Statistical Bioinformatics, Makerere
Basic Probability
Timo Koski

TK

02.08.2018

KTH Matematik

Why ?

Why should there be probabilistic modelling in analysis of biological databases ?
In the literature on data mining and *machine learning* (Baldi and Brunak 1998) we frequently find *Bayesian probability* advocated as the fundamental tool of uncertain reasoning.

- P. Baldi and S. Brunak (1998): *Bioinformatics. The Machine Learning Approach*. A Bradford Book. The MIT Press. Cambridge Massachusetts, London, England.

KTH Matematik

Why ?

The Cox-Jaynes axioms[1] of uncertain reasoning are often cited in this context, cf. (Arnborg and Sjödin 2000), and would thus seem to provide rationale for the infusion of probability in bioinformatics.

- S. Arnborg and G. Sjödin. (2000): On the Foundations of Baeysianism. *MaxEnt 2000: The Twentieth International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering.* American Institute of Physics,pp. 61−71.

[1] `http://omega.albany.edu:8008/JaynesBook.html`

KTH Matematik

Ewens and Grant

The text by Ewens and Grant does not subscribe to much of a Bayesian view of statistics. They do not refer to machine learning or data mining. We are going to learn about, e.g., significance of the BLAST score, about p-values , e.t.c. .

KTH Matematik

Alphabet, Sequence

$\mathcal{X}$ and $\mathcal{Y}$ are two discrete *alphabets*, whose generic elements are called *symbols* and denoted by x_i and y_j, respectively.

$$\mathcal{X} = \{x_1, \cdots, x_L\}, \mathcal{Y} = \{y_1, \cdots, y_J\}.$$

$| \mathcal{X} | \stackrel{def}{=}$ the number of elements in $\mathcal{X}$ = cardinality of $\mathcal{X} = L \leq \infty$, $| \mathcal{Y} | = J \leq \infty$. Unless otherwise stated the alphabets considered here are finite.

KTH Matematik

Alphabet, Sequence

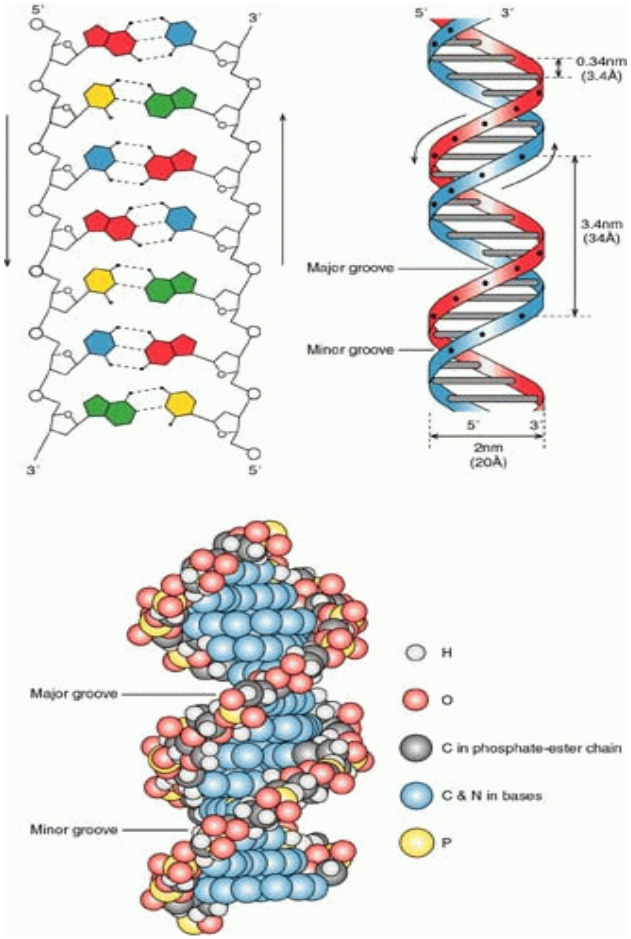
In analysis and modelling of DNA sequences the alphabet is

$$\mathcal{X} = \{A, T, C, G\}.$$

These are the four bases present in subunits of DNA called *nucleotides*: adenine (A), thymine (T), cytosine (C) and guanine (G). The particular order of these bases arranged along the sugarphosphate backbone is called the *DNA sequence*. □

KTH VETENSKAP OCH KONST
KTH Matematik

DNA



KTH VETENSKAP OCH KONST
KTH Matematik

Alphabet, Sequence

A *sequence* or a *string* $\mathbf{x}$ of length m is an ordered list of m symbols from an alphabet $\mathcal{X}$ written contiguously from left to right

$$\mathbf{x} = x_{l_1} x_{l_2} \dots x_{l_i} \dots x_{l_m}; x_{l_i} \in \mathcal{X}, i = 1, \dots, m.$$

The length of the symbol is denoted by $|\mathbf{x}| = m$. The *empty sequence* $\emptyset$ has length 0 and contains no symbols taken from $\mathcal{X}$.

KTH VETENSKAP OCH KONST
KTH Matematik

Alphabet, Sequence

DNA's genetic code can be represented as a single alphabetic sequence composed of these four symbols. It is by means of this code that the gene controls the formation of other substances in the cell. Progress of biotechnology as well as of computer hardware and software have made it possible to determine the nucleotide sequences for large and increasing number of organisms. Such data are now available in computer-readable form, so it is possible to look for and analyze patterns within sequences using probabilistic modelling and computer algorithms.

KTH Matematik

Probabilistic Models

There are numerous *model families* assigning a probability $P(\mathbf{x})$ to a string (or by extension to a set of strings) that have been applied in bioinformatics. The model families and the methods of probabilistic modeling often used in bioinformatics are presented in this course.

KTH VETENSKAP OCH KONST
KTH Matematik

Random Variables (R.V's)

X is a (**discrete**) random variable that assumes values in $\mathcal{X}$ and Y is a (**discrete**) random variable that assumes values in $\mathcal{Y}$.

KTH Matematik

R.V's, Distributions

Events are often simply written as

$$\{X = x_i\} \quad : X \text{ assumes the value } x_i$$

$$\{Y = y_j\} \quad : Y \text{ assumes the value } y_j.$$

Then the probability of the event $\{X = x_i\}$ is

$$f_X(x_i) \stackrel{def}{=} P(X = x_i).$$

The probability of the event $\{Y = y_j\}$ is

$$f_Y(y_j) \stackrel{def}{=} P(Y = y_j).$$

KTH VETENSKAP OCH KONST
KTH Matematik

R.V's, Distributions

Furthermore

$$f_X \stackrel{def}{=} (f_X(x_1), \cdots, f_X(x_L))$$

designates a discrete probability distribution on $\mathcal{X}$ and

$$f_Y \stackrel{def}{=} (f_X(y_1), \cdots, f_Y(y_J))$$

designates a discrete probability distribution on $\mathcal{Y}$.

KTH Matematik

$$f_X(x_i) \geq 0 \quad (1)$$

$$\sum_{i=1}^{L} f_X(x_i) = 1 \quad (2)$$

and similarly for f_Y.

If $A \subseteq X$ then

$$P_X(A) \stackrel{def}{=} \sum_{x_i \in A} f_X(x_i) \tag{3}$$

is the probability of the event that X assumes a value in A, a subset of $\mathcal{X}$. If $x_1 \leq x_1 \leq \ldots \leq x_L$, these are real numbers,

$$F_X(x) = \sum_{i:x_i \leq x} f_X(x_i),$$

is called the distribution function.

KTH VETENSKAP OCH KONST
KTH Matematik

R.V's, Distributions

If $x_1 \leq x_1 \leq \ldots \leq x_L$, real numbers,

$$F_X(x) = \sum_{i:x_i \leq x} f_X(x_i)$$

is called the distribution function.

- $0 \leq F_X(x) \leq 1$.
- $x \leq x^{'}$, $F_X(x) \leq F_X(x^{'})$

KTH VETENSKAP OCH KONST
KTH Matematik

R.V's, Distribution function

$A \subseteq X$

$$P_X(A) \stackrel{def}{=} \sum_{x_i \in A} f_X(x_i)$$

From this one easily finds the complement rule

$$P\left(A^c\right) = 1 - P(A), \tag{4}$$

where A^c is the complement of A, i.e. those outcomes that do not lie in A. Also

$$P\left(A \bigcup B\right) = P(A) + P(B) - P(A \bigcap B), \tag{5}$$

is immediate.

R.V's, Distributions

The notation f_X encompasses *tabular* probability distributions, which are simply tables of numbers with the stated properties, and *algorithmic* distributions, which are algorithms for computing probabilities. There are both sorts of probability distributions in bioinformatics as well as combinations thereof.

KTH VETENSKAP OCH KONST
KTH Matematik

R.V's, Distributions

A *parametric* distribution is a special kind of algorithmic distribution: it consists of a few numerical parameters and a relatively simple algorithm, usually a formula for computing probabilities given some specific values of these parameters. The examples immediately following belong to the parametric type.

KTH Matematik

Bernoulli R.V's

Consider the binary alphabet $\mathcal{X} = \{0, 1\}$. Let p be a number between zero and one. If X is a random variable assuming values in $\{0, 1\}$ and if

$$f_X(1) = P(X = 1) = p$$

then we call X a Bernoulli random variable with the *probability of success* p. We write

$$X \in Be(p).$$

We refer to p as a *parameter* of the distribution f_X.

Geometric R.V's

Consider an infinite sequence of Bernoulli random variables $\{X_n\}_{n=1}^{\infty}$ with a common probability of success p. This models an infinite sequence of zeros and ones. Let

$$X = \text{ the number of trials before the first failure not including this.}$$

Then $\mathcal{X} = \{0, 1, 2, 3, \ldots, \infty\}$. We say that X has *a geometric probability distribution* if

$$f_X(k) = (1-p) \cdot p^k; k = 0, 1, \ldots,$$

This is written as $X \in Ge(p)$.

Geometric-Like R.V's

Suppose $x \in \{0, 1, 2, 3, \ldots, \}$ and suppose that

$$1 - F_X(x-1) \sim Cp^x$$

for some fixed constant C, $0 < C < 1$. We say that X has *a geometric-like probability distribution*.

KTH VETENSKAP OCH KONST
KTH Matematik

Geometric-Like R.V's

$$1 - F_X(x-1) \sim Cp^x$$

We say that the function f is *asymptotic* to function g, as $t \to \infty$, and write $f \sim g$, if

$$\lim_{t\to\infty} \frac{f(t)}{g(t)} = 1.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Uniform R.V's

Let $\mathcal{X} = \{x_1, x_2, \ldots, x_L\}$ be a finite alphabet and

$$f_X(x_k) = \frac{1}{L}; k = 1, 2, \ldots, L.$$

This is called the *uniform probability distribution*.

KTH VETENSKAP OCH KONST
KTH Matematik

Uniform R.V's

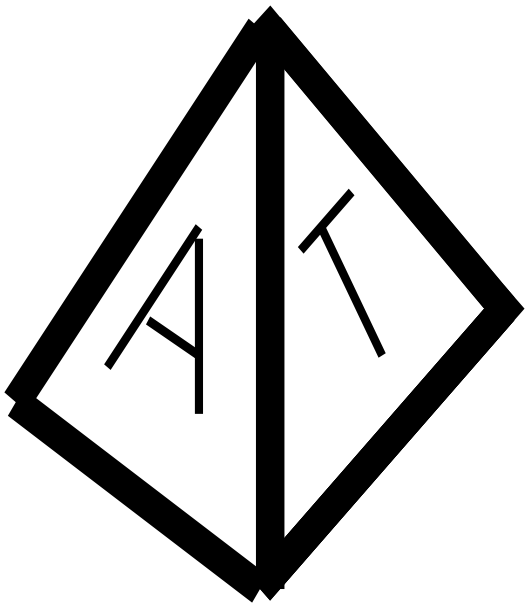
If X is a random variable assuming values in this alphabet and $P(X = k) = \frac{1}{L}$, then we say that

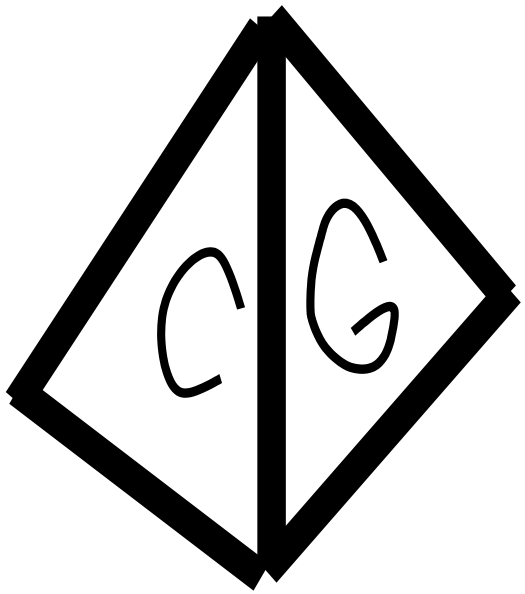
'X is selected at random'

and we write

$$X \in U(1, L).$$

Uniform R.V's: DNA Dice





are not generated like this !

$U(1,4)$: DNA sequences

Poisson distributed R.V's

Let

$$f_X(k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}, k = 0, 1, 2, \ldots.$$

This is a probability distribution on the non-negative integers called the *Poisson distribution*. A random variable X with this distribution is said to be Poisson distributed and we write

$$X \in Po(\lambda).$$

KTH Matematik

Binomial R.V's

Let A be some event defined in terms of some basic experiment. Introduce the random variable $X =$ *the number of times the event A occurs in n repetitions of some basic experiment*. Then the alphabet of X is $\mathcal{X} = \{0, 1, \cdots, n\}$. If the repetitions are modeled as independent (?), then X has the *binomial distribution*:

$$f_X(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

KTH Matematik

Binomial Distribution: The Conditions

A **binomial distribution** results from a procedure that meets all the following conditions:

The procedure has a fixed number of random events.

1.

The events have outcomes in two categories.

2.

The events are independent.

3.

The probabilities are constant for each event.

4.

Binomial R.V's

$$f_X(k) = P(X = k) = \begin{pmatrix} n \\ k \end{pmatrix} p^k (1-p)^{n-k}.$$

(Recall that $\begin{pmatrix} n \\ k \end{pmatrix} = \frac{n!}{(n-k)!k!}$ (the *binomial coefficients*), $0! \stackrel{def}{=} 1$). We say that X is a binomial random variable with parameters p and n and write

$$X \in Bin(n, p).$$

Binomial R.V's

Clearly the distibution of X is such that

$$X = X_1 + X_2 + \ldots + X_n,$$

as an equality in distribution, where $X_1, X_2, \ldots, X_n$ are independent $Be(p)$ - variables as in the preceding example.

KTH Matematik

Binomial R.V's in Bioinformatics

The number of matches in a sequence comparison ?

$$X = X_1 + X_2 + \ldots + X_n,$$

where $X_1, X_2, \ldots, X_n$ are independent $Be(p)$ - variables.

KTH VETENSKAP OCH KONST
KTH Matematik

Alignment of Sequences & series of random events

We wish to compare two sequences **x** and **y** with 15 nucleotides in each.

$$\begin{array}{cc} \mathbf{x} & G\ \overset{\downarrow}{A}\ T\ \overset{\downarrow}{A}\ \overset{\downarrow}{A}\ G\ \overset{\downarrow}{C}\ \overset{\downarrow}{C}\ \overset{\downarrow}{C}\ \overset{\downarrow}{C}\ T\ \overset{\downarrow}{G}\ \overset{\downarrow}{T}\ \overset{\downarrow}{C}\ \overset{\downarrow}{T} \\ = & \\ \mathbf{y} & C\ A\ A\ A\ A\ T\ C\ C\ C\ C\ A\ G\ T\ C\ T \end{array}$$

We say that we have a **match**, if the paired nucleotides are the same in both sequences. We have eleven matches indicated by $\downarrow$.

KTH Matematik

Alignment of Sequences & Binomial Distribution

$$\begin{array}{cc} \mathbf{x} & G\ \overset{\downarrow}{A}\ T\ \overset{\downarrow}{A}\ \overset{\downarrow}{A}\ G\ \overset{\downarrow}{C}\ \overset{\downarrow}{C}\ \overset{\downarrow}{C}\ \overset{\downarrow}{C}\ T\ \overset{\downarrow}{G}\ \overset{\downarrow}{T}\ \overset{\downarrow}{C}\ \overset{\downarrow}{T} \\ = & \\ \mathbf{y} & C\ A\ A\ A\ A\ T\ C\ C\ C\ C\ A\ G\ T\ C\ T \end{array}$$

If we are willing to assume that the nucleotides are random (DNA die !) and independent, and that the probabilities are the same at each site, then **the probabilities of the number of matches (=successes) follow a binomial distribution with parameters** $n = 15$, $p = \frac{1}{4}$. Then we can compute how probable or likely it is to get 11 matches in a sequence of 15 nucleotides. This is a case of the question of significance (**p-value**) in evaluating sequence alignments.

KTH Matematik

Rare Event Rule

If, under a given assumption, the probability of an observed event is extremely small (≈ 0), we conclude that the assumption is likely not correct.

In the example with alignment of sequences, the probability of 11 matches in two sequences of 15 nucleotides under the assumption of independent tosses of the DNA die is $1.0297e-04$. (A computation done using a Matlab function for the binomial probability `binopdf(x,n,p)` with $n = 15$, $p = \frac{1}{4}$, $x = 11$).

KTH Matematik

Binomial R.V's

If np is small, i.e. if the event A occurs 'rarely', the binomial distribution can be approximated by a Poisson distribution $Po(\lambda)$ with $\lambda = np$ (*the law of small numbers*). The approximation $Bin(n, p) \approx Po(np)$, or more appropriately its generalizations have turned out to be useful for a number of problems in bioinformatics.

KTH Matematik

Negative Binomial R.V's

Same setting as in Binomial R.V.'s but the number of successes is fixed in advance at m and the number of experiments up to and including the mth success is a random variable X.

$$f_X(n) = P(X = n) = \begin{pmatrix} n-1 \\ m-1 \end{pmatrix} p^m (1-p)^{n-m}.$$

$n = m, m+1, \ldots,$. X is a negative binomial random variable with parameters p and m.

KTH Matematik

Generalized Geometric R.V's

The number of failures is fixed in advance at $k+1$ and the number of experiments up to and not including the $k+1$st failure is a random variable X. Then as in the above

$$f_X(x) = P(X = x) = \left(\begin{array}{c} x \\ k \end{array} \right) p^{x-k}(1-p)^{k+1}.$$

$x = k, k+1, \ldots,$. This reduces to a geometric distribution when $k = 0$, hence this is a generalized geometric distribution.

KTH Matematik

Hypergeometric R.V's

n balls in an urn, r are red and $n-r$ are black. Introduce for $m \leq n$ the random variable $X =$ *the number of red balls is* k, *if* m *balls are chosen at random and with replacement*, X has the *hypergeometric distribution*:

$$P(X = k) = \frac{\binom{r}{k}\binom{n-r}{m-k}}{\binom{n}{m}}.$$

$$X \in \mathrm{Hyp}(n, m, r).$$

KTH VETENSKAP OCH KONST
KTH Matematik

Hypergeometric R.V's

R. Román-Roldán, P. Benaola-Galván and J.L. Oliver: Sequence Compositional Complexity of DNA through an entropic segmentation method. *Physical Review Letters*, 80 (6), 1344−1347.

KTH Matematik

Expectation

If $\mathcal{X}$ is a discrete alphabet and a finite set of real numbers we define the expectation, $E(X)$, of the random variable X as

$$E(X) \stackrel{def}{=} \sum_{i=1}^{L} x_i f_X(x_i).$$

In case $L = \infty$ we require absolute convergence of the series.

KTH VETENSKAP OCH KONST
KTH Matematik

Expectation

If $\mathcal{X}$ is an arbitrary alphabet and $g(\cdot)$ is a map (function) of $\mathcal{X}$ to R, $g : \mathcal{X} \longmapsto R$,

$$E(g(X)) \stackrel{def}{=} \sum_{i=1}^{L} g(x_i) f_X(x_i).$$

KTH Matematik

Variance

Let $\mathcal{X}$ be a real and finite alphabet and let the *variance*, $Var(X)$, of X be

$$\begin{aligned} Var(X) &\stackrel{def}{=} E\left([X - E(X)]^2\right) \\ &= \sum_{i=1}^{L} (x_i - E(X))^2 f_X(x_i) \end{aligned} \tag{6}$$

Hence $Var(X) = E(X^2) - (E(X))^2$.

Chebysjev's inequality

Chebysjev's inequality is another partial intuitive interpretation of variance. For $k > 0$,

$$P\left(\mid X - E(X) \mid > k\right) \leq \frac{Var(X)}{k^2}.$$

KTH Matematik

Chebysjev's inequality: proof

$$
\begin{aligned}
Var(X) &= \sum_{i=1}^{L}(x_i - E(X))^2 f_X(x_i) = \\
&= \sum_{\{x_i||x_i - E(X)|>k\}}(x_i - E(X))^2 f_X(x_i) + \\
&+ \sum_{\{x_i||x_i - E(X)|\leq k\}}(x_i - E(X))^2 f_X(x_i) \\
&\geq k^2 \sum_{\{x_i||x_i - E(X)|>k\}} f_X(x_i) \\
&= k^2 P(\mid X - E(X) \mid > k), \qquad (7)
\end{aligned}
$$

where we have used the fact that $\sum_{\{x_i||x_i - E(X)|\leq k\}}(x_i - E(X))^2 f_X(x_i) \geq 0$. □

KTH VETENSKAP OCH KONST
KTH Matematik

Factorial Moments

X is an integer-valued discrete R.V.,

$$\mu_{[r]} \stackrel{def}{=} E\left[X(X-1)\cdots(X-r+1)\right] =$$

$$= \sum_{x:\text{integer}} \left(x(x-1)\cdots(x-r+1)\right) f_X(x).$$

is called the r:th factorial moment.

Binomial Moments

X is an integer-valued discrete R.V..

$$E\left(\begin{array}{c} X \\ r \end{array}\right) = E\left[X(X-1)\cdots(X-r+1)\right]/r!$$

is called the binomial moment.

KTH Matematik

CONTINUOUS R.V's

Discrete random variables are in ways more relevant to bioinformatics than continuous random variables. But:

- Continuous variables are needed for approximations.
- Continuous variables are needed in Bayesian statistics.

KTH VETENSKAP OCH KONST
KTH Matematik

CONTINUOUS R.V's

A distribution function $F_X(x) \overset{def}{=} P(X \leq x)$ such that

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

where $f_X(x)$ is a density function,

- $f_X(x) \geq 0$.
- $\int_{-\infty}^{\infty} f_X(x)dx = 1$.

KTH VETENSKAP OCH KONST
KTH Matematik

CONTINUOUS R.V's

Expectation, $E(X)$, is defined as

$$E(X) = \int_{-\infty}^{+\infty} x f_X(x) dx$$

KTH Matematik

Variance $Var(X)$ is defined as

$$Var(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 f_X(x) dx$$

$$Var(X) = E(X^2) - (E(X))^2.$$

KTH VETENSKAP OCH KONST
KTH Matematik

CONTINUOUS R.V's

The random variable $Y = g(X)$ has the expectation

$$E[Y] = E[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x) dx.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Uniform R.V's

Let $a < b$ and

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{elsewhere.} \end{cases}$$

The $X \in \mathrm{U}(a, b)$, uniformly distributed.

KTH VETENSKAP OCH KONST
KTH Matematik

Uniform R.V's

If $a < b$ and $X \in \mathrm{U}(a, b)$,

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{elsewhere} \end{cases}$$

$$E(X) = \frac{a+b}{2}, V(X) = \frac{(b-a)^2}{12}$$

KTH VETENSKAP OCH KONST
KTH Matematik

Normal R.V's

X has the density $\phi(x;\mu,\sigma)$ defined by

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

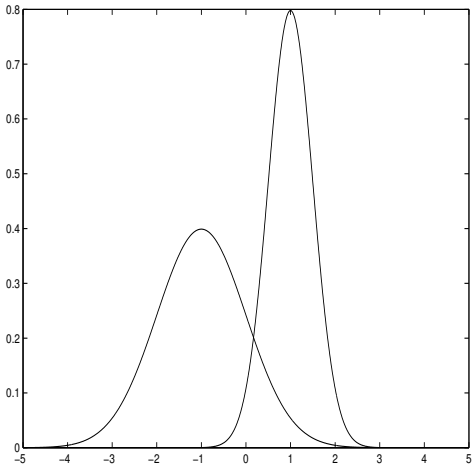
X has a normal distribution $N(\mu,\sigma^2)$.

KTH VETENSKAP OCH KONST
KTH Matematik

Normal random variables

X has a normal distribution $N(\mu, \sigma^2)$.

$$E(X) = \mu, V(X) = \sigma^2.$$



KTH VETENSKAP OCH KONST
KTH Matematik

Exponential R.V's

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & 0 < x < \infty \\ 0 & \text{elsewhere.} \end{cases}$$

$$F_X(x) = 1 - e^{-\lambda x}$$

$X \in \text{Exp}(\lambda)$.

KTH VETENSKAP OCH KONST
KTH Matematik

Exponential and Geometric R.V's

$X \in \mathrm{Exp}(\lambda)$

$$Y = \lfloor X \rfloor$$

$k = 0, 1, \ldots,,$

$$\begin{aligned} \mathrm{Prob}\left(Y = k\right) &= \mathrm{Prob}\left(k \leq X < k+1\right) \\ &= \left(1 - e^{-\lambda}\right) e^{-\lambda k}. \end{aligned}$$

Hence $Y \in \mathrm{Ge}\left(e^{-\lambda}\right)$.

KTH VETENSKAP OCH KONST
KTH Matematik

Density of the Fractional Part

$X \in \mathrm{Exp}(\lambda)$

$$D = X - \lfloor X \rfloor$$

D is the fractional part of X.

$$f_D(d) = \begin{cases} \frac{\lambda e^{-\lambda d}}{1-e^{-\lambda}} & 0 < d < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

KTH VETENSKAP OCH KONST
KTH Matematik

Gamma Density

$$f_X(x) = \begin{cases} \frac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)} & 0 < x < \infty \\ 0 & \text{elsewhere.} \end{cases}$$

λ and k are positive parameters. $\Gamma(k)$ is Euler's Gamma function. (See Westergren & Råde: **BETA. Mathematics Handbook** for facts and formulae about Euler's Gamma function.)

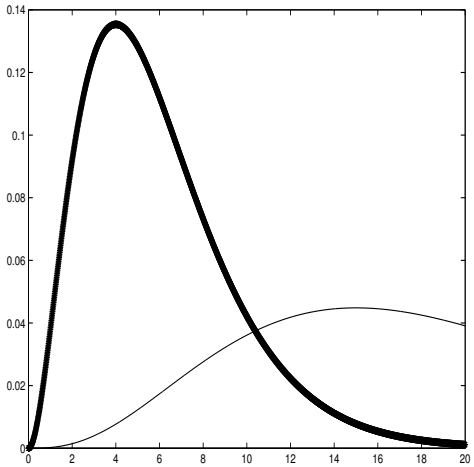
KTH VETENSKAP OCH KONST
KTH Matematik

Gamma Function

- $$\Gamma(z) = \int_1^\infty t^{z-1} e^{-t} dt$$
- $\Gamma(z+1) = z\Gamma(z)$
- $\Gamma(k) = (k-1)!$, if k is a positive integer.
- $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

KTH VETENSKAP OCH KONST
KTH Matematik

Gamma Density



$$f_X(x) = \begin{cases} \frac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)} & 0 < x < \infty \\ 0 & \text{elsewhere.} \end{cases}$$

KTH VETENSKAP OCH KONST
KTH Matematik

Chi-Square Density

$\lambda = 1/2$, $k = (1/2)\nu$, Gamma density becomes

$$f_X(x) = \begin{cases} \dfrac{x^{\frac{1}{2}\nu-1}\lambda e^{-\frac{1}{2}x}}{2^{\frac{1}{2}\nu}\Gamma\left(\frac{1}{2}\nu\right)} & 0 < x < \infty \\ 0 & \text{elsewhere.} \end{cases}$$

Chi-Square with ν degrees of freedom. This is important in statistical testing.

KTH VETENSKAP OCH KONST
KTH Matematik

Beta Density

$$f_X(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} x^{\beta-1} & 0 < x < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

$X \in \text{Beta}(\alpha, \beta)$. This density is important in Bayesian statistics.

KTH Matematik

Probability Generating Function

Let X have values $k = 0, 1, 2, \ldots,$.

$$G(t) = E\left(t^X\right) = \sum_{k=0}^{\infty} t^k f_X(k)$$

is called the probability generating function.

KTH VETENSKAP OCH KONST
KTH Matematik

Prob. Gen. Fnct: Properties

- $$\frac{d}{dt}G(1) = \sum_{k=1}^{\infty} kt^{k-1} f_X(k) \mid_{t=1}$$
$$= E[X]$$

- $$\mu_{[r]} = E[X(X-1)\cdots(X-r+1)] = \frac{d^r}{dt^r}G(1)$$

Prob. Gen. Fnct: Properties

$Z = X + Y$, X and Y integer valued, independent (?),

- $$G_Z(t) = E\left(t^Z\right) =$$

$$E\left(t^{X+Y}\right) = E\left(t^X\right) \cdot E\left(t^Y\right) = G_X(t) \cdot G_Y(t).$$

KTH VETENSKAP OCH KONST
KTH Matematik

Prob. Gen. Fnct: Examples

- $X \in \mathrm{Be}(p)$

$$G_X(t) = 1 - p + pt.$$

- $Y \in \mathrm{Bin}(n, p)$

$$G_Y(t) = (1 - p + pt)^n$$

Probability Generating Function

$$G(t) = E\left(t^X\right) = \sum_{k=0}^{\infty} t^k f_X(k)$$

is also known as called the *Mellin transform*.

Moment Generating Function

$$\phi_X(s) \stackrel{def}{=} E\left(e^{sX}\right) = \begin{cases} \sum_{x_i} e^{sx_i} f_X(x_i) & X \text{ discrete} \\ \int_{-\infty}^{\infty} e^{sx} f_X(x)dx & X \text{ continuous} \end{cases}$$

is called the moment generating function.

KTH VETENSKAP OCH KONST
KTH Matematik

Moment Gen. Fnctn: Properties

- $$\frac{d}{ds}\phi(0) = E[X]$$
- $$\phi(0) = 1$$
$$\frac{d^k}{ds^k}\phi(0) = E\left[X^k\right].$$

KTH Matematik

Moment Gen. Fnctn: Examples

- $X \in \mathrm{N}\left(\mu, \sigma^2\right)$

$$\phi_X(s) = e^{\mu s + \frac{1}{2}\sigma^2 s^2}$$

- $Y \in \exp\left(\lambda\right)$

$$\phi_Y(s) = \frac{\lambda}{\lambda - s}$$

KTH VETENSKAP OCH KONST
KTH Matematik

A Thm for Bioinformatics

Let X be a discrete R.V. with moment generating function $\phi_X(s)$. Say that X can take at least one negative value (say $-a$) with positive probability $f_X(-a)$ and at least one positive value (say b) with positive probability $f_X(b)$, and that the expectation of X is nonzero. Then there exists a unique nonzero value s^* such that

$$\phi_X(s^*) = 1.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Proof (1)

Proof: Assume $\phi_X(s)$ is defined for all s.

$$\phi_X(s) > f_X(-a)e^{-as}, \phi_X(s) > f_X(b)e^{bs}$$

Hence $\phi_X(s) \to \infty$, as $s \to \infty$ and also as $s \to -\infty$.

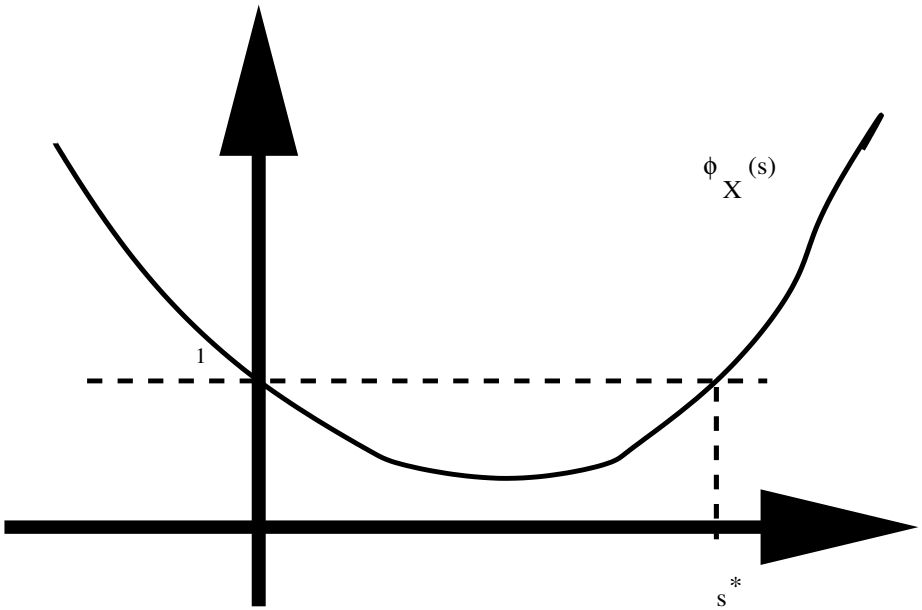
Proof (2)

$$\frac{d^2}{ds^2}\phi(s) = \sum x^2 e^{sx} f_X(x) \geq 0.$$

Hence $\phi(s)$ is convex as function of s. $\phi(0) = 1$ and the the expectation $\frac{d}{ds}\phi(0) = E[X]$ is nonzero by assumption. If $E[X]$ is negative, then the graph of $\phi_X(s)$ must be as in the Figure below:

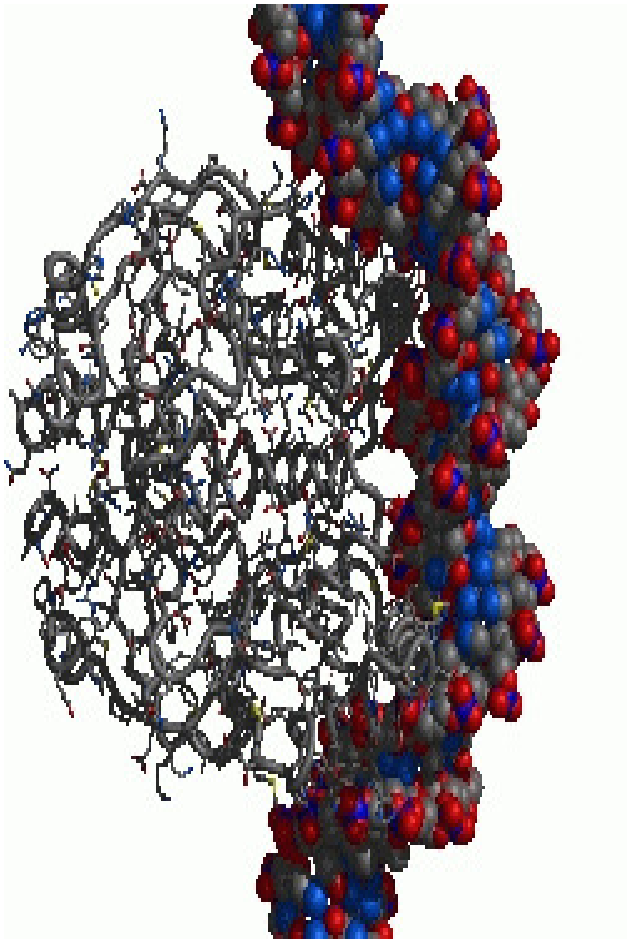
KTH VETENSKAP OCH KONST
KTH Matematik

Proof (3)



Similarly for $E[X]$ positive. □

End of Lecture



KTH Matematik