Some Probability and Statistics

David M. Blei

COS424
Princeton University

February 13, 2012

Card problem

- There are three cards
  - Red/Red
  - Red/Black
  - Black/Black
- I go through the following process.
  - Close my eyes and pick a card
  - Pick a side at random
  - Show you that side
- Suppose I show you red. What’s the probability the other side is red too?

Random variable

- Probability is about *random variables.*
- A random variable is any “probabilistic” outcome.
- For example,
    - The flip of a coin
    - The height of someone chosen randomly from a population
- We’ll see that it’s sometimes useful to think of quantities that are not strictly probabilistic as random variables.
    - The temperature on 11/12/2013
    - The temperature on 03/04/1905
    - The number of times “streetlight” appears in a document

Random variable

- Random variables take on values in a *sample space*.
- They can be *discrete* or *continuous*:
  - Coin flip: $\{H, T\}$
  - Height: positive real values $(0, \infty)$
  - Temperature: real values $(-\infty, \infty)$
  - Number of words in a document: Positive integers $\{1, 2, \ldots\}$
- We call the values *atoms*.
- Denote the random variable with a capital letter; denote a realization of the random variable with a lower case letter.
- E.g., X is a coin flip, x is the value (H or T) of that coin flip.

Discrete distribution

- A discrete distribution assigns a probability to every atom in the sample space
- For example, if X is an (unfair) coin, then

$$\begin{aligned} P(X=H) &= 0.7 \\ P(X=T) &= 0.3 \end{aligned}$$

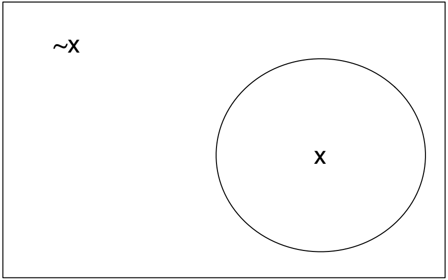
- The probabilities over the entire space must sum to one

$$\sum_x P(X=x)=1$$

- Probabilities of disjunctions are sums over part of the space. E.g., the probability that a die is bigger than 3:

$$P(D>3)=P(D=4)+P(D=5)+P(D=6)$$

A useful picture



- An *atom* is a point in the box
- An *event* is a subset of atoms (e.g., $d > 3$)
- The probability of an event is sum of probabilities of its atoms.

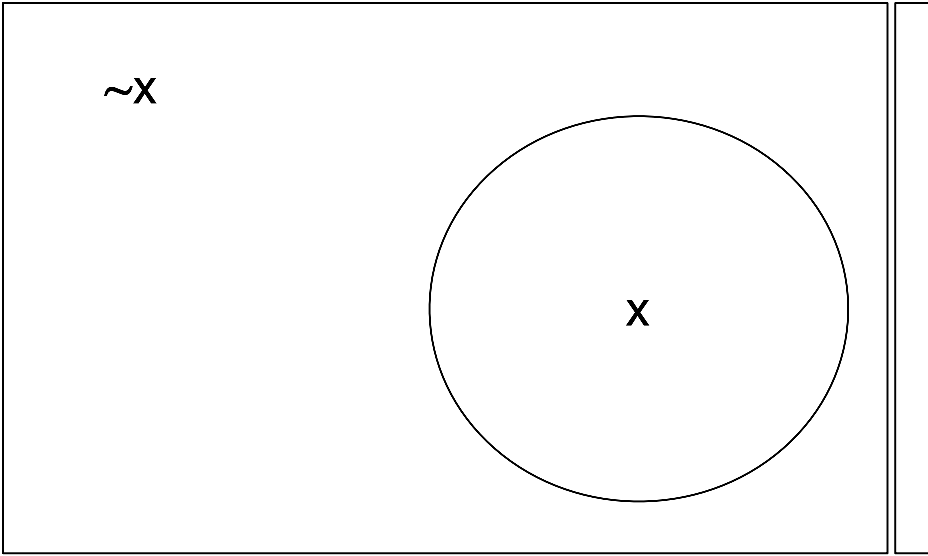
Joint distribution

- Typically, we consider collections of random variables.
- The joint distribution is a distribution over the configuration of all the random variables in the ensemble.
- For example, imagine flipping 4 coins. The joint distribution is over the space of all possible outcomes of the four coins.

$$
\begin{aligned}
P(HHHH) &= 0.0625 \\
P(HHHT) &= 0.0625 \\
P(HHTH) &= 0.0625 \\
&\cdots
\end{aligned}
$$

- You can think of it as a single random variable with 16 values.

Visualizing a joint distribution



Conditional distribution

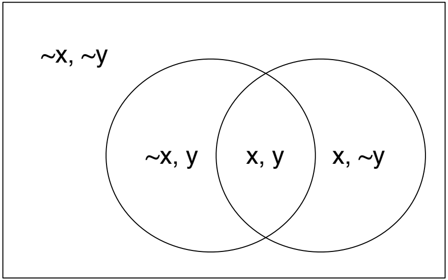
- A *conditional distribution* is the distribution of a random variable given some evidence.
- $P(X = x \mid Y = y)$ is the probability that $X = x$ when $Y = y$.
- For example,

$$
\begin{aligned}
P(\text{I listen to Steely Dan}) &= 0.5 \\
P(\text{I listen to Steely Dan} \mid \text{Toni is home}) &= 0.1 \\
P(\text{I listen to Steely Dan} \mid \text{Toni is not home}) &= 0.7
\end{aligned}
$$

- $P(X = x \mid Y = y)$ is a different distribution for each value of y

$$
\begin{aligned}
\sum_x P(X = x \mid Y = y) &= 1 \\
\sum_y P(X = x \mid Y = y) &\neq 1 \quad (\textit{necessarily})
\end{aligned}
$$

Definition of conditional probability



- Conditional probability is defined as:

$$P(X=x \mid Y=y) = \frac{P(X=x, Y=y)}{P(Y=y)},$$

  which holds when $P(Y) > 0$.
- In the Venn diagram, this is the relative probability of $X=x$ in the space where $Y=y$.

Returning to the card problem

- Now we can solve the card problem.
- Let X_1 be the random side of the random card I chose
- Let X_2 be the other side of that card
- Compute $P(X_2 = \mathrm{red} | X_1 = \mathrm{red})$

$$P(X_2 = \mathrm{red} | X_1 = \mathrm{red}) = \frac{P(X_1 = R, X_2 = R)}{P(X_1 = R)} \tag{1}$$

- Numerator is 1/3: Only one card has two red sides.
- Denominator is 1/2: There are three possible sides of the six that are red.

The chain rule

- The definition of conditional probability lets us derive the *chain rule*, which let's us define the joint distribution as a product of conditionals:

$$\begin{aligned} P(X,Y) &= P(X,Y)\frac{P(Y)}{P(Y)} \\ &= P(X|Y)P(Y) \end{aligned}$$

- For example, let Y be a disease and X be a symptom. We may know $P(X|Y)$ and $P(Y)$ from data. Use the chain rule to obtain the probability of having the disease and the symptom.
- In general, for any set of N variables

$$P(X_1,\ldots,X_N) = \prod_{n=1}^{N} P(X_n|X_1,\ldots,X_{n-1})$$

Marginalization

- Given a collection of random variables, we are often only interested in a subset of them.
- For example, compute $P(X)$ from a joint distribution $P(X, Y, Z)$
- Can do this with *marginalization*

$$P(X) = \sum_y \sum_z P(X, y, z)$$

- Derived from the chain rule:

$$\begin{aligned}
\sum_y \sum_z P(X, y, z) &= \sum_y \sum_z P(X) P(y, z | X) \\
&= P(X) \sum_y \sum_z P(y, z | X) \\
&= P(X)
\end{aligned}$$

- Note: now we can compute the probability that Toni is home.

Bayes rule

- From the chain rule and marginalization, we obtain *Bayes rule*.

$$P(Y|X) = \frac{P(X|Y)P(Y)}{\sum_y P(X|Y=y)P(Y=y)}$$

- Again, let Y be a disease and X be a symptom. From $P(X|Y)$ and $P(Y)$, we can compute the (useful) quantity $P(Y|X)$.
- Bayes rule is important in *Bayesian statistics*, where Y is a parameter that controls the distribution of X.

Independence

- Random variables are *independent* if knowing about X tells us nothing about Y.

$$P(Y|X) = P(Y)$$

- This means that their joint distribution factorizes,

$$X \perp\!\!\!\perp Y \iff P(X,Y) = P(X)P(Y).$$

- Why? The chain rule

$$\begin{aligned} P(X,Y) &= P(X)P(Y|X) \\ &= P(X)P(Y) \end{aligned}$$

Independence examples

- Examples of independent random variables:
  - Flipping a coin once / flipping the same coin a second time
  - You use an electric toothbrush / blue is your favorite color
- Examples of not independent random variables:
  - Registered as a Republican / voted for Bush in the last election
  - The color of the sky / The time of day

Are these independent?

- Two twenty-sided dice
- Rolling three dice and computing $(D_1 + D_2, D_2 + D_3)$
- # enrolled students and the temperature outside today
- # attending students and the temperature outside today

Two coins

- Suppose we have two coins, one biased and one fair,

$$P(C_1 = H) = 0.5 \quad P(C_2 = H) = 0.7.$$

- We choose one of the coins at random $Z \in \{1, 2\}$, flip C_Z twice, and record the outcome (X, Y).
- Question: Are X and Y independent?
- What if we knew which coin was flipped Z?

Conditional independence

- X and Y are *conditionally independent* given Z.

$$P(Y|X, Z = z) = P(Y|Z = z)$$

  for all possible values of z.

- Again, this implies a factorization

$$X \perp\!\!\!\perp Y|Z \iff P(X, Y|Z = z) = P(X|Z = z)P(Y|Z = z),$$

  for all possible values of z.

Continuous random variables

- We've only used discrete random variables so far (e.g., dice)
- Random variables can be continuous.
- We need a *density* $p(x)$, which *integrates* to one.
  E.g., if $x \in \mathbb{R}$ then
  $$\int_{-\infty}^{\infty} p(x)dx = 1$$
- Probabilities are integrals over smaller intervals. E.g.,
  $$P(X \in (-2.4, 6.5)) = \int_{-2.4}^{6.5} p(x)dx$$
- Notice when we use P, p, X, and x.

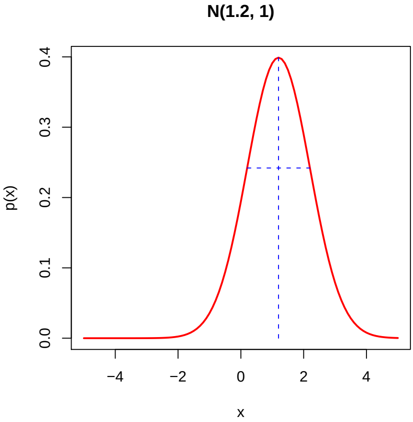
The Gaussian distribution

- The Gaussian (or Normal) is a continuous distribution.

$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

- The density of a point x is proportional to the negative exponentiated half distance to μ scaled by σ^2.
- μ is called the *mean*; σ^2 is called the *variance*.

Gaussian density



- The mean μ controls the location of the bump.
- The variance σ^2 controls the spread of the bump.

Notation

- For discrete RV's, p denotes the *probability mass function*, which is the same as the distribution on atoms.
- (I.e., we can use P and p interchangeably for atoms.)
- For continuous RV's, p is the density and they are not interchangeable.
- This is an unpleasant detail. Ask when you are confused.

Expectation

- Consider a function of a random variable, $f(X)$.
  (Notice: $f(X)$ is also a random variable.)
- The expectation is a weighted average of f,
  where the weighting is determined by $p(x)$,

$$\mathrm{E}[f(X)] = \sum_{x} p(x) f(x)$$

- In the continuous case, the expectation is an integral

$$\mathrm{E}[f(X)] = \int p(x) f(x) dx$$

Conditional expectation

- The conditional expectation is defined similarly

$$\mathrm{E}[f(X) \mid Y = y] = \sum_x p(x \mid y) f(x)$$

- Question: What is $\mathrm{E}[f(X) \mid Y = y]$? What is $\mathrm{E}[f(X) \mid Y]$?
- $\mathrm{E}[f(X) \mid Y = y]$ is a scalar.
- $\mathrm{E}[f(X) \mid Y]$ is a (function of a) random variable.

Iterated expectation

Let's take the expectation of $\mathrm{E}[f(X) \mid Y]$.

$$\begin{aligned}
\mathrm{E}[\mathrm{E}[f(X)] \mid Y]] &= \sum_y p(y) \mathrm{E}[f(X) \mid Y = y] \\
&= \sum_y p(y) \sum_x p(x \mid y) f(x) \\
&= \sum_y \sum_x p(x, y) f(x) \\
&= \sum_y \sum_x p(x) p(y \mid x) f(x) \\
&= \sum_x p(x) f(x) \sum_y p(y \mid x) \\
&= \sum_x p(x) f(x) \\
&= \mathrm{E}[f(X)]
\end{aligned}$$

Flips to the first heads

- We flip a coin with probability π of heads until we see a heads.
- What is the expected waiting time for a heads?

$$\begin{aligned}\mathrm{E}[N] &= 1\pi + 2(1-\pi)\pi + 3(1-\pi)^2\pi + \ldots \\ &= \sum_{n=1}^{\infty} n(1-\pi)^{(n-1)}\pi\end{aligned}$$

Let's use iterated expectation

$$\begin{aligned}
\mathrm{E}[N] &= \mathrm{E}[\mathrm{E}[N|X_1]] \\
&= \pi \cdot \mathrm{E}[N|X_1 = H] + (1-\pi)\mathrm{E}[N|X_1 = T] \\
&= \pi \cdot 1 + (1-\pi)(\mathrm{E}[N] + 1)] \\
&= \pi + 1 - \pi + (1-\pi)\mathrm{E}[N] \\
&= 1/\pi
\end{aligned}$$

Probability models

- Probability distributions are used as *models* of data that we observe.
- Pretend that data is drawn from an unknown distribution.
- *Infer* the properties of that distribution from the data
- For example
  - the bias of a coin
  - the average height of a student
  - the chance that someone will vote for H. Clinton
  - the chance that someone from Vermont will vote for H. Clinton
  - the proportion of gold in a mountain
  - the number of bacteria in our body
  - the evolutionary rate at which genes mutate
- We will see many models in this class.

Independent and identically distributed random variables

- Independent and identically distributed (IID) variables are:
  1. Independent
  2. Identically distributed
- If we repeatedly flip the same coin N times and record the outcome, then $X_1, \ldots, X_N$ are IID.
- The IID assumption can be useful in data analysis.

What is a parameter?

- Parameters are values that *index* a distribution.
- A coin flip is a *Bernoulli*. Its parameter is the probability of heads.

$$p(x \mid \pi) = \pi^{1[x=H]}(1-\pi)^{1[x=T]},$$

  where $1[\cdot]$ is called an *indicator function*. It is 1 when its argument is true and 0 otherwise.
- Changing π leads to different Bernoulli distributions.
- A Gaussian has two parameters, the mean and variance.

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

- A multinomial distribution has a vector parameter, on the *simplex*.

The likelihood function

- Again, suppose we flip a coin N times and record the outcomes.
- Further suppose that *we think* that the probability of heads is π. (This is distinct from whatever the probability of heads "really" is.)
- Given π, the probability of an observed sequence is

$$p(x_1, \ldots, x_N \,|\, \pi) = \prod_{n=1}^{N} \pi^{1[x_n=H]} (1-\pi)^{1[x_n=T]}$$

The log likelihood

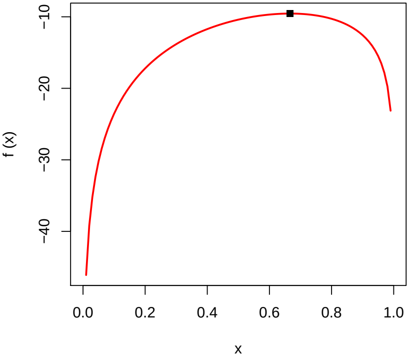
- As a function of π, the probability of a set of observations is called the likelihood function.

$$p(x_1, \ldots, x_N | \pi) = \prod_{n=1}^{N} \pi^{1[x_n=H]}(1-\pi)^{1[x_n=T]}$$

- Taking logs, this is the *log likelihood function*.

$$\mathscr{L}(\pi) = \sum_{n=1}^{N} 1[x_n = H] \log \pi + 1[x_n = T] \log(1-\pi)$$

Bernoulli log likelihood



- We observe *HHTHTHHTHHTHHTH*.
- The value of π that maximizes the log likelihood is 2/3.

The maximum likelihood estimate

- The *maximum likelihood estimate* is the value of the parameter that maximizes the log likelihood (equivalently, the likelihood).
- In the Bernoulli example, it is the proportion of heads.

$$\hat{\pi} = \frac{1}{N} \sum_{n=1}^{N} 1[x_n = H]$$

- In a sense, this is the value that best explains our observations.

Why is the MLE good?

- The MLE is *consistent*.
- Flip a coin N times with true bias π^*.
- Estimate the parameter from $x_1, \ldots x_N$ with the MLE $\hat{\pi}$.
- Then,

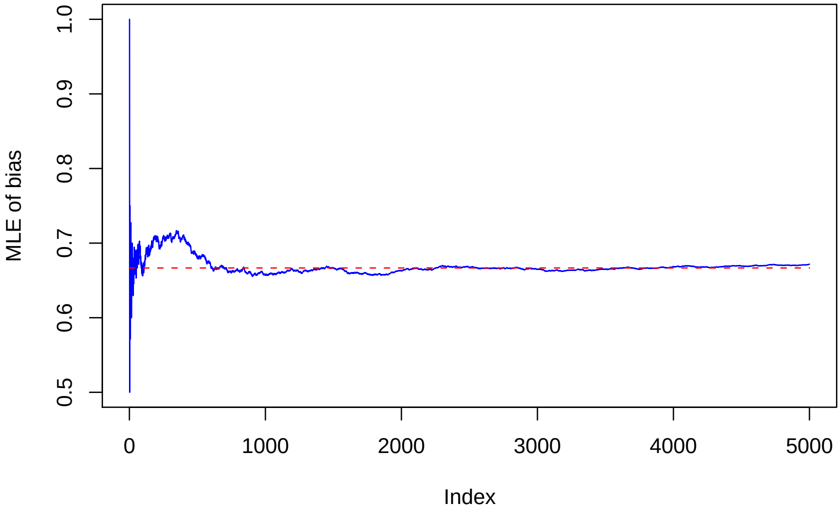
$$\lim_{N \to \infty} \hat{\pi} = \pi^*$$

- This is a good thing. It lets us sleep at night.

5000 coin flips

1 1 0 1 1 1 1 0 0 1 0 0 1 1 1 0 0 0 0 0 1 0 0 1 0 0 1 1 1 0 1 0 1 0 0 0 0 1 0 1 0 1 1
1 1 0 0 0 1 1 0 1 1 1 1 1 1 0 1 1 0 1 0 1 1 1 0 0 0 1 1 1 1 1 0 1 1 1 1 1 1 0 1 1 0 1
1 1 1 0 0 1 1 1 0 0 1 0 1 0 0 0 0 1 0 0 1 1 1 0 1 1 1 0 1 1 1 0 1 1 0 0 0 1 1 1 1 1 1
0 1 1 1 0 0 1 1 1 1 1 0 0 0 0 1 0 1 1 1 1 1 1 1 0 0 1 1 1 0 0 0 1 1 1 0 1 0 0 0 1 1 0
1 0 1 1 1 0 1 1 0 0 1 1 1 1 1 1 0 1 0 0 1 0 0 1 0 0 1 1 0 0 1 0 0 0 1 1 1 0 1 1 0 0 0
1 1 0 1 1 1 0 1 1 0 0 1 0 1 0 1 1 1 1 1 0 0 0 0 1 1 0 0 1 0 1 0 0 0 1 1 0 1 1 0 0 1 1
0 1 1 0 1 1 1 0 1 1 0 1 1 0 1 1 0 0 1 1 0 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 0 0 1 0 0
1 1 1 0 0 1 1 0 0 0 0 0 0 0 0 1 0 1 1 1 0 0 1 1 0 1 1 1 0 1 1 1 1 0 1 1 0 0 0 0 1 1 1
0 1 0 1 0 1 0 1 1 0 1 0 0 1 1 1 0 0 1 1 1 0 1 0 1 0 1 1 0 1 1 1 1 1 0 0 0 1 1 1 1 1 1
0 1 1 1 0 0 0 1 0 1 0 0 1 1 1 1 1 0 1 0 1 0 0 1 1 0 1 1 1 1 0 0 0 0 1 1 1 1 0 1 0 0 1
0 0 1 0 0 1 0 1 1 1 1 1 1 0 0 1 1 0 1 1 1 0 0 1 0 0 1 0 1 1 1 1 0 0 1 1 1 1 1 0 1 1 0
0 1 0 1 1 0 1 1 1 1 0 0 0 0 1 0 0 1 1 1 1 0 0 0 0 1 0 0 1 1 1 1 1 0 1 1 0 1 1 1 1 1 0
1 1 0 1 0 1 1 1 0 0 1 0 1 1 1 1 1 1 1 1 0 1 0 1 1 0 0 1 0 1 1 1 0 0 0 1 1 1 0 1 1 1 0
0 0 1 1 1...

Consistency of the MLE example



Gaussian log likelihood

- Suppose we observe $x_1, \ldots, x_N$ continuous.
- We choose to model them with a Gaussian

$$p(x_1, \ldots, x_N | \mu, \sigma^2) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ \frac{-(x_n - \mu)^2}{2\sigma^2} \right\}$$

- The log likelihood is

$$\mathscr{L}(\mu, \sigma) = -\frac{1}{2} N \log(2\pi\sigma^2) - \sum_{n=1}^{N} \frac{(x_n - \mu)^2}{2\sigma^2}$$

Gaussian MLE

- The MLE of the mean is the *sample mean*

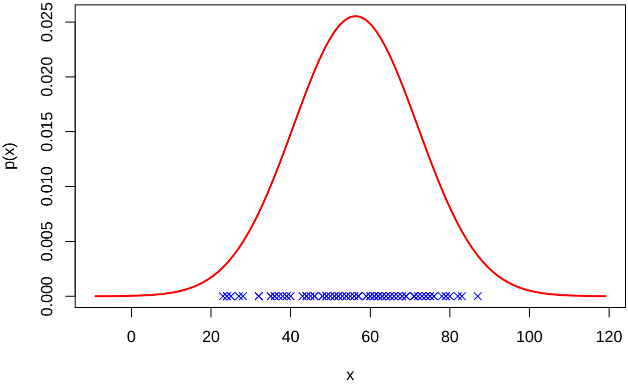
$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

- The MLE of the variance is the *sample variance*

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N} (x_n - \hat{\mu})^2$$

- E.g., approval ratings of the presidents from 1945 to 1975.

Gaussian analysis of approval ratings



Q: What's wrong with this analysis?

Model pitfalls

- What's wrong with this analysis?
  - Assigns positive probability to numbers < 0 and > 100
  - Ignores the sequential nature of the data
  - Assumes that approval ratings are IID!
- "All models are wrong. Some models are useful." (Box)

Graphical models

- Represents a joint distribution of random variables
- (Also called “Bayesian network”)
- Nodes are RVs; Edges denote possible dependence
- Shaded nodes are observed; unshaded nodes are hidden
- GMs with shaded nodes represent posterior distributions.
- Connects computing about models to graph structure (COS513)
- Connects independence assumptions to graph structure (COS513)
- Here we’ll use it as a schematic for factored joint distributions.

Some of the models we'll learn about

- Naive Bayes classification
- Linear regression and logistic regression
- Generalized linear models
- Hidden variables, mixture models, and the EM algorithm
- Factor analysis / Principal component analysis
- Sequential models
- Bayesian models