Quick Tour of Basic Probability Theory

CS224W: Social and Information Network Analysis
Fall 2012

Outline

Today's goal: A gentle refresher on probability

You should have seen this before

Outline

- Basic definitions
- Random variables
- Maximum likelihood estimation

Fundamentals of Probability

- Sample space Ω: Set of all possible outcomes
- Event space $\mathcal{F}$: 2^{Ω} (an event is a subset of the sample space)
- Probability measure: function $P : \mathcal{F} \to \mathbb{R}$ such that:
  - $P(A) \geq 0$ $(\forall A \in \mathcal{F})$
  - $P(\Omega) = 1$
  - For disjoint events A_i, $P(\cup_i A_i) = \sum_i P(A_i)$

In this session, I'll focus mostly on the **discrete** case (things are basically the same in the continuous case).

Example

Consider throwing a die twice:

- Sample space $\Omega = \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$
- Event space $\mathcal{F} = 2^{\Omega}$ (example events: let A be the event that the sum is even and let B be the event that we roll at least one 6).
- Probability measure: function P is simple counting in this simple discrete case. Example: $P(A) = 18/36 = 1/2$, $P(B) = 11/36$.

Note that multiple events can happen simultaneously. e.g. if we roll a 6 then a 2, the outcome is $\{6, 2\}$, and both A and B have occurred.

Union

For any two events A and B, the union of the two ("A or B") is:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

e.g. $P(A \cup B) = 18/36 + 11/36 - 5/36 = 24/36 = 2/3$

Conditional probability

Let A and B be two events. Then the conditional probability of A given B is:

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

"What's the probability of A once we know B has happened?"

Rewriting gives us the useful product rule:

$$P(A \cap B) = P(A|B)P(B)$$

Independence

Two events A and B are **independent** if

$$P(A \cap B) = P(A)P(B)$$

Equivalently: $P(A|B) = P(A)$ and $P(B|A) = P(B)$

Intuitively, knowing A doesn't tell you anything about B and vice-versa.

But beware of relying on your intuition: rolling two dice (x_a and x_b), events $x_a = 2$ and $x_a + x_b = k$ are independent if $k = 7$ and dependent otherwise.

Union bound

Recall that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ for any two events A and B.

If we're trying to upper bound the probability that A or B happens, the **worst case** is that A and B are disjoint (so $P(A \cap B) = 0$).

The surprisingly useful **union bound** now follows. Let A_i be some (not necessarily independent!) events, then:

$$P\left(\bigcup_i A_i\right) \leq \sum_I P(A_i)$$

Bayes' Rule

Most important basic rule of probability!

For two events A and B (such that $P(B) \neq 0$):

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

Often used to **update beliefs**:

$$\text{posterior} = \text{“support } B \text{ provides for } A\text{”} \times \text{“prior”}$$

Bayes' Rule Example

You friend told you she had a great conversation with someone on the Caltrain. Not knowing anything else, your prior belief that her conversation partner was a woman is 50%. Let W donate this event. Let L denote the event that her conversation partner has long hair. If you learn L to be true, how should you update your beliefs about W?

$P(W) = 0.5$ and suppose $P(L) = 0.6$, and $P(L|W) = 0.75$ are known.

Then $P(W|L) = \frac{P(L|W)P(W)}{P(L)} = 0.75 * 0.5/0.6 = 62.5\%$.

Random Variables

A *random variable* is a technically a function $X : \Omega \rightarrow \mathbb{R}$

Probabilities of random variable events come from underlying P function: $P(X = k) = P(\{\omega \in \Omega | X(\omega) = k\})$

It's called a random variable because it's a variable that doesn't take on a single, deterministic value, but it can take on a set of different values, each with an associated probability.

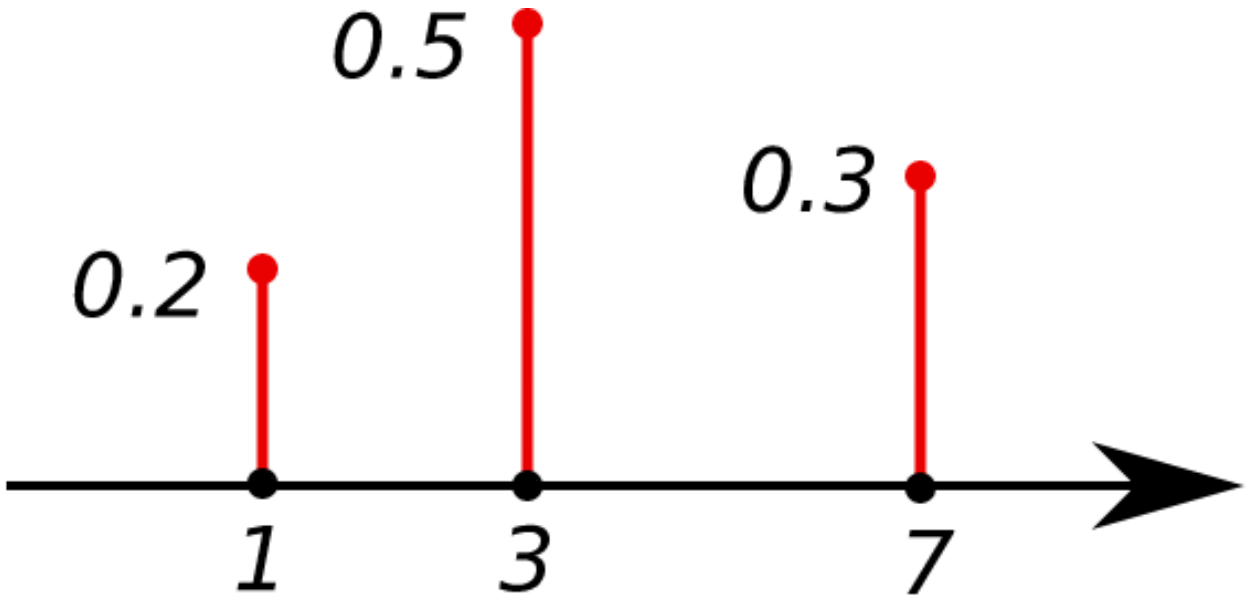
e.g. Let X be a random variable that counts the number of 6's we roll in 2 rolls of a die.

$P(X = 2) = P(\{6, 6\}) = 1/36$
$P(X = 1) =$
$P(\{1, 6\}) + \ldots + P(\{6, 6\}) + P(\{6, 1\}) + \ldots P(\{6, 5\}) = 10/36$
$P(X = 0) = 25/36$

Distributions

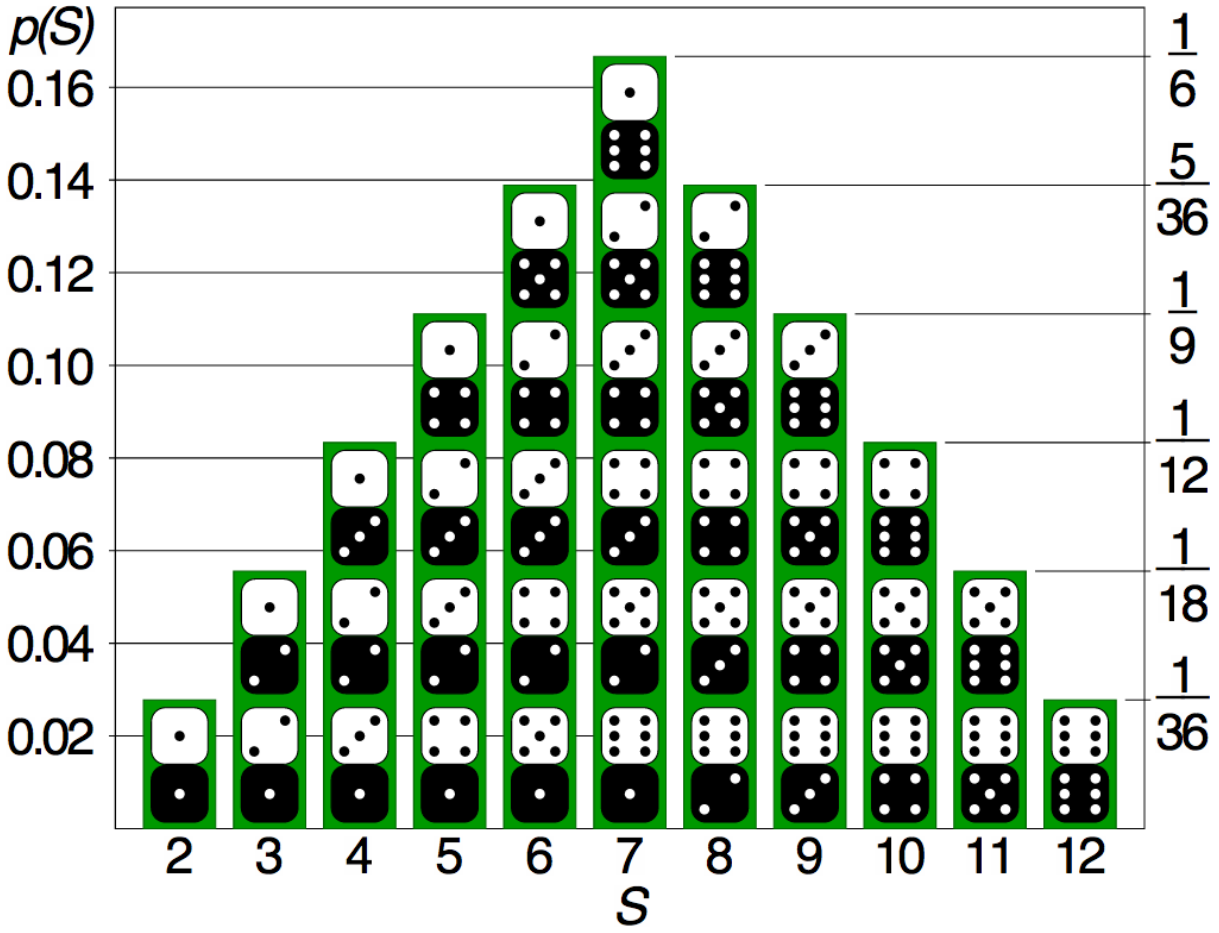
A probability mass function (pmf) assigns a probability to each possible value of a random variable (in the discrete case)

Example: funny die



Distributions

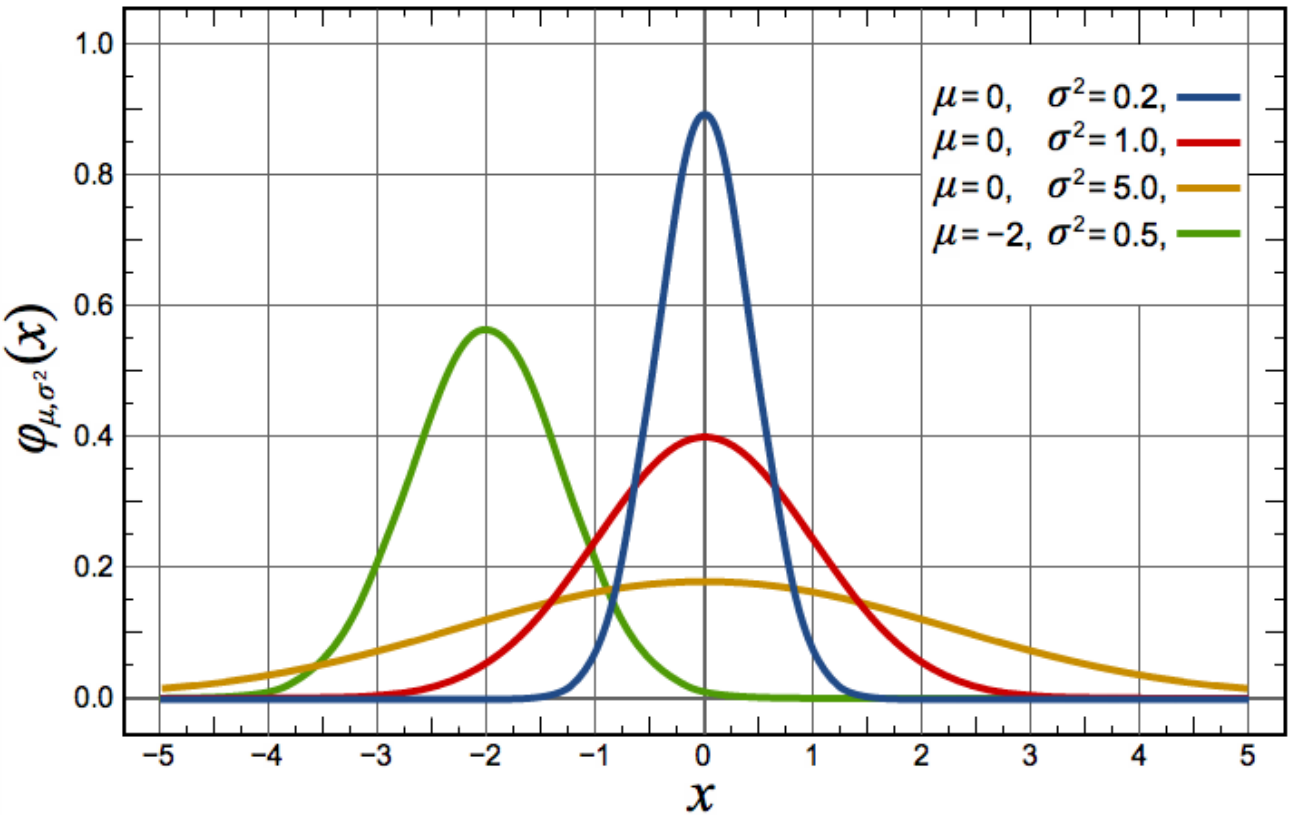
Another example: distribution over sum of two die rolls



Probability density functions

The PDF of a continuous random variable X is describes the relative likelihood for X to take on a given value:

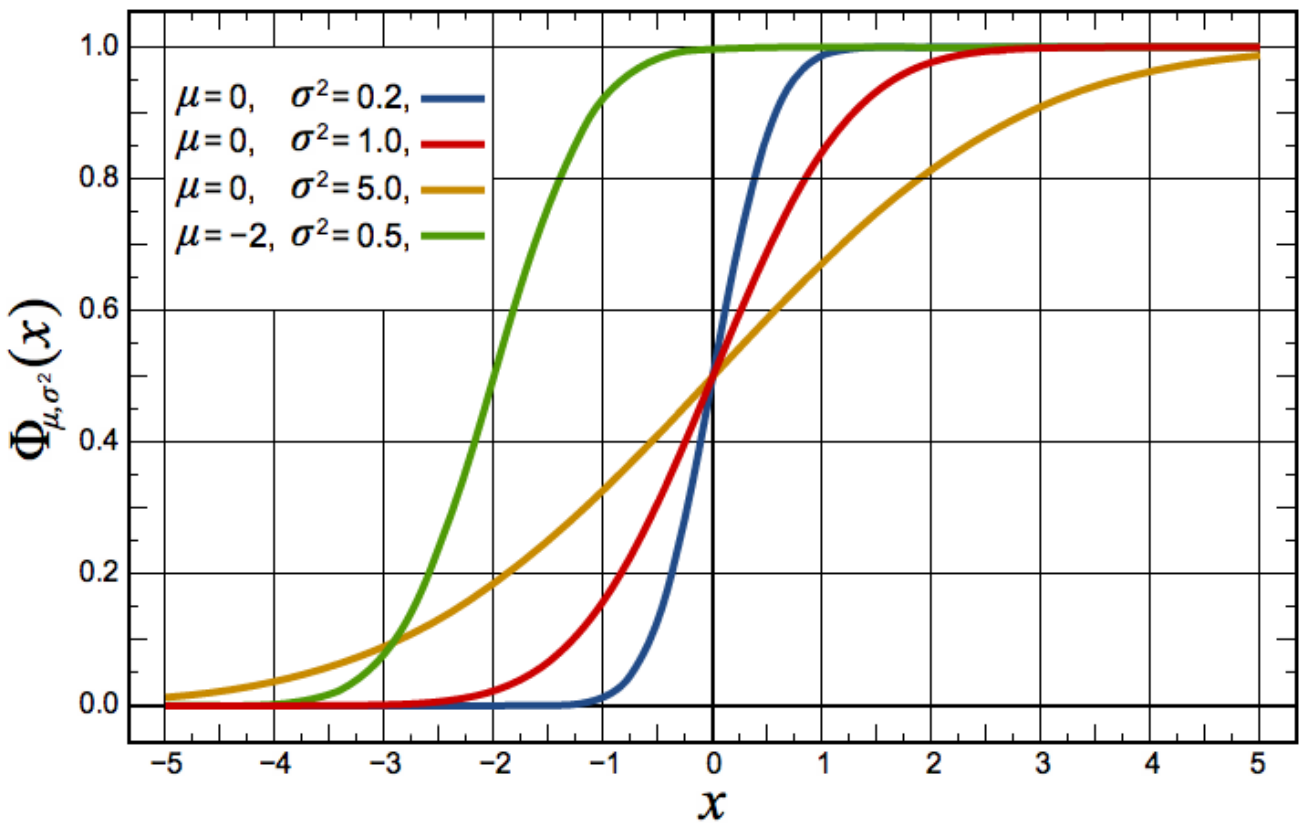
$$P[a \leq X \leq b] = \int_a^b f(x)dx$$



Cumulative Distributions

The CDF of a random variable X is:

$$F(x) = P(X \leq x)$$



Properties of Distribution Functions

- CDF (cumulative distribution function):
  - $0 \leq F_X(x) \leq 1$
  - F_X monotone increasing, with $lim_{x \to -\infty} F_X(x) = 0$, $lim_{x \to \infty} F_X(x) = 1$
- pmf:
  - $0 \leq p_X(x) \leq 1$
  - $\sum_x p_X(x) = 1$
  - $\sum_{x \in A} p_X(x) = p_X(A)$
- pdf:
  - $f_X(x) \geq 0$
  - $\int_{-\infty}^{\infty} f_X(x)dx = 1$
  - $\int_{x \in A} f_X(x)dx = P(X \in A)$

Some Common Random Variables

- $X \sim Bernoulli(p)$ $(0 \leq p \leq 1)$: $p_X(x) = \begin{cases} p & \text{x=1,} \\ 1-p & \text{x=0.} \end{cases}$
- $X \sim Geometric(p)$ $(0 \leq p \leq 1)$: $p_X(x) = p(1-p)^{x-1}$
- $X \sim Uniform(a, b)$ $(a < b)$: $f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases}$
- $X \sim Normal(\mu, \sigma^2)$: $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$

Expectation and Variance

- If the discrete random variable X has pmf $p(x)$, then the expectation is $E[X] = \sum_x x \cdot p(x)$
- Continuous case is similar: $E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x) dx$
- Expectation is linear:
  - for any constant $a \in \mathbb{R}$, $E[a] = a$
  - $E[a \cdot g(X) + b \cdot h(X)] = aE[g(X)] + bE[h(X)]$
- $Var[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$
- Variance is **not** linear

Example: expectation of rolling a die once:
$1 \cdot 1/6 + 2 \cdot 1/6 + 3 \cdot 1/6 + 4 \cdot 1/6 + 5 \cdot 1/6 + 6 \cdot 1/6 = 3.5$

Indicator variables

An indicator variable just indicates whether an event occurs or not:

$$I_A = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

They have a very useful property:

$$\begin{aligned} E[I_A] &= 1 \cdot P(I_A = 1) + 0 \cdot P(I_A = 0) \\ &= P(I_A = 1) \\ &= P(A) \end{aligned}$$

Method of indicators

Goal: find expected number of successes out of N trials

Method: define an indicator (Bernoulli) random variable for each trial, find expected value of the sum

Example: N professors are at dinner and take a random coat when they leave. Expected number of profs with the right coat?

Let G be the number of profs who get the right coat, and let G_i be an indicator for the event that professor i gets his own coat. Then

$$G = G_1 + G_2 + \ldots + G_n$$

These events are not independent!

But linearity of expectation saves us:

$$\begin{aligned} E[G] &= E[G_1 + G_2 + \ldots + G_n] \\ &= E[G_1] + E[G_2] + \ldots + E[G_n] \\ &= 1/n + 1/n + \ldots 1/n = 1 \end{aligned}$$

Remember: linearity of expectation does **not** assume independence!

Some Useful Inequalities

- Markov's Inequality: X random variable, and $a > 0$. Then:

$$P(X \geq a) \leq \frac{E[X]}{a}$$

Example: back to the professors and their coats. We know that E[G] = 1, so applying Markov's Inequality gives us:

$$P(G \geq a) \leq \frac{1}{a}$$

Plugging in $a = 5$, we get that the chance that at least 5 professors get the right coats is no higher than 20% (regardless of N).

- Chernoff bound: Let $X_1, \ldots, X_n$ independent Bernoulli with $P(X_i = 1) = p_i$. Denoting $\mu = E[\sum_{i=1}^n X_i] = \sum_{i=1}^n p_i$,

$$P(\sum_{i=1}^{n} X_i \geq (1+\delta)\mu) \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}} \right)^{\mu}$$

  for any δ. Multiple variants of Chernoff-type bounds exist, which can be useful in different settings

Parameter Estimation: Maximum Likelihood

- Say we have a parametrized distribution $f_X(x;\theta)$ and we don't know the parameter(s) θ.
- IID samples $x_1, \ldots, x_n$ observed.
- Goal: Estimate θ
- The maximum likelihood estimator (MLE) is the value $\hat{\theta}$ that maximizes the likelihood of observing the data you observed.

MLE Example

Say you flip a coin with unknown bias p of landing heads n times and get n_H heads and n_T tails. What's the MLE estimate for the coin's bias?

The likelihood of observing the data given a particular θ is $P(D|\theta) = \theta^{n_H}(1-\theta)^{n_T}$.

Take logs: $\log P(D|\theta) = n_H \log(\theta) + n_T \log(1-\theta)$.

MLE Example continued

Take the derivative and set to 0:

$$\begin{aligned}
\frac{d}{d\theta} \log P(D|\theta) &= 0 \\
\frac{d}{d\theta}[n_H \log(\theta) + n_T \log(1-\theta)] &= 0 \\
\frac{n_H}{\theta} - \frac{n_T}{1-\theta} &= 0 \\
\hat{\theta} &= \frac{n_H}{n_H + n_T}
\end{aligned}$$

Sometimes it is not possible to find the optimal estimate in closed form, in which case iterative methods must be used.

Interesting limits

- $\lim_{n\to\infty}(1+\frac{k}{n})^n \to e^k$
- $\lim_{n\to\infty} n! \to \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$ (lower bound)