Bioinformatics and Biostatistics BB2440: Biostatistics Lecture 1 :Fundamental Probability

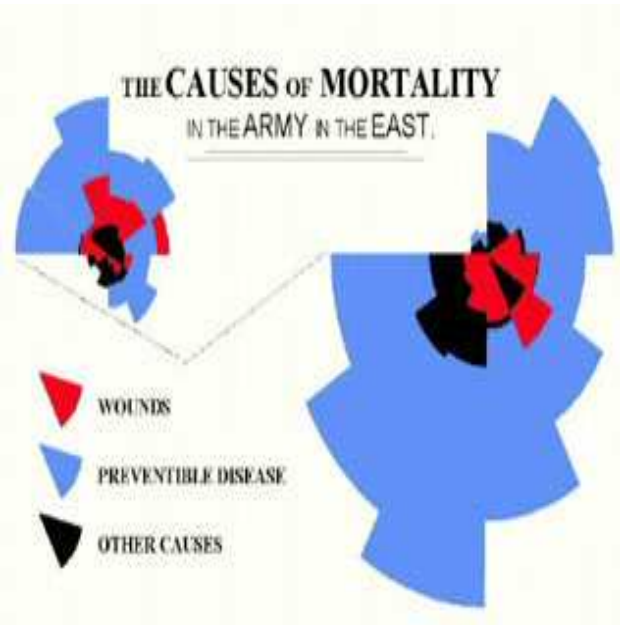
Timo Koski

TK

03.09.2013

KTH Matematik

Many people associate the word statistics [1] with diagrams, tables and charts. Here we mean by *statistics* the study of how one draws conclusions from observed data and/or draws correct conclusions under uncertainty and describes these conclusions appropriately.



[1] Florence Nightingale invented a form of the pie chart in order to illustrate seasonal sources of patient mortality in the military field hospital she managed.

From Data to Knowledge

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

Statistics is a tool for creating new understanding from a set of numbers

KTH VETENSKAP OCH KONST
KTH Matematik

Outline of Lecture 1.

- Absolute and Relative Frequencies, Histogram, Mean, Variance, Standard Deviation

KTH Matematik

Outline of Lecture 1.

- Absolute and Relative Frequencies, Histogram, Mean, Variance, Standard Deviation
- Boxplot

KTH Matematik

Outline of Lecture 1.

- Absolute and Relative Frequencies, Histogram, Mean, Variance, Standard Deviation
- Boxplot
- Empirical Probability, Theoretical Probability, Meaning of Probability

KTH Matematik

Outline of Lecture 1.

- Absolute and Relative Frequencies, Histogram, Mean, Variance, Standard Deviation
- Boxplot
- Empirical Probability, Theoretical Probability, Meaning of Probability
- Rules of Probability

KTH Matematik

Outline of Lecture 1.

- Absolute and Relative Frequencies, Histogram, Mean, Variance, Standard Deviation
- Boxplot
- Empirical Probability, Theoretical Probability, Meaning of Probability
- Rules of Probability
- Independent Events

KTH Matematik

Outline of Lecture 1.

- Absolute and Relative Frequencies, Histogram, Mean, Variance, Standard Deviation
- Boxplot
- Empirical Probability, Theoretical Probability, Meaning of Probability
- Rules of Probability
- Independent Events
- Binomial Probability Distribution

KTH VETENSKAP OCH KONST
KTH Matematik

Statistics

We check 35 boxes of matches and check in each case how many matches it contains.
Obviously this might be of importance for a producer of matches in industrial scale, but you might think that this is of no interest for you (counting matches in box is of no special hobby of mine either). This is, however, not the point. This is a simple example of observed data, quite clear for everyone, and to introduce the concepts of probability.

KTH Matematik

Observed Data

We obtain 35 boxes of matches and check in each case how many matches it contains. Result:

51	52	49	51	52	51	53
52	48	52	50	53	49	50
51	53	51	52	50	51	53
53	55	50	49	53	50	51
51	52	48	53	50	49	51

KTH Matematik

Frequencies

The table above can be summarized by the table of *frequencies*

- *absolute frequency* f_i for the observed values
- *relative frequency* $p_i = f_i / n$. ($i = 1$ lowest class, $i = 2$ next class e.t.c..) $n =$ number of data (=35).

KTH Matematik

Table of frequencies

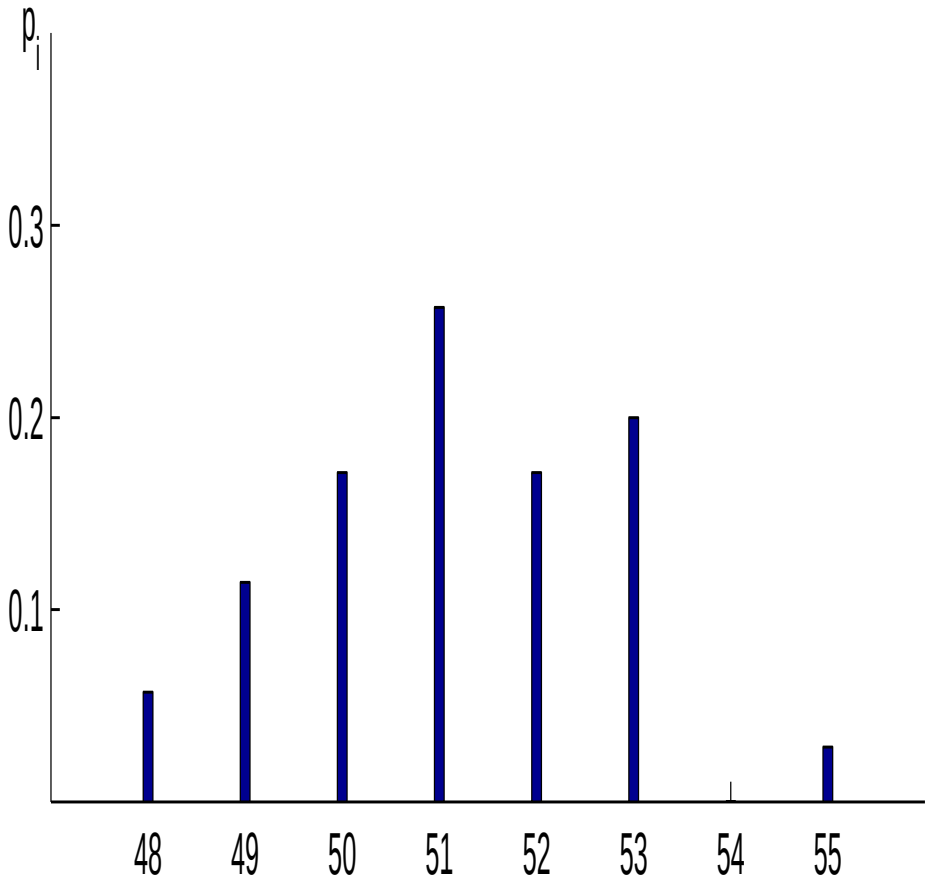
Table: Frequencies for the observed numbers of matches.

Class i	Absolute frequency f_i	Relative frequency (%) $100\, p_i$
48	2	5.7
49	4	11.4
50	6	17.1
51	9	25.7
52	6	17.1
53	7	20.0
54	0	0.0
55	1	2.9
Sum	35	100.0

KTH VETENSKAP OCH KONST
KTH Matematik

Histogram

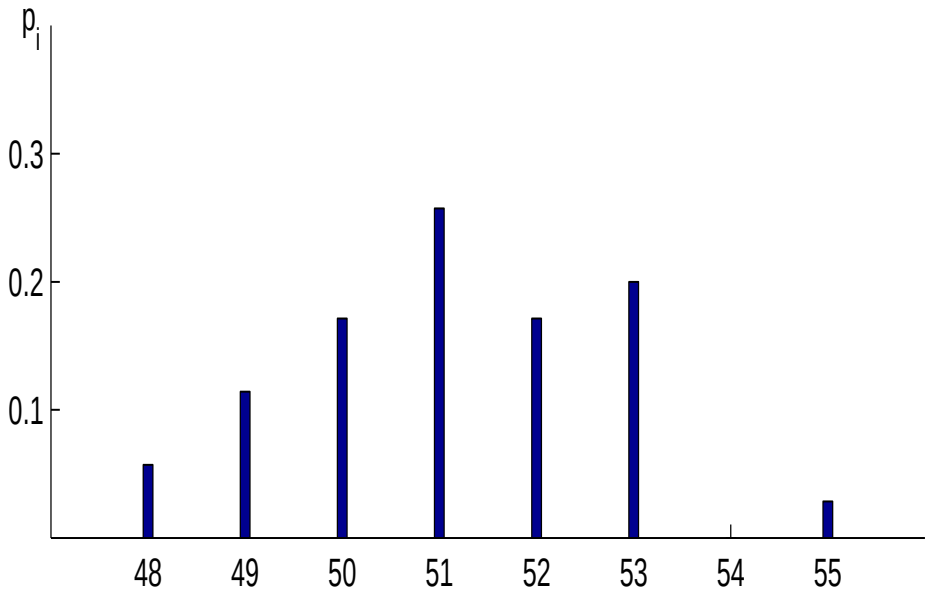
The data becomes easier to grasp if we plot the relative frequencies p_i in *histogram*.



KTH Matematik

Histogram for relative frequencies

A **histogram** *is a bar graph in which the horizontal scale represents classes of data values and the vertical scale represents the relative frequencies. The heights of the bars correspond to the relative frequency values.*



KTH Matematik

A Measure of Central Tendency: The Mean

Let $x_1, \ldots, x_n$ denote the data to be analysed. As a measure of central tendency one often uses the arithmetic *mean*

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

In the matchbox example we find $\sum_{j=1}^{35} x_j = 1789$ and $\bar{x} = 1789/35 = 51.1143$. This gives us the *typical value* of the data.

KTH Matematik

Measure of dispersion: Variance, Standard deviation

A measure of dispersion refers to how closely the data cluster around the measure of central tendency. Here we consider the *variance*

$$s^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2$$

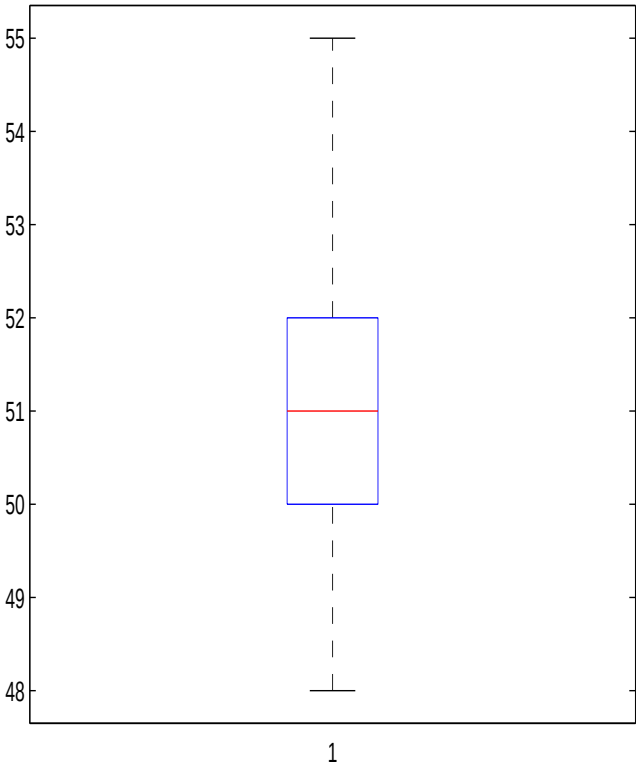
or *standard deviation*

$$s = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2},$$

We take the square root to get back to the original units.
In matchbox example we get the standard deviation $s \approx 1.62$.

KTH Matematik

BOXPLOT for the matchbox data

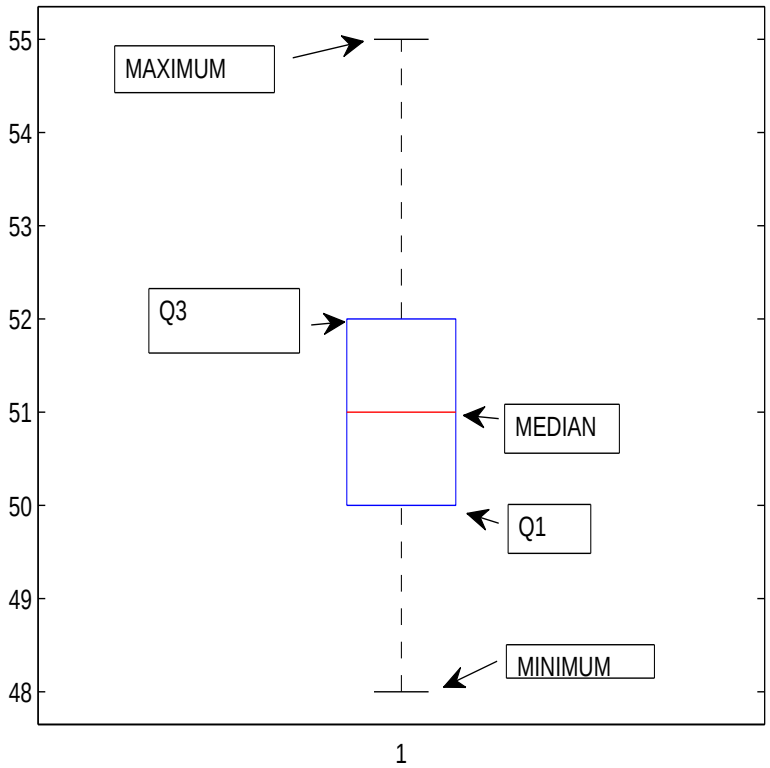


51	52	49	51	52	51	53
52	48	52	50	53	49	50
51	53	51	52	50	51	53
53	55	50	49	53	50	51
51	52	48	53	50	49	51

KTH VETENSKAP OCH KONST
KTH Matematik

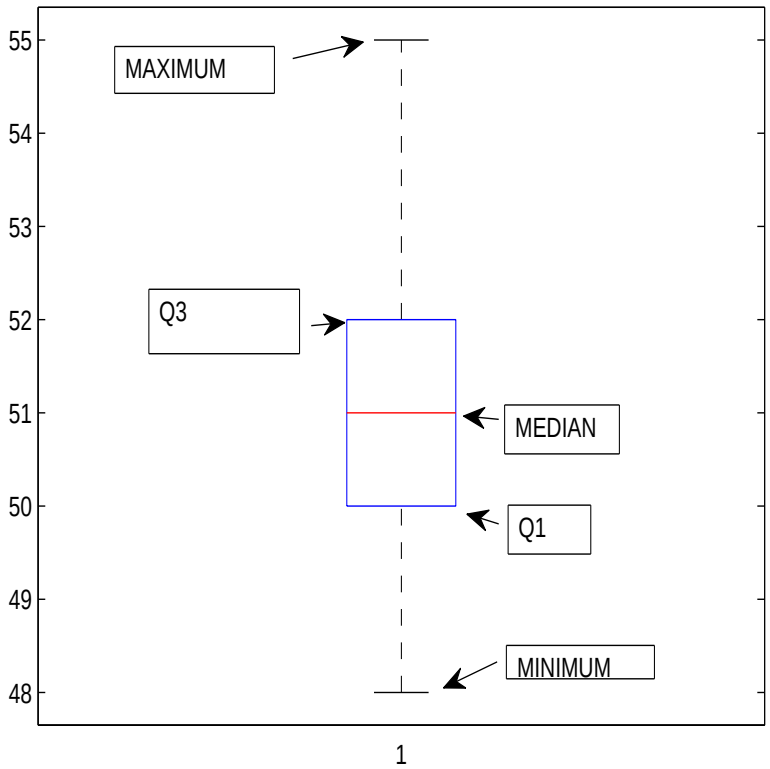
BOXPLOT a.k.a box-and-whisker diagram

A **boxplot** *is a graph of a data set that consists of a line from the minimum value to the maximum value and a box with lines drawn at the first quartile Q_1, the median and the third quartile Q_3*



BOXPLOT

first quartile Q_1= separates the bottom 25% of the sorted values, the median = midpoint, 50 % below, the third quartile Q_3 = separates the bottom 75% of the sorted values



New York Times 2009: Big Data

statistics . . . uses powerful computers and sophisticated mathematical models to hunt for meaningful patterns and insights in vast troves of data. The applications are as diverse as improving Internet search and online advertising, culling gene sequencing information for cancer research and analyzing sensor and location data to optimize the handling of food shipments.

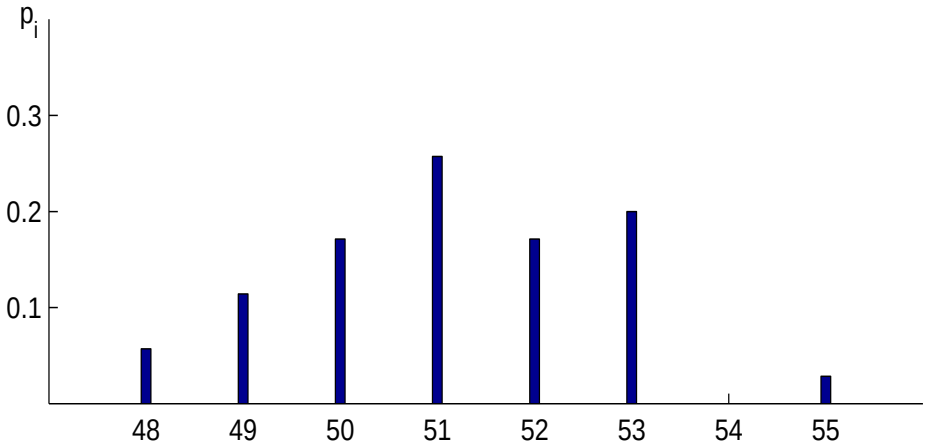
mathematical model $\leftrightarrow$ *probability*

KTH Matematik

Probability: What is it ?

The empirical answer: Let us think of the histogram of the matchbox data again. We assume that nothing changes (in the industrial production of matchboxes).

Then the **probability of the event** *"50 $\leq$ the number of matches in a box $\leq$53 " is taken as the relative frequency of "50 $\leq$ the number of matches in a box $\leq$53 ".*

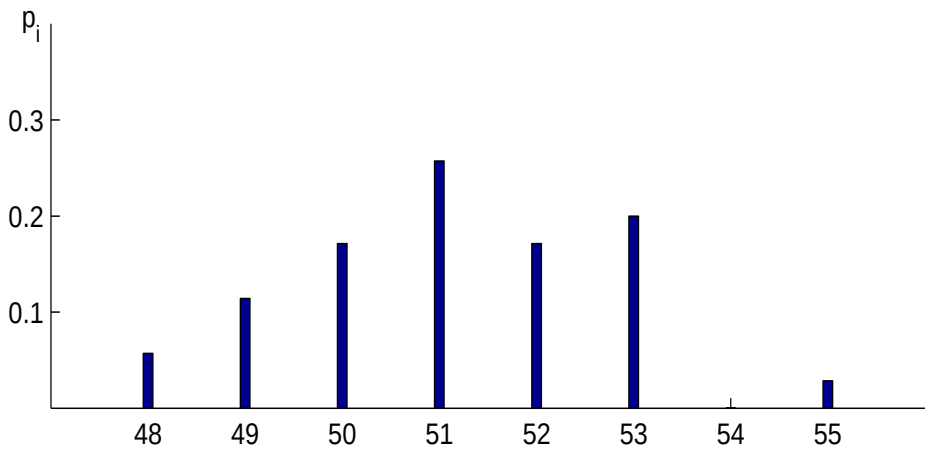


KTH Matematik

Probability: What is it ?

The empirical answer: Let us think of the histogram of the matchbox data again. We assume that nothing changes (in the industrial production of matchboxes). Then the **probability of the event** " $50 \leq$ the number of matches in a box ≤ 53 " is

$$p_{50} + p_{51} + p_{52} + p_{53} = 0.171 + 0.257 + 0.171 + 0.20 \approx 0.80$$



KTH Matematik

Probability

We write now the probability of **the event** " $50 \leq$ the number of matches in a box ≤ 53 " as

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

Then, of course, the probability of a simple event like for example " the number of matches in a box $= 53$" is its relative frequency

$$Pr\left(\text{ the number of matches in a box } = 53\right) = p_{53} = 0.20$$

KTH VETENSKAP OCH KONST
KTH Matematik

Probability

The probability of **the event** "$50 \leq$ the number of matches in a box ≤ 53 "

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

This is a probability based on past observations. This means, when we assume that nothing changes, that if we receive a new box of mathces and check the number of matches in it, there is 80% chance of the event "$50 \leq$ the number of matches in a box ≤ 53" occurring.

KTH VETENSKAP OCH KONST
KTH Matematik

Probability: empirical

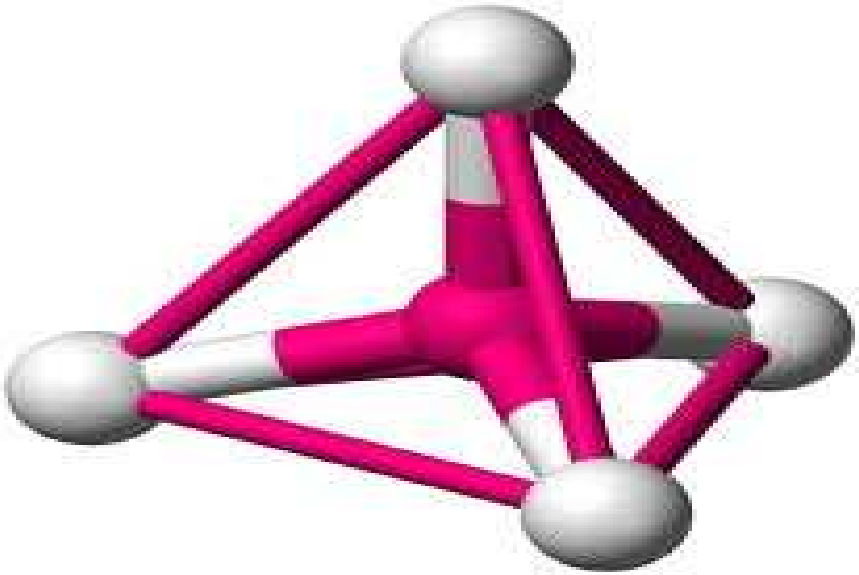
The probability

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

was based on checking 35 boxes. The relative frequencies will change if we observe new boxes, but they will stabilize.

KTH Matematik

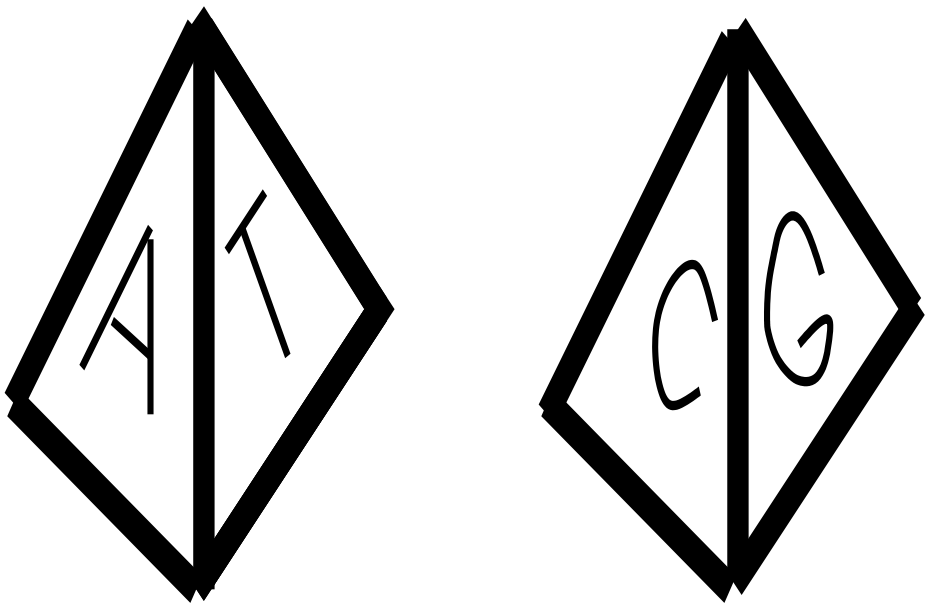
Probability



KTH Matematik

Probability: theoretical

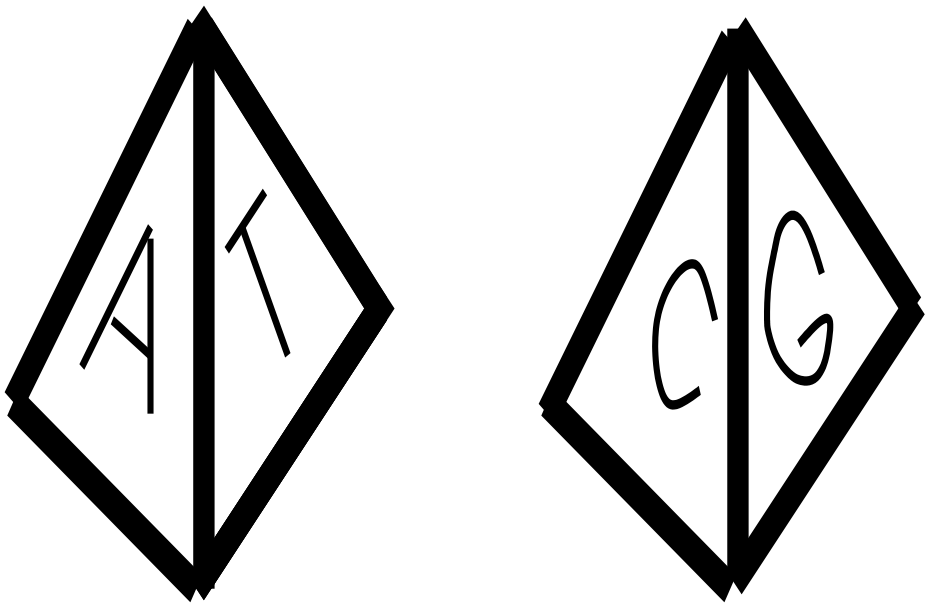
To each of the four sides of a tetrahedron there is assigned one of the letters A,T,C,G (nucleotides). All letters are used. We call this the DNA dice. We toss the dice in the air and note the side that it falls on.



KTH Matematik

Probability: theoretical

We toss the dice in the air and note the side that side that it falls on. Clearly we can thus produce a **random** DNA sequence. Such a sequence would seem to lack value for bioinformatics. Disregarding that for the moment, we might ask what is the probability of observing the sequence CAAGT in five tosses of the dice, or what is the probability of the event "CAAGT in five tosses".



KTH Matematik

Probability: theoretical

What is the probability of the event "CAAGT in five tosses" ? All the sides of the tetrahedron are of equal area, and the tetrahedron is balanced. The situation is such that we assign the probabilities

$$Pr(A) = Pr(T) = Pr(C) = Pr(G) = \frac{1}{4}$$

to the four possible outcomes of a single toss. In other words, all outcomes of the toss of the dice are equally **likely**. Note that we have here a very crude model of a DNA sequence and a way to compute the probability of such a sequence.

KTH Matematik

Probability: theoretical

In addition, we think that the individual tosses do not influence each other. This means that

$$Pr(A) = Pr(T) = Pr(C) = Pr(G) = \frac{1}{4}$$

are the probabilities of the letters at any toss. Hence we get that

$$Pr\,(\text{CAAGT}\,) = Pr\,(C)\,Pr\,(A)\,Pr\,(A)\,Pr\,(G)\,Pr\,(T)$$

$$= \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} = \left(\frac{1}{4}\right)^5 .$$

In probability terms one says that the outcomes at the different tosses are **independent events**.

KTH VETENSKAP OCH KONST
KTH Matematik

Probability: theoretical

The probability

$$Pr\left(\text{CAAGT }\right) = \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} = \left(\frac{1}{4}\right)^5 .$$

was obtained theoretically by reasoning about the symmetry of the tetrahedron and independence of the tosses. Of course, one may toss any given DNA dice, say 10^{10} times, and register the relative frequencies of the letters A,T,C,G. Then one can check whether the theoretical probabilities and the empirical probabilities seem to agree.

KTH VETENSKAP OCH KONST
KTH Matematik

Probability

Probability: a formal statement

We know continue by giving a general formal description and some rules of computation for $Pr(A)$, the probability of an event.

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

KTH VETENSKAP OCH KONST
KTH Matematik

Probability: a formal statement

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

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

KTH Matematik

Probability: a formal statement

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.

KTH VETENSKAP OCH KONST
KTH Matematik

Probability: a formal statement

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.

- For example, when tossing the DNA dice once, the event $S = \{A, T, C, G\}$ (= one of the letters A,T,C,G) is certain to happen. (Here we may call A, T, C, G simple events, because they cannot be broken down any further).

KTH Matematik

Probability

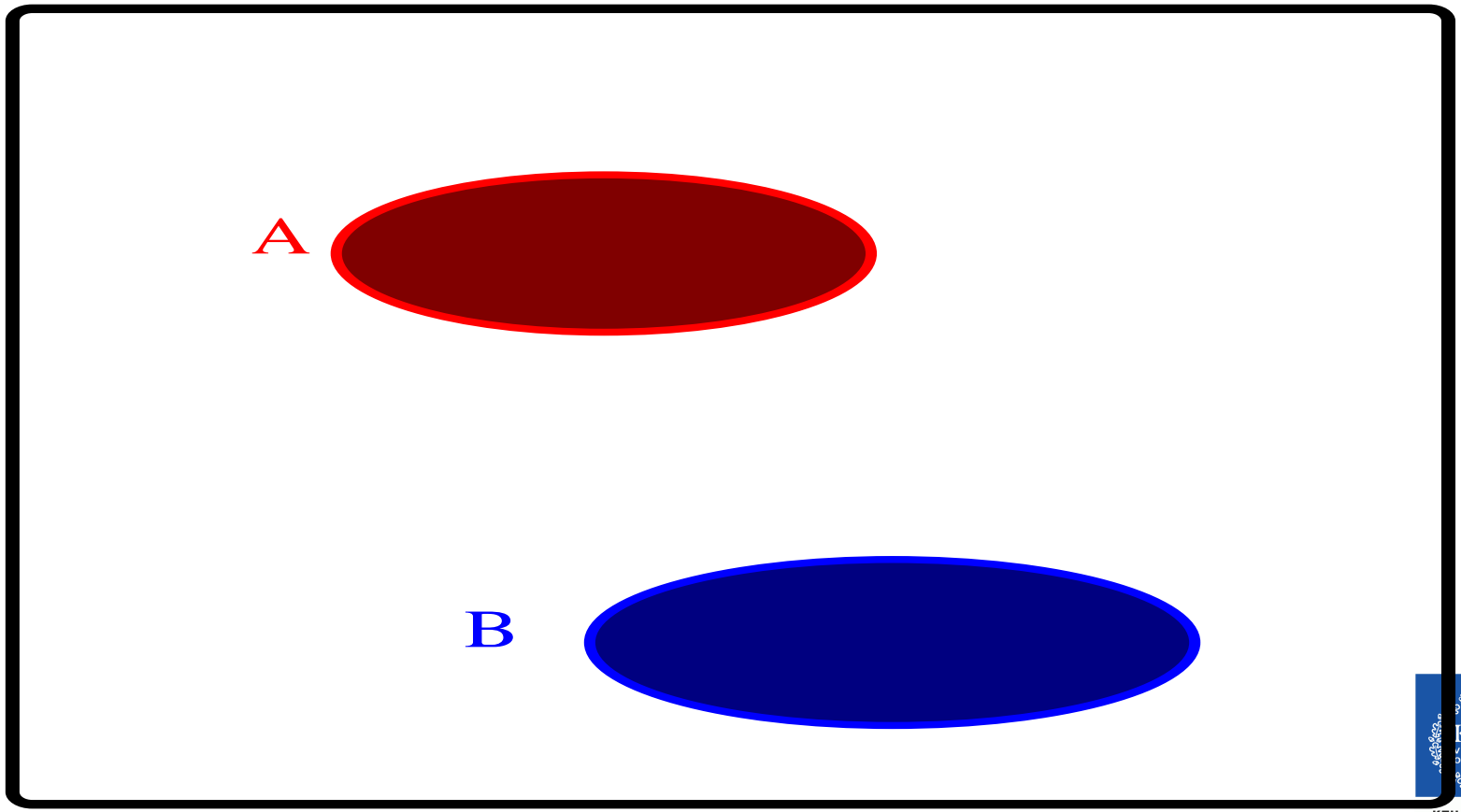
Rolling a 14

Heads

The sun will rise

Mutually exclusive events

We say that two events A and B are **mutually exclusive**, if the occurrence of one precludes the occurrence of the other. The diagram below illustrates this. For DNA dice, the events $\{A, T\}$, $\{C\}$ are mutually exclusive.



A or B

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

KTH Matematik

Probability: a formal statement of the additive law

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**

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

KTH Matematik

Example

The DNA dice again, the events $A_1 = \{A, T\}$, $A_2 = \{C\}$ are mutually exclusive. Also the events $\{A\}$ and $\{T\}$ are mutually exclusive. Thus the additive law together with the theoretical probability model from the above gives

$$Pr(A_1) = Pr(\{A, T\}) = Pr\,(\ \{A\} \text{ or } \{T\}\ ) = Pr(\{A\}) + Pr(\{T\})$$

$$= \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

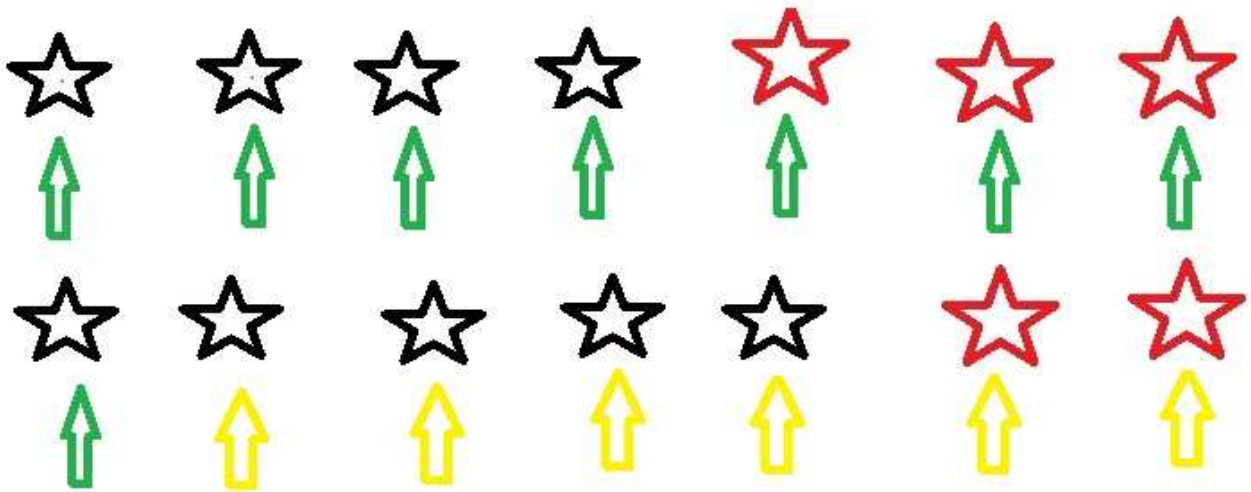
The additive law gives also

$$Pr\,(\ A_1 \text{ or } A_2\ ) = Pr(A_1) + Pr(A_2) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$$

KTH Matematik

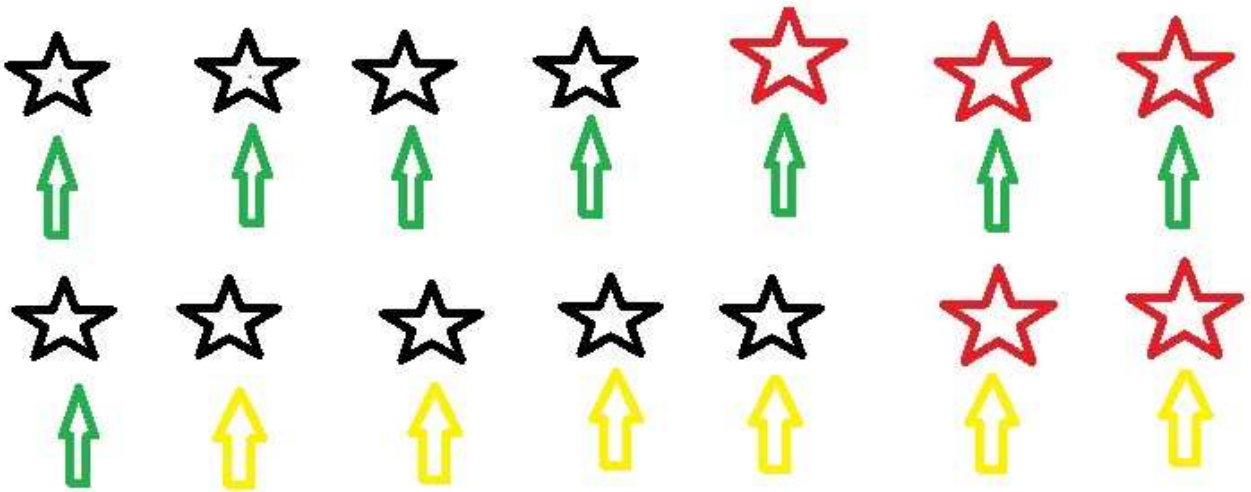
Probability in a Mendelian Spirit

Probabilities play a very big role in genetics. In the figure we have 14 (abstract) peas: nine have white flowers, five have red flowers, six have yellow pods, eight have green pods.

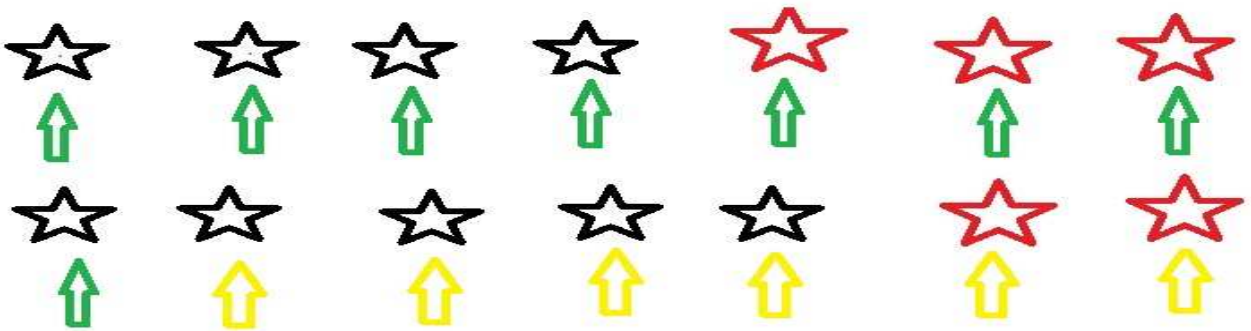


Probability in a Mendelian Spirit

What is $Pr(\text{green pods or white flowers})$?

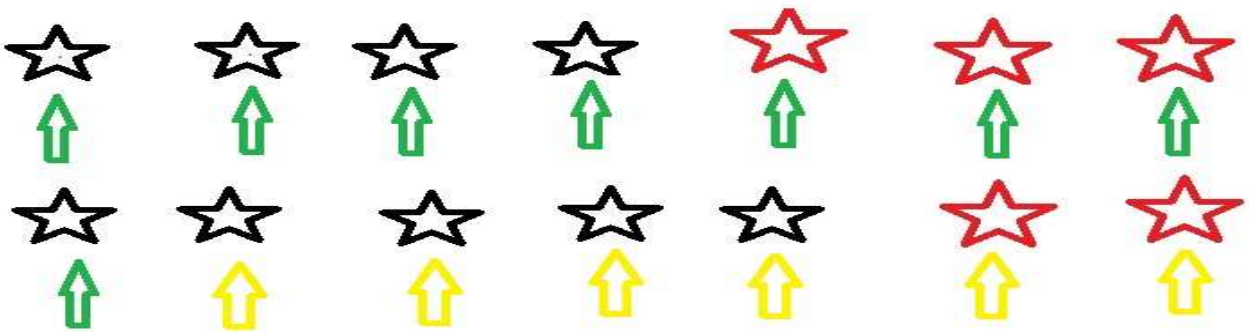


Probability in a Mendelian Spirit



- green pods or white flowers means green pods or white flowers or both

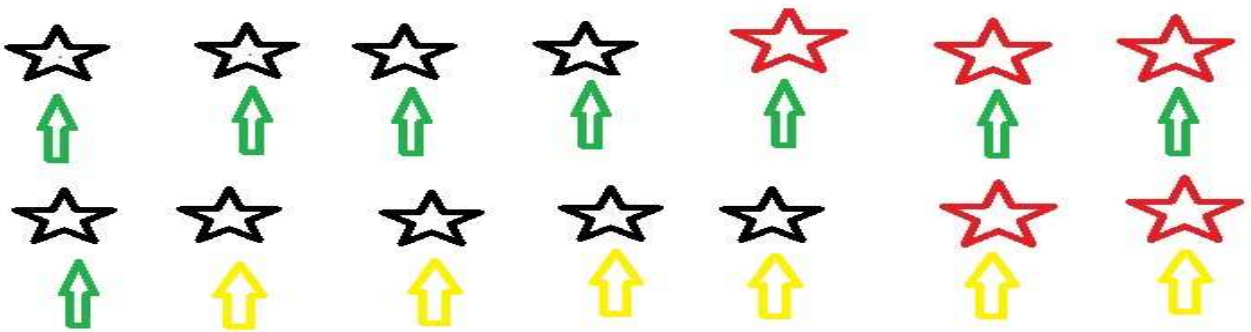
Probability in a Mendelian Spirit



- green pods or white flowers means green pods or white flowers or both
- it is wrong to add to the 8 peas with green pods the 9 peas with white flowers, because then you have counted 5 of the peas twice.

KTH VETENSKAP OCH KONST
KTH Matematik

Probability in a Mendelian Spirit



- green pods or white flowers means green pods or white flowers or both
- it is wrong to add to the 8 peas with green pods the 9 peas with white flowers, because then you have counted 5 of the peas twice.
- $Pr(\text{green pods or white flowers}) = \frac{12}{14} = \frac{6}{7}$

The General Rule

(d)

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

In the example above

$$Pr(\text{green pods or white flowers }\ )$$

$$= Pr(\text{green pods }\ ) + Pr(\ \text{white flowers }\ ) - Pr(\text{green pods and white flowers}$$

$$= \frac{8}{14} + \frac{9}{14} - \frac{5}{14} = \frac{12}{14} = \frac{6}{7}$$

KTH Matematik

Complementary event A^*

$$\text{"}A \text{ does not occur "} = A^*.$$

Example

$$S = \{A, T, C, G\}$$

$A = \{A, T\}$. $A^* = \{C, G\}$.

Probability of the Complementary event A^*

(e)

$$Pr\left(A^*\right) = 1 - Pr(A).$$

Probability of the Complementary event A^*

$$Pr\left(A^*\right) = 1 - Pr(A).$$

The DNA dice again, the event $A = \{A, T, C\}$, $A^* = \{G\}$ we have by the additive law

$$Pr(A) = Pr(\{A\}) + Pr(\{T\}) + Pr(\{C\}) = \frac{3}{4}$$

and

$$Pr\left(A^*\right) = 1 - Pr(A) = \frac{1}{4}$$

but this agrees with what we have by our model for the dice, namely that $Pr(\{G\}) = \frac{1}{4}$.

KTH Matematik

Probability: independent events

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

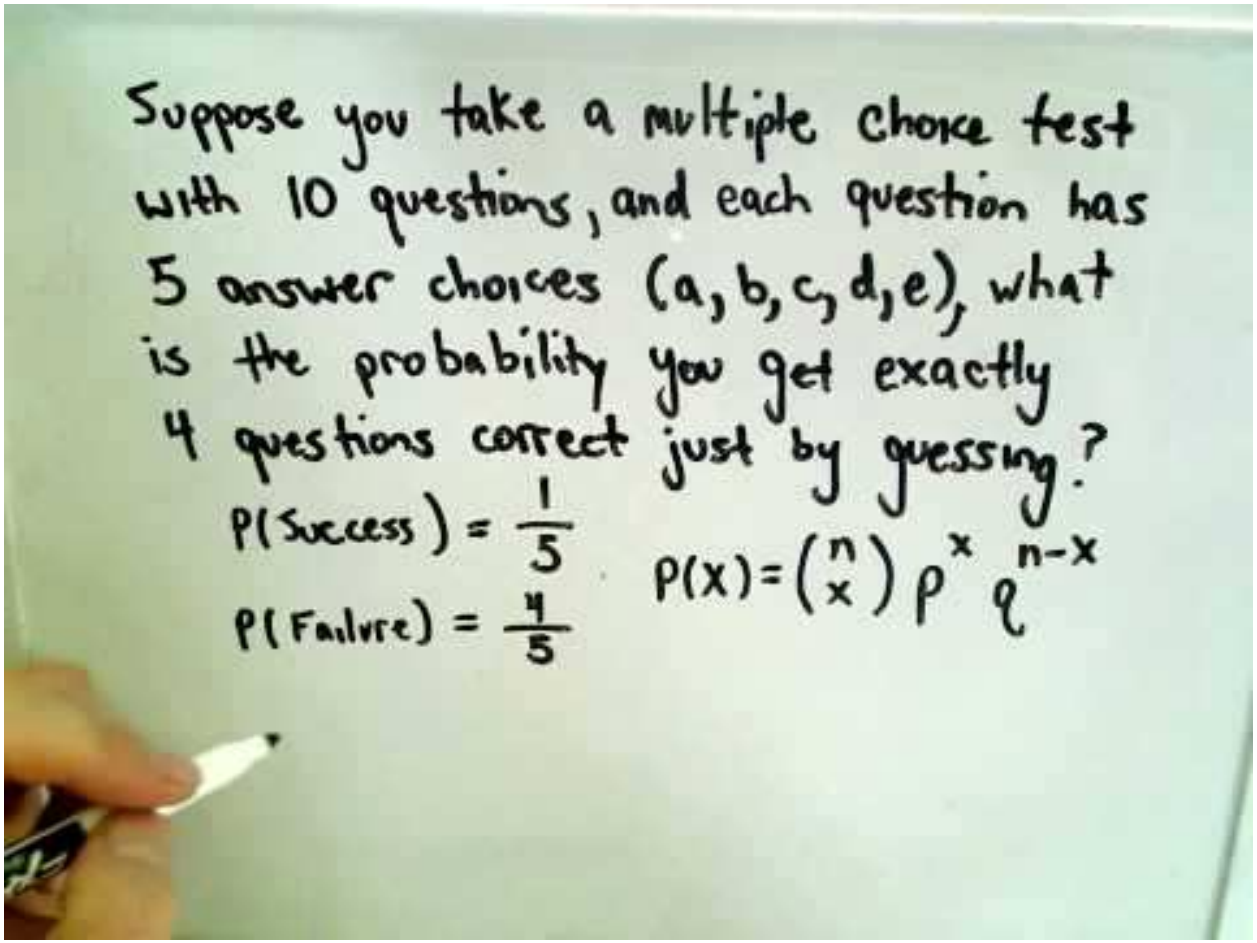
(f)

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

Probability

So far our examples been dealing with probability empirically as a relative frequency or by the rule 'simple events have an equal chance of occurring' (e.g., $Pr(A) = Pr(T) = Pr(C) = Pr(G) = \frac{1}{4}$). Next we check probabilities of different sort.

KTH Matematik


Suppose you take a multiple choice test with 10 questions, and each question has 5 answer choices (a, b, c, d, e), what is the probability you get exactly 4 questions correct just by guessing?
P(Success) = 1/5
P(Failure) = 4/5
P(X) = (n choose x) p^x q^(n-x)

Probability: series of independent events

We discuss finally the **binomial distribution** which shows the probabilities of different outcomes for a series of independent random events, each of which can have only one of two values, e.g., success and failure. Usually one presents here the toss of a coin, with outcomes heads and tails. I shall talk about tosses of a thumbtack.

Fair coin: $Pr(\text{heads}) = Pr(\text{tails}) = \frac{1}{2}$

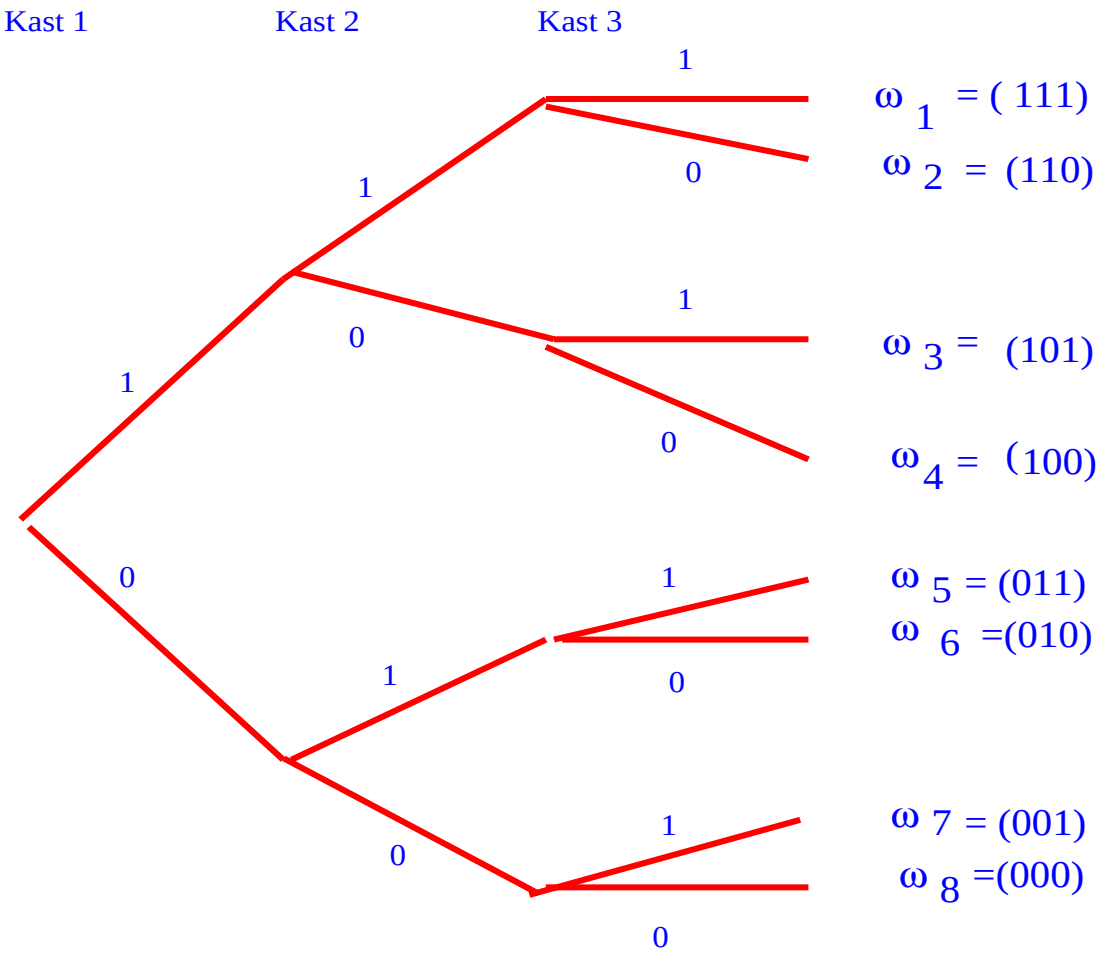
KTH Matematik

Tossing of a Thumbtack

Toss a thumbtack. If it falls on its point as in the picture above, we say that the event a digital 'one' (1), occurs. We say that att the event a digital zero (0) occurs if the thumbtack lands on its head.

KTH Matematik

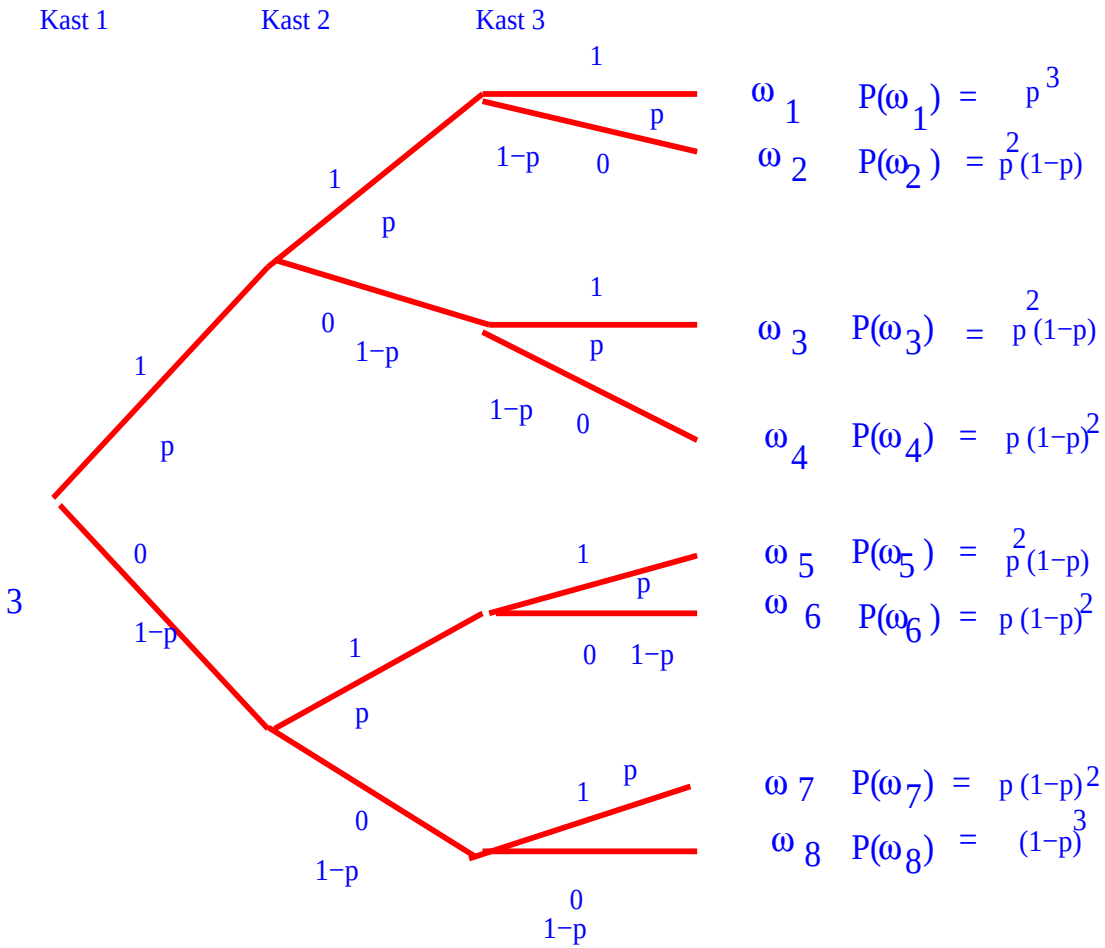
Three tosses of a thumbtack



Number of possible outcomes=

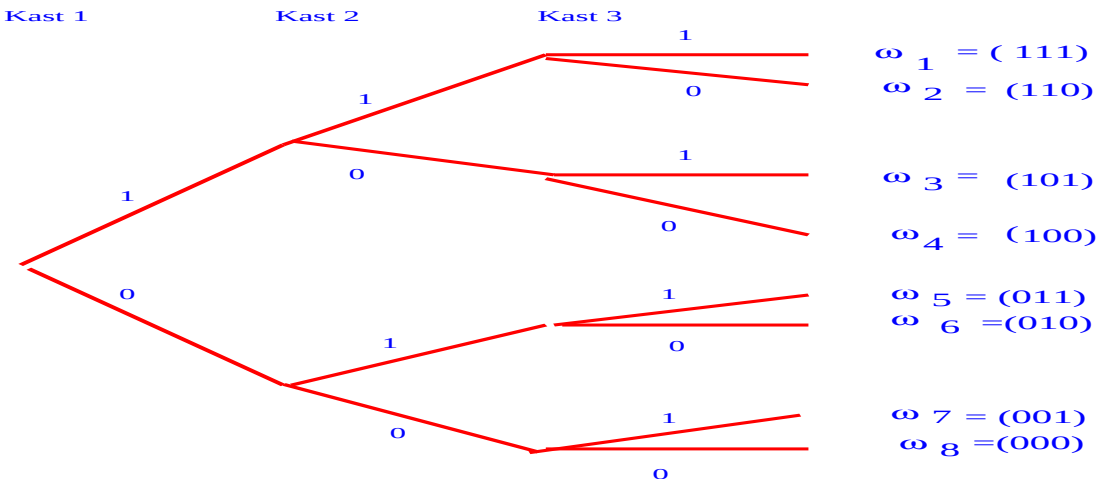
2^3

Three tosses of a thumbtack: probabilities by independence



$$Pr(\text{ zero } (0)) = 1 - p, Pr(\text{ one } (1)) = p, 0 < p < 1.$$

Three tosses of a thumbtack: probability of the number of ones



We consider the probabilities of the number of ones
$\omega_1 \to 3$, $\omega_2, \omega_3, \omega_5 \to 2$ $\omega_4, \omega_6, \omega_7 \to 1$, $\omega_8 \to 0$.

KTH Matematik

Three tosses of a thumbtack: probability of the number of ones

$$Pr(3) = Pr(\omega_1) = p^3,$$

$$Pr(2) = Pr(\omega_2) + Pr(\omega_3) + Pr(\omega_5) = 3p^2(1-p)$$

$$Pr(1) = Pr(\omega_4) + Pr(\omega_6) + Pr(\omega_7) = 3p(1-p)^2$$

$$Pr(0) = Pr(\omega_8) = (1-p)^3$$

KTH Matematik

Three tosses of a thumbtack: probability of the number of ones

We rewrite using the binomial coefficients:

$$Pr(3) = p^3 = \binom{3}{3} p^3 (1-p)^0$$

$$Pr(2) = 3p^2(1-p) = \binom{3}{2} p^2 (1-p)$$

$$Pr\,(1) = 3p(1-p)^2 = \binom{3}{1} p(1-p)^2$$

$$Pr\,(0) = \binom{3}{0} (1-p)^3$$

or with a single formula

$$Pr\,(k) = \binom{3}{k} p^k (1-p)^{3-k}, \quad , k = 0, 1, 2, 3.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Binomial coefficients

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

reads as 'n choose k',

$$n! = n \cdot (n-1) \ldots 2 \cdot 1 \quad \text{n factorial}$$

KTH VETENSKAP OCH KONST
KTH Matematik

Binomial distribution

Definition

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

This is called the binomial probability distribution with parameters n and p. The **mean** of this distribution is np and the **variance** is $np(1-p)$.

KTH Matematik

Binomial distribution

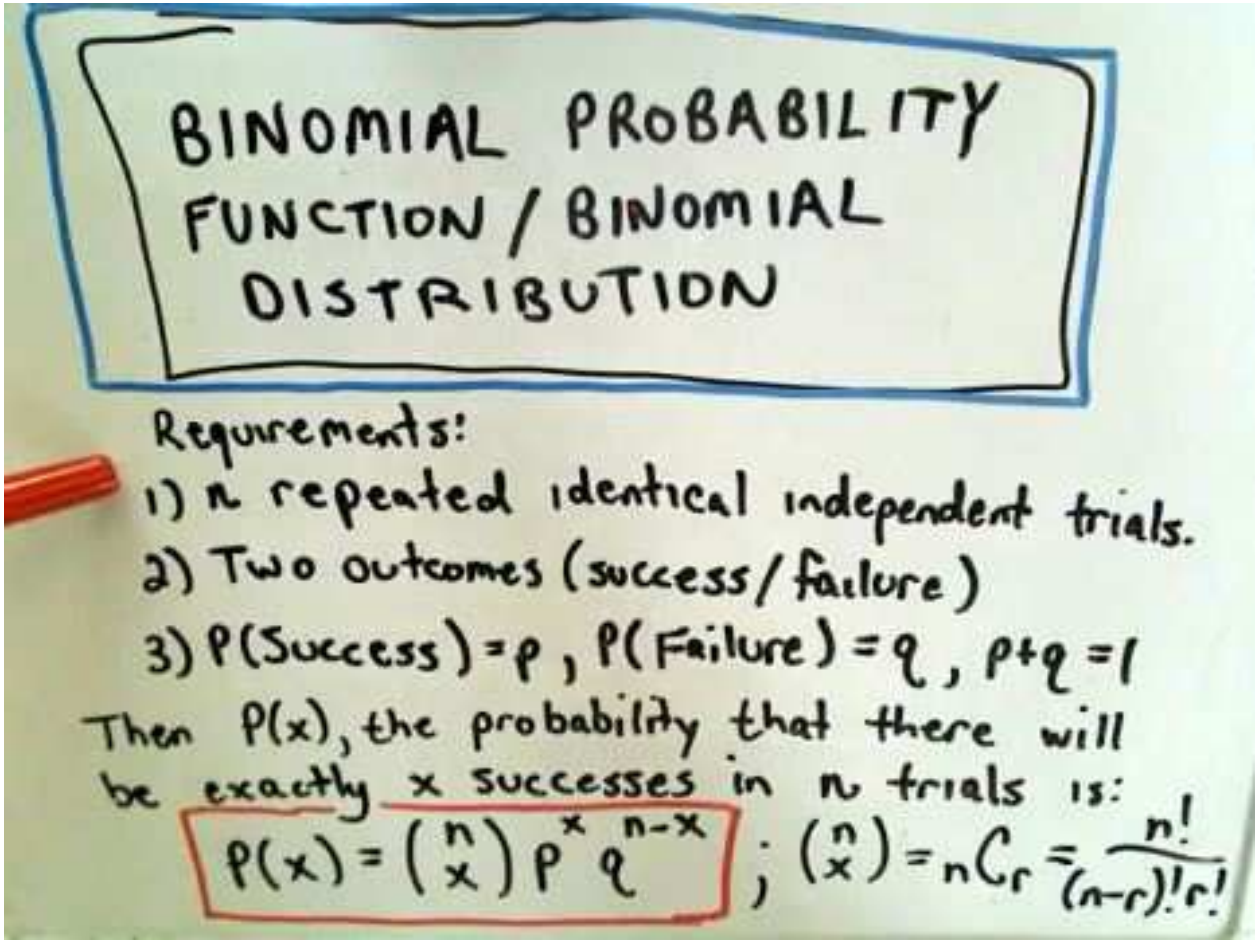
The binomial distribution shows the probabilities of different outcomes for a series of random events, each of which can have only one of two values.

Definition

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

KTH Matematik

,


BINOMIAL PROBABILITY FUNCTION / BINOMIAL DISTRIBUTION
Requirements:
1) n repeated identical independent trials.
2) Two outcomes (success/failure)
3) P(Success) = p, P(Failure) = q, p+q = 1
Then P(x), the probability that there will be exactly x successes in n trials is:
$P(x) = \binom{n}{x} p^x q^{n-x}$; $\binom{n}{x} = nC_r = \frac{n!}{(n-r)!r!}$

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 dice !) and independent, and that the probabilities are the same at each site, then the probabilities of the number of matches (=successes) are 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 in sequence alignment.

KTH Matematik

Rare Event Rule

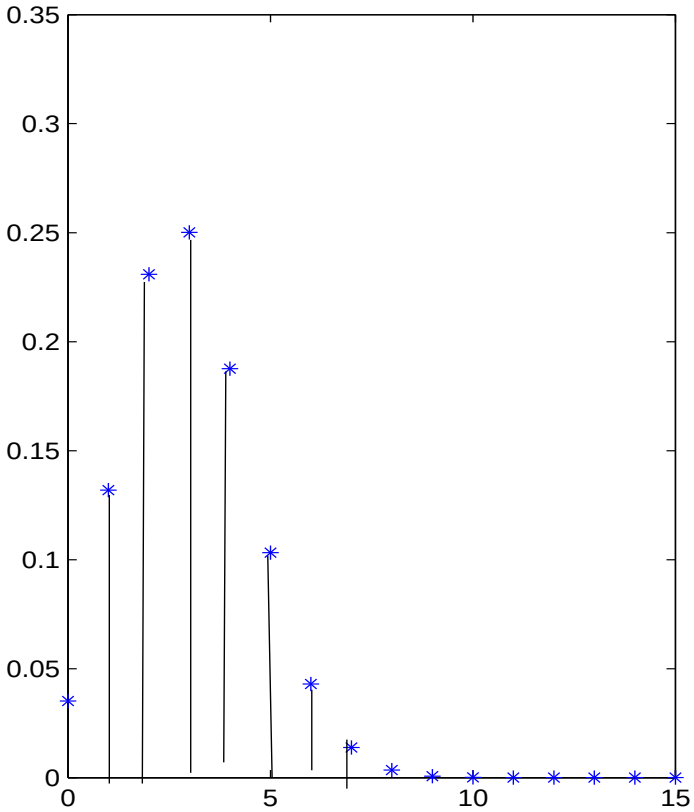
If, under a given assumption, the probability of an observed event is extremely small (≈ 0), we conclude that the assumption is likely not correct.

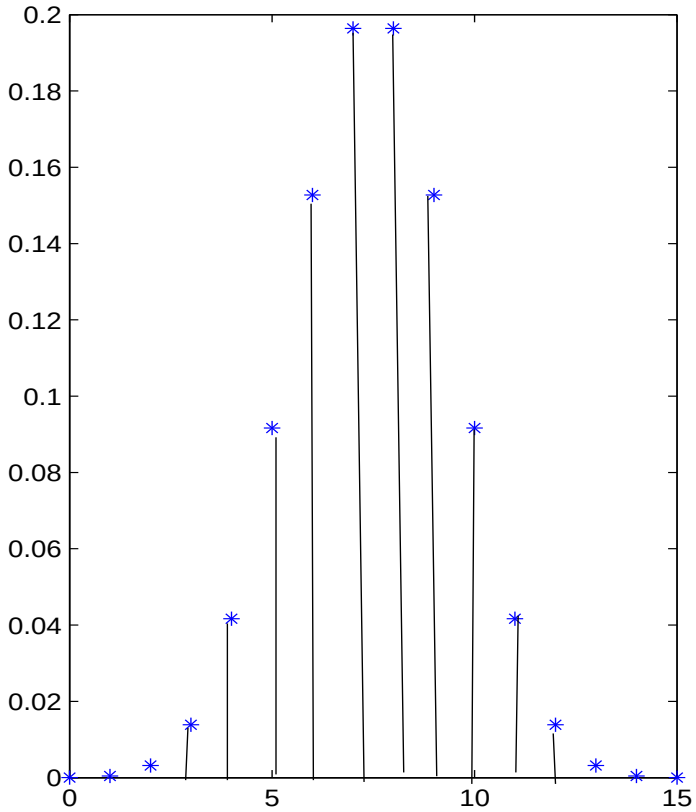
In the example with alignment of sequences, the probability of 11 matches in two sequences of 15 nucleotides under the assumption of independent tosses of the DNA dice is $1.0297e-04$. (A computation done using a Matlab function for the binomial probability with $n = 15$, $p = \frac{1}{4}$, $k = 11$).

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

KTH Matematik

Binomial distributions with $n = 15$ and $p = 0.2$, $p = 0.5$

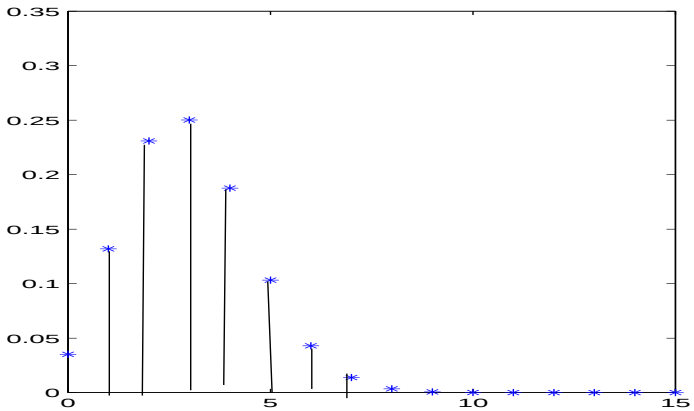


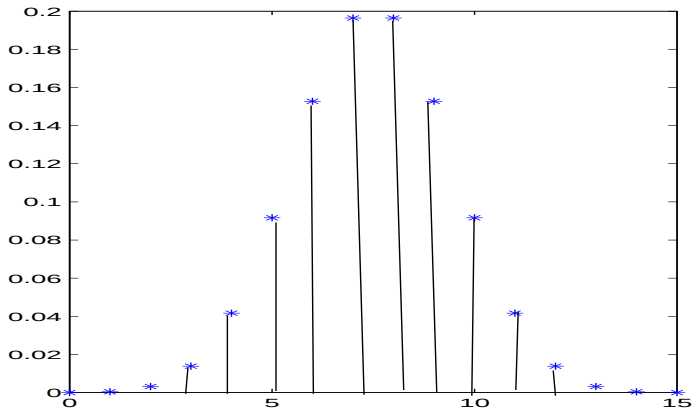


KTH VETENSKAP OCH KONST
KTH Matematik

Binomial distributions with $n = 15$ and $p = 0.2$, $p = 0.5$

For $p = 0.2$ the distribution is skewed, mean is $15 \cdot 0.2 = 3$, with variance $15 \cdot 0.2 \cdot 0.8 = 2.4$. For $p = 0.5$ the distribution is symmetric around its mean $15 \cdot 0.5 = 7.5$, with variance $15 \cdot 0.5 \cdot 0.5 = 3.75$.





KTH Matematik

Binomial distribution

Note that

$$\sum_{k=0}^{n} Pr(k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1$$

by the binomial formula

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

KTH Matematik

Thank You !

"How do you want it—the crystal mumbo-jumbo or statistical probability?"

KTH Matematik