  - Simple posterior form (engineer's approach)

- Uninformative priors:
  - Uniform distribution

- Conjugate priors:
  - Closed-form representation of posterior
  - P(q) and P(q|D) have the same algebraic form as a function of \theta

Conjugate Prior

- P(q) and P(q|D) have the same form as a function of theta

Eg. 1 Coin flip problem

Likelihood given Bernoulli model:

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

If prior is Beta distribution,

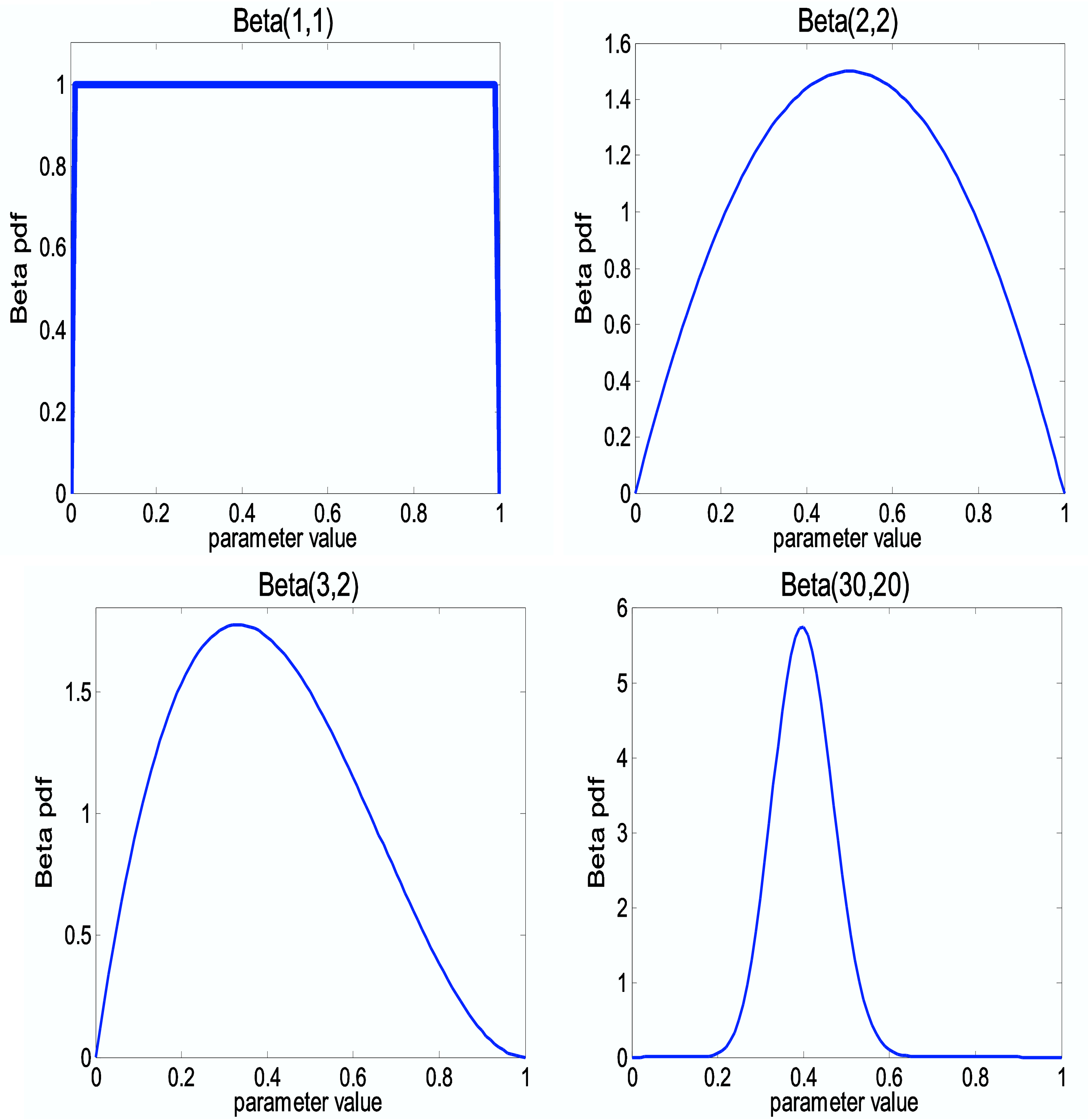
$$P(\theta) = \frac{\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

Then posterior is Beta distribution

$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

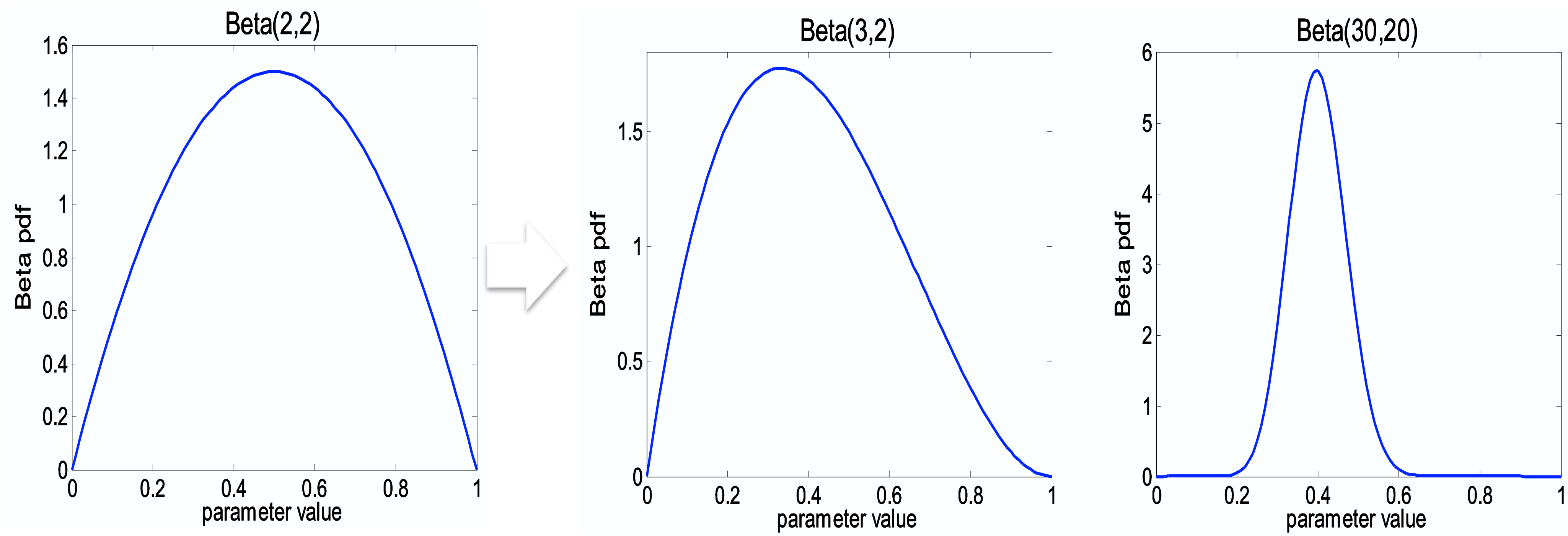
Beta distribution

$Beta(\beta_H, \beta_T)$ More concentrated as values of β_H, β_T increase



Beta conjugate prior

$$P(\theta) \sim Beta(\beta_H, \beta_T) \qquad P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$



As $n = \alpha_H + \alpha_T$ increases

As we get more samples, effect of prior is “washed out”

Conjugate Prior

- P(θ) and P(θ|D) have the same form

Eg. 2 Dice roll problem (6 outcomes instead of 2)

Likelihood is ~ Multinomial($\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$)

$$P(\mathcal{D} \mid \theta) = \theta_1^{\alpha_1} \theta_2^{\alpha_2} \ldots \theta_k^{\alpha_k}$$

If prior is Dirichlet distribution,

$$P(\theta) = \frac{\prod_{i=1}^{k} \theta_i^{\beta_i - 1}}{B(\beta_1, \ldots, \beta_k)} \sim \text{Dirichlet}(\beta_1, \ldots, \beta_k)$$

Then poste

$$P(\theta|D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \ldots, \beta_k + \alpha_k)$$

For Multinomial, conjugate prior is Dirichlet distribution.

Posterior Distribution

- The approach seen so far is what is known as a **Bayesian** approach
- Prior information encoded as a **distribution** over possible values of parameter
- Using the Bayes rule, you get an updated **posterior** distribution over parameters, which you provide with flourish to the Billionaire
- But the billionaire is not impressed
  - Distribution? I just asked for one number: is it 3/5, 1/2, what is it?
  - How do we go from a distribution over parameters, to a single estimate of the true parameters?

Maximum A Posteriori Estimation

Choose θ that maximizes a posterior probability

$$\widehat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid D)$$
$$= \arg\max_{\theta} P(D \mid \theta) P(\theta)$$

MAP estimate of probability of head:

$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

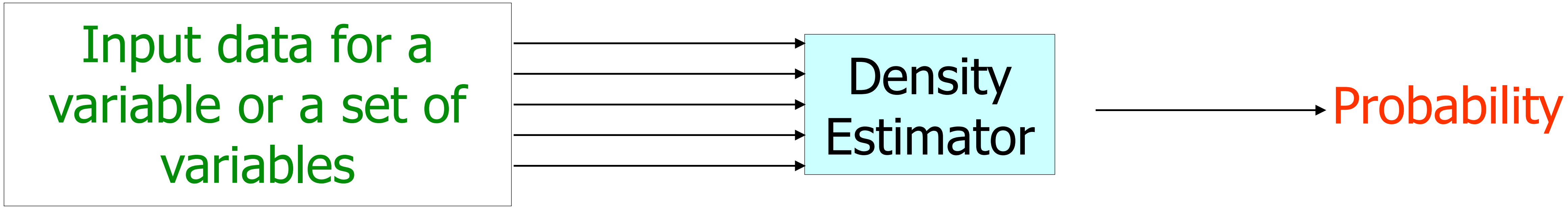
$$\widehat{\theta}_{MAP} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

Mode of Beta distribution

Density estimation

Density Estimation

- A Density Estimator learns a mapping from a set of attributes to a Probability



Density estimation

- Estimate the distribution (or conditional distribution) of a random variable
- Types of variables:

  - Binary

    coin flip, alarm

  - Discrete

    dice, car model year

  - Continuous

    height, weight, temp.,

When do we need to estimate densities?

- Density estimators are critical ingredients in several of the ML algorithms we will discuss
- In some cases these are combined with other inference types for more involved algorithms (i.e. EM) while in others they are part of a more general process (learning in BNs and HMMs)

Density estimation

- Binary and discrete variables:

Easy: Just count!

- Continuous variables:

Harder (but just a bit): Fit a model

Learning a density estimator for discrete variables

$$\hat{P}(x_i = u) = \frac{\#\text{records in which } x_i = u}{\text{total number of records}}$$

A trivial learning algorithm!

But why is this true?

Maximum Likelihood Principle

We can define the likelihood of the data given the model as follows:

$$\hat{P}(\text{dataset} \mid M) = \hat{P}(x_1 \wedge x_2 \ldots \wedge x_n \mid M) = \prod_{k=1}^{n} \hat{P}(x_k \mid M)$$

M is our model (usually a collection of parameters)

For example M is

- The probability of ‘head’ for a coin flip
- The probabilities of observing 1,2,3,4 and 5 for a dice
- etc.

Maximum Likelihood Principle

$$\hat{P}(\text{dataset} \mid M) = \hat{P}(x_1 \wedge x_2 \ldots \wedge x_n \mid M) = \prod_{k=1}^{n} \hat{P}(x_k \mid M)$$

- Our goal is to determine the values for the parameters in *M*
- We can do this by maximizing the probability of generating the observed samples
- For example, let Θ *be the probabilities for a coin flip*
- Then

$$L(x_1, \ldots, x_n \mid \Theta) = p(x_1 \mid \Theta) \ldots p(x_n \mid \Theta)$$

- The observations (different flips) are assumed to be independent
- For such a coin flip with $P(H)=q$ the best assignment for Θ_h is

$$\mathit{argmax}_q = \#H/\#\mathit{samples}$$

- Why?

Maximum Likelihood Principle: Binary variables

- For a binary random variable A with P(A=1)=q

  argmax$_q$ = #1/#samples

- Why?

Data likelihood: $P(D \mid M) = q^{n_1}(1-q)^{n_2}$

We would like to find: $\arg\max_q q^{n_1}(1-q)^{n_2}$

Omitting terms that do not depend on *q*

Maximum Likelihood Principle

Data likelihood: $P(D\,|\,M) = q^{n_1}(1-q)^{n_2}$

We would like to find: $\arg\max_q q^{n_1}(1-q)^{n_2}$

$$\frac{\partial}{\partial q} q^{n_1}(1-q)^{n_2} = n_1 q^{n_1-1}(1-q)^{n_2} - q^{n_1} n_2 (1-q)^{n_2-1}$$

$$\frac{\partial}{\partial q} = 0 \Rightarrow$$

$$n_1 q^{n_1-1}(1-q)^{n_2} - q^{n_1} n_2 (1-q)^{n_2-1} = 0 \Rightarrow$$

$$q^{n_1-1}(1-q)^{n_2-1}(n_1(1-q) - q n_2) = 0 \Rightarrow$$

$$n_1(1-q) - q n_2 = 0 \Rightarrow$$

$$n_1 = n_1 q + n_2 q \Rightarrow$$

$$q = \frac{n_1}{n_1 + n_2}$$

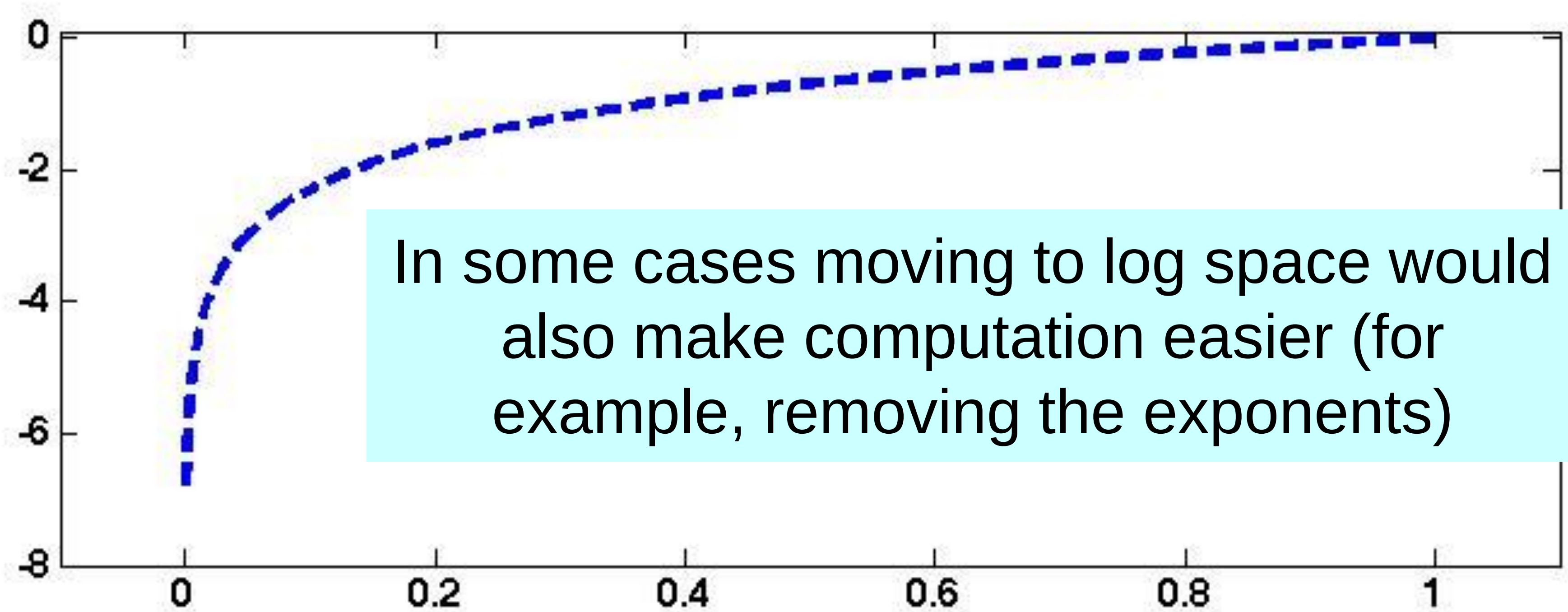
Log Probabilities

When working with products, probabilities of entire datasets often get too small. A possible solution is to use the log of probabilities, often termed 'log likelihood'

$$\log \hat{P}(\text{dataset} \mid M) = \log \prod_{k=1}^{n} \hat{P}(x_k \mid M) = \sum_{k=1}^{n} \log \hat{P}(x_k \mid M)$$

Maximizing this likelihood function is the same as maximizing P(dataset | M)

Log values between 0 and 1



In some cases moving to log space would also make computation easier (for example, removing the exponents)

How much do grad students sleep?

- Lets try to estimate the distribution of the time students spend sleeping (outside class).

Possible statistics

- **X**

Sleep time

- **Mean of X:**

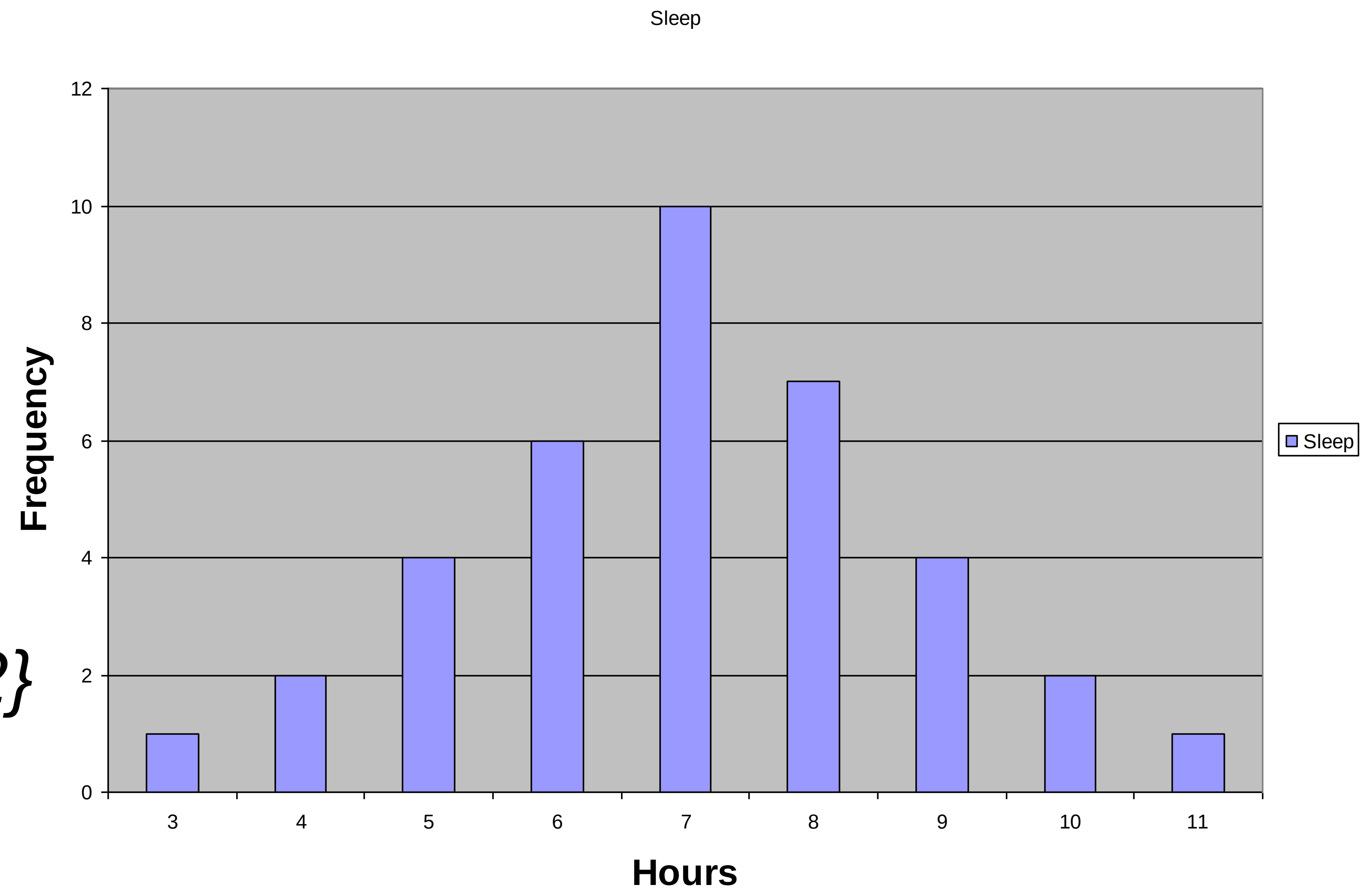
E{X}

7.03

- **Variance of X:**

Var{X} = E{(X-E{X})^2}

3.05

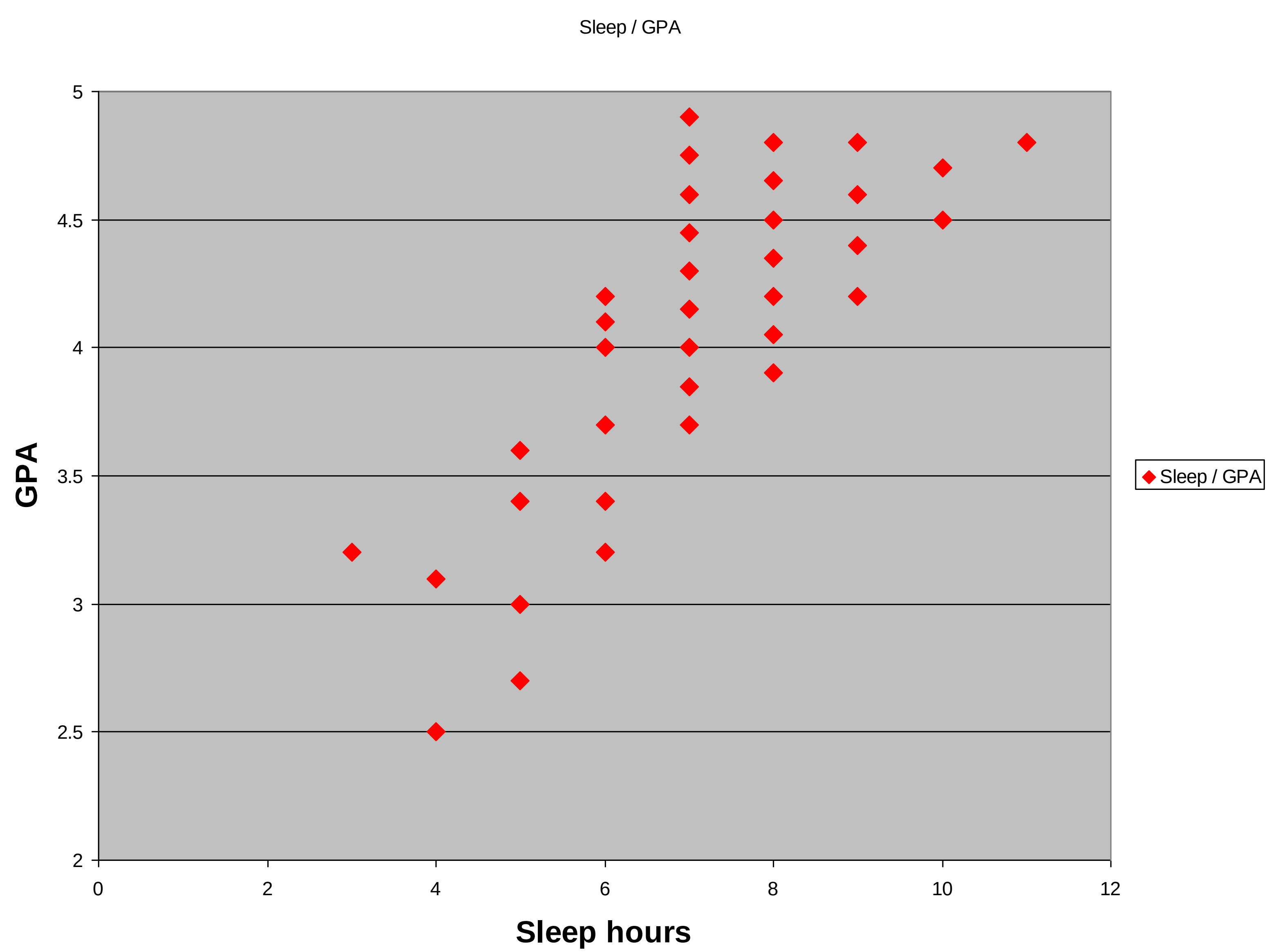


Covariance: Sleep vs. GPA

- **Co-Variance of X1, X2:**

Covariance{X1,X2} = E{(X1-E{X1})(X2-E{X2})}

= 0.88



Statistical Models

- Statistical models attempt to characterize properties of the population of interest

- For example, we might believe that repeated measurements follow a normal (Gaussian) distribution with some mean μ and variance σ^2 , x ~ N(μ,σ^2)

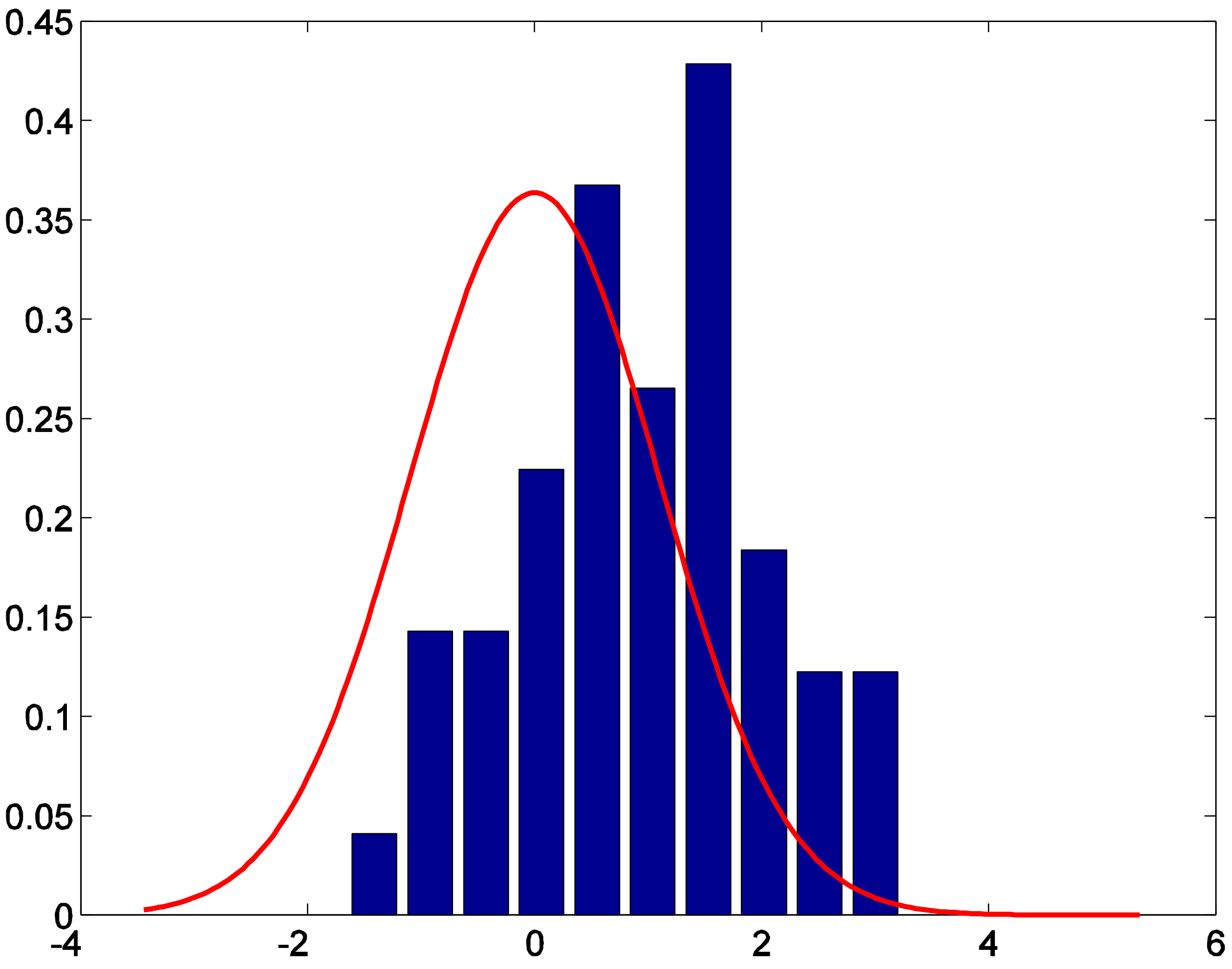
where

$$p(x \mid \Theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

and $\Theta = (\mu, \sigma^2)$ defines the parameters (mean and variance) of the model.

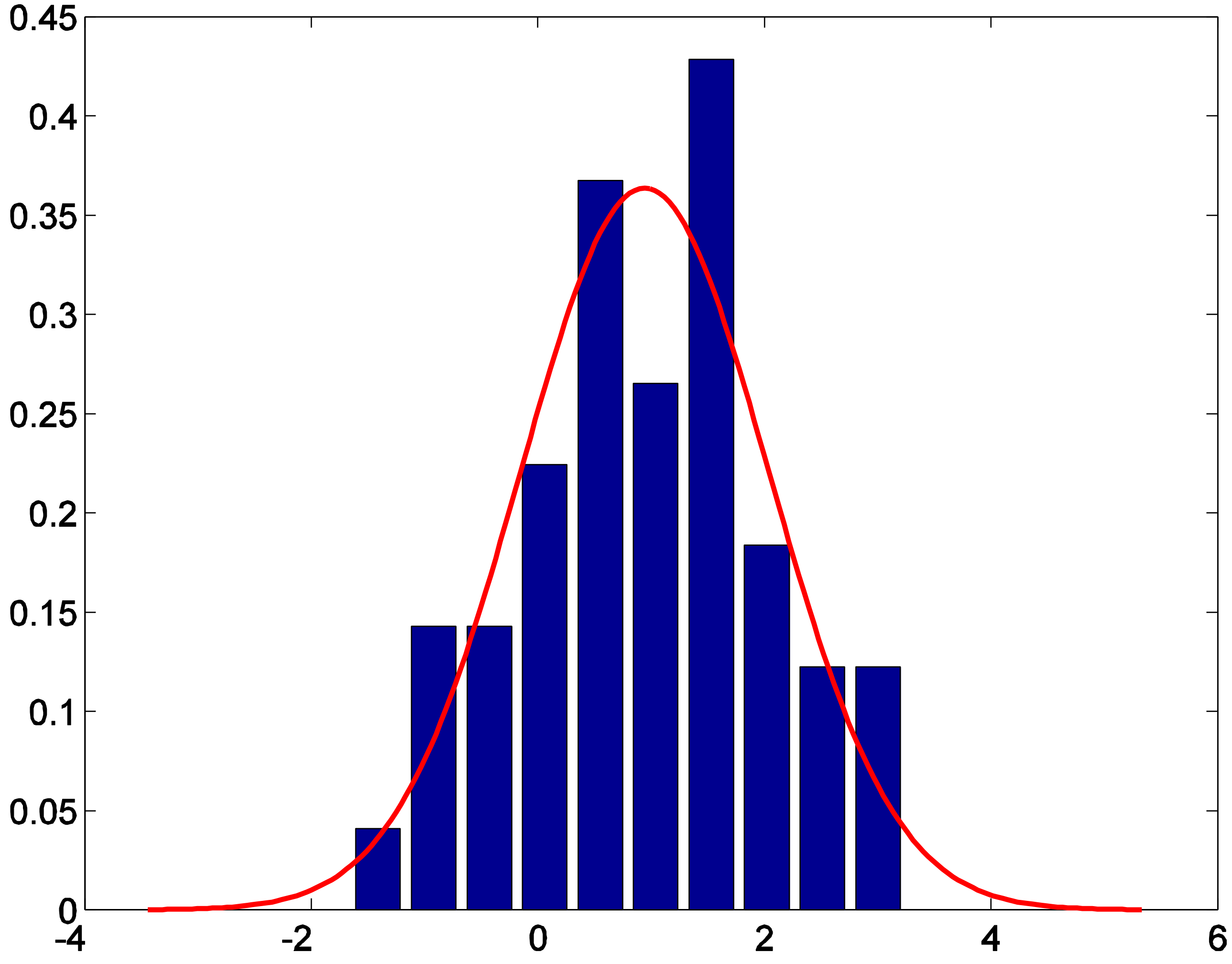
The Parameters of Our Model

- A statistical model is a **collection** of distributions; the **parameters** specify individual distributions x ~ N(μ,σ^2)
- We need to adjust the parameters so that the resulting distribution **fits** the data well



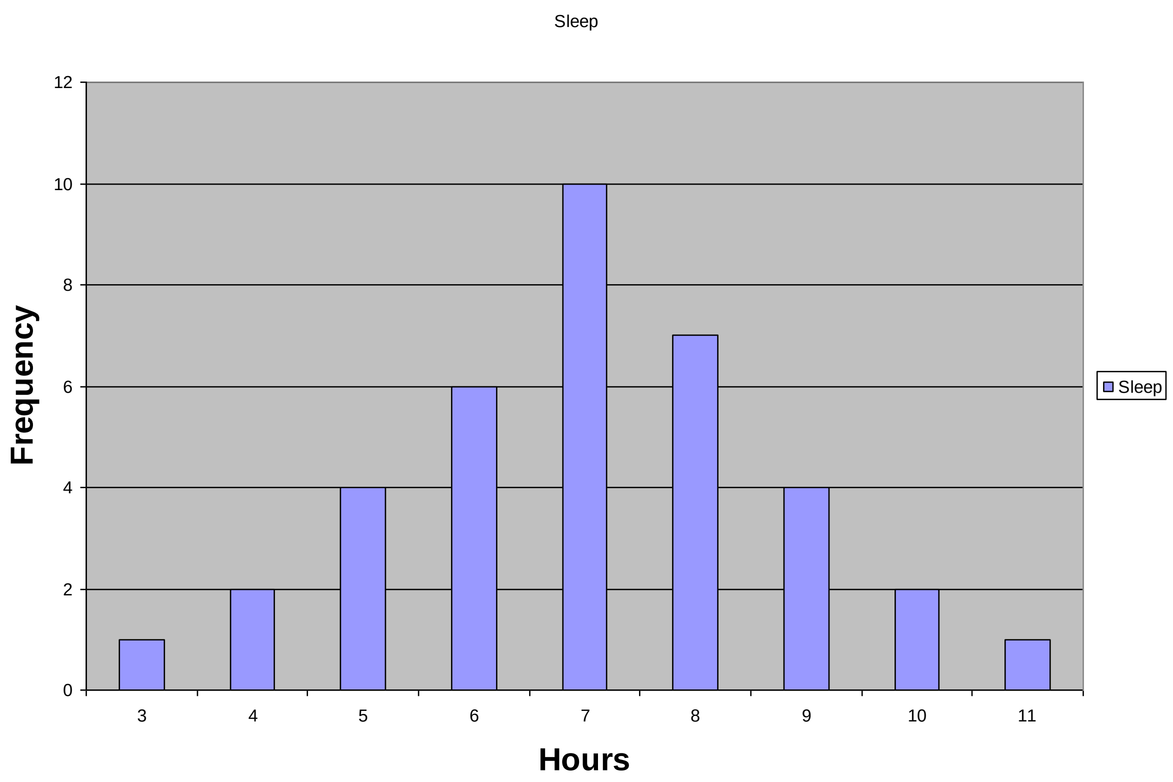
The Parameters of Our Model

- A statistical model is a **collection** of distributions; the **parameters** specify individual distributions x ~ N(μ,σ^2)
- We need to adjust the parameters so that the resulting distribution **fits** the data well



Computing the parameters of our model

- Lets assume a Guassian distribution for our sleep data
- How do we compute the parameters of the model?



Maximum Likelihood Principle

• We can fit statistical models by maximizing the probability of generating the observed samples:
$L(x_1, \ldots, x_n \mid \Theta) = p(x_1 \mid \Theta) \ldots p(x_n \mid \Theta)$
(the samples are assumed to be independent)

• In the Gaussian case we simply set the mean and the variance to the sample mean and the sample variance:

$$\overline{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \overline{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{\mu})^2$$

Why?

Density estimation

- Binary and discrete variables:

Easy: Just count!

- Continuous variables:

Harder (but just a bit): Fit a model

But what if we only have very few samples?

MLE vs. MAP

- Maximum Likelihood estimation (MLE)

  Choose value that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D|\theta)$$

- Maximum *a posteriori* (MAP) estimation

  Choose value that is most probable given observed data and prior belief

$$\begin{aligned}\widehat{\theta}_{MAP} &= \arg\max_{\theta} P(\theta|D) \\ &= \arg\max_{\theta} P(D|\theta)P(\theta)\end{aligned}$$

Important points

- Random variables
- Chain rule
- Bayes rule
- Joint distribution, independence, conditional independence
- MLE

Assume we performed *n* coin flips and used the outcome to learn the probability of heads, defined as *q*. In the questions below assume that $0 < q < 1$ unless stated otherwise.

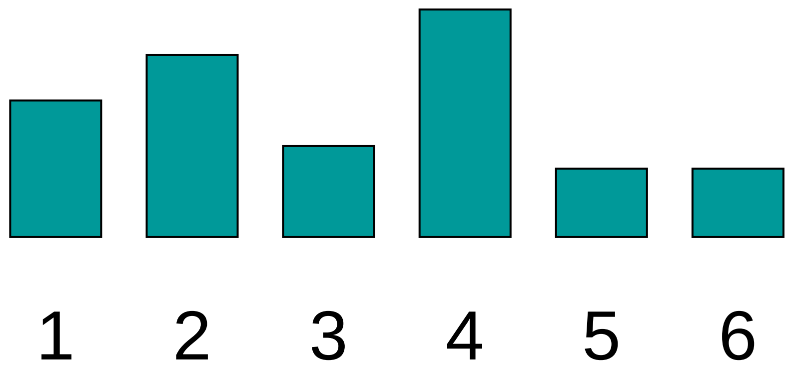
1. We have performed an additional coin flip and learned a new probability for heads, *q1*, based on the *n+1* observations. The following holds:
a. q1 = q
b. q1 ≠ q
c. it depends on q and the value of the new observation

2. We have performed *two* additional coin flips and learned a new probability for heads, *q1*, based on the *n+2* observations. The following holds:
a. q1 = q
b. q1 ≠ q
c. it depends on q and the values of the new observations

3. Now assume that $0.6 < q < 1$. Similar to (2) we have performed *two* additional coin flips and learned a new probability for heads, *q1*, based on the *n+2* observations. The following holds:
1. q1 = q
2. q1 ≠ q
3. it depends on q and the values of the new observations

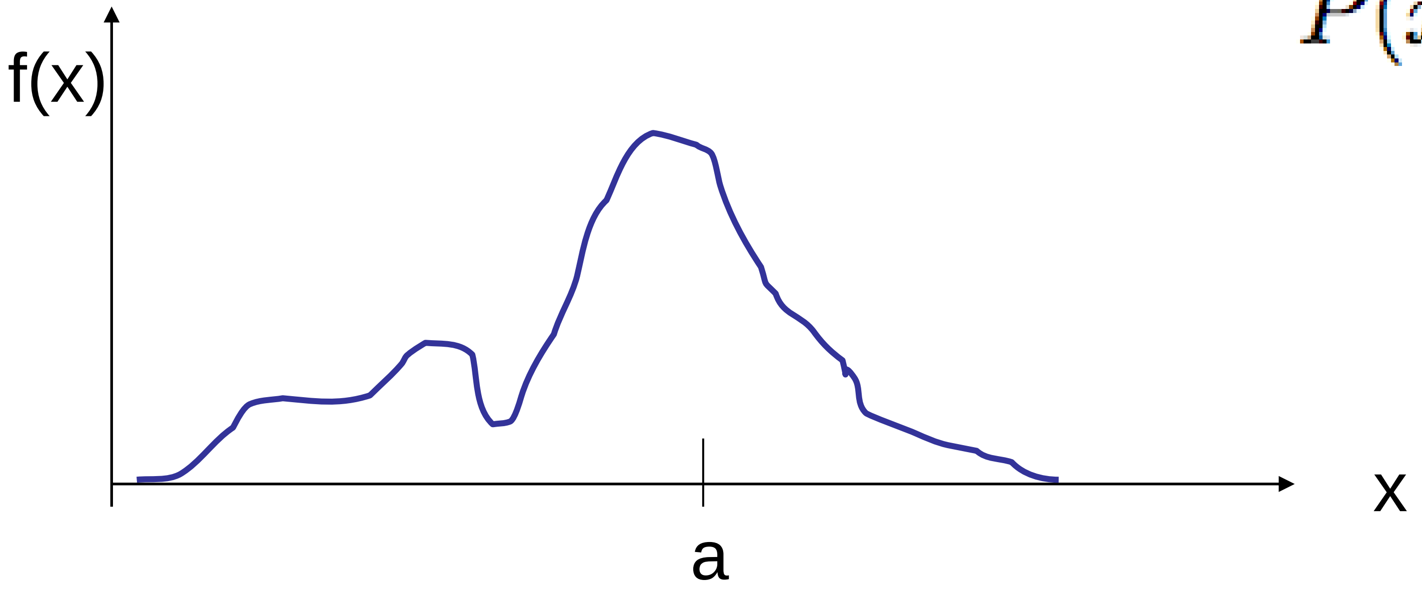
Probability Density Function

- Discrete distributions



$$\sum_i P(X = x_i) = 1$$

- Continuous: Cumulative Density Function (CDF): *F(a)*



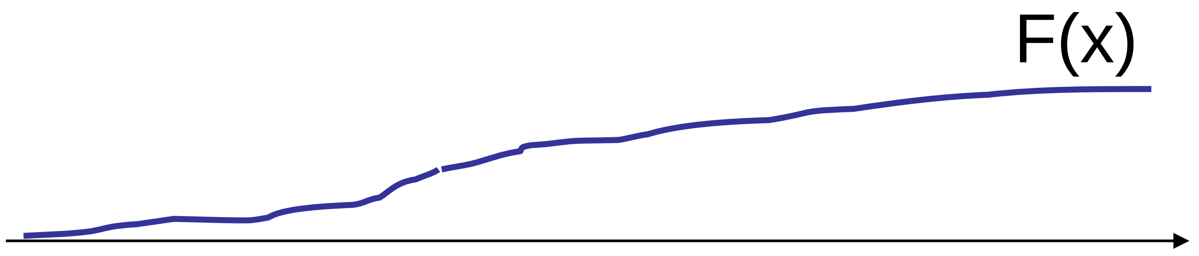
$$P(x \leq a) = \int_{-\infty}^{a} f(\tau) d\tau$$

Cumulative Density Functions

- Total probability
$$P(\Omega) = \int_{-\infty}^{\infty} f(x)dx = 1$$
- Probability Density Function (PDF)
$$\frac{d}{dx}F(x) = f(x)$$
- Properties:
$$P(a \leq x \leq b) = \int_{b}^{a} f(x)dx = F(b) - F(a)$$

$$\lim_{x \to -\infty} F(x) = 0$$

$$\lim_{x \to \infty} F(x) = 1$$



$$F(a) \geq F(b) \ \forall a \geq b$$

Expectations

- Mean/Expected Value:

$$E[x] = \bar{x} = \int x f(x) dx$$

- Variance:

$$Var(x) = E[(x - \bar{x})^2] = E[x^2] - (\bar{x})^2$$

- In general:

$$E[x^2] = \int x^2 f(x) dx$$

$$E[g(x)] = \int g(x) f(x) dx$$

Multivariate

- Joint for (x,y)

$$P((x,y) \in A) = \int \int_A f(x,y) dx dy$$

- Marginal:

$$f(x) = \int f(x,y) dy$$

- Conditionals:

$$f(x|y) = \frac{f(x,y)}{f(y)}$$

- Chain rule:

$$f(x,y) = f(x|y)f(y) = f(y|x)f(x)$$

Bayes Rule

- Standard form:

$$f(x|y) = \frac{f(y|x)f(x)}{f(y)}$$

- Replacing the bottom:

$$f(x|y) = \frac{f(y|x)f(x)}{\int f(y|x)f(x)dx}$$

Binomial

- Distribution:

$$x \sim Binomial(p, n)$$

$$P(x = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Mean/Var:

$$E[x] = np$$

$$Var(x) = np(1-p)$$

Uniform

- Anything is equally likely in the region [a,b]

- Distribution:

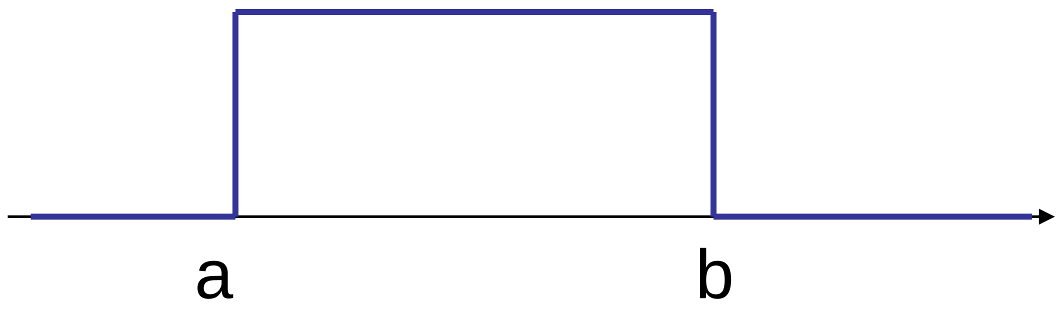
$$x \sim U(a,b)$$

- Mean/Var

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & otherwise \end{cases}$$

$$E[x] = \frac{a+b}{2}$$

$$Var(x) = \frac{a^2+ab+b^2}{3}$$



Gaussian (Normal)

- If I look at the height of women in country xx, it will look approximately Gaussian
- Small random noise errors, look Gaussian/Normal

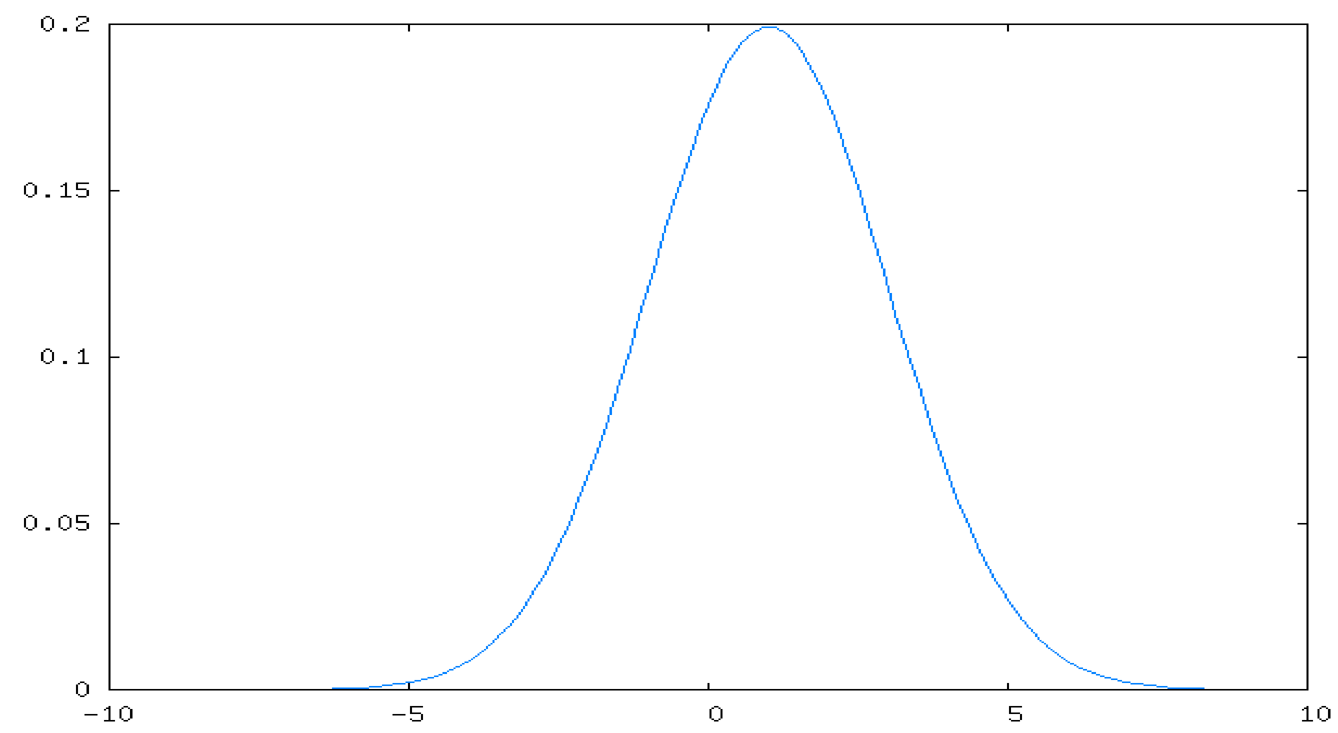
- Distribution:

$$x \sim N(\mu, \sigma^2) \qquad f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

- Mean/var

$$E[x] = \mu$$

$$Var(x) = \sigma^2$$



Why Do People Use Gaussians

- Central Limit Theorem: (loosely)
  - Sum of a large number of IID random variables is approximately Gaussian

Multivariate Gaussians

- Distribution for vector x

$$x = (x_1, \ldots, x_N)^T, \quad x \sim N(\mu, \Sigma)$$

- PDF:

$$f(x) = \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

$$E[x] = \mu = (E[x_1], \ldots, E[x_N])^T$$

$$Var(x) \rightarrow \Sigma = \begin{pmatrix} Var(x_1) & Cov(x_1, x_2) & \ldots & Cov(x_1, x_N) \\ Cov(x_2, x_1) & Var(x_2) & \ldots & Cov(x_2, x_N) \\ \vdots & & \ddots & \vdots \\ Cov(x_N, x_1) & Cov(x_N, x_2) & \ldots & Var(x_N) \end{pmatrix}$$

Multivariate Gaussians

$$f(x) = \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

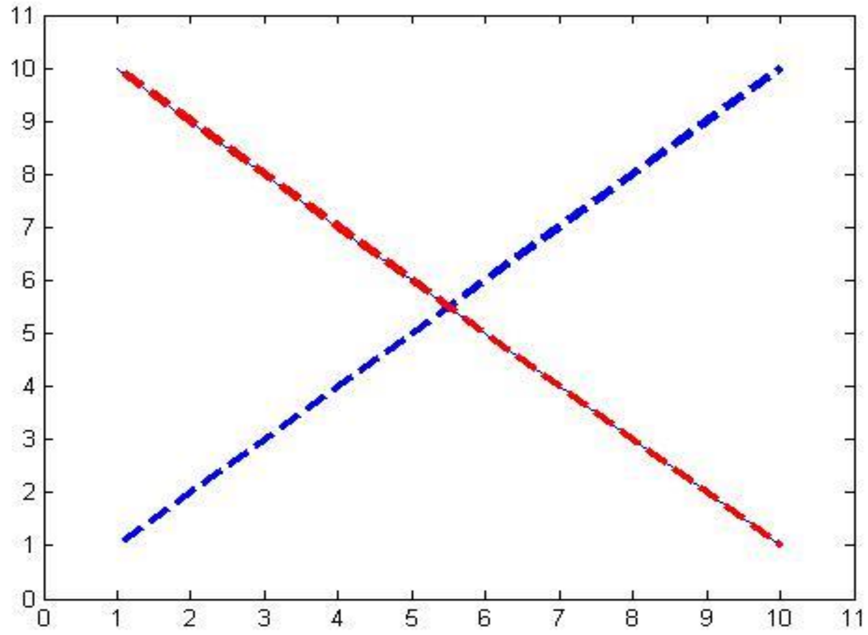
$$E[x] = \mu = (E[x_1], \ldots, E[x_N])^T$$

$$Var(x) \rightarrow \Sigma = \begin{pmatrix} Var(x_1) & Cov(x_1, x_2) & \ldots & Cov(x_1, x_N) \\ Cov(x_2, x_1) & Var(x_2) & \ldots & Cov(x_2, x_N) \\ \vdots & & \ddots & \vdots \\ Cov(x_N, x_1) & Cov(x_N, x_2) & \ldots & Var(x_N) \end{pmatrix}$$

$$\mathrm{cov}(\chi_1, \chi_2) = \frac{1}{n} \sum_{i=1}^{n} (x_{1,i} - \mu_1)(x_{2,i} - \mu_2)$$

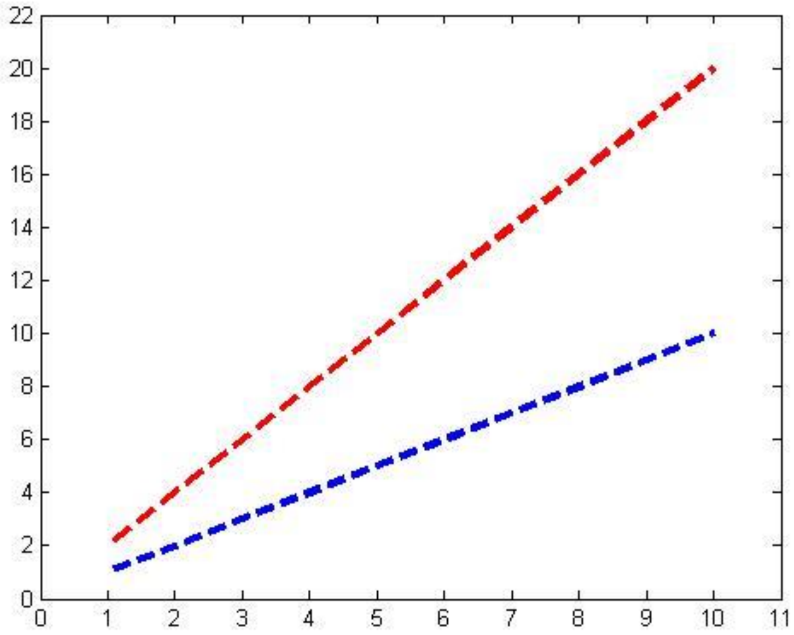
Covariance examples

Anti-correlated



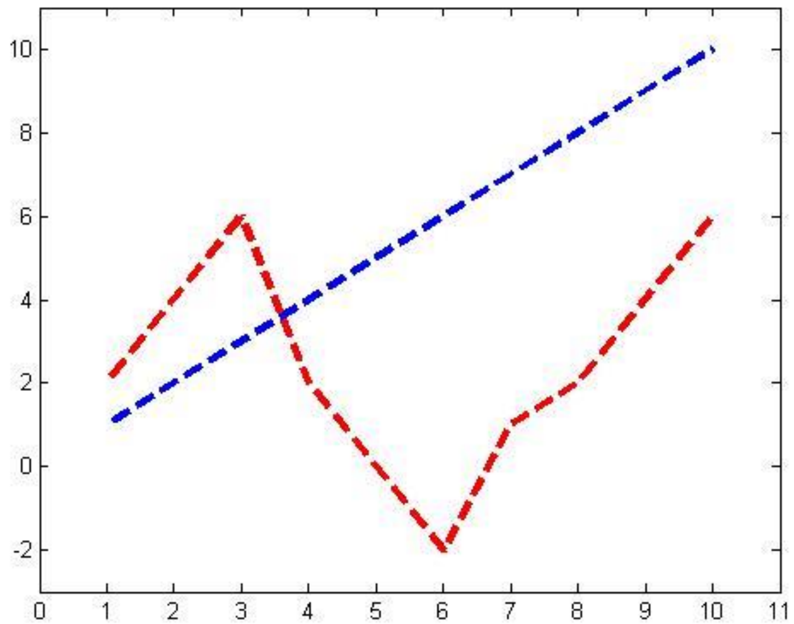
Covariance: -9.2

Correlated



Covariance: 18.33

Independent (almost)



Covariance: 0.6

Sum of Gaussians

- The sum of two Gaussians is a Gaussian:

$$x \sim N(\mu, \sigma^2) \quad y \sim N(\mu_y, \sigma_y^2)$$

$$ax + b \sim N(a\mu + b, (a\sigma)^2)$$

$$x + y \sim N(\mu + \mu_y, \sigma^2 + \sigma_y^2)$$