Bioinformatics and Biostatistics BB2440: Biostatistics Lecture 2: Bayes' Probability & Random Variables & Some Probability Distributions

Timo Koski

TK

04.09.2013

KTH Matematik

Outline of Lecture 2.

- A Quick Summary of the Formal Rules of Probability (from Lect 1.)

KTH Matematik

Outline of Lecture 2.

- A Quick Summary of the Formal Rules of Probability (from Lect 1.)
- Conditionally Probable Events

KTH Matematik

Outline of Lecture 2.

- A Quick Summary of the Formal Rules of Probability (from Lect 1.)
- Conditionally Probable Events
- Bayes' Rule of Probability (Inverse Conditional Probability)

KTH Matematik

Outline of Lecture 2.

- A Quick Summary of the Formal Rules of Probability (from Lect 1.)
- Conditionally Probable Events
- Bayes' Rule of Probability (Inverse Conditional Probability)
- Random Variables

KTH Matematik

Outline of Lecture 2.

- A Quick Summary of the Formal Rules of Probability (from Lect 1.)
- Conditionally Probable Events
- Bayes' Rule of Probability (Inverse Conditional Probability)
- Random Variables
- Probability Distributions

KTH Matematik

Outline of Lecture 2.

- A Quick Summary of the Formal Rules of Probability (from Lect 1.)
- Conditionally Probable Events
- Bayes' Rule of Probability (Inverse Conditional Probability)
- Random Variables
- Probability Distributions
  - Binomial Distribution

KTH Matematik

Outline of Lecture 2.

- A Quick Summary of the Formal Rules of Probability (from Lect 1.)
- Conditionally Probable Events
- Bayes' Rule of Probability (Inverse Conditional Probability)
- Random Variables
- Probability Distributions
  - Binomial Distribution
  - Geometric Distribution

KTH Matematik

Outline of Lecture 2.

- A Quick Summary of the Formal Rules of Probability (from Lect 1.)
- Conditionally Probable Events
- Bayes' Rule of Probability (Inverse Conditional Probability)
- Random Variables
- Probability Distributions
  - Binomial Distribution
  - Geometric Distribution
  - Poisson Distribution

KTH Matematik

Outline of Lecture 2.

- A Quick Summary of the Formal Rules of Probability (from Lect 1.)
- Conditionally Probable Events
- Bayes' Rule of Probability (Inverse Conditional Probability)
- Random Variables
- Probability Distributions
  - Binomial Distribution
  - Geometric Distribution
  - Poisson Distribution
  - Geometric Distribution

KTH Matematik

Outline of Lecture 2.

- A Quick Summary of the Formal Rules of Probability (from Lect 1.)
- Conditionally Probable Events
- Bayes' Rule of Probability (Inverse Conditional Probability)
- Random Variables
- Probability Distributions
  - Binomial Distribution
  - Geometric Distribution
  - Poisson Distribution
  - Geometric Distribution
  - Mean and Variance

KTH Matematik

Outline of Lecture 2.

- A Quick Summary of the Formal Rules of Probability (from Lect 1.)
- Conditionally Probable Events
- Bayes' Rule of Probability (Inverse Conditional Probability)
- Random Variables
- Probability Distributions
  - Binomial Distribution
  - Geometric Distribution
  - Poisson Distribution
  - Geometric Distribution
  - Mean and Variance
  - Normal Distribution

KTH Matematik

Outline of Lecture 2.

- A Quick Summary of the Formal Rules of Probability (from Lect 1.)
- Conditionally Probable Events
- Bayes' Rule of Probability (Inverse Conditional Probability)
- Random Variables
- Probability Distributions
  - Binomial Distribution
  - Geometric Distribution
  - Poisson Distribution
  - Geometric Distribution
  - Mean and Variance
  - Normal Distribution
  - Exponential Distribution

KTH Matematik

Probability Summary

Pr denotes probability
A,B, and C denote specific events.
Pr(A) denotes the probability of event A occurring

The probability of an event A, $Pr(A)$, has the following properties:

(a) $0 \leq Pr(A) \leq 1$;

(b) $Pr(S) = 1$, if S is an event that is certain to happen (e.g., the sun rises).

KTH Matematik

Probability Summary

We say that two events A and B are **mutually exclusive**, if the occurrence of one precludes the occurrence of the other.

The event A or B means that A occurs or both A and B occur or B occurs.

(c) if A and B are mutually exclusive events then we have the **additive law**

$$Pr(\ A \text{ or } B\ ) = Pr(A) + Pr(B).$$

(d)

$$Pr(\ A \text{ or } B\ ) = Pr(A) + Pr(B) - Pr(\ A \text{ and } B\ )$$

Note that if A and B are mutually exclusive, then the event "A and B" is impossible, and thus $Pr(\ A \text{ and } B\ ) = 0$.

Complement *"A does not occur"* $= A^*$.

(e)

$$Pr(A^*) = 1 - Pr(A).$$

KTH VETENSKAP OCH KONST
KTH Matematik

Probability Summary : independent events

We say that A and B are **independent** events if

(f)

$$Pr\,(\ A \text{ and } B\ ) = Pr(A) \cdot Pr(B).$$

Now we introduce a new element: **Conditionally Probable Events**

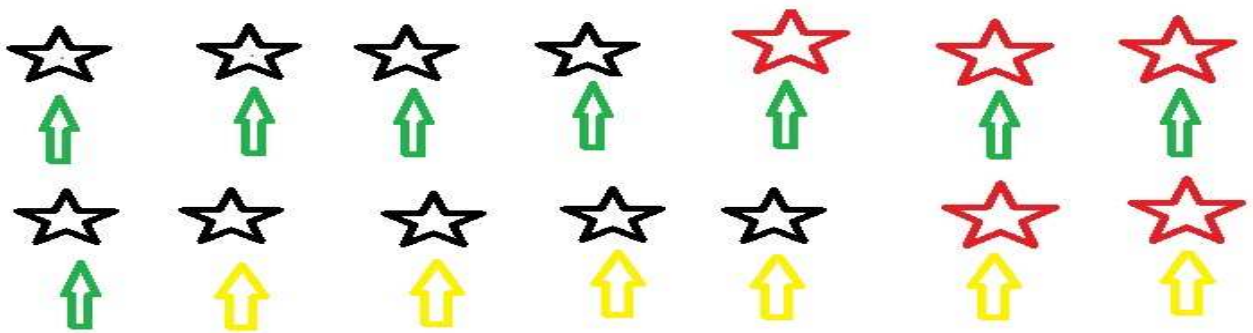
KTH Matematik

Conditionally Probable Events: a Mendelian introduction

KTH Matematik

Conditionally Probable Events: a Mendelian introduction

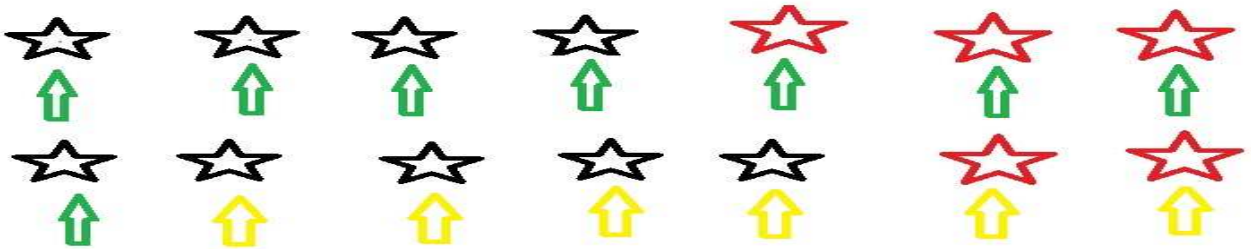
In the figure we have 14 (abstract) peas: nine have white flowers, five have red flowers, six have yellow pods, eight have green pods.



Two peas are selected at **random** and **without replacement**. **Q**: What is the probability for the event that the first selection has a green pod and the second selection has a yellow pod ?

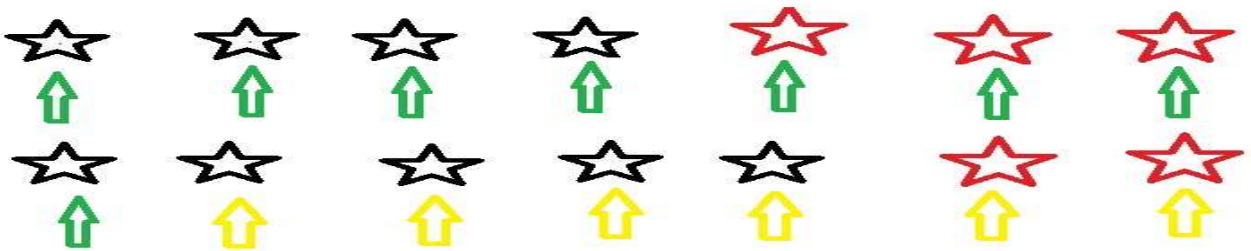
KTH Matematik

Conditionally Probable Events: a Mendelian introduction



- First selection: $Pr(\text{green pod}) = \frac{8}{14}$

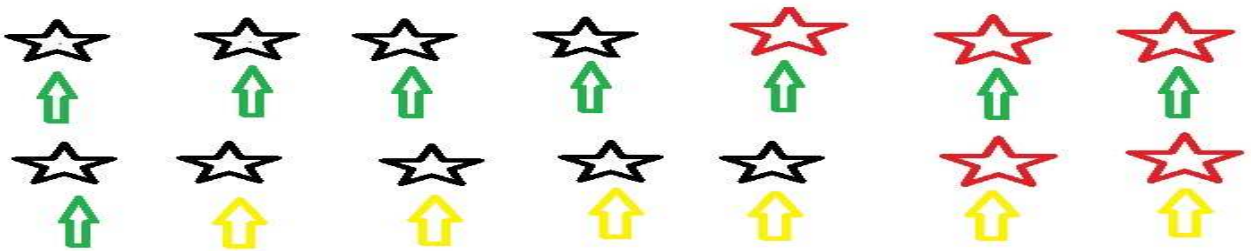
Conditionally Probable Events: a Mendelian introduction



- First selection: $Pr(\text{green pod}) = \frac{8}{14}$
- Second selection: $Pr(\text{yellow pod}) = \frac{6}{13}$

KTH VETENSKAP OCH KONST
KTH Matematik

Conditionally Probable Events: a Mendelian introduction



- First selection: $Pr(\text{green pod}) = \frac{8}{14}$
- Second selection: $Pr(\text{yellow pod}) = \frac{6}{13}$
- $Pr(\text{ 1st pea with green pod and 2nd pea yellow pod}) = \frac{8}{14} \cdot \frac{6}{13} \approx 0.264$.

KTH Matematik

Conditionally Probable Events

Two peas are selected at **random** and **without replacement**.

$$Pr(\text{ 1st pea green pod and 2nd pea with yellow pod}) = \frac{8}{14} \cdot \frac{6}{13} \approx 0.264.$$

The key point here is that *we must adjust the probability of the second event to reflect the outcome of the first event*. One pea was removed in the first selection, there are only 13 left to choose for the second selection.

KTH Matematik

Conditional Probability Events: The Principle & Notation

(Notation for Conditional Probability)

$Pr(B \mid A)$ represents the probability of the event B occurring after it is assumed that the event A has already occurred .

We read $B \mid A$ as " B given A " or as " event B after that event A has already occurred " .

KTH Matematik

Multiplication Rule

$$Pr\,(\ A \text{ and } B\ ) = Pr(A) \cdot Pr(B \mid A)$$

But also

$$Pr\,(\ A \text{ and } B\ ) = Pr(B) \cdot Pr(A \mid B)$$

Confusing !

KTH VETENSKAP OCH KONST
KTH Matematik

Conditionally Probable Events

Now we have in fact already used the multiplication rule

$$Pr(\text{ 1st pea green pod and 2nd pea yellow pod}) =$$

$$Pr(\text{ 1st pea green pod}) \cdot Pr(\text{2nd pea yellow pod} \mid \text{1st pea green pod }) =$$

$$\frac{8}{14} \cdot \frac{6}{13}$$

KTH VETENSKAP OCH KONST
KTH Matematik

Multiplication Rule

If $Pr(A) > 0$

$$\frac{Pr\left(\ A \text{ and } B\ \right)}{Pr(A)} = Pr(B \mid A)$$

and if $Pr(B) > 0$

$$\frac{Pr\left(\ A \text{ and } B\ \right)}{Pr(B)} = Pr(A \mid B)$$

KTH Matematik

Probability: independent events

$$Pr\,(\ A \text{ and } B\ ) = Pr(A) \cdot Pr(B \mid A)$$

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

$$Pr\,(\ A \text{ and } B\ ) = Pr(A) \cdot Pr(B).$$

Hence, if A and B are independent events

$$P(B \mid A) = Pr(B),$$

i.e., we need not adjust the probability of B to reflect the outcome of A !

KTH Matematik

Health Test

We consider a syndrom of disease and a diagnostic test. With positive result we mean that the test indicates the presence of the syndrom, in case of negative result the test indicates the absence of the syndrom. The table gives the probabilities of various outcomes. To explain it we have for example

$$Pr(\text{ Positive Result and Syndrom }) = 0.81$$

	Positive Result	Negative Result
Syndrom	0.81	0.05
No syndrom	0.03	0.11

KTH Matematik

Health Test

	Positive Result	Negative Result
Syndrom	0.81	0.05
No syndrom	0.03	0.11

$$Pr(\text{ Positive Result} \mid \text{Syndrom }) = \frac{Pr(\text{ Positive Result and Syndrom })}{Pr(\text{ Syndrom })}$$

By the additive law $Pr(\text{ Syndrom }) = 0.81 + 0.05 = 0.86$. Thus

$$Pr(\text{ Positive Result} \mid \text{Syndrom }) = \frac{0.81}{0.86} \approx 0.94$$

KTH Matematik

Health Test

	Positive Result	Negative Result
Syndrom	0.81	0.05
No syndrom	0.03	0.11

$$Pr(\text{ Syndrom } | \text{ Positive Result }) = \frac{Pr(\text{ Positive Result and Syndrom })}{Pr(\text{ Positive Result })}$$

By the additive law $Pr(\text{Positive Result } = 0.81 + 0.03 = 0.84$. Thus

$$Pr(\text{ Syndrom } | \text{ Positive Result }) = \frac{0.81}{0.84} \approx 0.96$$

Hence
$Pr(\text{ Positive Result } | \text{ Syndrom }) \neq Pr(\text{ Syndrom } | \text{ Positive Result }).$

KTH Matematik

Inversion: Th. Bayes 1702 - 1761

We shall next relate $Pr(A \mid B)$ and $Pr(B \mid A)$ using **Bayes' Rule**

KTH Matematik

Bayes' Rule

Bayes' rule goes from probability of $B \mid A$ to probability of $A \mid B$.(Inverts the roles of A and B as conditionally probably events

Bayes' Rule

$$Pr(A \mid B) = \frac{Pr(A)Pr(B \mid A)}{Pr(A)Pr(B \mid A) + Pr(A^*)Pr(B \mid A^*)} .$$

This is a formidable expression, but is relatively easy to use.

KTH VETENSKAP OCH KONST
KTH Matematik

Bayes' Rule

You can check (exercise)

$$Pr(\text{ Syndrom } | \text{ Positive Result }) \approx 0.96$$

by using Bayes' Rule and the probailities above

$$\frac{Pr(\text{ Syndrom })Pr(\text{ Pos Res } | \text{ Syndrom })}{Pr(\text{Syndrom })Pr(\text{ Pos. Res } | \text{ Syndrom }) + Pr(\text{ No Syndrom })Pr(\text{ Pos. Res } | \text{ No Syndrom }}$$

KTH Matematik

Bayes' Rule & Cancer Test

Let us assume that 0.4 % of the population has a form of cancer. In biostatistics 0.4 % is called the **prevalence** of this disease, we write $Pr(\text{ Cancer }) = 0.004$. Biomedical studies have shown that for a diagnostic test $Pr(\text{ Positive Result } | \text{ Cancer }) = 0.995$ and $Pr(\text{ Positive Result } | \text{ No Cancer }) = 0.01$.
We take at random an individual from the population and the diagnostic test gives a positive result. We want to find

$$Pr(\text{ Cancer } | \text{ Positive Result })$$

The rule of the complement gives
$Pr(\text{ No Cancer }) = 1 - Pr(\text{ Cancer }) = 0.996$.

KTH Matematik

Bayes' Rule: Numbers

By Bayes' rule

$$Pr(\text{ Cancer } | \text{ Positive Result }) =$$

$$\frac{Pr(\text{ Canc })Pr(\text{ Pos Res } | \text{ Canc })}{Pr(\text{ Canc })Pr(\text{ Pos. Res } | \text{ Canc }) + Pr(\text{ No Canc })Pr(\text{ Pos. Res } | \text{ No Canc })}.$$

We insert the numbers from the above

$$Pr(\text{ Cancer } | \text{ Positive Result }) = \frac{0.004 \cdot 0.995}{0.004 \cdot 0.995 + 0.996 \cdot 0.01}$$

$$\approx 0.29$$

What does this mean !?

Data $\rightarrow$ *Statistics* $\rightarrow$ *Information/Knowledge*

KTH VETENSKAP OCH KONST

KTH Matematik

Bayes' rule : screening tests

John Allen Paulos: The Math behind Screening Tests. What a positive result really means. Scientific American, *January 2012.*

`http://www.scientificamerican.com/article.cfm?id=weighing-the-positives`

KTH Matematik

Bayes' rule : vocabulary

$$Pr(A \mid B) = \frac{Pr(A)Pr(B \mid A)}{Pr(A)Pr(B \mid A) + Pr(A^*)Pr(B \mid A^*)}.$$

The probability $Pr(A)$ is called the **prior probability** for A. The probability $Pr(A \mid B)$ is called the **posterior probability** for A. Bayes' rule shows the priori probability is transformed to the posterior probability.

Data $\rightarrow$ *Statistics* $\rightarrow$ *Information/Knowledge*

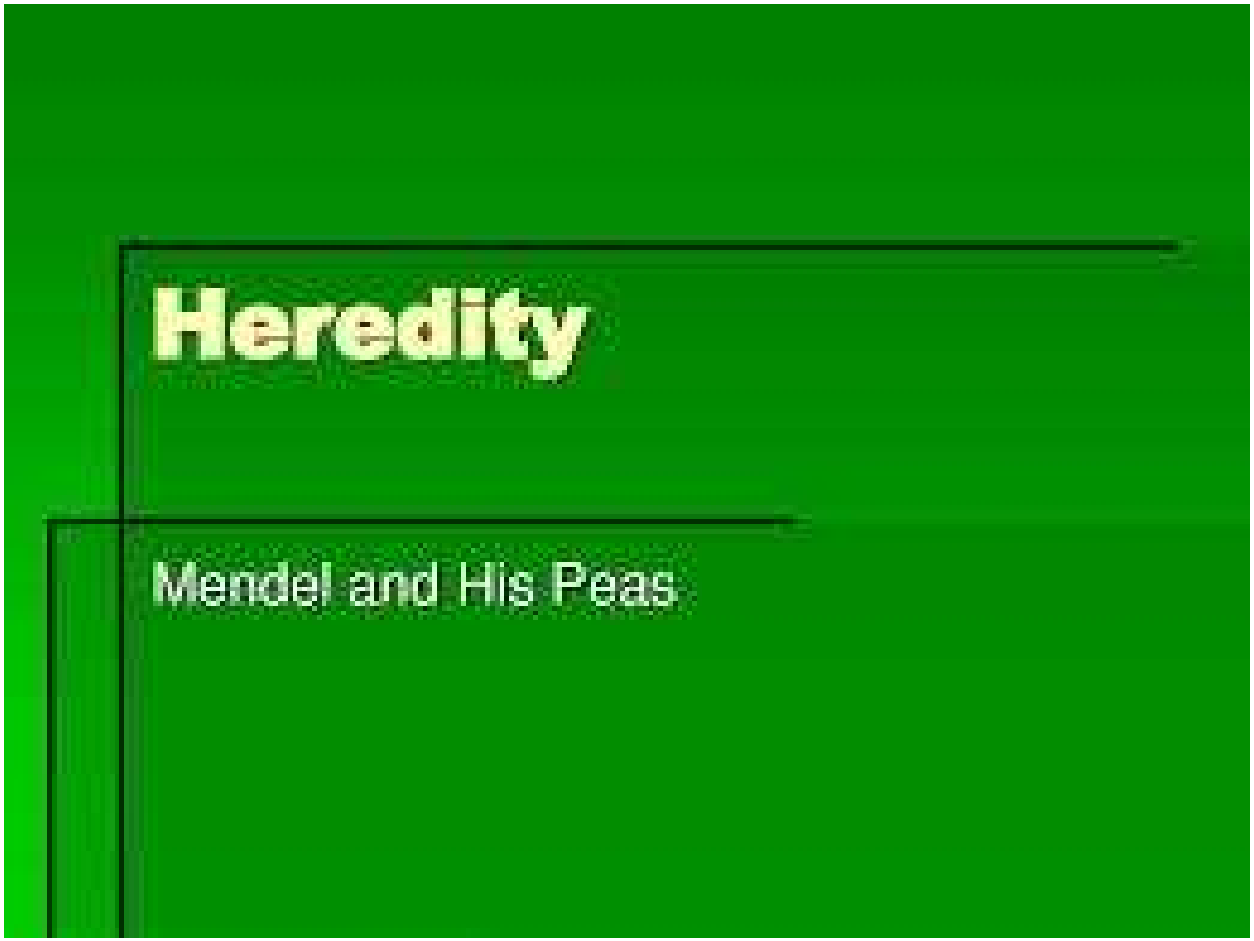
KTH VETENSKAP OCH KONST
KTH Matematik

Bayes' rule & Bioinformatics

Wilkinson, D.J. :Bayesian methods in bioinformatics and computational systems biology. Briefings in bioinformatics, *pp. 109–116, 8, 2007.*

KTH Matematik

Random Variables



KTH Matematik

Random Variables

Recall from Lect. 1:

$$Pr\,(50 \leq \text{the number of matches in a box} \leq 53\ ) = 0.80$$

(here match= tändsticka in Swedish) or

$$Pr\,(\ 11 \text{ matches in two sequences of 15 nucleotides} = 11\ )$$

(here match= något som matchar in Swedish)

KTH Matematik

Random Variables: a technical device

Things become easier if we define

$$X = \text{ the number of matches in a box}$$

and then write

$$Pr\,(50 \leq X \leq 53\ ) = 0.80$$

and

$$Pr\,(\ X = 50\ )$$

or any other probability connected to the counting of matches in the matchbox experiment in Lect. 1..

KTH VETENSKAP OCH KONST
KTH Matematik

Random Variables: Definition

In the same way

$$X = \text{ number of matches in two sequences of 15 nucleotides}$$

and write

$$Pr\left(X = 11\right)$$

Random Variables & Probability Distributions

Definition

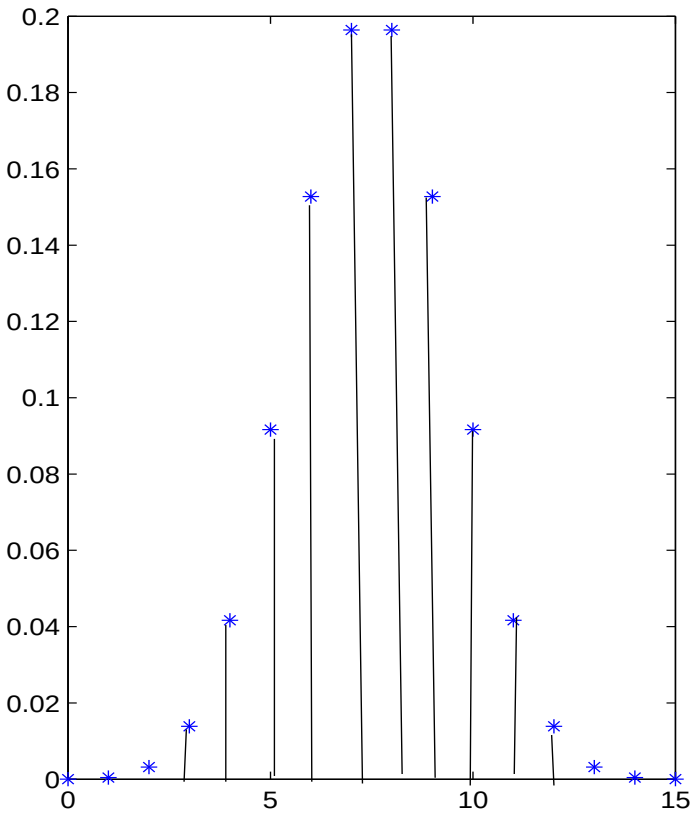
A **random variable** *is a variable represented by X (or Y, Z,...) that has a single numerical value, determined by chance, for each outcome of a procedure.*

Definition

A **probability distribution** *is graph, table, formula or algorithm that gives the probability of each value of the random variable.*

KTH VETENSKAP OCH KONST
KTH Matematik

Binomial probability distribution with $n = 15$ and $p = 0.2$, $p = 0.5$ as a graph



KTH Matematik

Definition

A **probability distribution** *is graph, table, formula or algorithm that gives the probability of each value of the random variable.*

A Formula

$$Pr(X = k), \quad , k = 0, 1, 2, \ldots,$$

Note that

$$0 \leq Pr(X = k) \leq 1, \sum_{k=0}^{\infty} Pr(X = k) = 1.$$

KTH Matematik

Binomial Distribution

Definition

A random variable X has the binomial distribution with the parameters n och p, if

$$Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \text{ för } k = 0, 1, \ldots, n.$$

We write this as $X \sim \text{Bin}(n, p)$ and say that X is $\text{Bin}(n, p)$- distributed.

KTH VETENSKAP OCH KONST
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 Distribution: Example

$X =$ number of matches in two sequences of 15 nucleotides

If we are willing to assume that the nucleotides are random (DNA dice !) and independent, and that the probabilities are the same at each site, then the probability distribution of the number of matches (=successes) in an alignment of two sequences is a binomial distribution with parameters $n = 15$, $p = \frac{1}{4}$.

$$X \sim \text{Bin}(15, 1/4)$$

KTH VETENSKAP OCH KONST
KTH Matematik

Geometric Distribution

If a procedure meets all the conditions of a binomial distribution except that the number of events is not fixed in advance, then the **geometric distribution** can be used.

$$Pr(X = k) = p \cdot (1 - p)^{k-1}, k = 1, 2, \ldots,$$

$X \sim \mathrm{Ge}(p)$.

KTH VETENSKAP OCH KONST
KTH Matematik

Geometric Distribution: an Example

Think of tossing the DNA dice with $p = Pr(G)$. If $X =$ the number of tosses of before you get G for the first time **including** the successful toss. Then $X \sim \mathrm{Ge}(1/4)$ and

$$Pr(X = k) = \frac{1}{4} \cdot \left(\frac{3}{4}\right)^{k-1}, k = 1, 2, \ldots,$$

KTH Matematik

Geometric Distribution: an Example

Usually in statistics p is the probability of success in $\mathrm{Ge}(p)$.
In bioinformatics (e.g., in the theory underlying BLAST, c.f., Lect. 10) $(1-p)$ is the probability of success.

Then

$$Pr(X = k) = p \cdot (1-p)^{k-1}, k = 1, 2, \ldots,$$

is the probability that there were $k-1$ successes (a **success run** *of length $k-1$) before the first failure at event k. X = number of successes plus the first failure.*

KTH Matematik

Binomial Distribution for small p and large n

In bioinformatics it happens often that $X \sim \text{Bin}(n, p)$ with p being small and n being large, or $p = \lambda/n$. Then

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{n(n-1)\dots(n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

$$\approx \frac{\lambda^k}{k!} e^{-\lambda}.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Poisson distribution

Definition

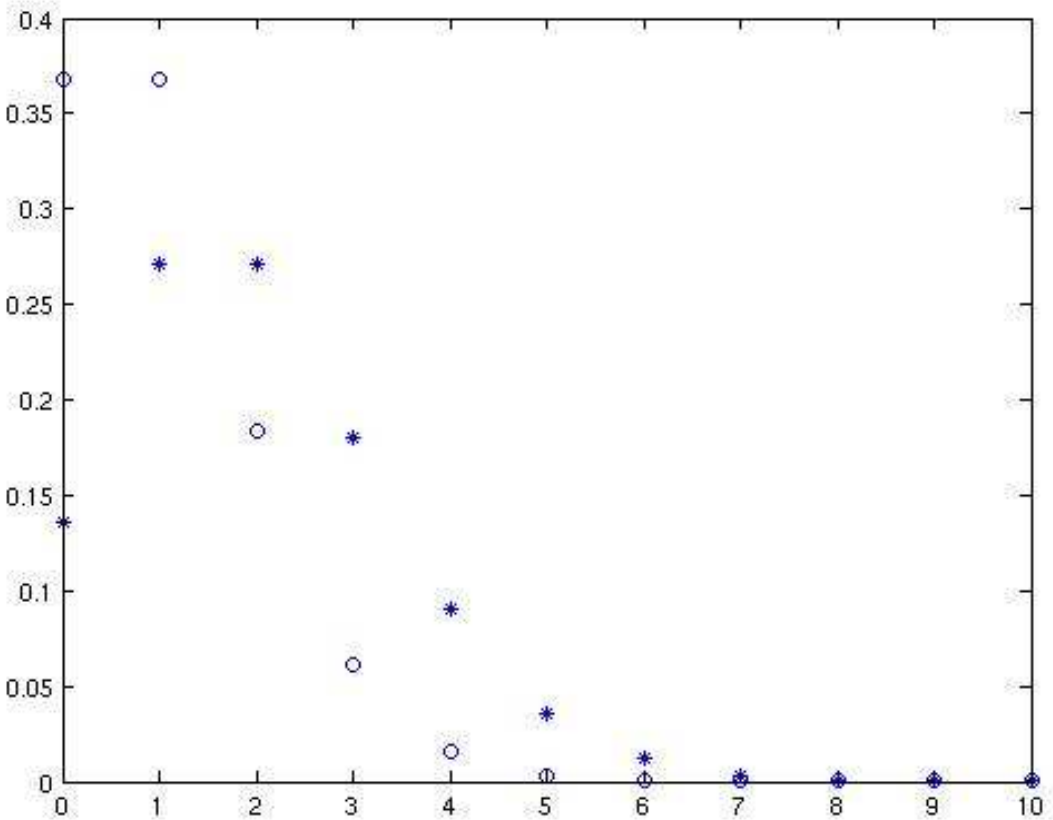
A random variable X has a Poisson distribution with parameter $\lambda > 0$, if

$$Pr(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \text{ för } k = 0, 1, 2 \ldots .$$

We write this with $X \sim \text{Po}(\lambda)$ and say that X is $\text{Po}(\lambda)$-distributed.

KTH Matematik

$Pr(X = k)$ for Po(2) och Po(1)



$* \leftrightarrow \text{Po}(2)$, $\circ \leftrightarrow \text{Po}(1)$

KTH Matematik

Poisson distribution

Poisson distribution applies to occurrences of of some event over a specified " unit ". The random variable X is the number of random occurrences of the event in that " unit ". The " unit " can be time, distance, area, volume or similar.

Example: The number of patients arriving the emergency room on Fridays between 10.00 p.m. and 11.00 p.m.

KTH Matematik

Poisson distribution

Note that

$$\sum_{k=0}^{\infty} Pr(X = k) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda}$$

$$= e^{-\lambda} \underbrace{\sum_{k=0}^{\infty} \frac{\lambda^k}{k!}}_{=e^{\lambda}} = e^{-\lambda} \cdot e^{\lambda} = 1.$$

Poisson distribution & Sequencing

Genomic projects are generally guided by probabilistic models. In shotgun sequencing a large DNA molecule is broken into a collection of *fragments* which are sampled as randomly as possible.
The fragments are cloned and sequenced individually. The sampled fragments are then used for assembly by determining their relative orientations and overlaps and aligned to form a column-by-column matrix.

KTH Matematik

Poisson distribution & Sequencing

Suppose that a collection of fragment sequences $\mathbf{x}^1, \ldots, \mathbf{x}^m$ has been aligned by some procedure. The symbol $\emptyset$ is used as a place holder for nonaligned positions beyond the ends of the fragment. The depth of coverage of position i is defined as the number of fragments contributing sequence information at position i

$$d = \sum_{l=1}^{t} I\left(x_i^l \neq \emptyset\right),$$

where $I\left(x_i^l \neq \emptyset\right) = 1$ if $x_i^l \neq \emptyset$ and zero otherwise. In addition

$$\overline{p} = \sum_{x_j \neq x_l, x_j \in \mathcal{X}} Pr\left(x_i^l = x_j \mid x_l\right),$$

where $Pr\left(x_i^l = x_j \mid x_i\right)$ is the conditional probability that the shotgun sequenced fragment $\mathbf{x}^l$ is equal to x_j in position i, given that the true sequence symbol is x_l. It is assumed that $\overline{p}$ is the same for all x_l and i.

Poisson distribution & Sequencing

For a position covered at depth d, we compute the probability that at most one half of the bases are correct. This is the binomial probability given by

$$Pr\left(\text{error}|d\right) = \sum_{k=0}^{\lfloor d/2 \rfloor} \left(\begin{array}{c} d \\ k \end{array} \right) \overline{p}^{d-k} (1-\overline{p})^{k},$$

where the probability of success is taken as $1-\overline{p}$ and $\lfloor d/2 \rfloor$ is largest integer smaller than or equal to $d/2$.

KTH Matematik

Poisson distribution & Sequencing

In shotgun sequencing the distribution of d can be taken as $\mathrm{Po}(\lambda)$. Positions not covered by any fragments are ignored and thus d is at least 1. Then we obtain

$$Pr(\text{error}) = \frac{1}{1 - e^{-\lambda}} \sum_{d=1}^{\infty} \frac{\lambda^d}{d!} e^{-\lambda} Pr(\text{error}|d) \tag{1}$$

as a practical measure of the accuracy of shotgun sequencing.

KTH Matematik

Poisson distribution & Sequencing

The strategy for the shotgun approach to sequencing follows

Lander E.S., Waterman M.S. Genomic mapping by fingerprinting random clones: a mathematical analysis *Genomics 2 (3): 231-239 (1988)*

The "Lander-Waterman model" provides statistical estimates for the read depth in a whole genome shotgun (WGS) sequencing experiment via the Poisson approximation to the Binomial distribution. Although originally intended for estimating the redundancy when mapping by fingerprinting random clones, the Lander-Waterman model has served as an essential tool for estimating sequencing requirements for modern WGS experiments.

KTH Matematik

Discrete Data, Discrete Random Variables

Binomial, Geometric and Poisson distributions are examples of probability distributions for discrete data (represented by respective discrete random variables)

A **discrete random variable** *has either a finite number of values or countable number of values, where " countable " refers to the fact that there might be infinitely many values, but they can be associated with a counting process.*

KTH Matematik

Mean and Variance of Discrete Random Variables

The **mean** of a discrete random variable X is denoted by $E(X)$ and is defined by

$$E(X) \stackrel{\text{def}}{=} \sum_{k=0}^{\infty} kPr(X = k).$$

The **variance** of a discrete random variable X is denoted by $V(X)$ and is defined by

$$V(X) \stackrel{\text{def}}{=} \sum_{k=0}^{\infty} (k - E(X))^2 Pr(X = k).$$

The mean is a measure of the central tendency of the random variable and the variance is a measure of dispersion

KTH VETENSKAP OCH KONST
KTH Matematik

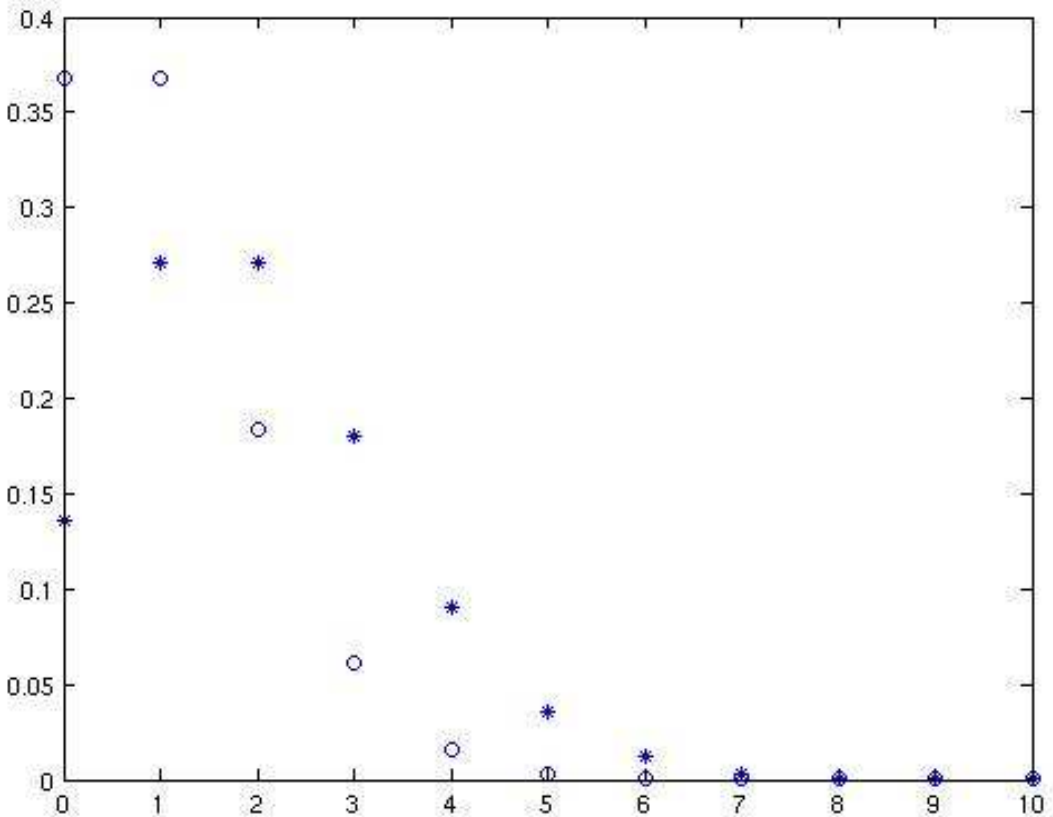
Mean and Variance of Discrete Random Variables

- $X \sim \text{Bin}(n, p)$: $E(X) = np$, $V(X) = np(1-p)$
- $X \sim \text{Ge}(p)$: $E(X) = 1/p$, $V(X) = (1-p)/p^2$
- $X \sim \text{Po}(\lambda)$: $E(X) = \lambda$, $V(X) = \lambda$

KTH VETENSKAP OCH KONST
KTH Matematik

$Pr(X = k)$ for Po(2) och Po(1)

We see that λ influences the location and the shape of the mass of



probability.

$* \leftrightarrow \text{Po}(2)$, $\circ \leftrightarrow \text{Po}(1)$

KTH Matematik

Continuous Random Variables

A **continuous random variable** *has infinitely many values, and those values can be associated with measurements on a continuous scale without gaps or interruptions.*

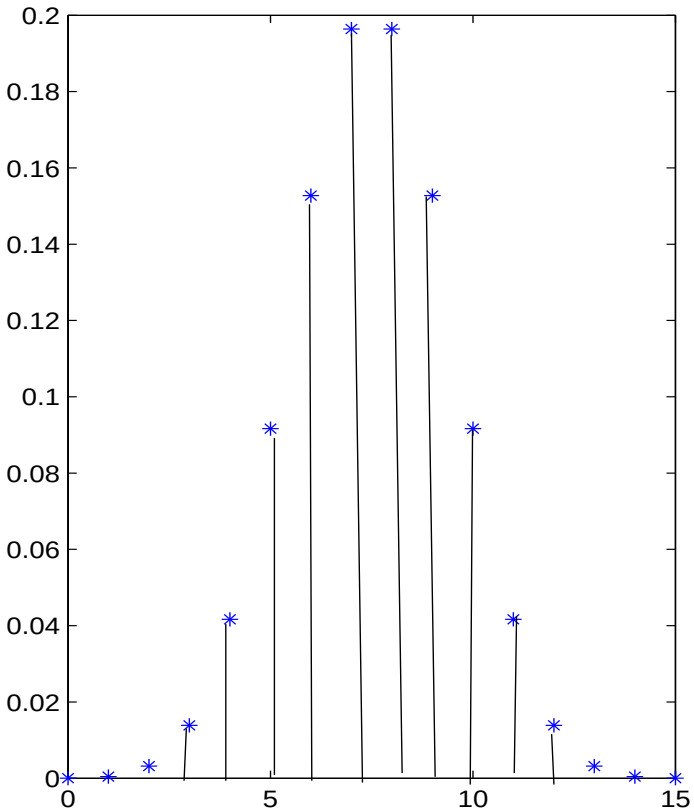
Continuous random variables can often be obtained as approximative forms of discrete distributions in a natural way.

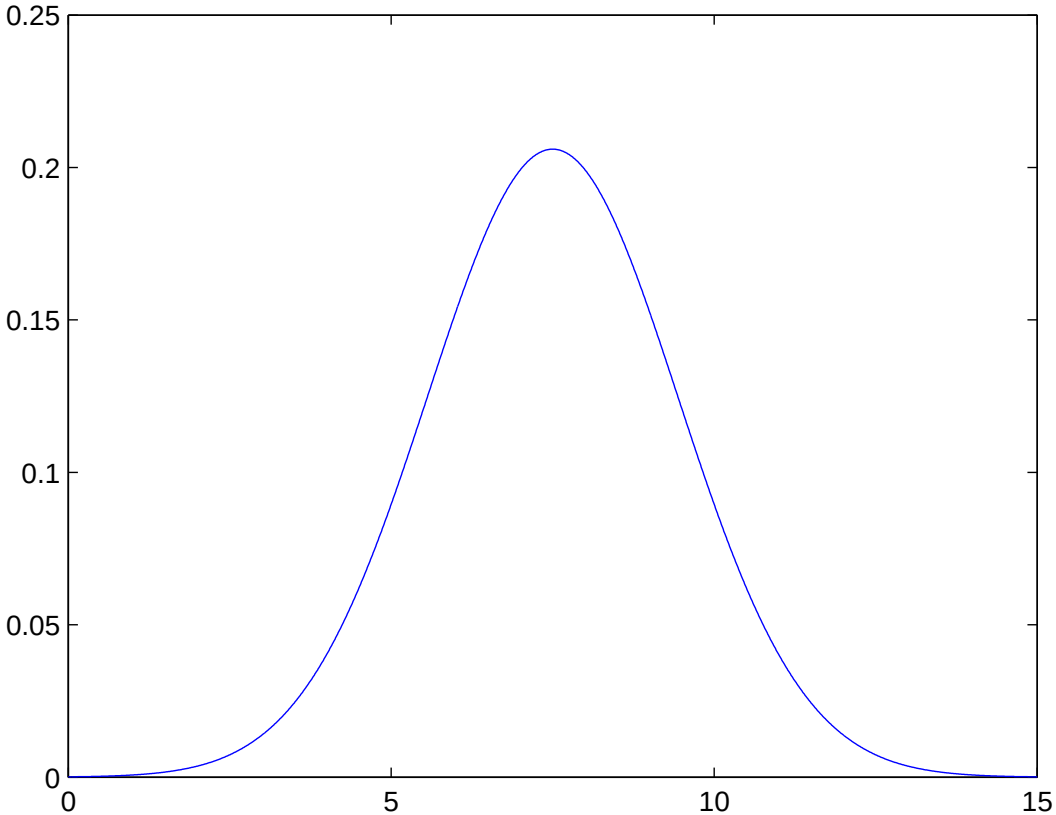
KTH Matematik

Binomial distribution: approximation

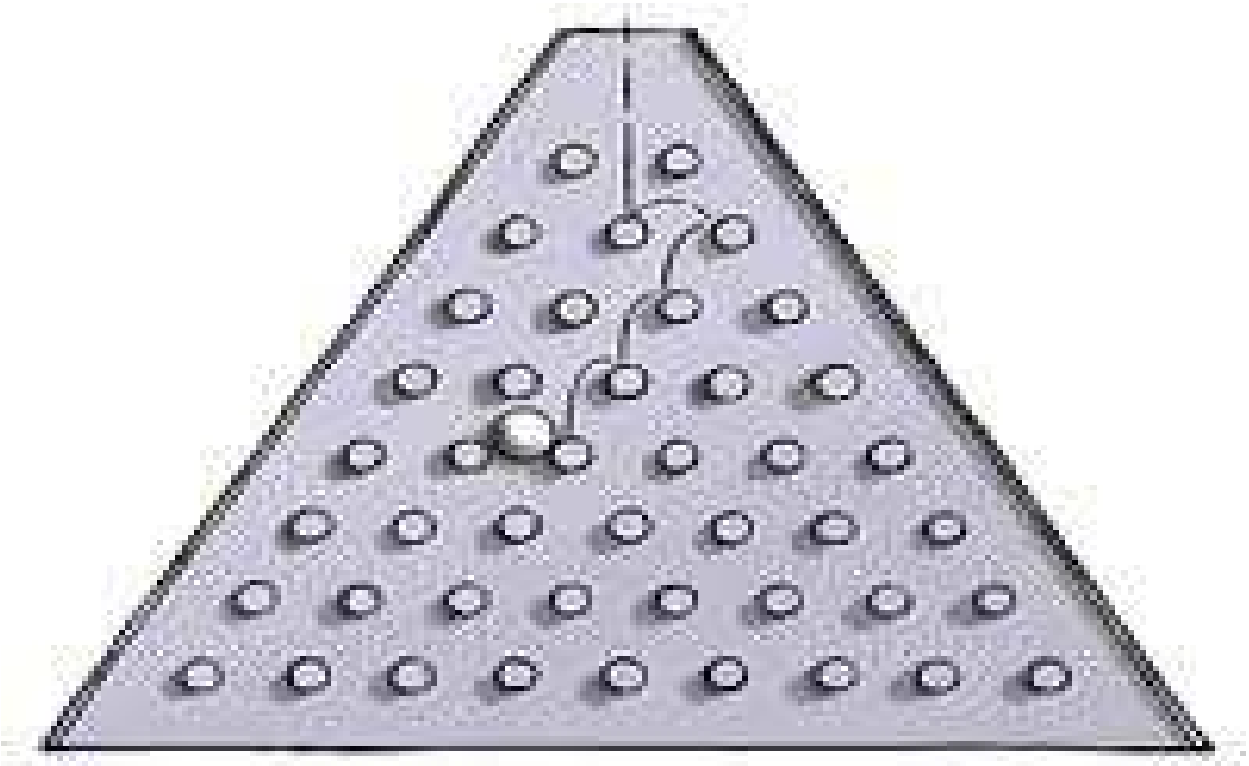
For $p = 0.5$ and $n = 15$ the Binomial distribution with mean $15 \cdot 0.5 = 7.5$, and variance $15 \cdot 0.5 \cdot 0.5 = 3.75$. The curve in the right hand figure $\mu = 7.5$ and $\sigma^2 = 3.75$.

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





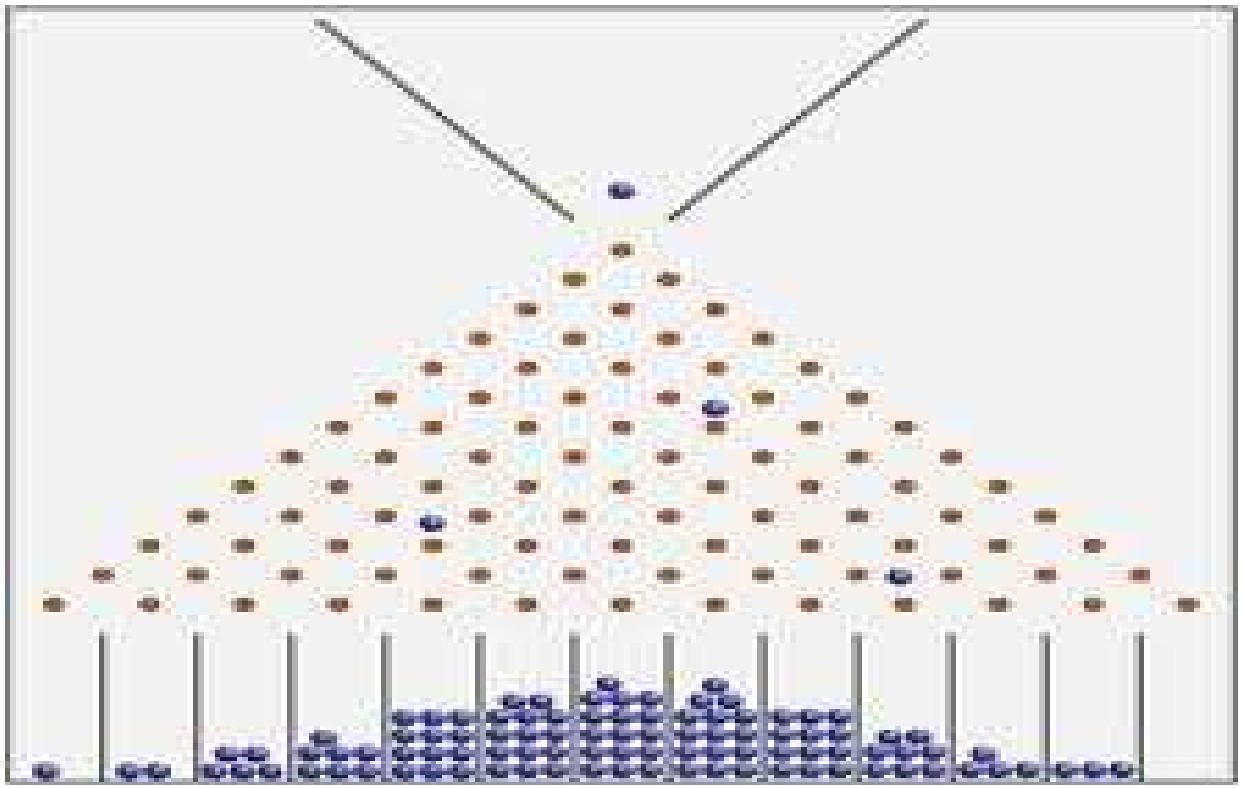
The Galton Board



KTH Matematik

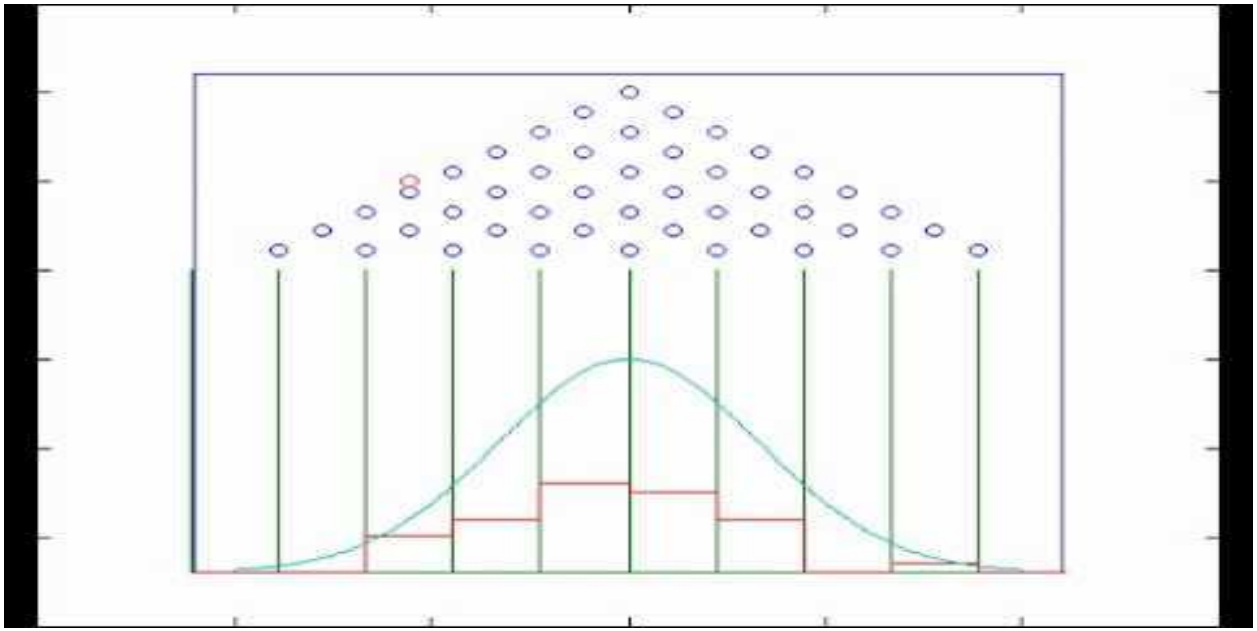
The Galton Board Experiment

The Galton board experiment consists of performing n trials with probability of success p. The trial outcome are represented graphically as a path in the Galton board: success corresponds to a bounce to the right and failure to a bounce to the left.



KTH Matematik

The Galton Board Experiment: Bell Shape

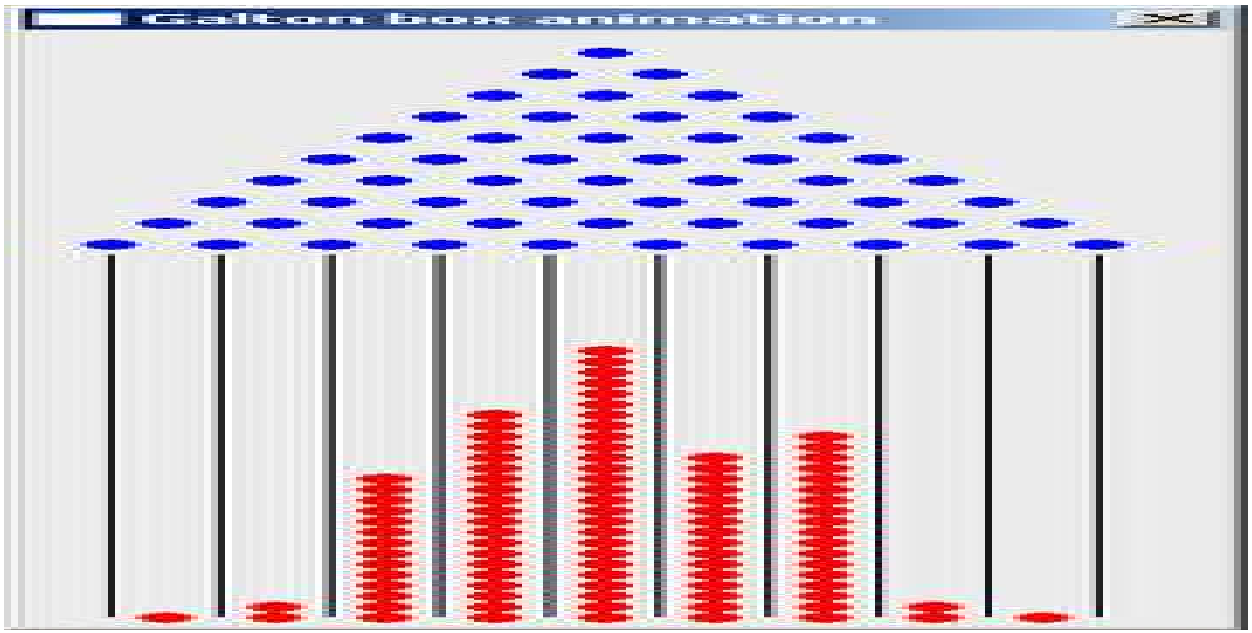


KTH Matematik

The Galton Board Experiment

An applet for the Galton board experiment

`http://www.math.uah.edu/stat/applets/GaltonBoardExperiment.html`



KTH Matematik

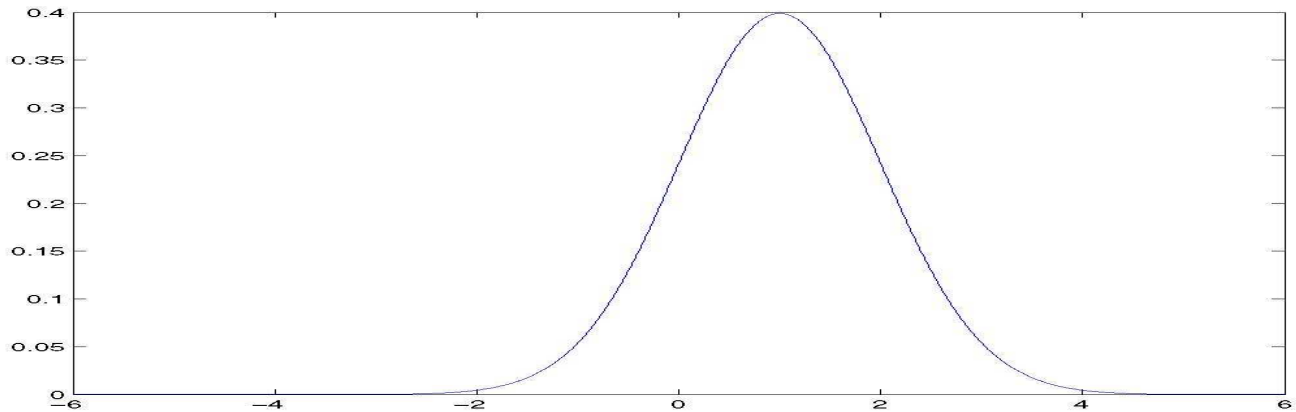
Normal (Probability) Density Curve, the Bell Shape

We call the function $f(x)$, $\sigma > 0$,

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

the **normal (probability) density** *or the* **normal (probability) density curve**.

In the figure $\mu = 1, \sigma = 1$



KTH Matematik

Normal Density Curve

We have

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

i.e.,

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx = 1.$$

KTH Matematik

A Density Curve

A density curve $f(x)$ is graph of a continuous probability distribution. It must satisfy the following properties

- Total area under the curve must equal 1:

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

KTH Matematik

A Density Curve

A density curve $f(x)$ is graph of a continuous probability distribution. It must satisfy the following properties

- Total area under the curve must equal 1:

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

- Every point on the curve must have a vertical height that is 0 or greater (that is, the curve cannot fall under the x-axis.

KTH Matematik

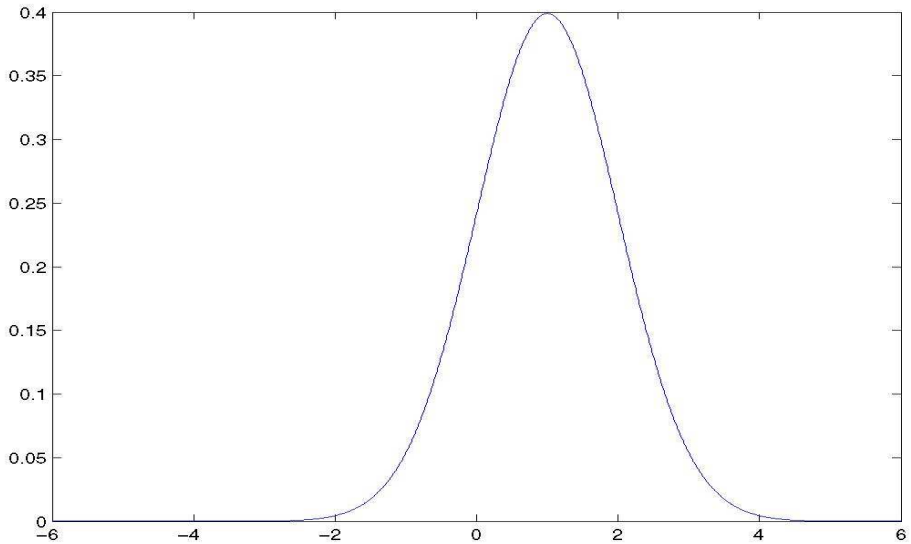
A Density Curve

A density curve $f(x)$ is graph of a continuous probability distribution. It must satisfy the following properties

- Total area under the curve must equal 1:

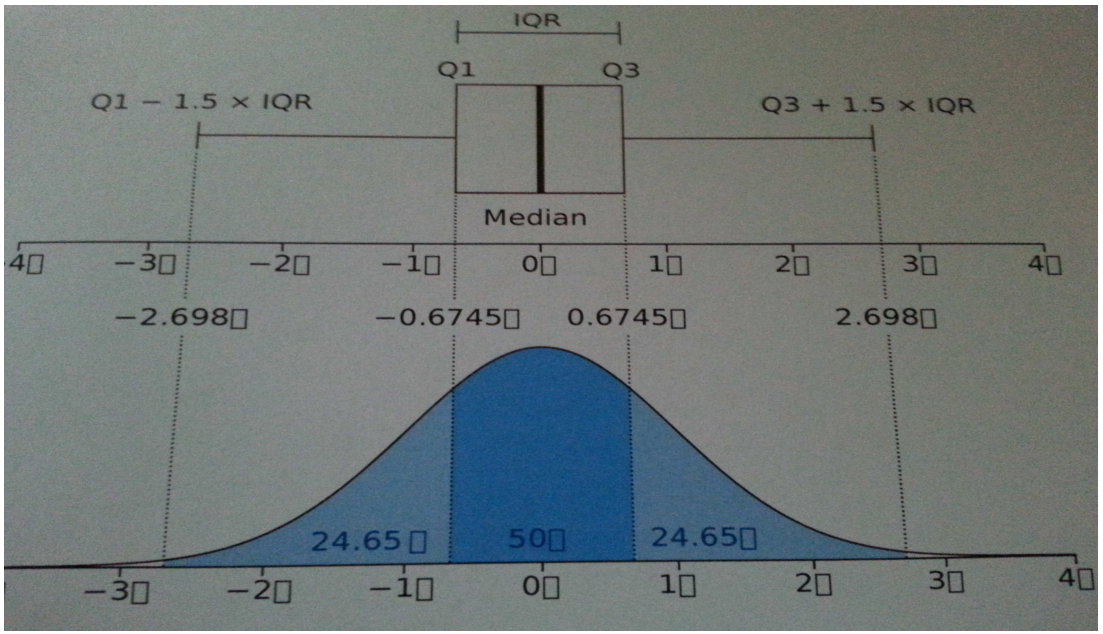
$$\int_{-\infty}^{\infty} f(x)dx = 1$$

- Every point on the curve must have a vertical height that is 0 or greater (that is, the curve cannot fall under the x-axis.



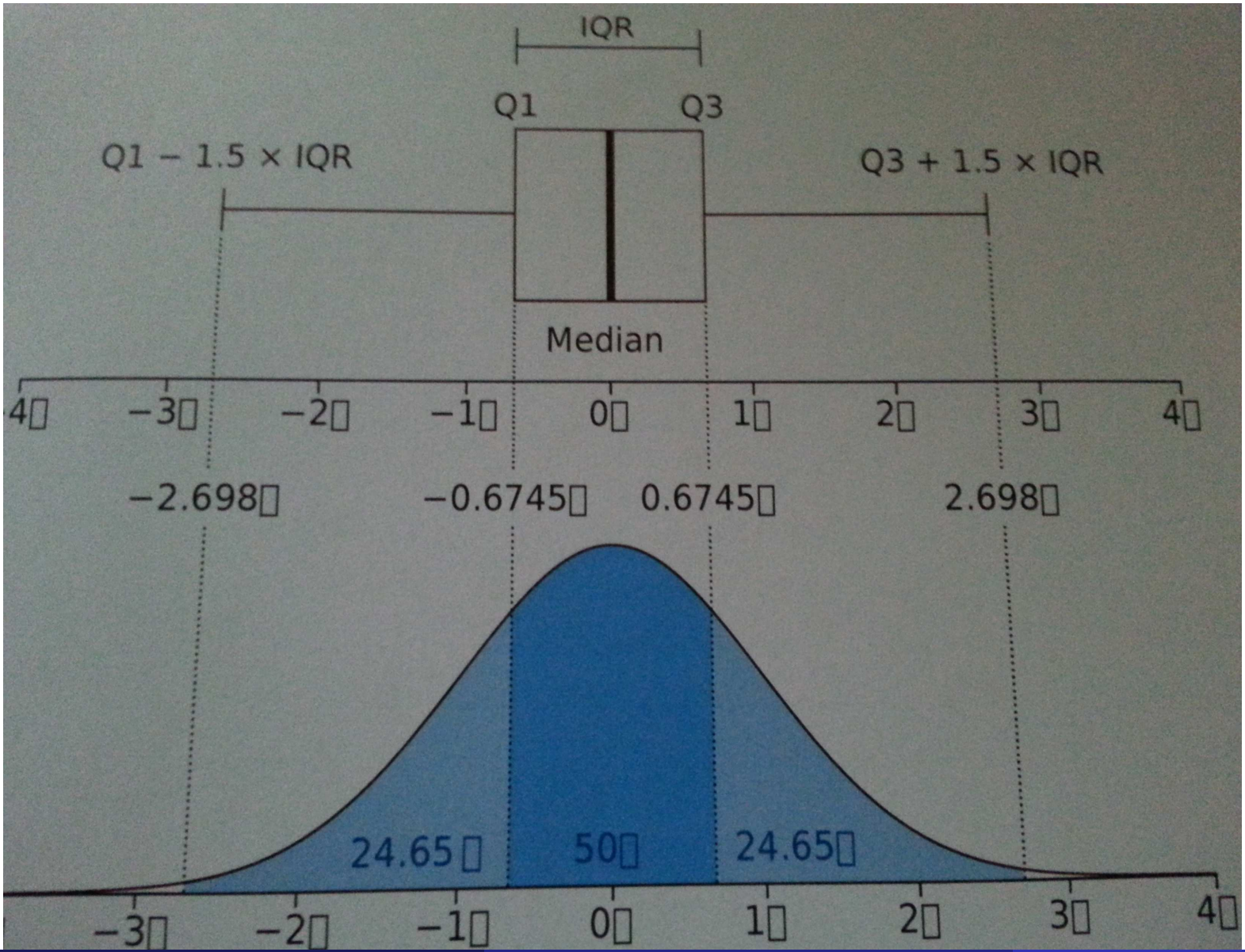
A Density Curve

Because the total area under a density curve is equal to 1, there is correspondence between area and probability.



A Density Curve

In this figure the symbol □ should be read as σ.



Probability as Area under Density Curve

X is continuous and represented by a density curve $f(x)$. Then

$$Pr(a < X \leq b) = \int_a^b f(x)dx$$

Normal Distribution (a.k.a. Gaussian Distribution)

Definition

A continuous random variable X is said to have the $N(\mu, \sigma)$- distribution, where $\sigma > 0$, if

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

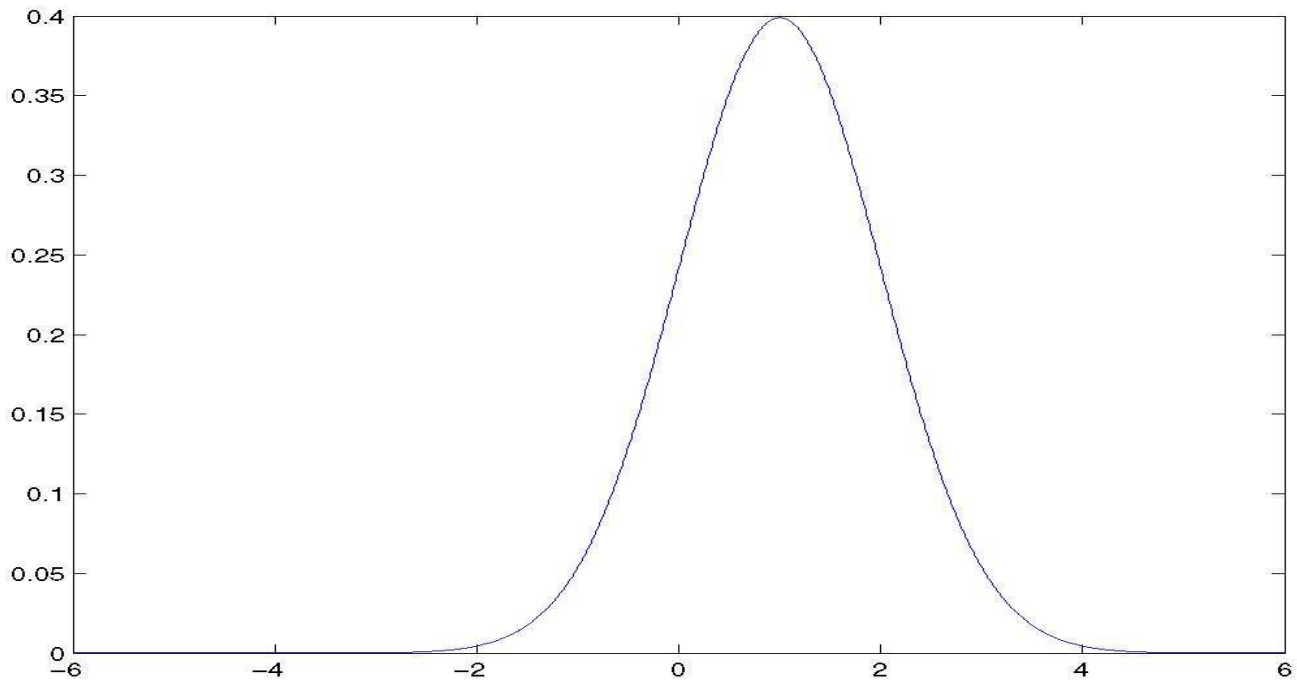
We write $X \sim N(\mu, \sigma)$.

KTH VETENSKAP OCH KONST
KTH Matematik

Normal Distribution (a.k.a. Gaussian Distribution)

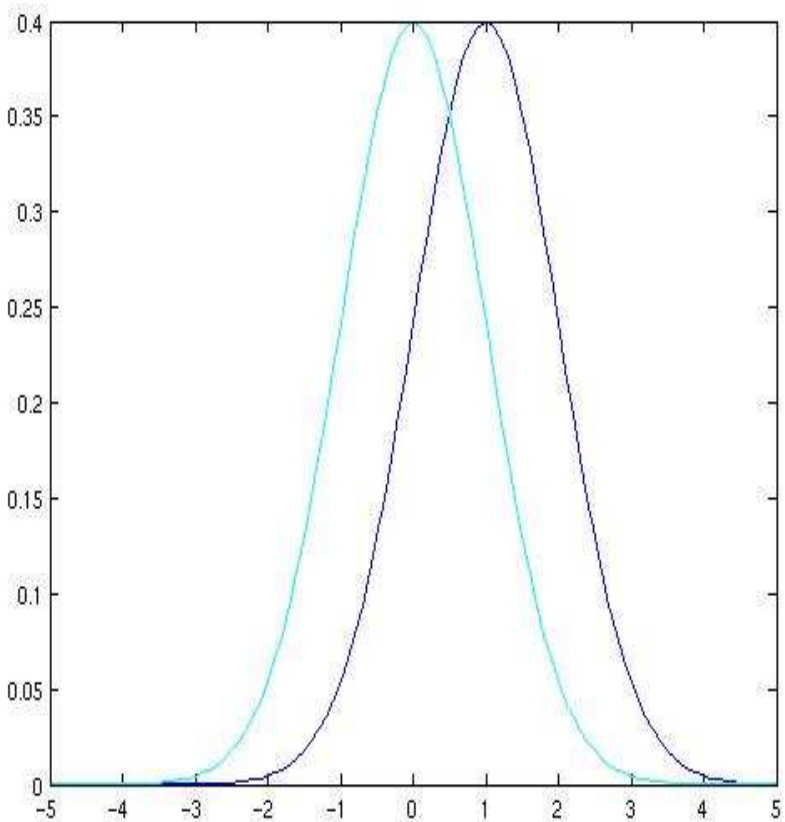
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

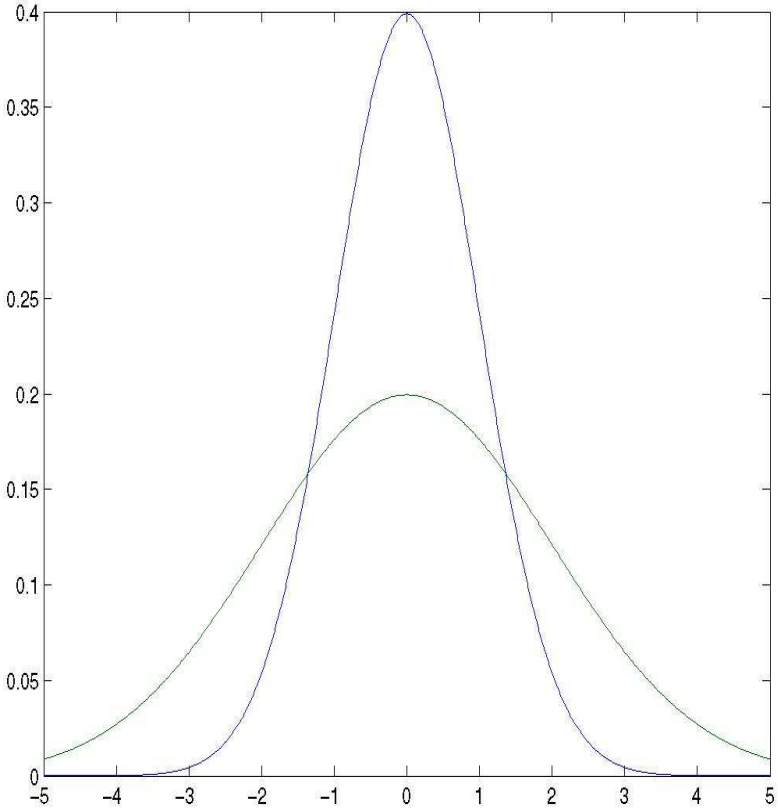
In the figure $\mu = 1, \sigma = 1$



KTH Matematik

$N(0,1)$ och $N(1,1)$ och $N(0,1)$ och $N(0,2)$ (from left to the right)

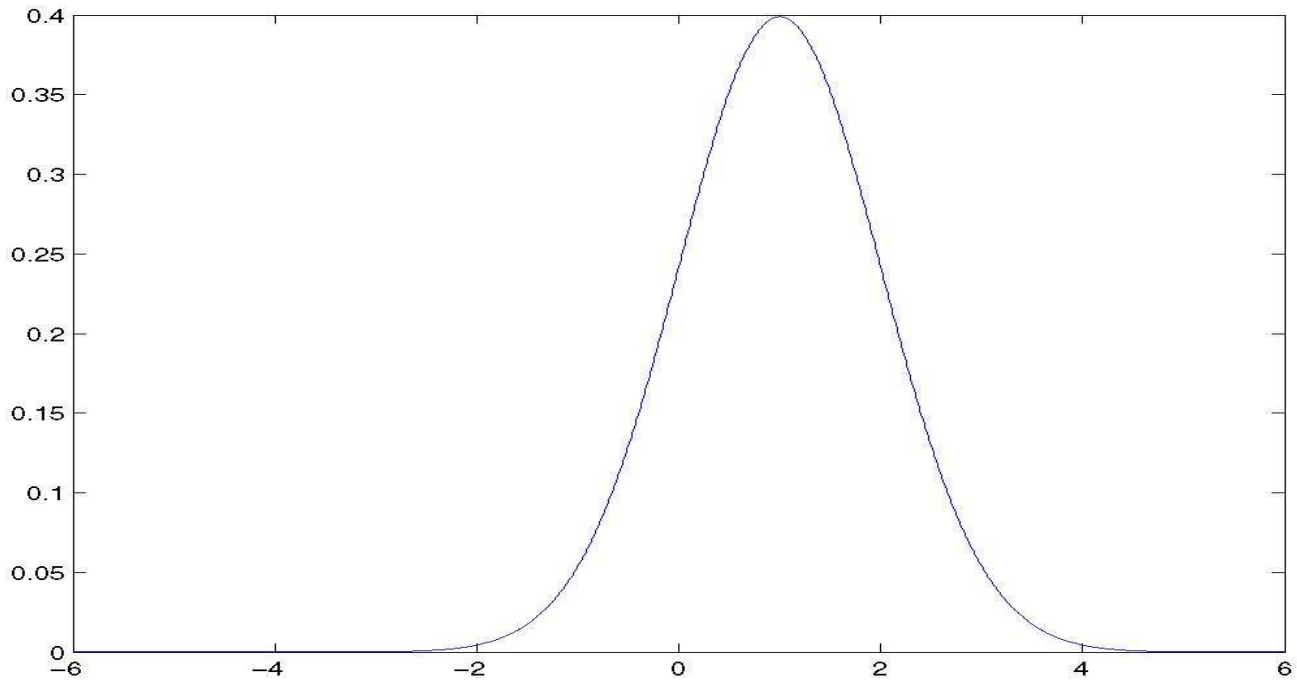




KTH Matematik

Normal Distribution (a.k.a. Gaussian Distribution)

In the figure $\mu = 1, \sigma = 1$



KTH Matematik

Normal Distribution: Probability as Area under Density Curve

$X \sim N(\mu, \sigma)$. Then

$$Pr(a < X \leq b) = \int_a^b f(x)dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-(x-\mu)^2/2\sigma^2} dx$$

This integral must be evaluated by calculator, computer software or old-fashioned tables.

KTH Matematik

Normal Distribution: Probability as Area under Density Curve

In fact the method of tables corresponds to computing values of

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{(a-\mu)/\sigma}^{(b-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

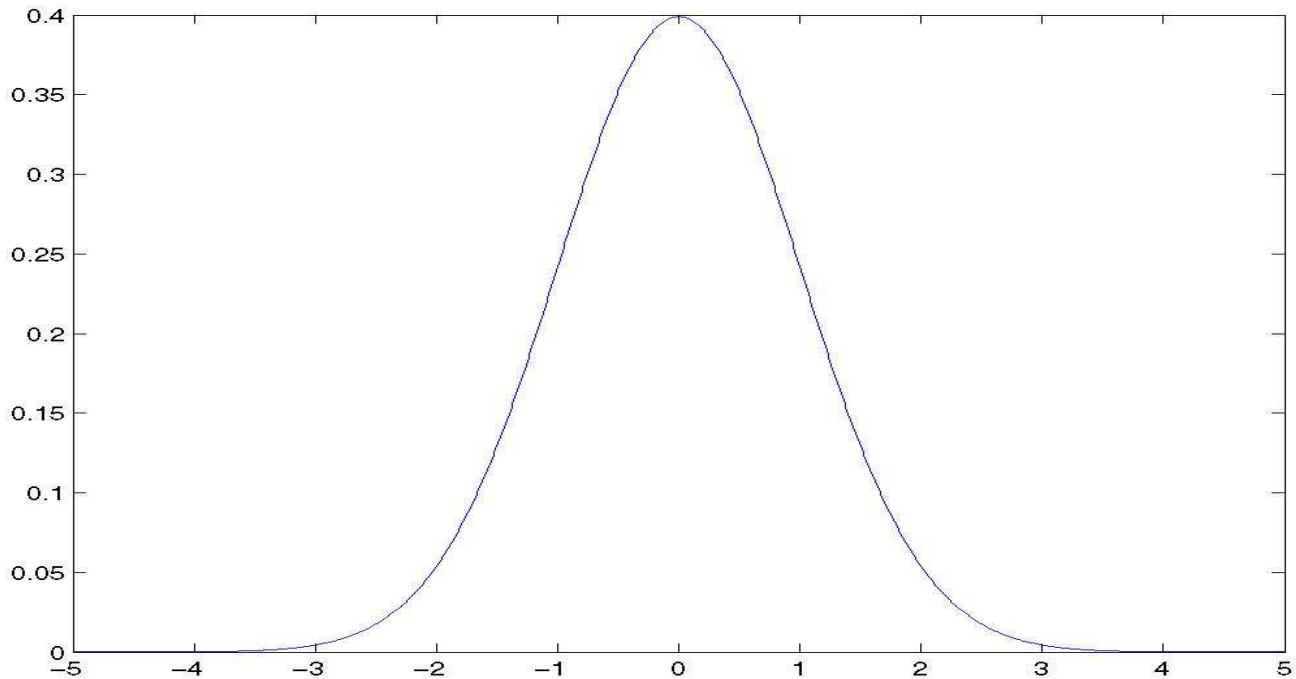
How is this ?

Standard Normal distribution, $N(0,1)$

The distribution $N(0,1)$ density

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

is called standard normal distribution



KTH VETENSKAP OCH KONST
KTH Matematik

Standard Normal distribution, $N(0,1)$

Definition

En random variable Z is said to have standardizerad normal distribution if $Z \sim N(0,1)$, i.e., if it has the density/density curve

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

We have

$$Pr(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx.$$

$$Pr(Z \leq 0) = \frac{1}{2}.$$

KTH Matematik

A Table for $Pr(Z \leq z)$

Standard normal, cumulative density, $P(Z<z)$

	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07
0	0.500	0.504	0.508	0.512	0.516	0.520	0.524	0.52
0.1	0.540	0.544	0.548	0.552	0.556	0.560	0.564	0.56
0.2	0.579	0.583	0.587	0.591	0.595	0.599	0.603	0.60
0.3	0.618	0.622	0.626	0.629	0.633	0.637	0.641	0.64
0.4	0.655	0.659	0.663	0.666	0.670	0.674	0.677	0.68
0.5	0.691	0.695	0.698	0.702	0.705	0.709	0.712	0.71
0.6	0.726	0.729	0.732	0.736	0.739	0.742	0.745	0.74
0.7	0.758	0.761	0.764	0.767	0.770	0.773	0.776	0.77
0.8	0.788	0.791	0.794	0.797	0.800	0.802	0.805	0.80
0.9	0.816	0.819	0.821	0.824	0.826	0.829	0.831	0.83
1	0.841	0.844	0.846	0.848	0.851	0.853	0.855	0.85
1.1	0.864	0.867	0.869	0.871	0.873	0.875	0.877	0.87
1.2	0.885	0.887	0.889	0.891	0.893	0.894	0.896	0.89
1.3	0.903	0.905	0.907	0.908	0.910	0.911	0.913	0.91

KTH VETENSKAP OCH KONST

KTH Matematik

Normal Distribution: Probability as Area under Density Curve

$X \sim N(\mu, \sigma)$. Set

$$Z = \frac{X - \mu}{\sigma}.$$

Then $Z \sim N(0, 1)$ and thus

$$Pr(a < X \leq b) = Pr(\frac{a - \mu}{\sigma} < Z \leq \frac{b - \mu}{\sigma}) =$$

$$= \frac{1}{\sqrt{2\pi}} \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} e^{-x^2/2} dx$$

KTH Matematik

Normal Distribution: Probability as Area under Density Curve

$X \sim N(1,1)$. Set

$$Z = \frac{X-1}{1}.$$

Then

$$Pr(X \leq 1.1) = Pr(Z \leq 1.1-1) = Pr(Z \leq 0.1) = 0.54$$

Standard normal, cumulative density, P(Z<z)

	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07
0	0.500	0.504	0.508	0.512	0.516	0.520	0.524	0.52
0.1	0.540	0.544	0.548	0.552	0.556	0.560	0.564	0.56
0.2	0.579	0.583	0.587	0.591	0.595	0.599	0.603	0.60
0.3	0.618	0.622	0.626	0.629	0.633	0.637	0.641	0.64
0.4	0.655	0.659	0.663	0.666	0.670	0.674	0.677	0.68
0.5	0.691	0.695	0.698	0.702	0.705	0.709	0.712	0.71
0.6	0.726	0.729	0.732	0.736	0.739	0.742	0.745	0.74
0.7	0.758	0.761	0.764	0.767	0.770	0.773	0.776	0.77
0.8	0.788	0.791	0.794	0.797	0.800	0.802	0.805	0.80
0.9	0.816	0.819	0.821	0.824	0.826	0.829	0.831	0.83

KTH VETENSKAP OCH KONST
KTH Matematik

Exponential distribution $\mathrm{Exp}(\lambda)$

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

$$f(x) = \begin{cases} \lambda\, e^{-\lambda x} & \text{för } x \geq 0, \\ 0 & \text{för } x < 0. \end{cases}$$

KTH VETENSKAP OCH KONST
KTH Matematik

Exponential distribution Exp(λ)

$X \in \text{Exp}(\lambda)$. Then for $x > 0$

$$Pr(X \leq x) = \int_0^x f(u)du = \int_0^x \lambda\, e^{-\lambda u} du = 1 - e^{-\lambda x}.$$

By the law of the complement

$$Pr(X \geq x) = 1 - Pr(X \leq x) = e^{-\lambda x}.$$

In bioinformatics (sequence alignment) one often approximates the distribution of the alignment score S by the exponential distribution so that

$$Pr(S \geq x) = e^{-\lambda x}.$$

KTH Matematik

End of Lecture 2