ANLP

02 - Review Probability Theory

David Schlangen
University of Potsdam, MSc Cognitive Systems
Winter 2019 / 2020

some more remarks on course

- our expectations:
- implement algorithms, use / build practical applications, learn how to analyse language, build programming skills

some more remarks on course

- obstacles in the way of success:
  - workload too high, too many other courses, not enough time to review slides (& do background reading)
  - missing background knowledge (& not enough effort to get it):
    - linguistics
    - theoretical CS
    - maths
- commuting takes too much time / badly planned
- English skills
- programming experience

today: review of basic probability theory

- random experiment, sample space, events, event space
- probability [mass, function]
- conditional probability, statistical independence
- Bayes' theorem
- random variable
- maximum likelihood estimation

probability theory

It all started with games of chance…

Pierre de Fermat

November 1654: Antoine Gombaud asks Fermat for help: "Coin toss game. 5 rounds, best wins. Game is 2:1 for Robert, when the police busts us & ends the game. How shall we split the pool?"

In a long exchange of letters, Fermat and Pascal lay the foundations of probability theory..

Blaise Pascal

probability theory

It all started with games of chance…

Pierre de Fermat

R R C
C R R
R C R

R < R C
C < R C

R: 3/4, C: 1/4

Blaise Pascal

probability theory

- 1657. Christian Huyghens writes a 16-page paper that lays out pretty well all of modern probability theory, including the notion of expectation, which he introduces.
- 1662. John Graunt, an English haberdasher, publishes an analysis of the London mortality tables, and in so doing establishes the beginnings of modern statistical inference.
- 1669. Huyghens uses his new probability theory to re-compute Graunt's mortality tables with greater precision.
- 1709. Nikolas Bernoulli writes a book describing applications of the new methods in the law. One problem he shows how to solve is how long must elapse after an individual goes missing before the court can declare him dead and allow his estate to be divided among his heirs.
- 1713. Jakob Bernoulli writes a book showing how the new probability theory can be used to predict the future in the everyday world. This is the first time the word "probability" is used in the precise, mathematical sense we use it today. He also proves the law of large numbers.
- 1732. The first American insurance company begins in Charleston, S.C., restricted to fire insurance.
- 1732. Edward Lloyd starts the precursor of what in 1734 becomes Lloyd's List, and eventually gives birth to the insurance company Lloyds of London.
- 1733. Abraham de Moivre discovers the bell curve, the icon of modern data collection.
- 1738. Daniel Bernoulli introduces the concept of utility to try to get a better handle on human decision making under uncertainty.
- 1760s. The first life insurance companies begin.

https://web.archive.org/web/20100228052123/http://www.maa.org/devlin/devlin_11_08.html

random experiment

- or trial
- repeatable procedure with well-defined possible **outcomes**
- *outcome*: the result of a single experiment run

sample space

- use set theory to make this notion precise
- one run of the experiment = one basic outcome
- the set $\Omega = \{e_1, \ldots, e_n\}$ of the n basic outcomes $e_1, \ldots, e_n$ of the experiment is called the *sample space* of the experiment
- can allow finite and infinite sample spaces (here: mostly finite); discrete and continuous (here: discrete)
- e.g.: throwing a die, possible outcomes $\Omega = \{1,2,3,4,5,6\}$

event

- and *event* is a subset of Ω
- e.g., experiment of throwing two coins, Ω = {HH, HT, TH, TT}
  - event “at least once H” =
  - {HH, HT, TH}

events

- so events are *sets*.

- we can do the usual things with them. Assuming $A,B \subseteq \Omega$ (so A and B are events, relative this sample space), we also have the events:

  - $A \cup B$

  - $A \cap B$

  - $\bar{A}$

events

- some special events:
  - the *certain event* Ω
  - the *impossible event* $\varnothing$

events

- how many events are there?
- events are subsets of Ω
- how many subsets are there?
  - the set of all subsets is the powerset, 2^{Ω} so $|2^{\Omega}|$
- (formally, we only want that the set of events $\mathscr{F}$ forms a σ-algebra, a lattice closed against the Boolean set operations)

probability function

- the (discrete) probability function assigns each event a value in [0,1]

- P: $\mathscr{F} \rightarrow [0,1]$, s.t.
    - $P(A) \geq 0$, for any A in $\mathscr{F}$
    - $P(\Omega) = 1$
    - for *disjoint sets* $A_j \in \mathscr{F}$: $P(\cup_{j=1}^{\infty}) = \sum P(A_j)$

- these are the Kolmogorov axioms

consequences of the axioms

- $P(\varnothing) = 0$
- if A ⊆ B then P(A) ≤ P(B)
- 0 ≤ P(A) ≤ 1, for all A ∈ $\mathscr{F}$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  (the sum rule)
- (try to prove that from the axioms as an exercise)

probability space

- is a triple (Ω, $\mathscr{F}$, P)

- but which concrete P??

determining P

- a fair coin is tossed 3 times
- what is the sample space?
  - {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}
- what probability would we want to assign to the elementary outcomes?
- 1/8
- *uniform distribution*
- what is P(A)? (where A is the event “(exactly) two heads”)
  - A = {HHT, HTH, THH}
  - P(A) = 3/8

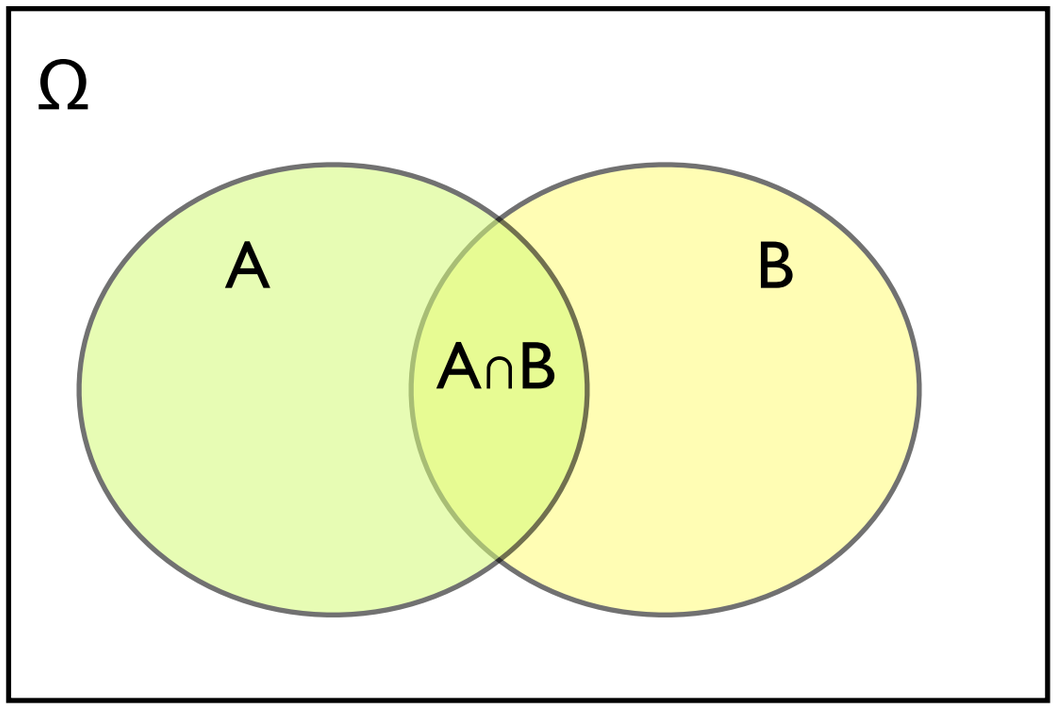
taking stock

- we gave a logical interpretation to the set ops.:
  - ∪ as "or": P(A∪B) is probability that A *or* B will happen
    - ∩ as "and": P(A∩B) is prob that A *and B will happen*
- additivity, generally: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- additivity, disjunct A and B: $P(A \cup B) = P(A) + P(B)$

conditional probability

- a fair coin is tossed 3 times.
- what's the chance that we get exactly 2 heads? (A = {HHT, HTH, THH})
- what's the chance that we get exactly 2 heads (A), when the first coin came up heads (B)?
  - {HHH, HHT, HTH, HTT, ~~THH, THT, TTH, TTT~~}
  - $P(A \mid B) = 1/2$

conditional probability



P(A|B) : Probability of A, given that we know that B has happened.
Resulting event is in the intersection, and since we know that B has happened, it's is proportion of B's probability mass that interests us. (If it if != 0)
P(A|B) = P(A∩B) / P(B).
[it follows that: P(A∩B) = P(A|B) * P(B) *multiplication rule*]

the chain rule

- the generalisation of the multiplication rule, the *chain rule,* is super important in statistical NLP:

$$P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n| \cap_{i=1}^{n-1} A_i)$$

conditional probability

- some special events:
  - the *certain event* Ω
  - the *impossible event* $\varnothing$

independence

- two events are independent of each other, if $P(A \cap B) = P(A)\,P(B)$
- equivalently (if $P(B) \neq 0$): $P(A \mid B) = P(A)$
- (claiming independence is a way of getting rid of a conditioning)

Bayes' theorem

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

If $A \subseteq \cup_{i=1}^{n} B_i$, $P(A) > 0$, and $B_i \cap B_j = \emptyset$ for $i \neq j$ then

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A)} = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

naming convention

likelihood
(how likely observation is, given that A)

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

prior probability
(what we knew about A)

posterior probability
(what we've learned about A,
given B)

likelihood and prior can often be estimated from data!

Example of Bayes' Rule

- 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?

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)}$$

$$= \frac{0.5 \times 1/50{,}000}{1/20} = 0.0002$$

random variables

- a *random variable* is a function X: $\Omega \rightarrow \mathbb{R}^n$
- lets us talk about events using numbers
- lets us abstract away from the concrete probability space
- we can then define a *probability mass function* (for discrete spaces):
  $p(x) = P(X = x) = P(A_x)$, $A_x = \{\omega \in \Omega : X(\omega) = x\}$
- we will often abuse this notation and use non-numerical x, e.g. words (P(X=‘the’))

random variables

- example:
  - experiment is 2 throws of die
  - random variable X is sum of values of dice
  - so 11:2, 12:3, 13:4, 14:5, 15:6, …

joint and conditional distributions

- $p(x,y) = P(X=x, Y=y)$ … “,” means “and”

- marginalisation: $p_X(x) = \sum_y p(x,y)$

- $p_{X|Y}(x|y) = p(x,y) / p_y(y)$

expectation

- the *expectation* is the mean or average of a random variable:
  $E(X) = \sum_x x * p(x)$

- e.g., experiment is rolling one die, Y is the value on its face, what is the expectation of Y?

  - $E(Y) = 1/6 * 1 + 1/6 * 2 + \ldots. 1/6 * 6 = 3.5$

Expected values / Expectation

- Frequentist interpretation of probability: if P(X = a) = p, and we repeat the experiment N times, then we see outcome "a" roughly p N times.
- Now imagine each outcome "a" comes with reward R(a). After N rounds of playing the game, what reward can we (roughly) expect?
- Measured by *expected value*:

$$E_P[R] = \sum_{a \in A} P(X = a) \cdot R(a)$$

Language Model (predicting next word)

- In general, for language events, P is unknown
- We need to *estimate* P, (or model M of the language)
- We'll do this by looking at evidence about what P must be based on a sample of data (*observations*)

Example: model estimation

- Example: we flip a coin 100 times and observe H 61 times. Should we believe that it is a fair coin?
- observation: 61x H, 39x T
- model: assume rv X follows a *Bernoulli* distribution, i.e. X has two outcomes, and there is a value p such that P(X = H) = p and P(X = T) = 1 - p.
- want to estimate the *parameter* p of this model

Estimation of P

- Frequentist statistics
  - parametric methods
  - non-parametric (distribution-free)
- Bayesian statistics

Frequentist Statistics

- Relative frequency: proportion of times an outcome *u* occurs

  $f_U = C(u) / N$

- C(u) is the number of times u occurs in N trials
- For N approaching infinity, the relative frequency tends to stabilize around some number: probability estimates

Non-Parametric Methods

- No assumption about the underlying distribution of the data
- For ex, simply estimate P empirically by counting a large number of random events is a distribution-free method
- Less prior information, more training data needed

Parametric Methods

- Assume that some phenomenon in language is acceptably modeled by one of the well-known families of distributions (such as binomial, normal)
- We have an explicit probabilistic model of the process by which the data was generated, and determining a particular probability distribution within the family requires only the specification of a few parameters (less training data)

Binomial Distribution

- Series of trials with only two outcomes, each trial being independent from all the others
- Number *r* of successes out of *n* trials given that the probability of success in any trial is *p*:

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

Bin. Distribution – Examples

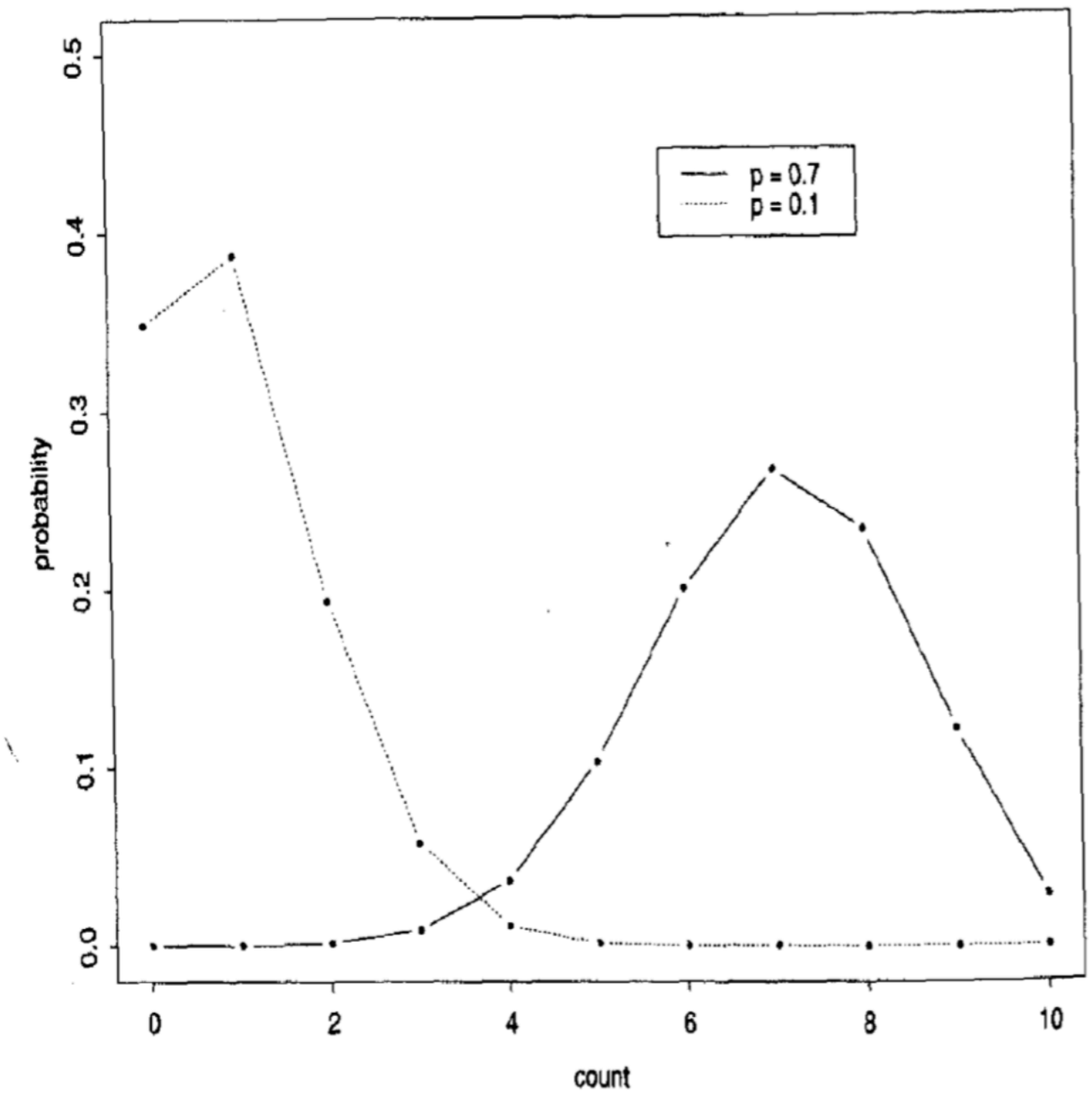


Figure 2.3 Two examples of binomial distributions: b(r; 10, 0.7) and b(r; 10, 0.1).

M/S, Fig. 2.3

Normal (Gaussian) Distribution

- Continuous
- Two parameters: mean μ and standard deviation σ

$$n(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Used in clustering

Maximum Likelihood Estimation

- We want to estimate the parameters of our model from frequency observations. There are many ways to do this. For now, we focus on *maximum likelihood estimation*, MLE.

- *Likelihood* L(O ; p) is the probability of our model generating the observations O, given parameter values p.

- Goal: Find value for parameters that maximizes the likelihood.

ML Estimation

- For Bernoulli and multinomial models, it is extremely easy to estimate the parameters that maximize the likelihood:
  - P(X = a) = f(a)
  - in the coin example above, just take p = f(H)
- Why is this?

Bernoulli model

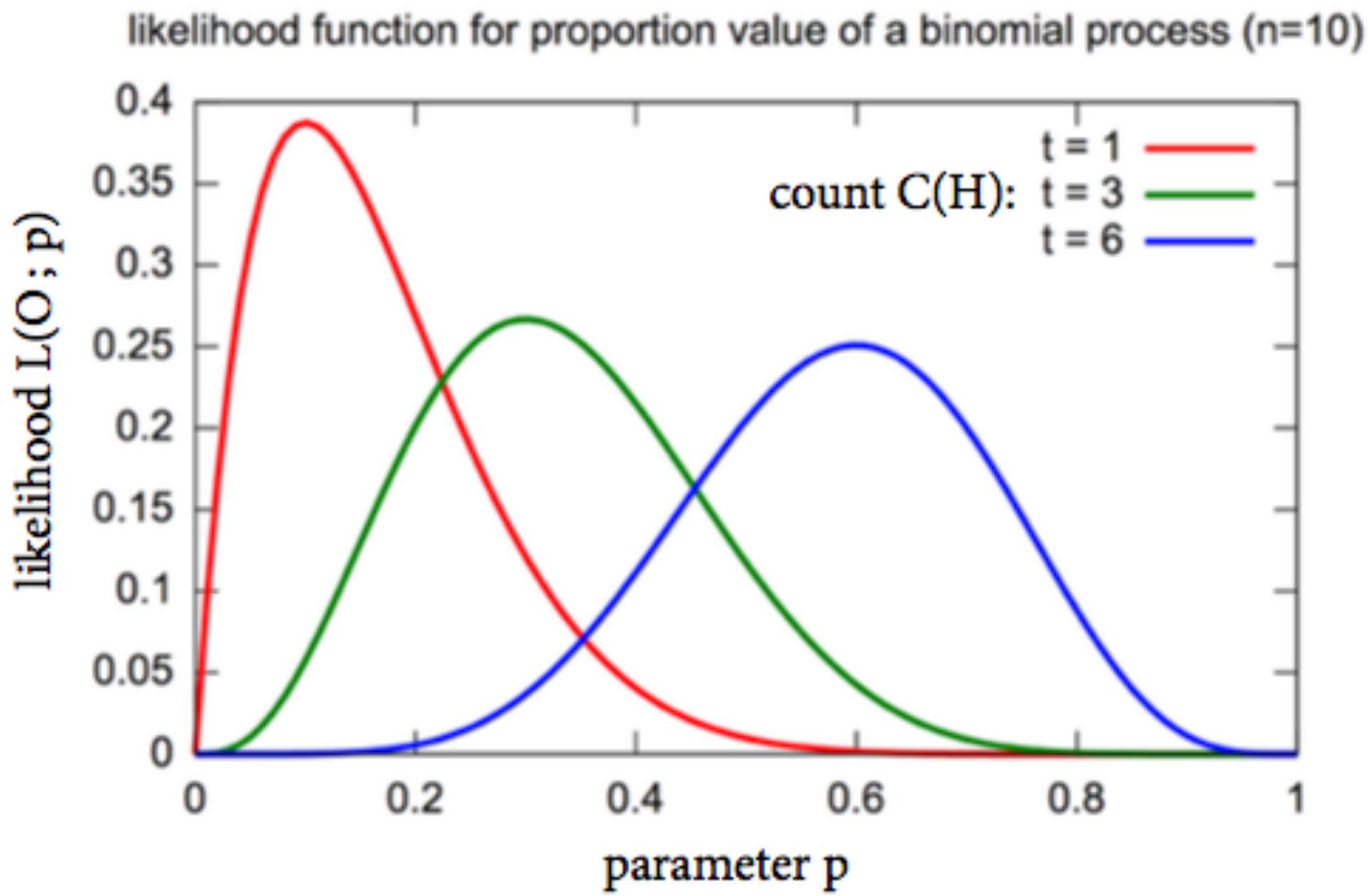
- Let's say we had training data C of size N, and we had N_H observations of H and N_T observations of T.

likelihood $L(C) = \prod_{i=1}^{N} P(w_i \mid p) = \prod_{i=1}^{N} p^{N_H}(1-p)^{N_T}$

log-likelihood

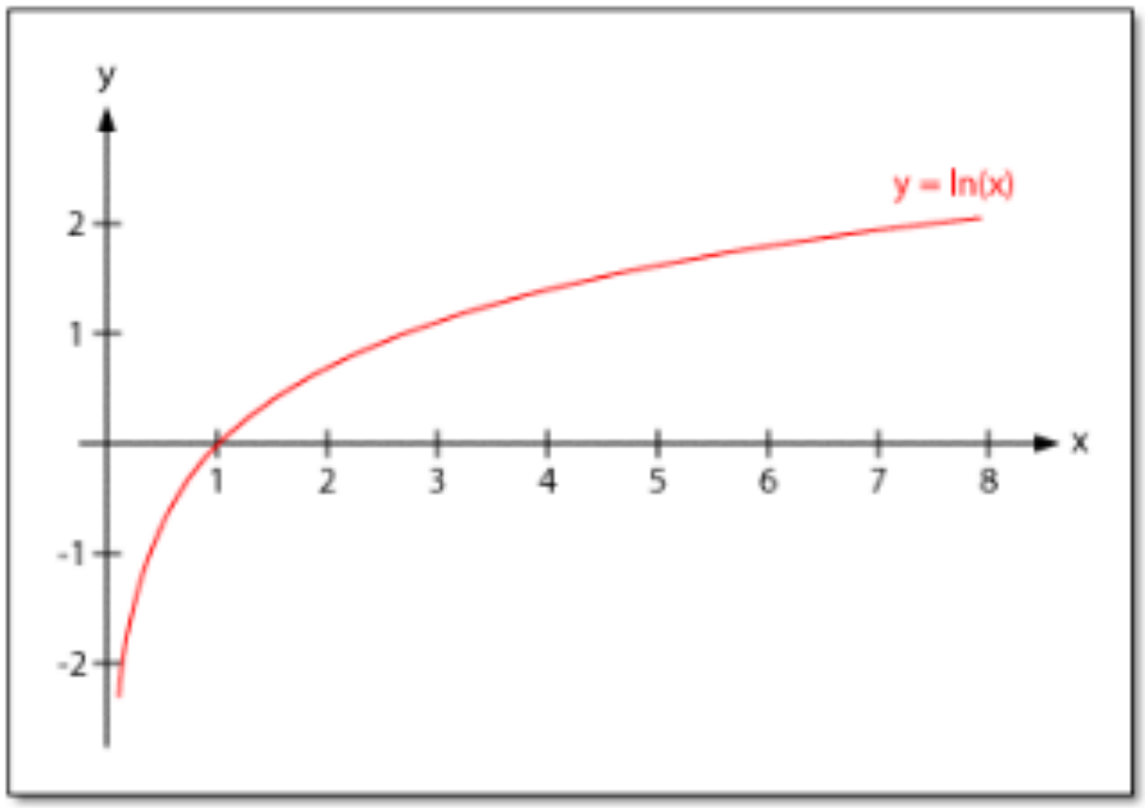
$$\ell(C) = \log L(C) = \sum_{i=1}^{N} \log P(w_i \mid p) = N_H \log p + N_T \log(1-p)$$

Likelihood functions



(Wikipedia page on MLE; licensed from Casp11 under CC BY-SA 3.0)

Logarithm is monotonic



- Observation: If $x_1 > x_2$, then $\ln(x_1) > \ln(x_2)$.
- Therefore, argmax L(C) = argmax l(C)

p p

Maximizing the log-likelihood

- Find maximum of function by setting derivative to zero:

$$\ell(C) = N_H \log p + N_T \log(1-p)$$

$$\frac{d\ell(C)}{dp} = \frac{N_H}{p} - \frac{N_T}{1-p}$$

- Solution is $p = N_H / N = f(H)$.

More complex models

- Many, many models we use in NLP are *multinomial* probability distributions. More than two outcomes possible; think dice rolling.
- MLE result generalizes to multinomial models:
  $P(X = a) = f(a)$.
- Maximizing log-likelihood uses technique called *Lagrange multipliers* to ensure parameters sum to 1.
- If you want to see the details, see Murphy paper on the website.

Conclusion

- Probability theory is essential tool in modern NLP.
- Important concepts today:
  - random variable, probability distribution
  - joint and conditional probs; Bayes' rule; independence
  - expected values
  - statistical models; parameters; likelihood; MLE
- We will use all of these concepts again and again in this course. If you have questions, ask me early.

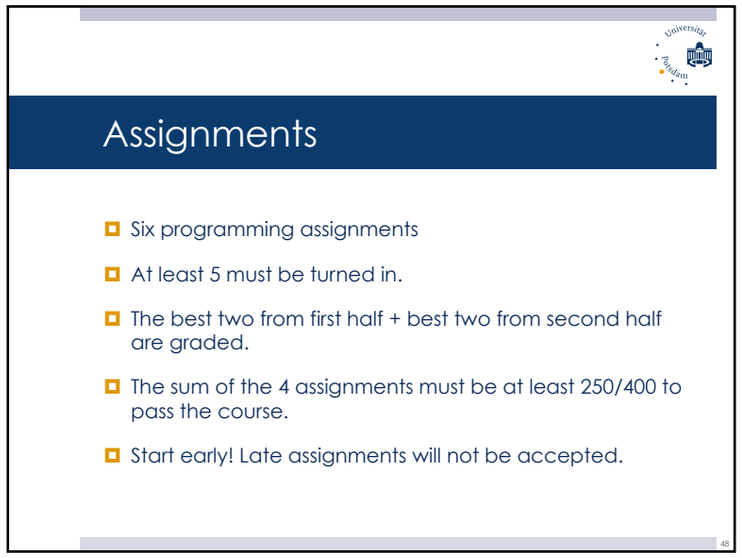
Questions, Queries, Comments?

background reading

- Sharon Goldwater's tutorial (do the exercises!): https://homepages.inf.ed.ac.uk/sgwater/teaching/general/probability.pdf

slide credits

slides that look like this



come from

earlier editions of this class (ANLP), given by Tatjana Scheffler and Alexander Koller

and their use is gratefully acknowledged. I try to make any modifications that I made obvious, but if there are errors on a slide, assume that I added them.