

Basic Concepts

Experiment

Sample Space: Venn Diagram

Basic Outcome

Event

Probability Function

Probability Distribution

Disjoint events

Conditional Probability

Independence

Bayes' Rule

Random Variable

Mean and variance

Bayesian Inference

Experiment

We observe **outcome** of some situation, assuming that it is always clear which of a range of things actually happened.

e.g. pocketful of change; pick one coin.

- Is it bi-metal? Binary.
- How old is it? 4 possibilities.
- How heavy is it? Infinite possibilities.

Different outcomes have different **probabilities**.

In some simple situations, there are a finite number of **basic outcomes**, which are often **equi-probable**.

Games of chance are often constructed like this; toss a coin, pick a card, throw a die, et.c.

Now define an **event** as a combination of basic outcomes, e.g. “throw an even number with a die.” (2 or 4 or 6).

Very easy to discuss probabilities in such cases.

In reality, usually cannot easily deduce probabilities from idealised model of world, so need to be able to estimate and calculate probabilities.

Probability

Sample space is a **set** Ω . It contains all possible outcomes of an experiment. The **probability** $P(A)$ of an event A is a number between 0 and 1 that tells us how likely event A is to occur.

What does that mean?

- $P(A) = 0$ if A never occurs. (**warning: simplification**)
- $P(A) = 1$ if A always occurs. (**warning: simplification**)
- $P(A) = 0.5$ if A happens half the time.

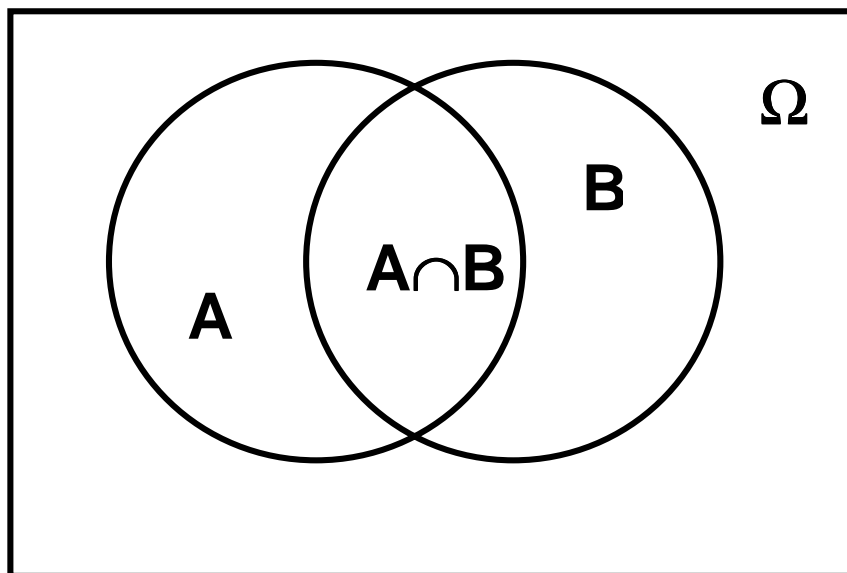
(Probability theory does **NOT** supply the numbers!)

A ***probability function*** is a mathematical formula (or table) which allocates a single non-negative number to each basic outcome.

In any experiment, the probabilities of each of the basic outcomes (all disjoint) must add up to 1; “nothing can’t happen!”

There are various more complicated types of probability, but the basic rule applies; the sum of all the (disjoint) probabilities must be 1.

Consider 2 events A and B:



A and B *could* be identical, ($A=B$) in which case probability of A is probability of B:

$$(A=B) \Rightarrow P(A) = P(B).$$

A could be a **subset** of B ($A \subset B$) in which case:

$$(A \subset B) \Rightarrow P(A) \leq P(B).$$

In general, A and B **may** overlap to an extent, so that the new event $A \cap B$ may have non-zero probability:

$$P(A \cap B) \geq 0.$$

If $P(A \cap B) = 0$, we say A and B are **disjoint**.

Conditional Probability

In many cases we can make initial (“a priori”) guess as to probability of an event e.g. the occurrence of “the” in a text.

If we then find that some other event occurs, we may be able to improve our guess of probability. e.g.:

text so far	Probability of “the”
<i>“..and that is how I caught...”</i>	Moderate
<i>“.. the book is slightly...”</i>	Low

Conditional probability of A ***conditioned*** on B is **$P(A|B)$** .

We **define** $P(A|B)$ as:

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

Or:

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

What if there is no connection between A and B?

If A and B are ***independent***:

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

NOTE: **independent** does **NOT** mean **disjoint**.

Bayes' Rule/Theorem

In many situations, we can actually estimate $P(B|A)$ but cannot directly estimate $P(A|B)$, which we really want.

Using the conditional probability definition, we can work around the problem, noting that $(A \cap B)$ is the same as $(B \cap A)$:

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

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

As $P(A \cap B) = P(B \cap A)$, we can write:

$$P(A|B) \times P(B) = P(B|A) \times P(A).$$

So (Bayes' theorem):

$$\boxed{P(A|B) = P(B|A) \times P(A) / P(B).}$$

This *looks* trivial, but is enormously important.

Chain Rule

Suppose several events A_1, A_2, \dots, A_n . What is probability $P(A_1 \cap A_2 \cap \dots \cap A_n)$ of **all** of the events happening? E.g. for $n=4$:

$$P(A_1 \cap A_2 \cap A_3 \cap A_4) =$$

$$P(A_1) \times P(A_2|A_1) \times P(A_3|A_1 \cap A_2) \times P(A_4|A_1 \cap A_2 \cap A_3)$$

(from Rosario's slide set.)

S:stiff neck, M: meningitis

$P(S|M) = 0.5$, $P(M) = 1/50,000$, $P(S) = 1/20$

I have stiff neck; should I worry?

$$\begin{aligned} P(M | S) &= \frac{P(S | M)P(M)}{P(S)} \\ &= \frac{0.5 \times 1/50,000}{1/20} = 0.0002 \end{aligned}$$

Random Variable

There are strict rules about the values that the probability function can take.

We may also wish to associate other numbers with elements of the set of basic outcomes. (e.g. in poker we associate different stakes with different combinations of cards.)

A ***random variable*** is a mathematical function that uniquely associates a real number with each basic outcome, and hence with each event.

A ***stochastic process*** is a process that does experiments and emits a random number for each outcome.

Probability Mass Function (pmf)

For a random variable X , we define $p(x)$ as $P(X=x)$, often written as:

$$X \sim p(x).$$

The individual probabilities must all add up to 1:

$$\sum_x p(x) = 1$$

Expectation

The **expectation** $E(X)$ of a random variable X is essentially the expected or **mean** value of X as observed over very many trials.

$$E(X) = \sum_x x \cdot p(x)$$

e.g. expected value of die throw:

$$\mu_x = E(Y) = \sum_x y \cdot p(y)$$

$$= \sum_{y=1}^6 y \cdot \frac{1}{6}$$

$$= 21/6 = 3 \frac{1}{2}$$

So we can work out mean or average values of a random variable. Can we characterise the spread of values it can take on?

Variance σ_X^2 of random variable X is:

$$\sigma_X^2 = \text{Var}(X) = E[X - E(X)]^2 = E(X^2) - E^2(X).$$

Standard Probability Distributions

We defined a probability mass function so:

For a random variable X ,

$$p(x) \equiv P(X=x),$$

x being a particular value of X .

There are several standard forms of $p(x)$.

Binomial Distribution

“Bernoulli trial”; e.g. tossing a coin many times. Let p (between 0 and 1) be probability of heads, so probability of tails is $(1 - p)$. **Assume** that each result is **independent** of any other (true here.)

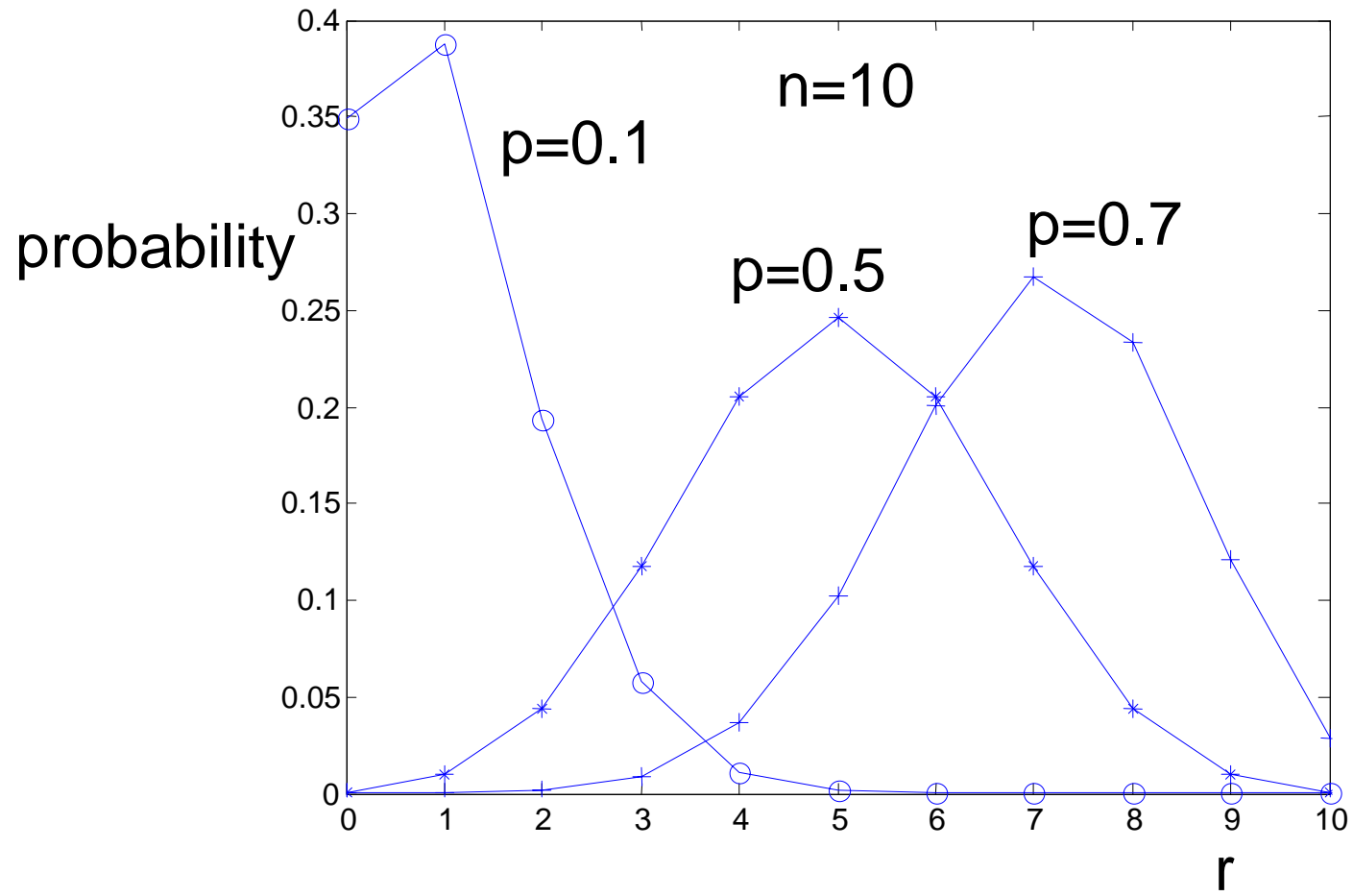
If we do N tosses, probability of exactly r heads is:

$$b(r; n, p) = \binom{n}{r} p^r (1 - p)^{n-r}$$

where :

$$\binom{n}{r} = \frac{n!}{(n-r)! r!}$$

Probability for Statistical NLP



Several linguistic applications where a **binary** decision is made: e.g. looking through text; does each individual sentence contain “the” or not? However remember that **it is assumed** that each “trial” is **independent**.

For binomial distribution there are 2 **parameters**, n and p . The probability of r successes depends only on r , n and p .

All other standard distributions can be characterised by a small number of parameters.

Questions to ask about data:

- What model best describes variability of data?
- What are optimum parameters of (model x) to explain data?
- How sure am I of the values of the parameters?
- Given a model (and parameters) how likely is this datum to have been produced by the model?
- If I get more data, how much more sure can I be about the model parameters?