Basic Statistics and Probability Theory

Based on

"Foundations of Statistical NLP"

C. Manning & H. Schütze, ch. 2, MIT Press, 2002

"Probability theory is nothing but common sense reduced to calculation."

Pierre Simon, Marquis de Laplace (1749-1827)

PLAN

1. Elementary Probability Notions:

- Sample Space, Event Space, and Probability Function
- Conditional Probability
- Bayes' Theorem
- Independence of Probabilistic Events

2. Random Variables:

- Discrete Variables and Continuous Variables
- Mean, Variance and Standard Deviation
- Standard Distributions
- Joint, Marginal and and Conditional Distributions
- Independence of Random Variables

PLAN (cont'd)

3. Limit Theorems

- Laws of Large Numbers
- Central Limit Theorems

4. Estimating the parameters of probabilistic models from data

- Maximum Likelihood Estimation (MLE)
- Maximum A Posteriori (MAP) Estimation

5. Elementary Information Theory

- Entropy; Conditional Entropy; Joint Entropy
- Information Gain / Mutual Information
- Cross-Entropy
- Relative Entropy / Kullback-Leibler (KL) Divergence
- Properties: bounds, chain rules, (non-)symmetries, properties pertaining to independence

1. Elementary Probability Notions

- **sample space:** Ω **(either discrete or continuous)**
- **event:** $A \subseteq \Omega$
  - **the certain event:** Ω
  - **the impossible event:** $\emptyset$
  - **elementary event: any** $\{\omega\}$**, where** $\omega \in \Omega$
- **event space:** $\mathcal{F} = 2^{\Omega}$ **(or a subspace of** 2^{Ω} **that contains** $\emptyset$ **and is closed under complement and countable union)**
- **probability function/distribution:** $P : \mathcal{F} \to [0, 1]$ **such that:**
  - $P(\Omega) = 1$
  - **the "countable additivity" property:** $\forall A_1, ..., A_k$ **disjoint events,** $P(\cup A_i) = \sum P(A_i)$

Consequence: for a uniform distribution in a finite sample space:

$$P(A) = \frac{\#\textit{favorable elementary events}}{\#\textit{all elementary events}}$$

Conditional Probability

- $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$

  Note: $P(A \mid B)$ is called the a posteriory probability of A, given B.

- **The "multiplication" rule:**
  $P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$

- **The "chain" rule:**
  $P(A_1 \cap A_2 \cap \ldots \cap A_n) =$
  $P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1, A_2) \ldots P(A_n \mid A_1, A_2, \ldots, A_{n-1})$

- **The "total probability" formula:**

  $P(A) = P(A \mid B)P(B) + P(A \mid \neg B)P(\neg B)$

  More generally:

  if $A \subseteq \cup B_i$ **and** $\forall i \neq j \; B_i \cap B_j = \emptyset$**, then**

  $P(A) = \sum_i P(A \mid B_i)P(B_i)$

- **Bayes' Theorem:**

  $$P(B \mid A) = \frac{P(A \mid B)\; P(B)}{P(A)}$$

  or $$P(B \mid A) = \frac{P(A \mid B)\; P(B)}{P(A \mid B)P(B) + P(A \mid \neg B)P(\neg B)}$$

  or ...

Independence of Probabilistic Events

- **Independent events:** $P(A \cap B) = P(A)P(B)$

  Note: When $P(B) \neq 0$, the above definition is equivalent to $P(A|B) = P(A)$.

- **Conditionally independent events:**
  $P(A \cap B \mid C) = P(A \mid C)P(B \mid C)$**, assuming, of course, that** $P(C) \neq 0$**.**

  Note: When $P(B \cap C) \neq 0$, the above definition is equivalent to $P(A|B, C) = P(A|C)$.

2. Random Variables

2.1 Basic Definitions

Let Ω be a sample space, and $P : 2^{\Omega} \to [0, 1]$ a probability function.

- **A random variable of distribution P is a function**

$$X : \Omega \to \mathbb{R}^n$$

  - **For now, let us consider $n = 1$.**

- **The cumulative distribution function of X is $F : \mathbb{R} \to [0, \infty)$ defined by**

$$F(x) = P(X \leq x) = P(\{\omega \in \Omega \mid X(\omega) \leq x\})$$

2.2 Discrete Random Variables

Definition: Let $P : 2^{\Omega} \to [0,1]$ be a probability function, and X be a random variable of distribution P.

- If $\boldsymbol{Val}(X)$ is either finite or unfinite countable, then X is called a **discrete random variable.**
  - For such a variable we define the **probability mass function** (pmf) $p : \mathbb{R} \to [0,1]$ as $p(x) \overset{not.}{=} p(X = x) \overset{def.}{=} P(\{\omega \in \Omega \mid X(\omega) = x\})$.

    (Obviously, it follows that $\sum_{x_i \in Val(X)} p(x_i) = 1$.)

Mean, Variance, and Standard Deviation:

- **Expectation / mean** of X:

  $E(X) \overset{not.}{=} E[X] = \sum_x xp(x)$ if X is a discrete random variable.
- **Variance** of X: $\boldsymbol{Var}(X) \overset{not.}{=} \boldsymbol{Var}[X] = E((X - E(X))^2)$.
- **Standard deviation**: $\sigma = \sqrt{\boldsymbol{Var}(X)}$.

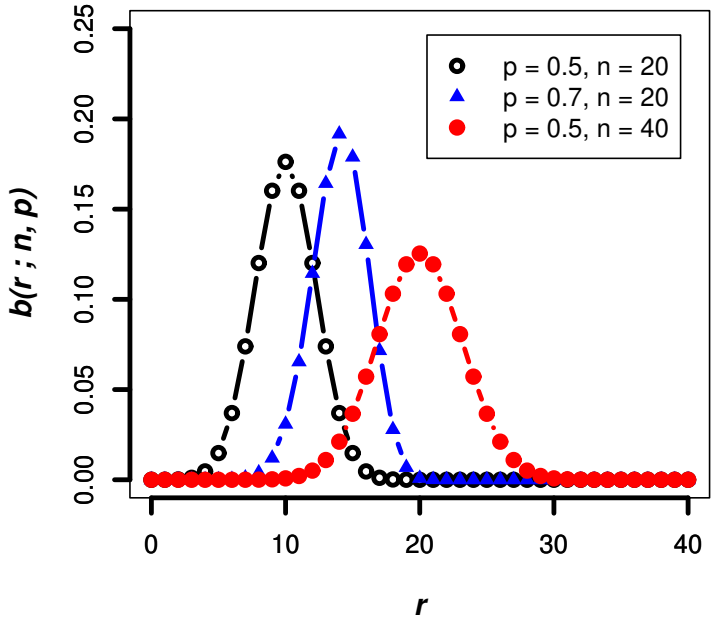
Covariance of X and Y, two random variables of distribution P:

- $\boldsymbol{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$

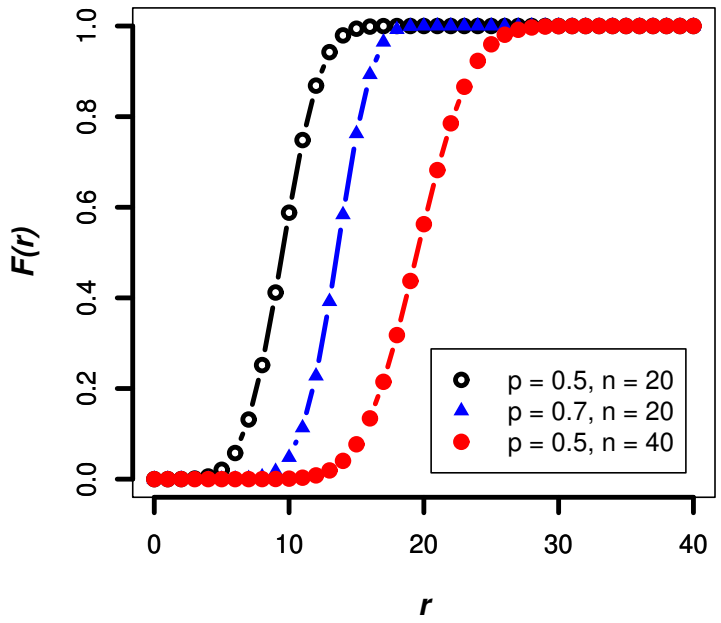
Exemplification:

- **the Bernoulli distribution:** $b(r;1,p)$
  mean: p, **variance:** $p(1-p)$, **entropy:** $-p\log_2 p-(1-p)\log_2(1-p)$
  - **the Binomial distribution:** $b(r;n,p)=C_n^r\,p^r(1-p)^{n-r}$ **for** $r=0,\ldots,n$
    mean: np, **variance:** $np(1-p)$

Binomial probability mass function



Binomial cumulative distribution function



2.3 Continuous Random Variables

Definitions:

Let $P : 2^{\Omega} \to [0,1]$ be a probability function, and $X : \Omega \to \mathbb{R}$ be a random variable of distribution P.

- **If *Val*(X) is unfinite non-countable set, and F, the cumulative distribution function of X is continuous, then X is called a continuous random variable.**
  (It follows, naturally, that $P(X = x) = 0$, for all $x \in \mathbb{R}$.)
- **If there exists $p : \mathbb{R} \to [0, \infty)$ such that $F(x) = \int_{-\infty}^{x} p(t)dt$, then X is called absolutely continuous.**
  In such a case, p is called the probability density function (pdf) of X.
  - **For $B \subseteq \mathbb{R}$ for which $\int_B p(x)dx$ exists, $P(X^{-1}(B)) = \int_B p(x)dx$, where $X^{-1}(B) \overset{not.}{=} \{\omega \in \Omega \mid X(\omega) \in B\}$.**
    In particular, $\int_{-\infty}^{+\infty} p(x)dx = 1$.
- **Expectation / mean of X: $E(X) \overset{not.}{=} E[X] = \int xp(x)dx$.**

Exemplification:

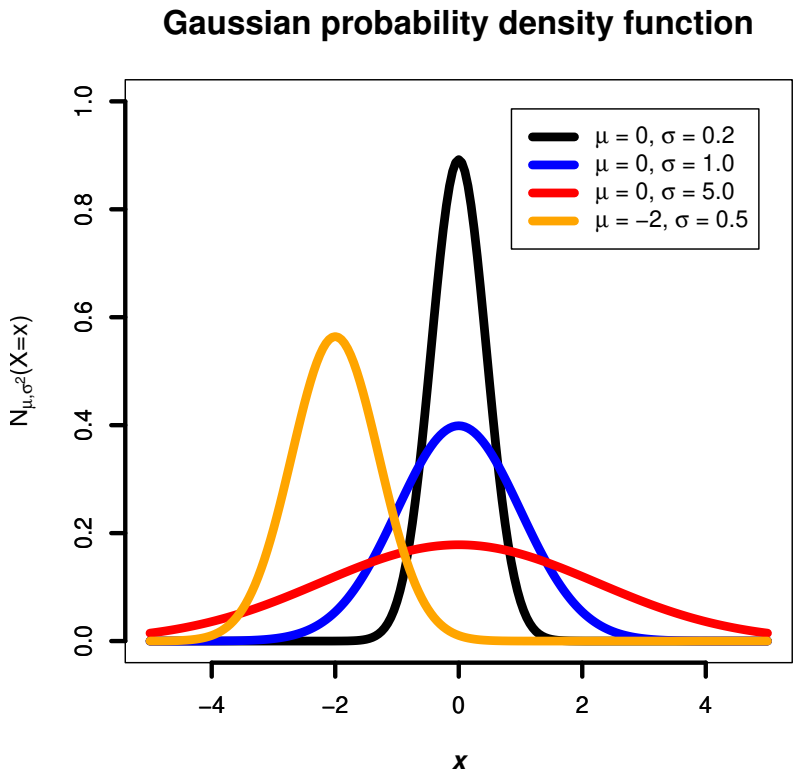
- **Normal (Gaussean) distribution:** $N(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

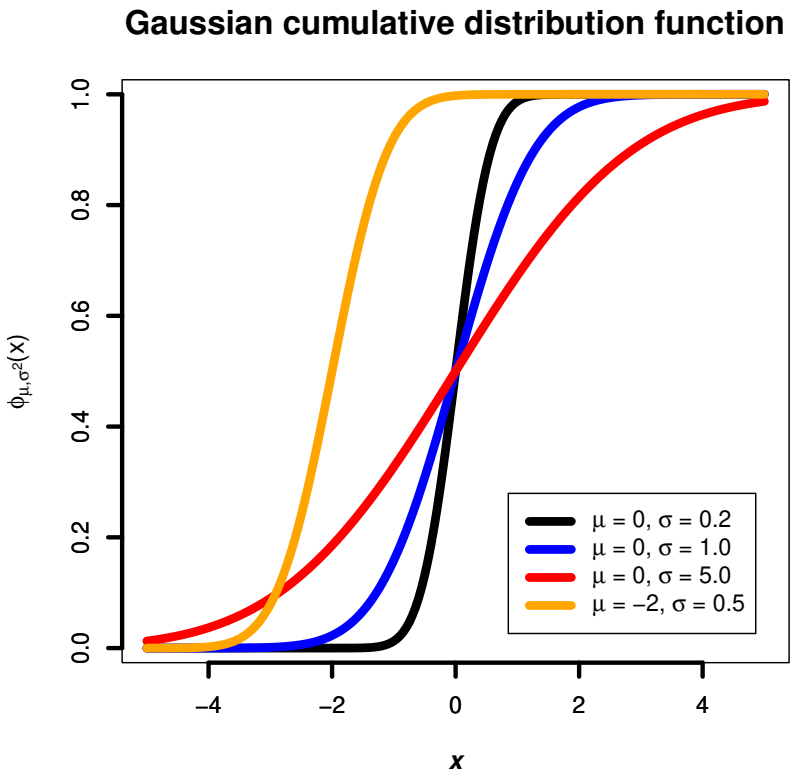
  mean: μ**, variance:** σ^2

  - **Standard Normal distribution:** $N(x;0,1)$

- **Remark:**

  For n, p **such that** $np(1-p) > 5$**, the Binomial distributions can be approximated by Normal distributions.**


Gaussian probability density function
$N_{\mu,\sigma^2}(X=x)$
0.0
0.2
0.4
0.6
0.8
1.0
−4
−2
0
2
4
x
μ = 0, σ = 0.2
μ = 0, σ = 1.0
μ = 0, σ = 5.0
μ = −2, σ = 0.5


Gaussian cumulative distribution function
$\phi_{\mu,\sigma^2}(x)$
0.0
0.2
0.4
0.6
0.8
1.0
−4
−2
0
2
4
x
μ = 0, σ = 0.2
μ = 0, σ = 1.0
μ = 0, σ = 5.0
μ = −2, σ = 0.5

2.4 Basic Properties of Random Variables

Let $P : 2^{\Omega} \to [0,1]$ be a probability function,
$X : \Omega \to \mathbb{R}^n$ be a random discrete/continuous variable of distribution P.

- **If $g : R^n \to R^m$ is a function, then $g(X)$ is a random variable.**

  If $g(X)$ is discrete, then $E(g(X)) = \sum_x g(x)p(x)$.

  If $g(X)$ is continuous, then $E(g(X)) = \int g(x)p(x)dx$.

  - **If g is non-linear $\not\Rightarrow E(g(X)) = g(E(X))$.**

- $E(aX) = aE(X)$.

- $E(X+Y) = E(X) + E(Y)$, **therefore** $E[\sum_{i=1}^n a_i X_i] = \sum_{i=1}^n a_i E[X_i]$.

- $\boldsymbol{Var}(aX) = a^2\,\boldsymbol{Var}(X)$.
- $\boldsymbol{Var}(X+a) = \boldsymbol{Var}(X)$.

- $\boldsymbol{Var}(X) = E(X^2) - E^2(X)$.

- $\boldsymbol{Cov}(X,Y) = E[XY] - E[X]E[Y]$.

2.5 Joint, Marginal and Conditional Distributions

Exemplification for the bi-variate case:

Let Ω be a sample space, $P : 2^{\Omega} \to [0, 1]$ a probability function, and $V : \Omega \to \mathbb{R}^2$ be a random variable of distribution P.
One could naturally see V as a pair of two random variables $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$. (More precisely, $V(\omega) = (x, y) = (X(\omega), Y(\omega))$.)

- **the joint pmf/pdf of X and Y is defined by**
  $p(x, y) \overset{not.}{=} p_{X,Y}(x, y) = P(X = x, Y = y) = P(\omega \in \Omega \mid X(\omega) = x, Y(\omega) = y)$.
- **the marginal pmf/pdf functions of X and Y are:**

  for the discrete case:
  $p_X(x) = \sum_y p(x, y), \quad p_Y(y) = \sum_x p(x, y)$
  for the continuous case:
  $p_X(x) = \int_y p(x, y)\, dy, \quad p_Y(y) = \int_x p(x, y)\, dx$
- **the conditional pmf/pdf of X given Y is:**
  $$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

2.6 Independence of Random Variables

Definitions:

- **Let X, Y be random variables of the same type (i.e. either discrete or continuous), and $p_{X,Y}$ their joint pmf/pdf.**

  X and Y are said to be independent if

  $$p_{X,Y}(x, y) = p_X(x) \cdot p_Y(y)$$

  for all possible values x and y of X and Y respectively.

- **Similarly, let X, Y and Z be random variables of the same type, and p their joint pmf/pdf.**

  X and Y are conditionally independent given Z if

  $$p_{X,Y|Z}(x, y \mid z) = p_{X|Z}(x \mid z) \cdot p_{Y|Z}(y \mid z)$$

  for all possible values x, y and z of X, Y and Z respectively.

Properties of random variables pertaining to independence

- **If X, Y are independent, then $\boldsymbol{Var}(X+Y) = \boldsymbol{Var}(X) + \boldsymbol{Var}(Y)$.**

- **If X, Y are independent, then $E(XY) = E(X)E(Y)$, i.e. $\boldsymbol{Cov}(X, Y) = 0$.**
  - **$\boldsymbol{Cov}(X, Y) = 0 \not\Rightarrow X, Y$ are independent.**

  - **The covariance matrix corresponding to a vector of random variables is symmetric and positive semi-definite.**
- **If the covariance matrix of a multi-variate Gaussian distribution is diagonal, then the marginal distributions are independent.**

3. Limit Theorems

[Sheldon Ross, A first course in probability, 5th ed., 1998]

"The most important results in probability theory are limit theorems. Of these, the most important are...

laws of large numbers, concerned with stating conditions under which the average of a sequence of random variables converges (in some sense) to the expected average;

central limit theorems, concerned with determining the conditions under which the sum of a large number of random variables has a probability distribution that is approximately normal."

Two Probability Bounds

Markov's inequality:

If X is a random variable that takes only non-negative values, then for any value $a > 0$,

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

Chebyshev's inequality:

If X is a random variable with finite mean μ and variance σ^2, then for any value $a > 0$,

$$P(|\, X - \mu \,| \geq a) \leq \frac{\sigma^2}{a^2}.$$

Note: As Chebyshev's inequality is valid for all distributions of the random variable X, we cannot expect the bound of the probability to be very close to the actual probability in most cases. (See ex. 2b, pag. 397 in Ross' book.)

The weak law of large numbers

[Bernoulli; Khintchine]

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables, each having a finite mean $E[X_i] = \mu$. Then, for any value $\epsilon > 0$,

$$P\left(\left|\frac{X_1 + \ldots + X_n}{n} - \mu\right| \geq \epsilon\right) \to 0 \text{ as } n \to \infty.$$

The central limit theorem for i.i.d. random variables

[Pierre Simon, Marquis de Laplace; Liapunoff in 1901-1902]

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables, each having finite mean μ and finite variance σ^2.
Then the distribution of

$$\frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}$$

tends to be the standard normal (Gaussian) as $n \to \infty$.

That is, for $-\infty < a < \infty$,

$$P\left(\frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \leq a\right) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx \text{ as } n \to \infty$$

The central limit theorem for independent random variables

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables having respective means μ_i and variances σ_i^2.

If

(a) the variables X_i are uniformly bounded,
i.e. for some $M \in \mathbb{R}^+$ $P(\mid X_i \mid < M) = 1$ for all i,

and

(b) $\sum_{i=1}^{\infty} \sigma_i^2 = \infty$,

then

$$P\left(\frac{\sum_{i=1}^{n}(X_i - \mu_i)}{\sqrt{\sum_{i=1}^{n}\sigma_i^2}} \leq a\right) \to \Phi(a) \text{ as } n \to \infty$$

where Φ is the cumulative distribution function for the standard normal (Gaussian) distribution.

The strong law of large numbers

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables, each having a finite mean $E[X_i] = \mu$. Then, with probability 1,

$$\frac{X_1 + \ldots + X_n}{n} \to \mu \text{ as } n \to \infty$$

That is,

$$P\left(\lim_{n\to\infty} (X_1 + \ldots + X_n)/n = \mu\right) = 1$$

Other Probability Bounds

One-sided Chebyshev inequality:

If X is a random variable with mean 0 and finite variance σ^2, then for any $a > 0$,

$$P(X \geq a) \leq \frac{\sigma^2}{\sigma^2 + a^2}$$

Corollary:

If $E[X] = \mu$, $\mathbf{Var}(X) = \sigma^2$, then for $a > 0$

$$P(X \geq \mu + a) \leq \frac{\sigma^2}{\sigma^2 + a^2}$$

$$P(X \leq \mu - a) \leq \frac{\sigma^2}{\sigma^2 + a^2}$$

Other Probability Bounds (cont'd)

Chernoff bounds:

If X is a random variable, then $M(t) \stackrel{not}{=} E[e^{tX}]$, is called the moment generating function of X.
It can be shown that

$$P(X \geq a) \leq e^{-ta} M(t) \quad \text{for all } t > 0$$
$$P(X \leq a) \leq e^{-ta} M(t) \quad \text{for all } t < 0.$$

Chernoff bounds for the standard normal distribution:

If Z is a standard normal random variable,
then $M(t) \stackrel{not.}{=} E[e^{tX}] \stackrel{calculus}{=} e^{t^2/2}$.
It can be shown that

$$P(Z \geq a) \leq e^{-a^2/2} \quad \text{for all } a > 0$$
$$P(Z \leq a) \leq e^{-a^2/2} \quad \text{for all } a < 0.$$

Other Probability Bounds (cont'd)

Hoeffding bounds:

Let $X_1, \ldots, X_n$ be some independent random variables, each X_i being bounded by the interval $[a_i, b_i]$.
If $\bar{X} \stackrel{not.}{=} \frac{1}{n}\sum_{i=1}^{n} X_i$, then it follows that
for any $t \geq 0$

$$P(\bar{X} - E[\bar{X}] \geq t) \leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

$$P(E[\bar{X}] - \bar{X} \geq t) \leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

$$\Rightarrow P(|\bar{X} - E[\bar{X}]| \geq t) \leq 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

4. Estimation/inference of the parameters of probabilistic models from data

(based on [Durbin et al, *Biological Sequence Analysis*, 1998], p. 311-313, 319-321)

A probabilistic model can be anything from a simple distribution to a complex stochastic grammar with many implicit probability distributions. Once the type of the model is chosen, the parameters have to be inferred from data.

We will first consider the case of the categorical distribution, and then we will present the different strategies that can be used in general.

A case study: Estimation of the parameters of a categorical distribution from data

Assume that the *observations* — for *example*, when rolling a die about which we don't know whether it is fair or not, or when counting the number of times the amino acid i occurs in a column of a multiple sequence alignment — can be expressed as counts n_i for each outcome i $(i = 1, l \ldots, K)$, and we want to estimate the probabilities θ_i of the underlying distribution.

Case 1:

When we have plenty of data, it is natural to use the maximum likelihood (ML) solution, i.e. the observed frequency $\theta_i^{ML} = \dfrac{n_i}{\sum_j n_j} \overset{not.}{=} \dfrac{n_i}{N}$.

Note: it is easy to show that indeed $P(n \mid \theta^{ML}) > P(n \mid \theta)$ for any $\theta \neq \theta^{ML}$.

$$\ln \frac{P(n \mid \theta^{ML})}{P(n \mid \theta)} = \ln \frac{\Pi_i (\theta_i^{ML})^{n_i}}{\Pi_i \theta_i^{n_i}} = \sum_i n_i \ln \frac{\theta_i^{ML}}{\theta_i} = N \sum_i \theta_i^{ML} \ln \frac{\theta_i^{ML}}{\theta_i} > 0$$

The inequality follows from the fact that the relative entropy is always positive except when the two distributions are identical.

Case 2:

When the data is scarce, it is not clear what is the best estimate.
In general, we should use prior knowledge, via Bayesian statistics.
For instance, one can use the Dirichlet distribution with parameters α.

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

It can be shown (see calculus on R. Durbin et. al. BSA book, pag. 320) that the posterior mean estimation (PME) of the parameters is

$$\theta_i^{\boldsymbol{PME}} \stackrel{def.}{=} \int \theta P(\theta \mid n) d\theta = \frac{n_i + \alpha_i}{N + \sum_j \alpha_j}$$

The $\alpha's$ are like pseudocounts added to the real counts. (If we think of the $\alpha's$ as extra observations added to the real ones, this is precisely the ML estimate!) This makes the Dirichlet regulariser very intuitive.

How to use the pseudocounts: If it is fairly obvious that a certain residue, let's say i, is very common, than we should give it a very high pseudocount α_i; if the residue j is generaly rare, we should give it a low pseudocount.

Strategies to be used in the general case

A. The Maximum Likelihood (ML) Estimate

When we wish to infer the parameters $\theta = (\theta_i)$ for a model M from a set of data D, the most obvious strategy is to maximise $P(D \mid \theta, M)$ over all possible values of θ. Formally:

$$\theta^{ML} = \underset{\theta}{\operatorname{argmax}}\, P(D \mid \theta, M)$$

Note: Generally speaking, when we treat $P(x \mid y)$ as a function of x (and y is fixed), we refer to it as a probability. When we treat $P(x \mid y)$ as a function of y (and x is fixed), we call it a likelihood. Note that a likelihood is not a probability distribution or density; it is simply a function of the variable y.

A serious drawback of maximum likelihood is that it gives poor results when data is scarce. The solution then is to introduce more prior knowledge, using Bayes' theorem. (In the Bayesian framework, the parameters are themselves seen as random variables!)

B. The Maximum A posteriori Probability (MAP) Estimate

$$\begin{aligned} \theta^{MAP} &\stackrel{def.}{=} \operatorname*{argmax}_{\theta} P(\theta \mid D, M) = \operatorname*{argmax}_{\theta} \frac{P(D \mid \theta, M)P(\theta \mid M)}{P(D \mid M)} \\ &= \operatorname*{argmax}_{\theta} P(D \mid \theta, M)P(\theta \mid M) \end{aligned}$$

The prior probability $P(\theta \mid M)$ has to be chosen in some reasonable manner, and this is the art of Bayesian estimation (although this freedom to choose a prior has made Bayesian statistics controversial at times...).

C. The Posterior Mean Estimator (PME)

$$\theta^{\boldsymbol{PME}} = \int \theta P(\theta \mid D, M) d\theta$$

where the integral is over all probability vectors, i.e. all those that sum to one.

D. Yet another solution is to use the posterior probability $P(\theta \mid D, M)$ to sample from it (see [Durbin et al, 1998], section 11.4) and thereby locate regions of high probability for the model parameters.

5. Elementary Information Theory

Definitions:

Let X and Y be discrete random variables.

- **Entropy:** $H(X) \stackrel{def.}{=} \sum_x p(x) \log_2 \frac{1}{p(x)} = -\sum_x p(x) \log_2 p(x) = E_p[-\log_2 p(X)]$.

  Convention: if $p(x) = 0$ then we shall consider $p(x) \log_2 p(x) = 0$.

- **Specific Conditional entropy:** $H(Y \mid X = x) \stackrel{def.}{=} -\sum_{y \in Y} p(y \mid x) \log_2 p(y \mid x)$.

- **Average conditional entropy:**
  $H(Y \mid X) \stackrel{def.}{=} \sum_{x \in X} p(x) H(Y \mid X = x) \stackrel{imed.}{=} -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(y \mid x)$.

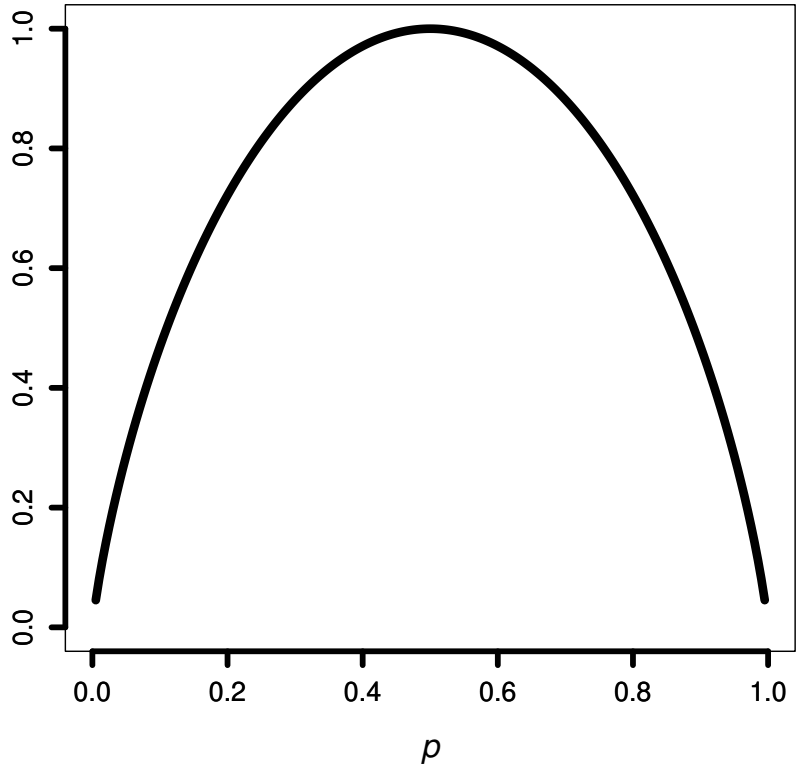
- **Joint entropy:**
  $H(X, Y) \stackrel{def.}{=} -\sum_{x,y} p(x, y) \log_2 p(x, y) \stackrel{dem.}{=} H(X) + H(Y \mid X) \stackrel{dem.}{=} H(Y) + H(X \mid Y)$.

- **Information gain (or: Mutual information):**

$$\begin{aligned} IG(X;Y) &\stackrel{def.}{=} H(X) - H(X \mid Y) \stackrel{imed.}{=} H(Y) - H(Y \mid X) \\ &\stackrel{imed.}{=} H(X, Y) - H(X \mid Y) - H(Y \mid X) = IG(Y;X). \end{aligned}$$

Exemplification: Entropy of a Bernoulli Distribution

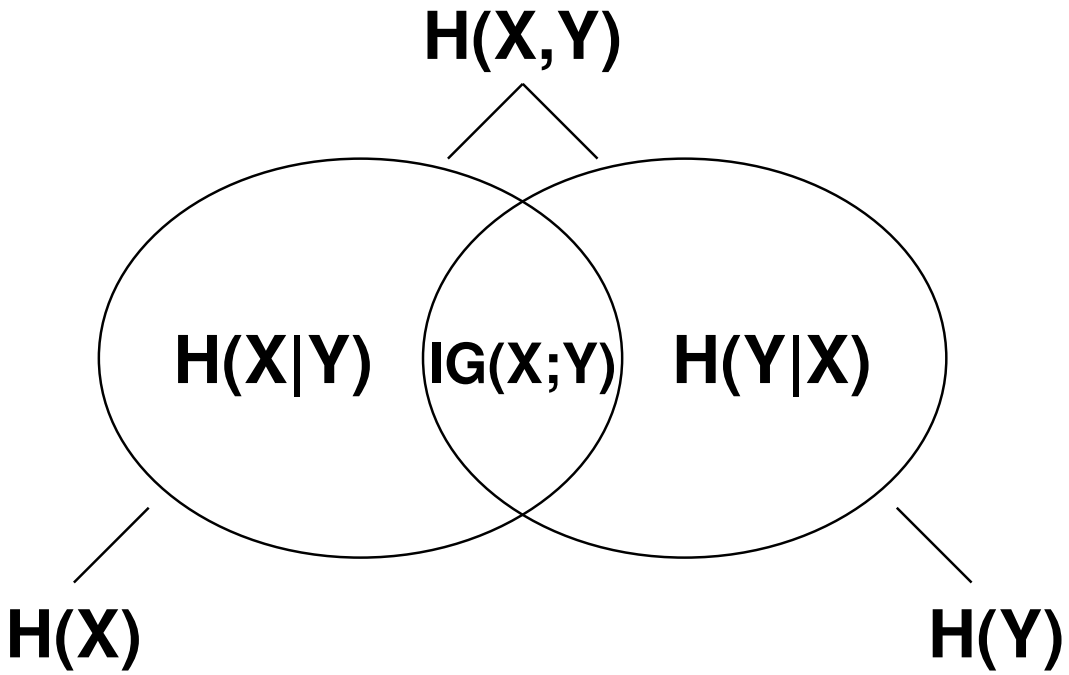
$$H(p) = -p \log_2 p - (1-p) \log_2 (1-p)$$



Basic properties of Entropy, Conditional Entropy, Joint Entropy and Information Gain / Mutual Information

- $0 \leq H(p_1, \ldots, p_n) \leq H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) = \log_2 n$;
  $H(X) = 0$ **iff** X **is a constant random variable.**
- $IG(X;Y) \geq 0$;
  $IG(X;Y) = 0$ **iff** X **and** Y **are independent;**
  $IG(X;X) = H(X)$**.**
- $H(X \mid Y) \leq H(X)$
  $H(X \mid Y) = H(X)$ **iff** X **and** Y **are independent.**
- $H(X,Y) \leq H(X) + H(Y)$;
  $H(X,Y) = H(X) + H(Y)$ **iff** X **and** Y **are independent;**
  $H(X,Y|A) = H(X|A) + H(Y|X,A)$ **(a conditional form).**
- **a chain rule:** $H(X_1, \ldots, X_n) = H(X_1) + H(X_2|X_1) + \ldots + H(X_n|X_1, \ldots, X_{n-1})$**.**

The Relationship between Entropy, Conditional Entropy, Joint Entropy and Information Gain



Other definitions

- **Let X be a discrete random variable, p its pmf and q another pmf (usually a *model* of p).**

  Cross-entropy:

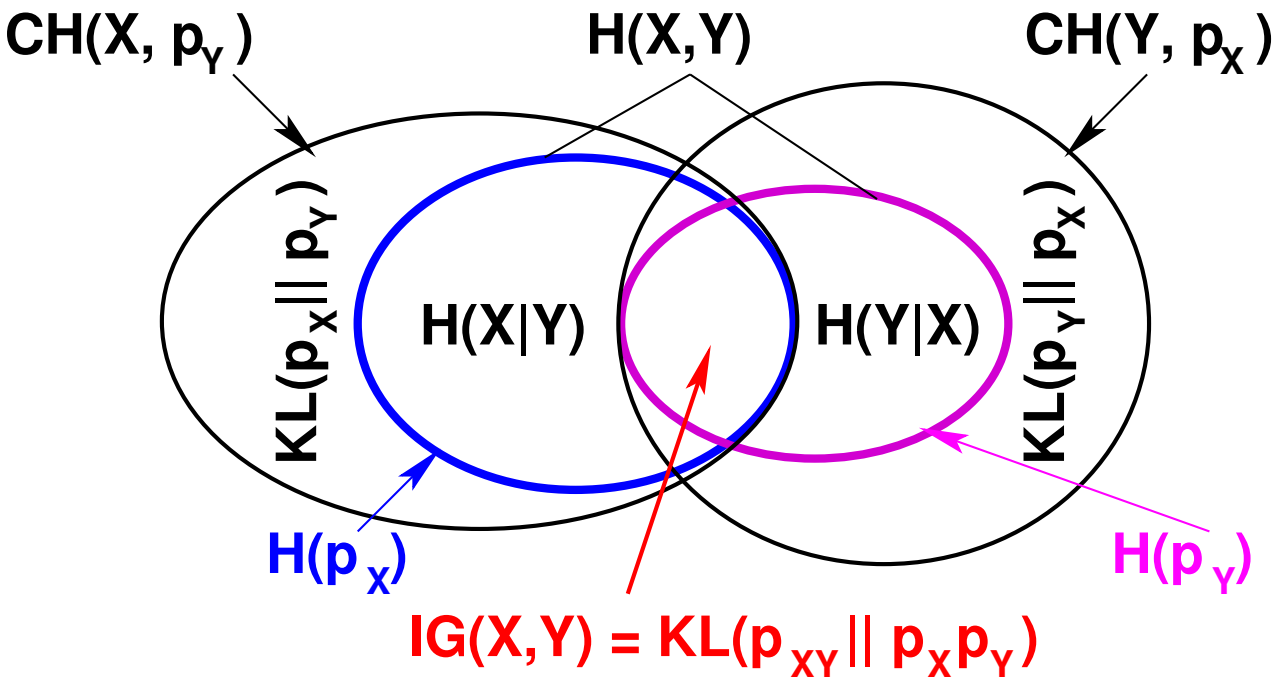
$$CH(X,q) = -\sum_{x\in X} p(x)\log_2 q(x) = E_p\left[\log_2 \frac{1}{q(X)}\right]$$

- **Let X and Y be discrete random variables, and p and q their respective pmf's.**

  Relative entropy (or, Kullback-Leibler *divergence*):

$$\begin{aligned} KL(p \,||\, q) &= -\sum_{x\in X} p(x)\log_2 \frac{q(x)}{p(x)} = E_p\left[\log_2 \frac{p(X)}{q(X)}\right] \\ &= CH(X,q) - H(X). \end{aligned}$$

The Relationship between
Entropy, Conditional Entropy, Joint Entropy, Information Gain,
Cross-Entropy and Relative Entropy (or KL divergence)



Basic properties of cross-entropy and relative entropy

- $CH(X, q) \geq 0$
- **$KL(p \,||\, q) \geq 0$ for all p and q;**
  $KL(p \,||\, q) = 0$ iff p and q are identical.
- **[Consequence:]**
  If X is a discrete random variable, p its pmf, and q another pmf, then $CH(X, q) \geq H(X) \geq 0$.

  The first of these two inequations is also known as Gibbs' inequation: $-\sum_{i=1}^{n} p_i \log_2 p_i \leq -\sum_i p_i \log_2 q_i$.
- **Unlike H of a discrete n-ary variable, which is bounded by $\log_2 n$, there is no (general) upper bound for CH. (However, KL is upper-bounded.)**
- **Unlike $H(X, Y)$, which is symmetric in its arguments, CH and KL are not! Therefore KL is NOT a distance metric! (See the next slide.)**
- $IG(X; Y) = KL(p_{X,Y} \,||\, p_X \, p_Y) = -\sum_x \sum_y p(x, y) \log_2 \left(\dfrac{p(x)p(y)}{p(x, y)} \right)$.

Remark

- **The quantity**

$$\begin{aligned} VI(X,Y) &\stackrel{def}{=} H(X,Y) - IG(X;Y) = H(X) + H(Y) - 2IG(X;Y) \\ &= H(X \mid Y) + H(Y \mid X) \end{aligned}$$

known as variation of information, is a *distance metric*, i.e. it is nonengative, symmetric, implies indiscernability, and satisfies the triangle inequality.

- **Consider** $M(p,q) = \frac{1}{2}(p+q)$**.**

The function $JSD(p||q) = \frac{1}{2}KL(p||M) + \frac{1}{2}KL(q||M)$ **is called the *Jensen-Shannon divergence*.**

One can prove that $\sqrt{JSD(p||q)}$ **defines a distance metric (the *Jensen-Shannon distance*).**

6. Recommended Exercises

- **From [Manning & Schütze, 2002 , ch. 2:]**

  Examples 1, 2, 4, 5, 7, 8, 9
  Exercises 2.1, 2.3, 2.4, 2.5

- **From [Sheldon Ross, 1998 , ch. 8:]**

  Examples 2a, 2b, 3a, 3b, 3c, 5a, 5b

Addenda:
Other Examples of Probabilistic Distributions

Multinomial distribution:

generalises the binomial distribution to the case where there are K independent outcomes with probabilities θ_i, $i = 1, \ldots, K$ such that $\sum_{i=1}^{K} \theta_i = 1$. The probability of getting n_i occurrence of outcome i is given by

$$P(n \mid \theta) = \frac{n!}{\Pi_{i=1}^{K}(n_i!)} \Pi_{i=1}^{K} \theta_i^{n_i},$$

where $n = n_1 + \ldots + n_K$, and $\theta = (\theta_1, \ldots, \theta_K)$.

Note: The particular case $n = 1$ represents the categorical distribution. This is a generalisation of the Bernoulli distribution.

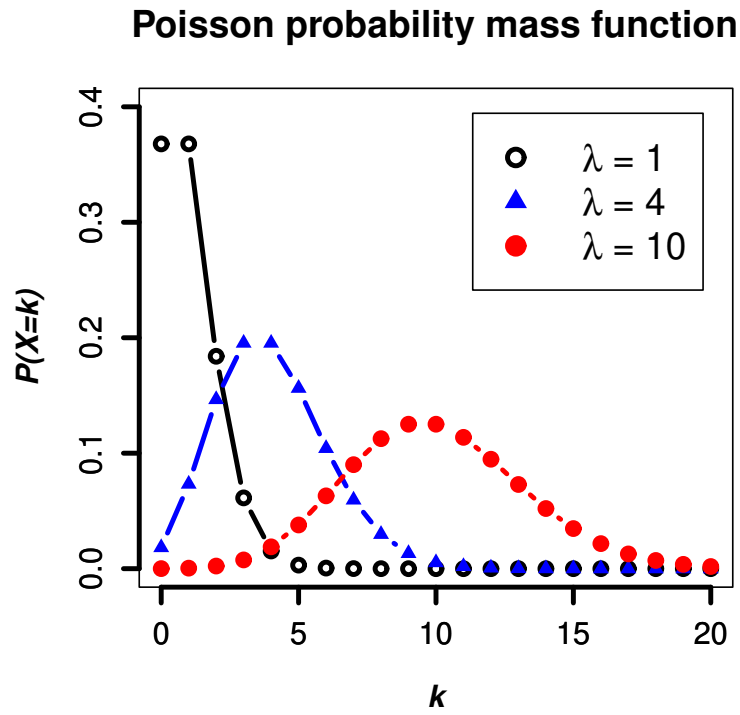
Example: The outcome of rolling a die n times is described by a categorical distribution. The probabilities of each of the 6 outcomes are $\theta_1, \ldots, \theta_6$. For a fair die, $\theta_1 = \ldots = \theta_6$, and the probability of rolling it 12 times and getting each outcome twice is:

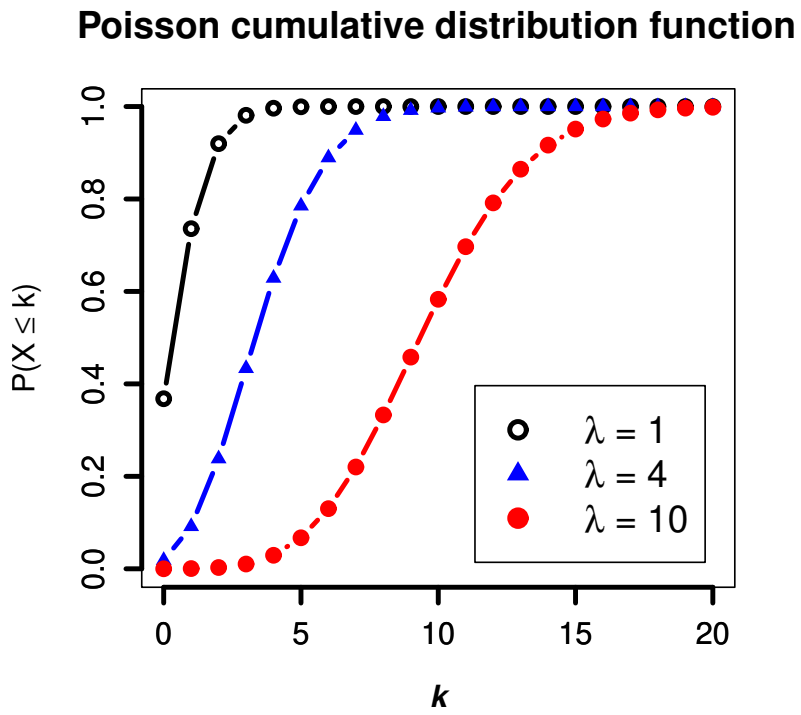
$$\frac{12!}{(2!)^6} \left(\frac{1}{6}\right)^{12} = 3.4 \times 10^{-3}$$

Poisson distribution (or, Poisson law of small numbers):

$p(k;\lambda) = \dfrac{\lambda^k}{k!} \cdot e^{-\lambda}$, **with** $k \in \mathbb{N}$ **and parameter** $\lambda > 0$.

Mean = variance = λ.

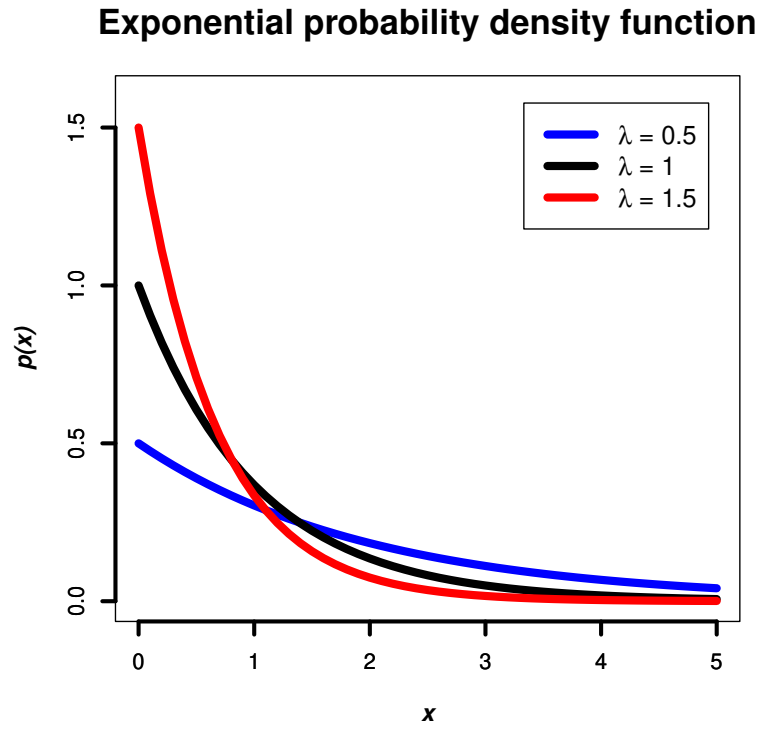


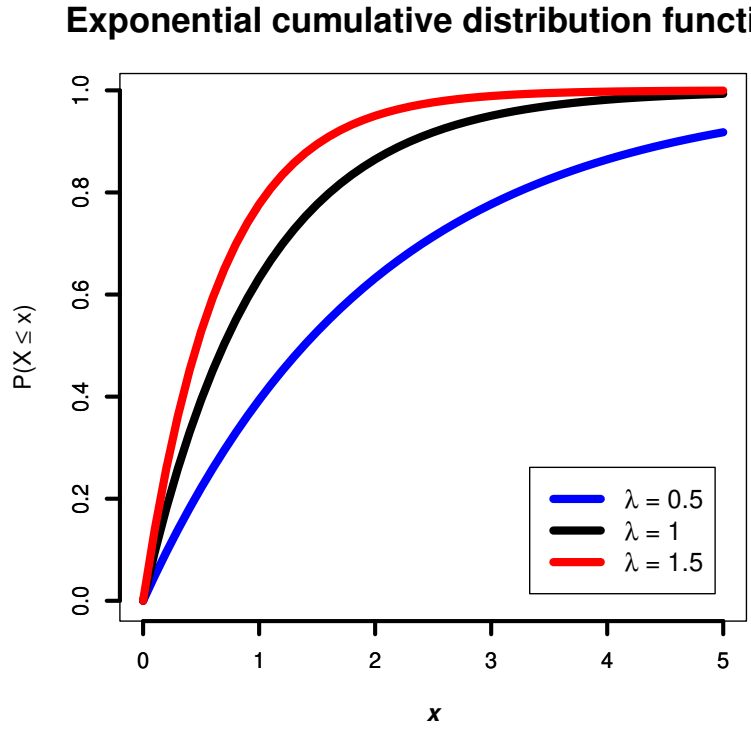


Exponential distribution (a.k.a. negative exponential distribution):

$p(x;\lambda) = \lambda e^{-\lambda x}$ for $x \geq 0$ and parameter $\lambda > 0$.

Mean $= \lambda^{-1}$, variance $= \lambda^{-2}$.





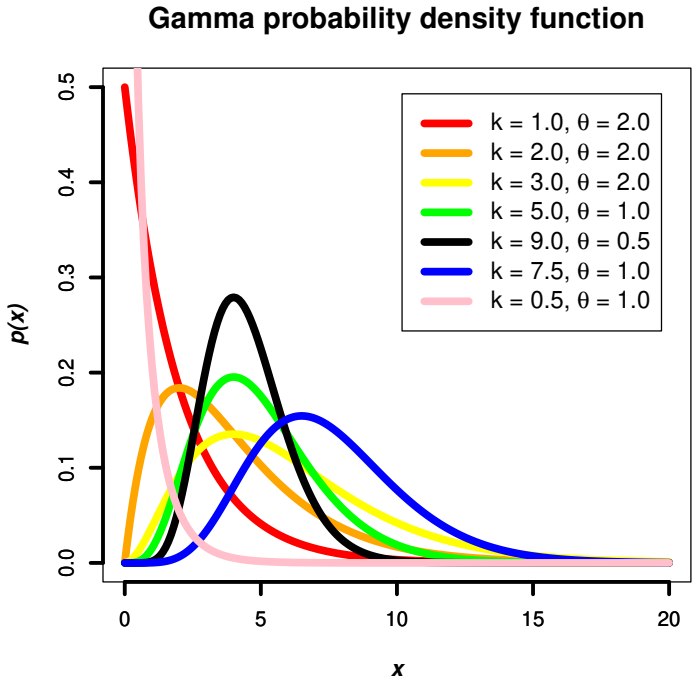
Note: The Exponential distribution is a particular case of the Gamma distribution (take $k = 1$ in the next slide).

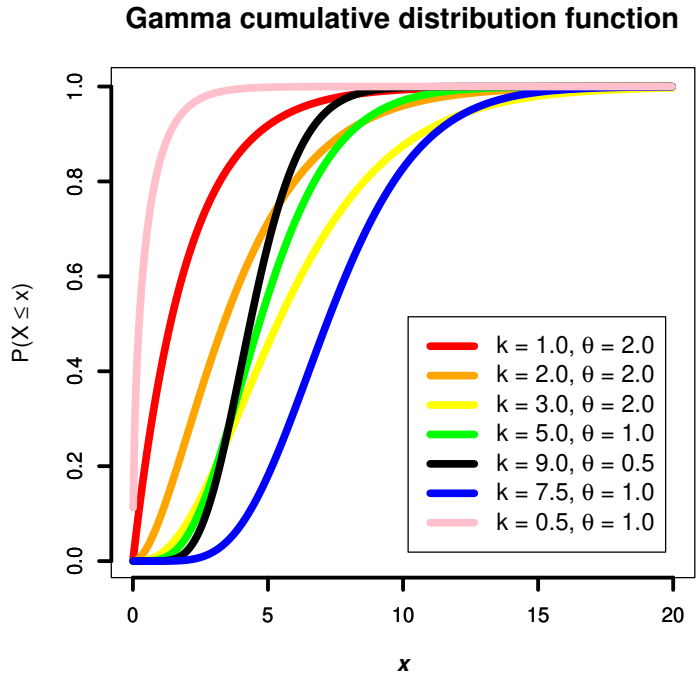
Gamma distribution:

$p(x;k,\theta) = x^{k-1}\frac{e^{-x/\theta}}{\Gamma(k)\theta^k}$ **for** $x \geq 0$ **and parameters** $k > 0$ **(shape) and** $\theta > 0$ **(scale).**

Mean $= k\theta$**, variance** $= k\theta^2$**.**

The gamma function is a generalisation of the factorial function to real values. For any positive real number x**,** $\Gamma(x+1) = x\Gamma(x)$**. (Thus, for integers** $\Gamma(n) = (n-1)!$**.)**



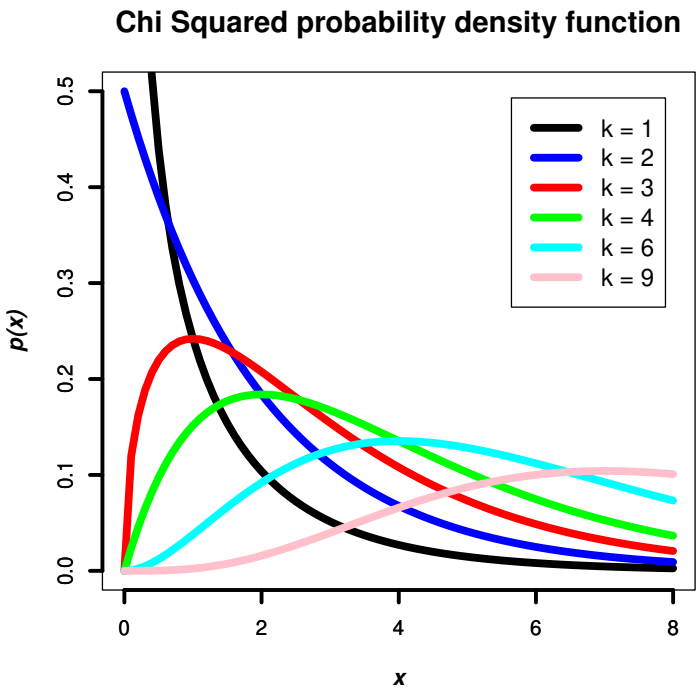


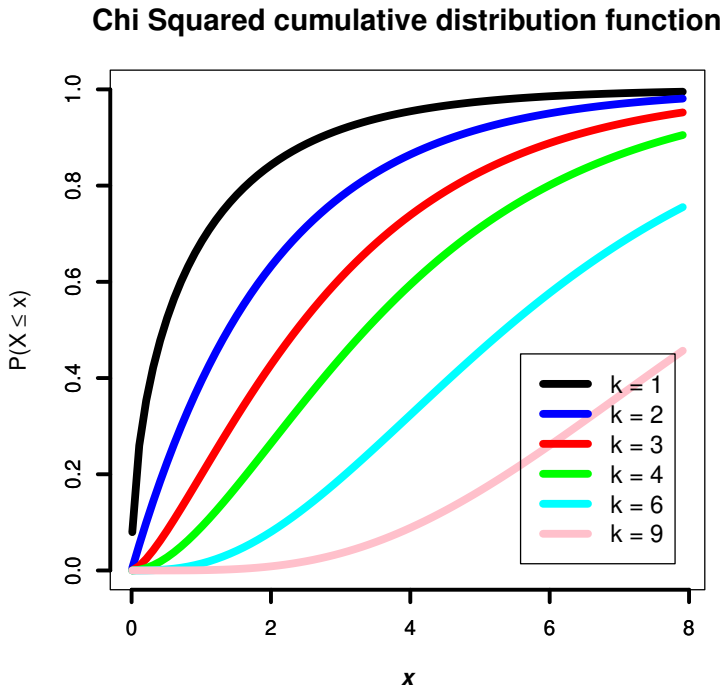
χ^2 distribution:

$p(x;\nu) = \dfrac{1}{\Gamma(\nu/2)} \left(\dfrac{1}{2}\right)^{\nu/2} x^{\nu/2-1} e^{-\frac{1}{2}x}$ **for** $x \geq 0$ **and** ν **a positive integer.**

It is obtained from Gamma distribution by taking $k = \nu/2$ **and** $\theta = 2$.

Mean $= \nu$, **variance** $= 2\nu$.



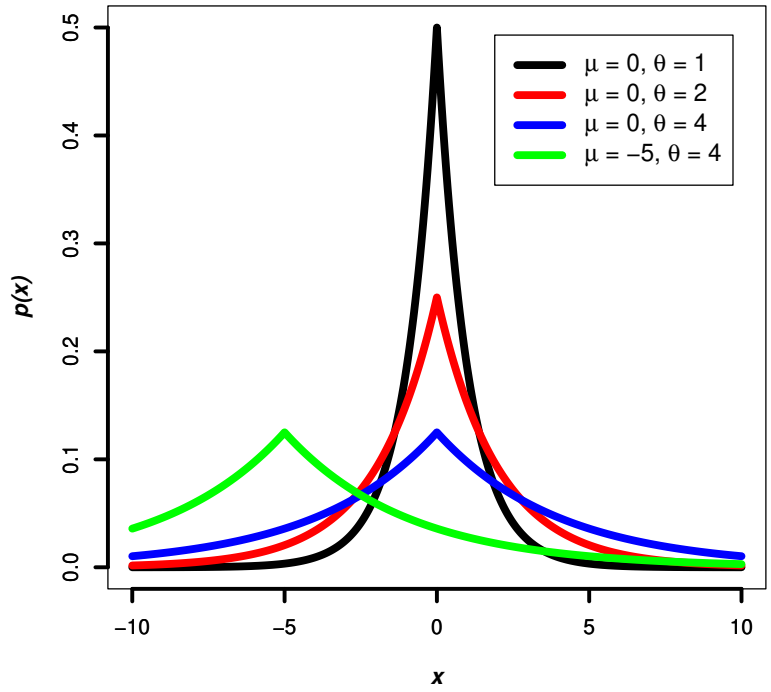


Laplace distribution:

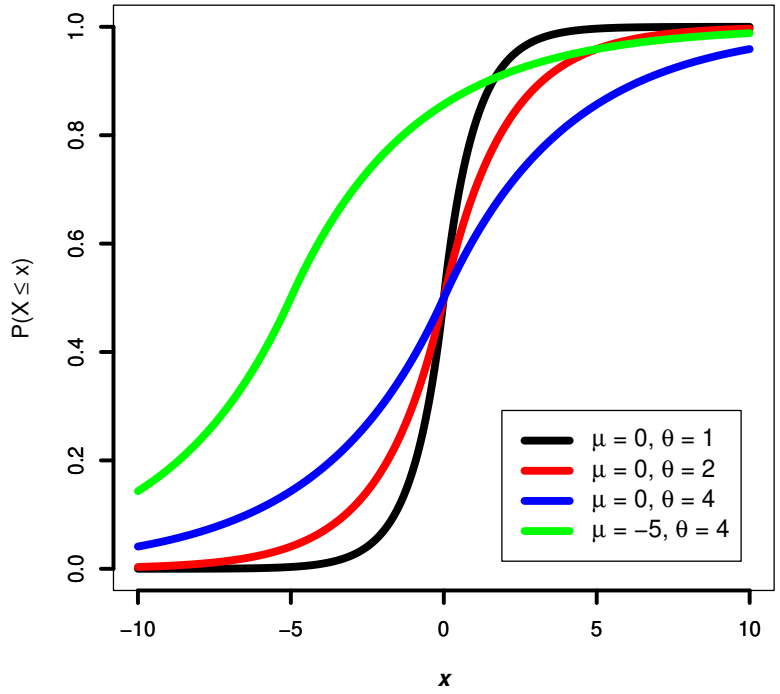
$$p(x;\mu,\theta) = \frac{1}{2\theta} e^{-\frac{|\mu - x|}{\theta}}, \text{ with } \theta > 0.$$

Mean $= \mu$, **variance** $= 2\theta^2$.

Laplace probability density function



Laplace cumulative density function



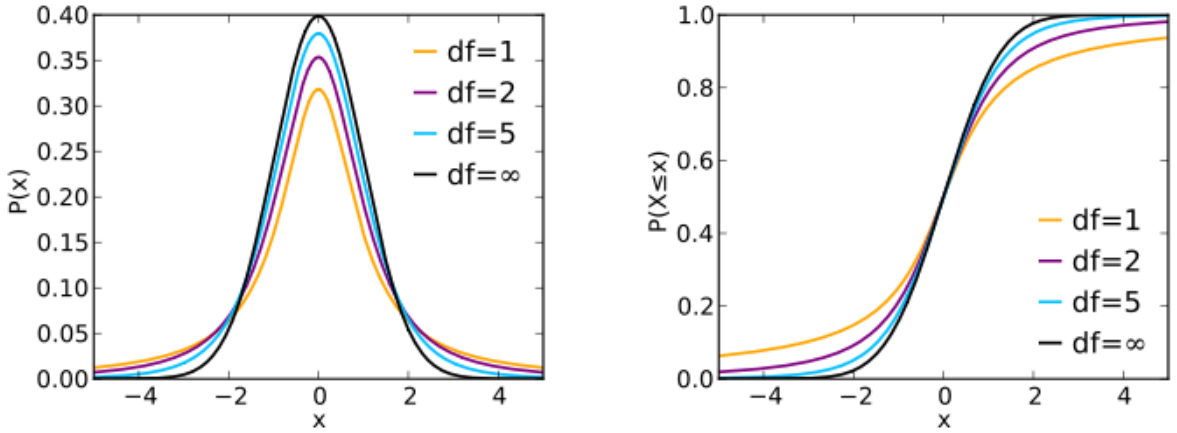
Student's distribution:

$$p(x;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$ **for** $x \in \mathbb{R}$ **and** $\nu > 0$ **(the "degree of freedom" param.)**

Mean = 0 for $\nu > 1$**, otherwise undefined.**

Variance = $\frac{\nu}{\nu-2}$ **for** $\nu > 2$**,** ∞ **for** $1 < \nu \leq 2$**, otherwise undefined.**

The probability density function and the cumulative distribution function:



Note [from Wiki]: The t**-distribution is symmetric and bell-shaped, like the normal distribution, but it has havier tails, meaning that it is more prone to producing values that fall far from its mean.**

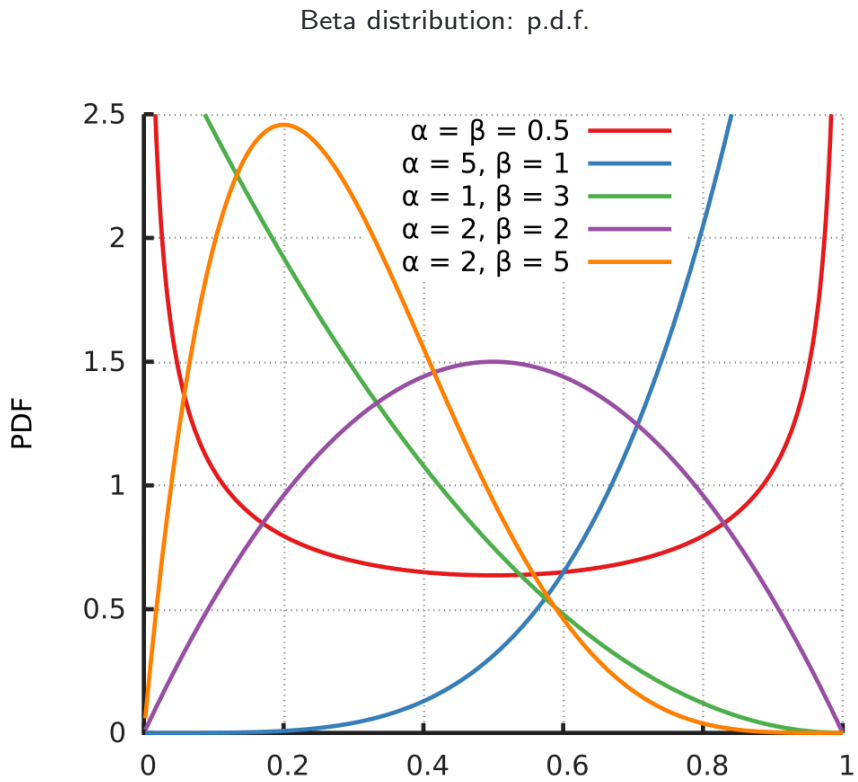
Beta distribution:

$$p(\theta;\alpha,\beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)},$$

where $B(\alpha,\beta)$ is the Beta function of arguments $\alpha,\beta \in \mathbb{R}_+$

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)},$$

with $\Gamma(x) = (x-1)!$ for any $x \in \mathbb{N}^*$.



Dirichlet distribution:

$\mathcal{D}(\theta \mid \alpha) = \frac{1}{Z(\alpha)} \Pi_{i=1}^{K} \theta_i^{\alpha_i - 1} \delta(\sum_{i=1}^{K} \theta_i - 1)$

where

$\alpha = \alpha_1, \ldots, \alpha_K$ **with** $\alpha_i > 0$ **are the parameters,**

θ_i **satisfy** $0 \leq \theta_i \leq 1$ **and sum to 1, this being indicated by the delta function term** $\delta(\sum_i \theta_i - 1)$**, and**

the normalising factor can be expressed in terms of the gamma function:

$Z(\alpha) = \int \Pi_{i=1}^{K} \theta_i^{\alpha_i - 1} \delta(\sum_i -1) d\theta = \frac{\Pi_i \Gamma(\alpha_i)}{\Gamma(\sum_i \alpha_i)}$

Mean of θ_i**:** $\frac{\alpha_i}{\sum_j \alpha_j}$**.**

For $K = 2$**, the Dirichlet distribution reduces to the Beta distribution.**

Remark:

Concerning the multinomial and Dirichlet distributions:

The algebraic expression for the parameters θ_i is similar in the two distributions.
However, the multinomial is a distribution over its exponents n_i, whereas the Dirichlet is a distribution over the numbers θ_i that are exponentiated.

The two distributions are said to be **conjugate distributions** and their close formal relationship leads to a harmonious interplay in many estimation problems.

Similarly,
the Beta distribution is the conjugate of the Bernoulli distribution, and
the Gamma distribution is the conjugate of the Poisson distribution.