# SF1911: Statistik för bioteknik
# Föreläsning 5.

TK

9.11.2017

# Lärandemål

- Stokastiska modeller för kvantitativa diskreta datatyper
  - Diskreta stokastiska variabler
    - Sannolikhetsfunktion
    - Fördelningsfunktion
    - Oberoende stokatiska variabler
    - Väntevärde och varians
- Speciella sannolikhetsfunktioner
  - Binomialfördelning
  - Poissonfördelning
  - Geometrisk fördelning
  - Negativ Binomialfördelning
  - Hypergeometrisk fördelning
- Väntevärde och varians för linjära kombinationer av oberoende stokastiska variabler

KTH Matematik

Vita blodkroppar är blodceller som försvarar kroppen mot infektioner. För många eller för få vita blodkroppar kan vara en indikation på ett antal saker. Vi kan ta ett urinprov och kolla antalet vita blodkroppar. 4,000 till 11,000 vita blodkroppar per 1 $mm^3$ (kubikmillimeter) av blod är normalvärde. . Vi söker

Sannolikheten för normalvärdet vita blodkroppar per 1 $mm^3$ (kubikmillimeter) blod i ett prov.

KTH Matematik

Låt oss införa

$X =$ Antalet vita blodkroppar per 1 $mm^3$ (kubikmillimeter) blod i ett prov .

Då kan vi skriva

Sannolikheten för normalvärdet vita blodkroppar per 1 $mm^3$ (kubikmillimeter) blod i ett prov.

som

$$P(4,000 \leq X \leq 11000)$$

$X$ är ett exempel på en stokastisk variabel (med kvantitativa antalsdata som värden).

KTH Matematik

# Random Variables (stokastiska variabler) & 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.*

Procedure: t.ex. urinprovet ovan.

KTH VETENSKAP OCH KONST
KTH Matematik

# Stokastiska variabler (Random variables): **Diskret stokastisk variabel**

Vi betraktar fallet med kvantitativa diskreta data.

**Definition**

*En stokastisk variabel (st.v.) $X$ säges vara diskret om den kan anta ett ändligt eller uppräkneligt oändligt antal olika värden.*

KTH VETENSKAP OCH KONST
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

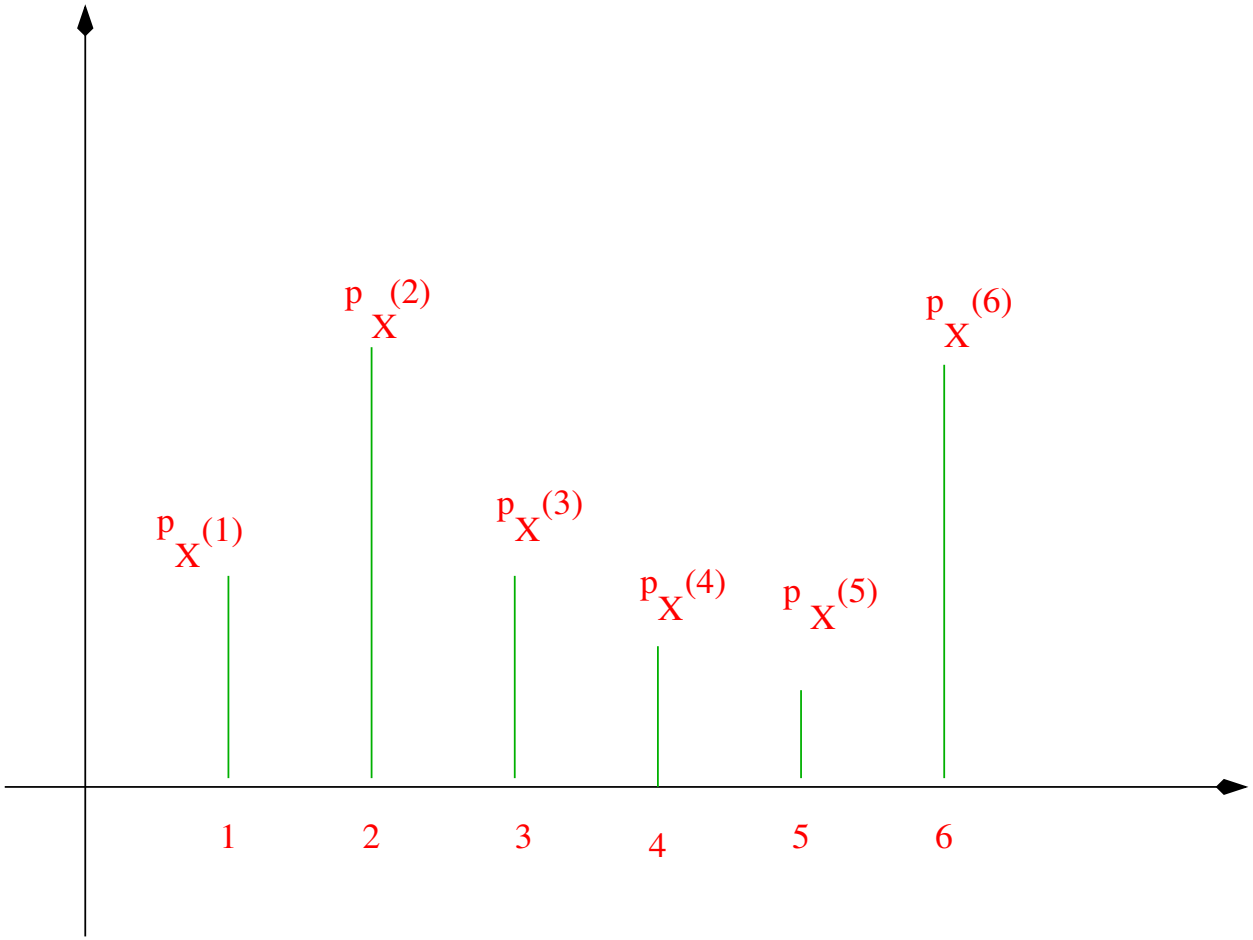
$$p_X(x) \stackrel{\text{def}}{=} P(X = x), \quad x \in \{x_1, x_2, \ldots, \}$$

Note that

$$0 \leq p_X(x_i) \leq 1, \sum_{x_i} p_X(x_i) = 1.$$

KTH VETENSKAP OCH KONST
KTH Matematik

# En sannolikhetsfunktion



# Stokastiska variabler: inledning

Låt $X$ vara en stokastisk variabel. Det mest allmänna sättet att beskriva $X$, dvs. hur $X$ varierar, är att ange dess fördelningsfunktion.

## Definition

*Fördelningsfunktionen $F_X(x)$ till en s.v. $X$ definieras av*

$$F_X(x) = P(X \leq x).$$

KTH VETENSKAP OCH KONST
KTH Matematik

- Fördelningsfunktion $F_X(x)$;

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a)$$

- diskreta st.v:er, $F_X(x) = \sum_{x_j \leq x} p_X(x_j)$,

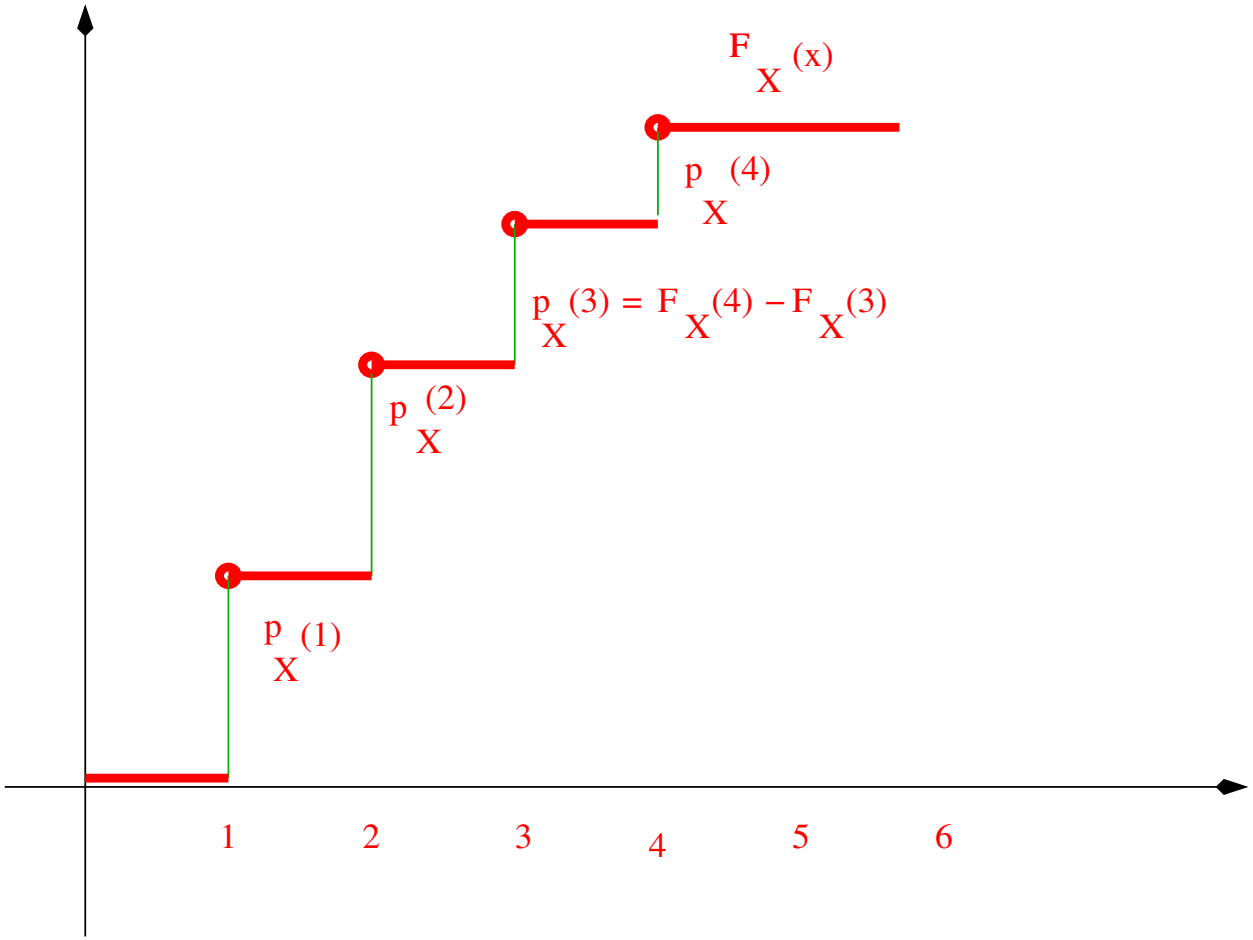
# Stokastiska variabler: inledning

En fördelningsfunktion $F_X(x)$ har följande egenskaper:

1) $F_X(x)$ är icke-avtagande[1] ;
2) $F_X(x) \to 1$ då $x \to \infty$;
3) $F_X(x) \to 0$ då $x \to -\infty$;
4) $F_X(x)$ är högerkontinuerlig (om denna tanke är obekant, oroa Dig inte).

[1] d.v.s. $x_1 \leq x_2 \Rightarrow F_X(x_1) \leq F_X(x_2)$

# En fördelningsfunktion



# Beräkning av sannolikheter

Låt $X$ vara en stokastisk variabel.

$$F_X(x) = P(X \leq x).$$

Då fås

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a)$$

## Sats

$$P(a < X \leq b) = F_X(b) - F_X(a)$$

KTH Matematik

# Diskreta stokastiska variabler

Relationen mellan sannolikhetsfunktionen och fördelningsfunktionen för en diskret stokastisk variabel fås av sambanden

$$F_X(x) = \sum_{x_j \leq x} p_X(x_j)$$

och

$$p_X(x_k) = F_X(x_k) - F_X(x_{k-1}).$$

KTH VETENSKAP OCH KONST
KTH Matematik

# Väntevärden

Vi ska nu införa begreppet väntevärde för en st.v. Detta är den teoretiska motsvarigheten till begreppet medelvärde för en datamängd.
Antag att vi har $\mathcal{X} = x_1, \ldots, x_n$. Medelvärdet definierades av

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k.$$

Vi kan återkalla i minnet frå övningarna omskrivningen av $\bar{x}$ med empirisk frekvens

$$\bar{x} = \sum_{x} x \cdot f_{\mathcal{X}}(x),$$

där

$$f_{\mathcal{X}}(x) = \frac{\text{antalet } \{k;\ x_k = x\}}{n}.$$

KTH Matematik

# Väntevärden

## Definition

*Väntevärdet $\mu$ för en diskret s.v. $X$ är*

$$\mu = E(X) \stackrel{\text{def}}{=} \sum_{x_k} x_k p_X(x_k)$$

# Discrete Uniform Distribution

$X \sim U(1, 2, \ldots, 6) \leftarrow$ en förkortande beteckning för $X$ **har en likformig (uniform) fördelning på heltalen** $1, 2, 3, 4, 5, 6$.

$$p_X(x) = \begin{cases} \frac{1}{6} & \text{för } x = 1, 2, 3, 4, 5, 6 \\ 0 & \text{för övriga värden på } k. \end{cases}$$

KTH VETENSKAP OCH KONST
KTH Matematik

# Väntevärden

$X \sim U(1, \ldots, 6)$.

$$p_X(x) = \begin{cases} \frac{1}{6} & \text{för } x = 1, 2, 3, 4, 5, 6 \\ 0 & \text{för övriga värden på } k, \end{cases}$$

vilket ger

$$E(X) = \sum_{x=0}^{6} x_k p_X(x_k) = \sum_{k=1}^{6} k \frac{1}{6} = \frac{1}{6} \frac{6(6+1)}{2} = 3.5.$$

KTH VETENSKAP OCH KONST
KTH Matematik

# Väntevärde för $Y = g(X)$

Antag att vi känner förd. för $X$, och vill beräkna $E(Y)$ där $Y = g(X)$.

## Sats

*Väntevärdet för $g(X)$ är*

$$E(g(X)) = \sum_{x_k} g(x_k) p_X(x_k)$$

**Momentgenerande funktion** $m_X(t)$: *Tag* $g(x) = e^{tx}$.

$$m_X(t) \stackrel{\text{def}}{=} E\left(e^{tX}\right) = \sum_{x_k} e^{x_k} p_X(x_k).$$

*Första derivatan i $t = 0$ genererar väntevärdet.*

$$m_X^{(1)}(t) \mid_{t=0} = E(X).$$

$$m_X^{(2)}(t) \mid_{t=0} = E(X^2).$$

# Väntevärde för $Y = g(X)$

$$E(h(X) + g(X)) = E(h(X)) + E(g(X))$$

med det viktiga specialfallet

**Sats**

$$E(aX + b) = aE(X) + b.$$

# Varians

Väntevärdet säger inget om hur $X$ varierar.
Betrakta följande:

$$(X - E(X))^2$$

Vi leds nu till följande definition.

## Definition

*Variansen $\sigma^2$ för en s.v. $X$ är*

$$\sigma^2 = Var(X) = E[(X - E(X))^2] = \sum_{x_k} (x_k - E(X))^2 \, p_X(x_k).$$

# Varians

Följande räkneregel är mycket användbar:

**Sats**

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

**Bevis.**

$$Var(X) = E[(X - E(X))^2] = E[X^2 + E(X)^2 - 2E(X)X]$$

$$= E[X^2] + E(X)^2 - 2E(X)E[X] = E(X^2) - E(X)s^2.$$

□

# Varians

I exemplet med $X \sim U(1, \ldots, 6)$ har vi $\mu = 3.5 = \frac{21}{6}$. Vidare har vi

$$E(X^2) = \sum_{x=1}^{6} x^2 p_X(x) = \sum_{k=1}^{6} k^2 \frac{1}{6} = \frac{91}{6} = 15.16$$

Här utnyttjade vi, se Beta, $\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$ för $n = 6$,.
Enligt räkneregeln fås

$$Var(X) = \frac{91}{6} - \left(\frac{21}{6}\right)^2 = \frac{546 - 441}{36} = 2.92.$$

# Discrete Uniform Distribution

$X \sim U(1, 2, \ldots, n)$, where $n > 1$. The p.m.f. is

$$p_X(x) = \begin{cases} \frac{1}{n} & x = 1, 2, \ldots, n \\ \\ 0 & \text{else.} \end{cases} \tag{1}$$

$$E[X] = \frac{n+1}{2}, \operatorname{Var}[X] = \frac{n^2 - 1}{12}.$$

KTH Matematik

# Varians

## Sats

$Var(aX + b) = a^2 Var(X)$.

Detta gäller, ty om $\mu = E(X)$,

$$Var(aX + b) = E[(aX + b - E(aX + b))^2] = E[(aX + b - a\mu - b)^2]$$

$$= E[(aX - a\mu)^2] = a^2 E[(X - \mu)^2] = a^2 Var(X).$$

□

# Standardavvikelsen

## Definition

*Standardavvikelsen $\sigma$ för en s.v. $X$ är*

$$\sigma = D(X) = \sqrt{Var(X)}.$$

## Sats

$$D(aX + b) = |a|D(X).$$

Allmänt gäller:

$D$ – rätt sort.

$Var$ – lättare att räkna med.

KTH VETENSKAP OCH KONST

KTH Matematik

# Oberoende stokastiska variabler

## Definition

*$X$ och $Y$ är oberoende stokastiska variabler om*

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$$

*för alla $x$ och $y$.*

KTH VETENSKAP OCH KONST
KTH Matematik

En linjär kombination av $n$ s.v. $X_1, \ldots, X_n$

$$a_1X_1 + a_2X_2 + \cdots + a_nX_n + b$$

Konstanterna $a_1, \ldots, a_n$ och $b$ kan vara positiva eller negativa tal.
För alla s.v. $X_1, \ldots, X_n$ gäller att

$$E\Big(\sum_{i=1}^{n} a_iX_i + b\Big) = \sum_{i=1}^{n} a_iE(X_i) + b \qquad (2)$$

KTH VETENSKAP OCH KONST
KTH Matematik

För oberoende s.v. $X_1, \ldots, X_n$ gäller att

$$Var\Big(\sum_{i=1}^{n} a_i X_i + b\Big) = \sum_{i=1}^{n} a_i^2 Var(X_i).$$

Om $X_1, \ldots, X_n$ är s.v. med samma väntevärde $\mu$ så gäller att

$$E\Big(\sum_{i=1}^{n} X_i\Big) = n\mu. \tag{3}$$

Om $X_1, \ldots, X_n$ är oberoende och har samma standardavvikelse $\sigma$ gäller även att

$$Var\Big(\sum_{i=1}^{n} X_i\Big) = n\sigma^2 \quad \text{och} \quad D\Big(\sum_{i=1}^{n} X_i\Big) = \sigma\sqrt{n}. \tag{4}$$

# Bernoulli Distribution

Let $X \sim \mathcal{B}er\,(p)$. $X$ has two values, usually numerically coded as 0 and 1. The p.m.f. is

$$p_X(x) = \begin{cases} p & x = 1 \\ q = 1 - p & x = 0. \end{cases} \tag{5}$$

$$E\,[X] = p, \mathrm{Var}\,[X] = p(1-p).$$

## Definition

*En diskret s.v. $X$ säges vara binomialfördelad med parametrarna $n$ och $p$, $Bin(n, p)$-fördelad, om*

$$p_X(x) = \binom{n}{k} p^x (1-p)^{n-x}, \ \textit{för } x = 0, 1, \ldots, n.$$

Vi skriver detta med $X \sim \mathcal{B}in(n, p)$.
`binopdf(x,n,p)` $\leftarrow$ sannolikhetsfunktion $p_X(x)$ , `binocdf(x, p, n)` $\leftarrow$ fördelningssfunktion (cumulative distribution function) $F_X(x)$

KTH Matematik

# Diskreta st. v:er: Binomialfördelningen

De generella villkoren för detta:

- *n oberoende upprepningar av ett försök.*
- *varje försök har två utfall, 0 och 1.*
- *sannolikheten för lyckat försök (=1) är densamma = p vid varje försök.*

$X$ = antalet lyckade försök.

$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \text{ för } x = 0, 1, \ldots, n.$$

Vi skriver detta med $X \sim \mathcal{B}in(n, p)$.

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.

# 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 $\downarrow$.

# 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.

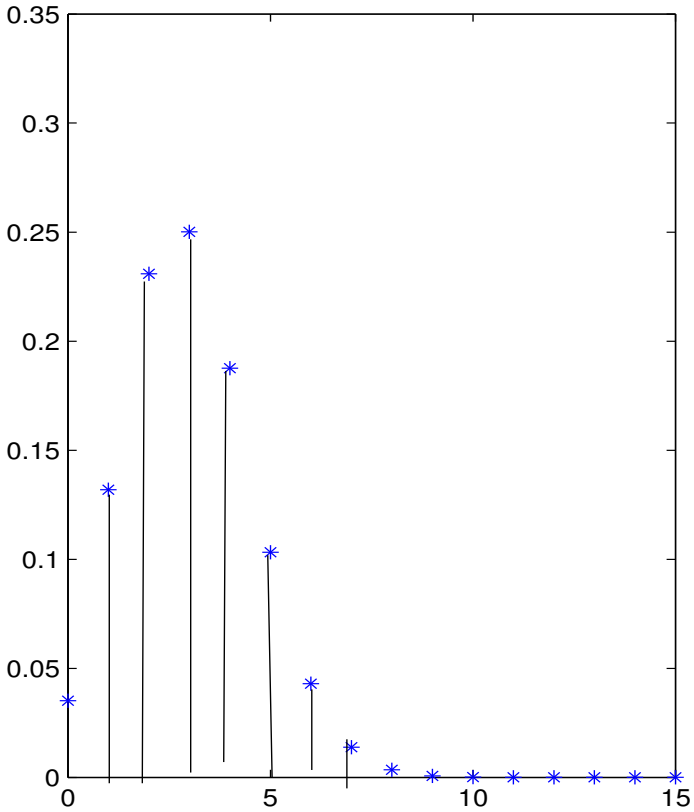
KTH Matematik

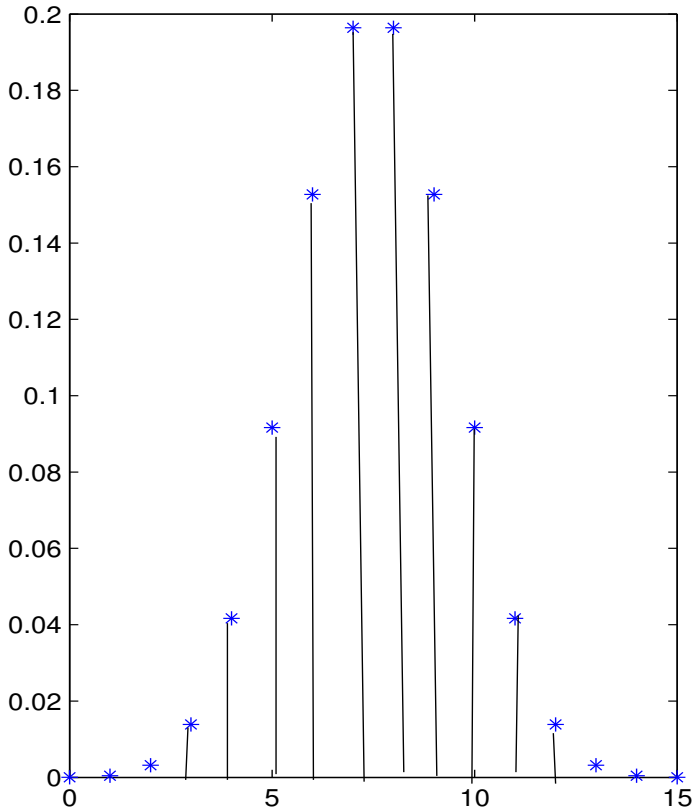
# Rare Event Rule

*If, under a given assumption, the probability of an observed event is extremely small ($\approx 0$), 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 $n = 15$, $p = \frac{1}{4}$, $x = 11$).

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

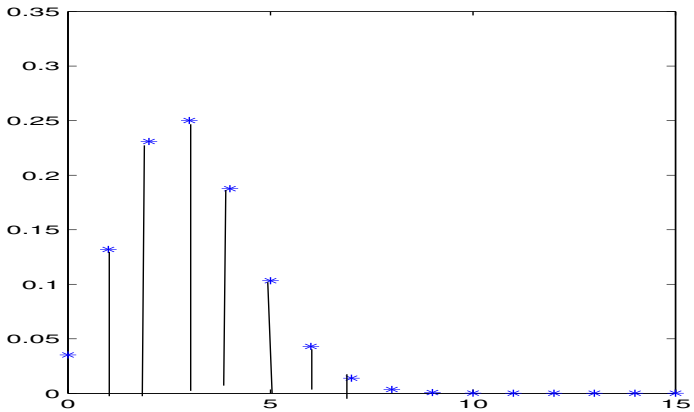


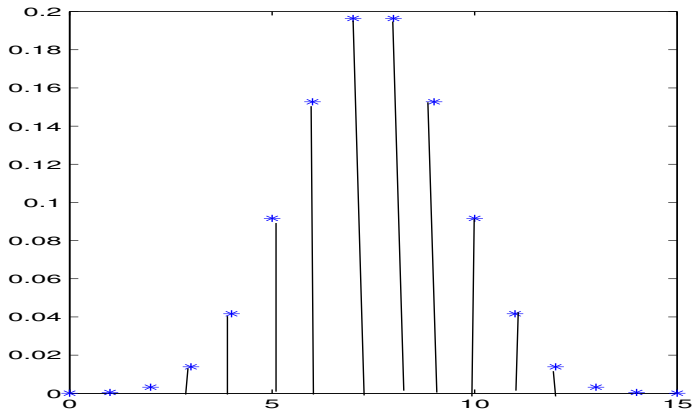




KTH VETENSKAP OCH KONST
KTH Matematik

# 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 VETENSKAP OCH KONST
KTH Matematik

# Binomial distribution

Note that

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

by the binomial formula

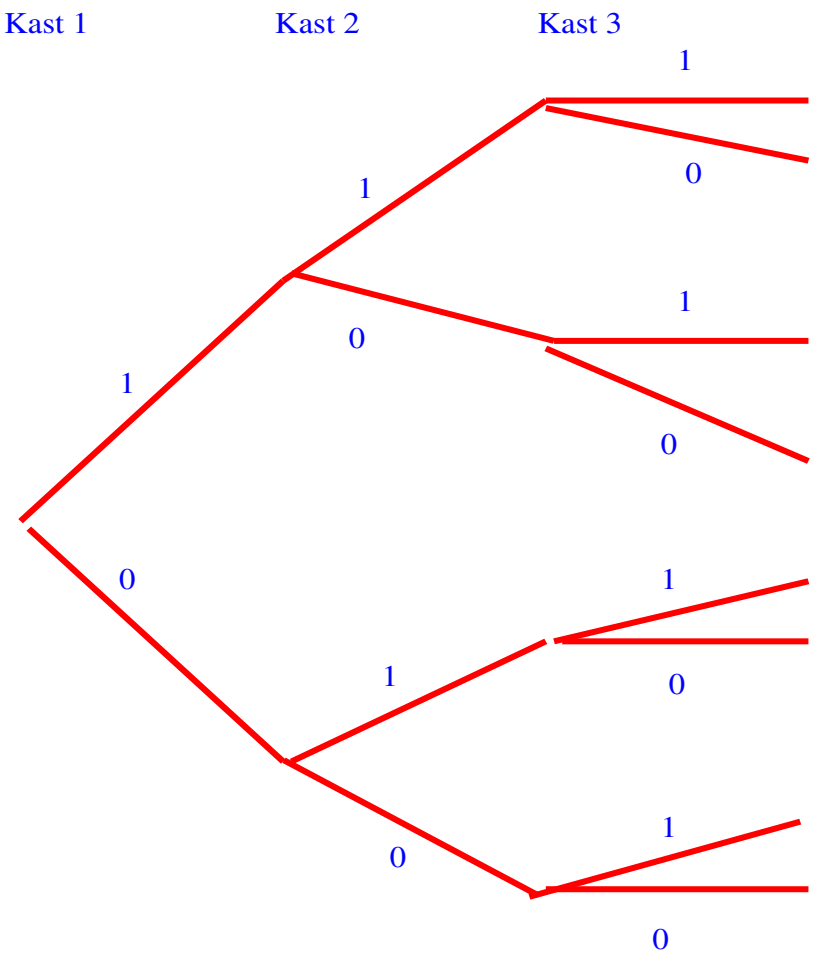
$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

# Kast av ett häftstift (= Thumbtack på (amerikansk) engelska)


Slumpexperiment (procedur): Kast av ett häftstift. Om det landar på spetsen som i bilden ovan, säger vi att en 'etta' (1) inträffar. Vi säger att en nolla (0) inträffar om häftstiftet landar på hatten.

KTH Matematik

# Tre kast av ett häftstift: utfallsrummet som ett träddiagram




$\omega_1 = (111)$

$\omega_2 = (110)$

$\omega_3 = (101)$

$\omega_4 = (100)$

$\omega_5 = (011)$

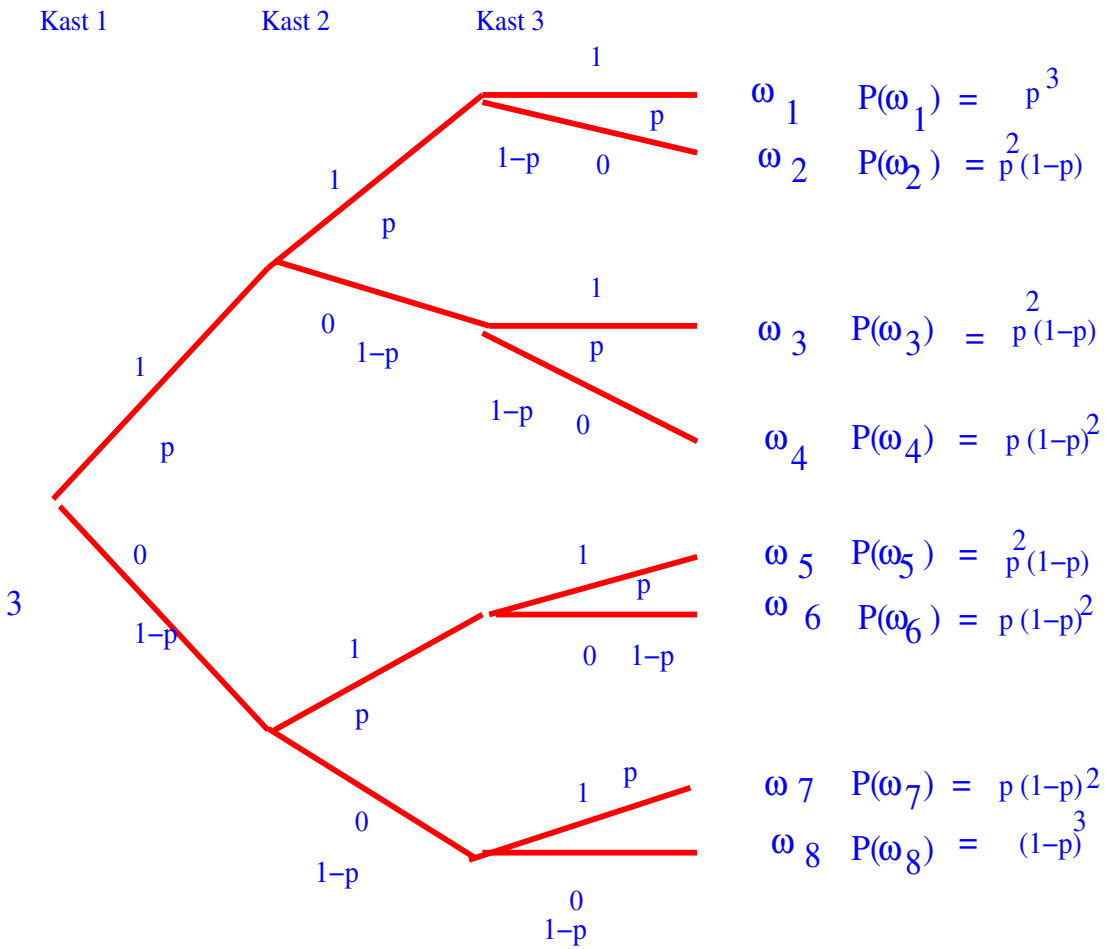
$\omega_6 = (010)$

$\omega_7 = (001)$

$\omega_8 = (000)$

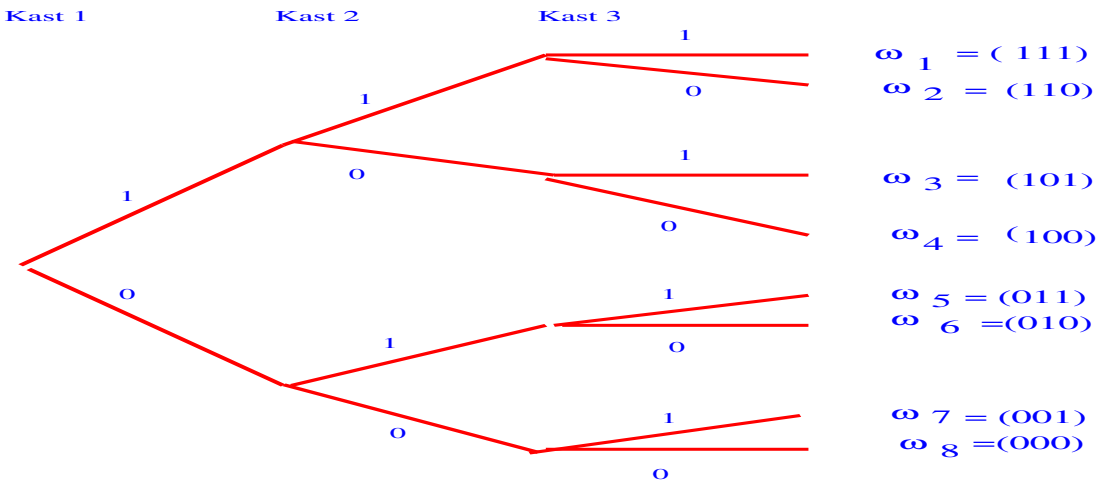
Antalet utfall$= 2^3$

# Tre oberoende kast av ett häftstift: sannolikheter




$$P(\text{en nolla } (0)) = 1 - p, P(\text{en etta } (1)) = p.$$

# Tre oberoende kast av ett häftstift: en stokastisk variabel




$X$ = 'antalet ettor i tre oberoende kast av ett häftstift'.

$X(\omega_1) = 3, X(\omega_2) = X(\omega_3) = X(\omega_5) = 2,$

$X(\omega_4) = X(\omega_6) = X(\omega_7) = 1,\ X(\omega_8) = 0.$

# sannolikhetsfunktion för antalet ettor

$X(\omega_1) = 3, X(\omega_2) = X(\omega_3) = X(\omega_5) = 2,$
$X(\omega_4) = X(\omega_6) = X(\omega_7) = 1,\ X(\omega_8) = 0.$

$$P(X = 3) = P(\omega_1) = p^3,$$

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

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

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

KTH VETENSKAP OCH KONST
KTH Matematik

# Tre oberoende kast av ett häftstift: sannolikhetsfunktion för antalet ettor

Vi omskriver dessa med hjälp av binomialkoefficienterna:

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

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

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

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

eller med en enda formel

$$P(X=x) = \binom{3}{x} p^x (1-p)^{3-x}, \quad , x = 0,1,2,3.$$

Denna är sannolikhetsfunktionen för binomialfördelningen med parametrarna 3 och $p$.

KTH Matematik

# Binomial DistributionExpectation and variance

$X \sim \mathcal{B}in\,(n, p)$, $0 \leq p \leq 1$, $q = 1 - p$, and the p.m.f. is

$$E\,[X] = np, \mathrm{Var}\,[X] = nqp.$$

KTH VETENSKAP OCH KONST
KTH Matematik

# Diskreta st. v:er: Poissonfördelningen

## Definition

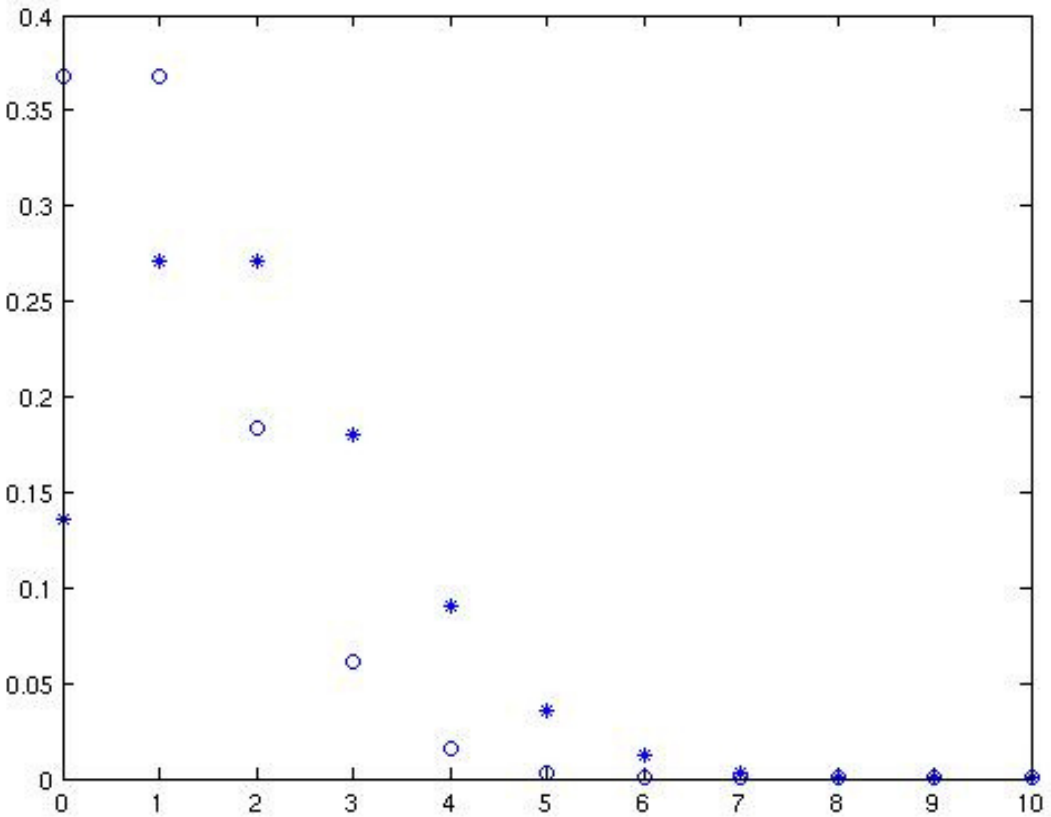
*En diskret s.v. $X$ säges vara Poissonfördelad med parameter $\mu$, $\mathcal{P}oi(\mu)$-fördelad, om*

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

Vi skriver detta med $X \sim \mathcal{P}oi(\mu)$. `poisspdf(x,`$\mu$`)`, `poisscdf(x,` $\mu$`)`.

KTH VETENSKAP OCH KONST
KTH Matematik

# Sannolikhetsfunktionerna för $\mathcal{P}oi(2)$ och $\mathcal{P}oi(1)$




$* \leftrightarrow \mathcal{P}oi(2)$, $\circ \leftrightarrow \mathcal{P}oi(1)$

Observera hur sannolikhetsmassan flyttas som funktion av $\mu$.

KTH VETENSKAP OCH KONST
KTH Matematik

# Binomial Distribution for small $p$ and large $n$ becomes Poisson

In calculations for molecular biology and biotechnology it happens often that $X \sim Bin(n, p)$ with $p$ being small and $n$ being large, or $p = \mu/n$. Then

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

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

KTH VETENSKAP OCH KONST
KTH Matematik

# Poisson distribution

Poisson distribution applies to occurrences 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 of a hospital on Fridays between 10.00 p.m. and 11.00 p.m.*

KTH Matematik

# Poisson

$X =$ Antalet vita blodkroppar per 1 $mm^3$ (kubikmillimeter) blod i ett prov .

Kanske

$$X \sim \mathcal{P}oi(\mu)$$

Men är $\mu$ densamma för alla individer?

# Poisson distribution

Note that

$$\sum_{x=0}^{\infty} p_X(x) = \sum_{x=0}^{\infty} \frac{\mu^x}{x!} e^{-\mu}$$

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

# Genome Coverage Distributions & Poisson distribution

Genome coverage is the number of sequencing reads[a] mapped to a position in a genome.

[a] In shotgun sequencing, sequences are obtained from each cloned fragment of DNA. Each nucleotide sequences is called a "read". The reads are used later to reconstruct the original sequence.

When reads are mapped to a reference genome, the per-base coverage for each position in the reference genome is given as the number of reads covering that position and yields the histogram over all per-base coverages.

KTH VETENSKAP OCH KONST
KTH Matematik

# Genome Coverage Distributions & Poisson distribution

Lander ES, Waterman MS.(1988) Genomic mapping by fingerprinting ran dom clones: a mathematical analysis, Genomics 2(3): 231-239.

KTH Matematik

# Genome Coverage Distributions & Poisson distribution

Lander and Waterman made two assumptions about sequencing:

- Reads will be distributed randomly across the genome
- Overlap detection does not vary between reads.

Based upon these two assumptions, they reached the conclusion that the number of times a base is sequenced follows a Poisson distribution.

KTH Matematik

# Genome Coverage Distributions & Poisson distribution

The Poisson distribution can be used to model any discrete occurrence given an average number of occurrences.The probability function is the following:

$$P(X = x) = e^{-C} C^x / x!$$

$x$ is the number of times a base is read and $C$ stands for coverage.
The Lander-Waterman model has served as an essential tool for estimating sequencing requirements for modern WGS (=whole genome sequence) experiments.

KTH VETENSKAP OCH KONST
KTH Matematik

# Väntevärde för Poissonfördelningen: detaljerna kommer inte på en tenta

$$p_X(x) = \frac{\mu^x}{x!} e^{-\mu}, \text{ för } x = 0, 1, 2 \ldots.$$

$$E(X) = \sum_{x=0}^{\infty} x \cdot \frac{\mu^x}{x!} e^{-\mu} = \sum_{x=1}^{\infty} x \cdot \frac{\mu^x}{x!} e^{-\mu} = \sum_{x=1}^{\infty} \frac{\mu^x}{(x-1)!} e^{-\mu}$$

$$= \mu \sum_{x=1}^{\infty} \frac{\mu^{x-1}}{(x-1)!} e^{-\mu} = \mu \sum_{x=0}^{\infty} \frac{\mu^x}{x!} e^{-\mu} = \mu.$$

$X \sim \mathcal{P}oi(\mu), E(X) = \mu$

KTH VETENSKAP OCH KONST
KTH Matematik

# Poissonfördelningen **EXTRA** detaljerna kommer inte på en tenta

$$E(X(X-1)) = \sum_{x=0}^{\infty} x(x-1) \cdot \frac{\mu^x}{xk!} e^{-\mu} = \sum_{x=2}^{\infty} x(x-1) \cdot \frac{\mu^x}{x!} e^{-\mu}$$

$$= \sum_{x=2}^{\infty} \frac{\mu^x}{(x-2)!} e^{-\mu} = \mu^2 \sum_{x=2}^{\infty} \frac{\mu^{x-2}}{(x-2)!} e^{-\mu} = \mu^2 \sum_{x=0}^{\infty} \frac{\mu^x}{x!} e^{-\mu} = \mu^2.$$

Detta ger $\mu^2 = E(X(X-1)) = E(X^2) - \mu$, eller $E(X^2) = \mu^2 + \mu$, vilket ger

$$Var(X) = E(X^2) - \mu^2 = \mu^2 + \mu - \mu^2 = \mu.$$

$X \sim \mathcal{P}oi(\mu), Var(X) = \mu$

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. $0 < p < 1$, $q = 1 - p$. The p.m.f. of $X \sim \mathcal{G}e(p)$ is

$$p_X(x) = q^x p, \quad x = 0, 1, 2, \ldots$$

$$E[X] = \frac{q}{p}, \mathrm{Var}[X] = \frac{q}{p^2}.$$

KTH Matematik

# First Success Distribution

$X \sim \mathcal{G}eom(p)$, $0 < p < 1$, $q = 1 - p$. The p.m.f. is

$$p_X(x) = q^{x-1}p, \quad x = 1, 2, \ldots.$$

$$E[X] = \frac{1}{p}, \mathrm{Var}[X] = \frac{q}{p^2}.$$

KTH VETENSKAP OCH KONST
KTH Matematik

# First Success: an Example

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

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

KTH Matematik

# First Success: another Example

In bioinformatics one often renames $1-p$ as 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.*

# Negative Binomial Distribution

A procedure meets all the conditions of a binomial distribution. You set $X$ = number of failures until you have got $n$ successes. Then $X$ follows the Negative Binomial distribution, $X \sim \mathcal{NB}(n, p)$, $0 < p < 1$.

# Negative Binomial Distribution

$X$ is said to follow the Negative Binomial distribution, $X \sim \mathcal{NB}(n, p)$, $0 < p < 1$, $q = 1 - p$, if its p.m.f. is

$$p_X(x) = \binom{n+x-1}{x} p^n q^x, \quad x = 0, 1, 2, \ldots \tag{6}$$

$$E[X] = n\frac{q}{p}, \mathrm{Var}[X] = n\frac{q}{p^2}.$$

Obs! $\mathcal{G}e(p) = \mathcal{NB}(1, p)$.

KTH VETENSKAP OCH KONST
KTH Matematik

# Hypergeometric Distribution

The hypergeometric distribution is a discrete probability distribution that describes the probability of $x$ successes in $n$ draws, without replacement, from a finite population of size $m$ that contains exactly $k$ successes, wherein each draw is either a success or a failure.

$$X = \text{number of succcesses}$$

# Hypergeometric Distribution

$X \sim \mathcal{HG}(m, k, n)$, the hypergeometric distribution with parameters $(m, k, n)$ and $x \in \{o, \ldots, \min(n, k)\}$

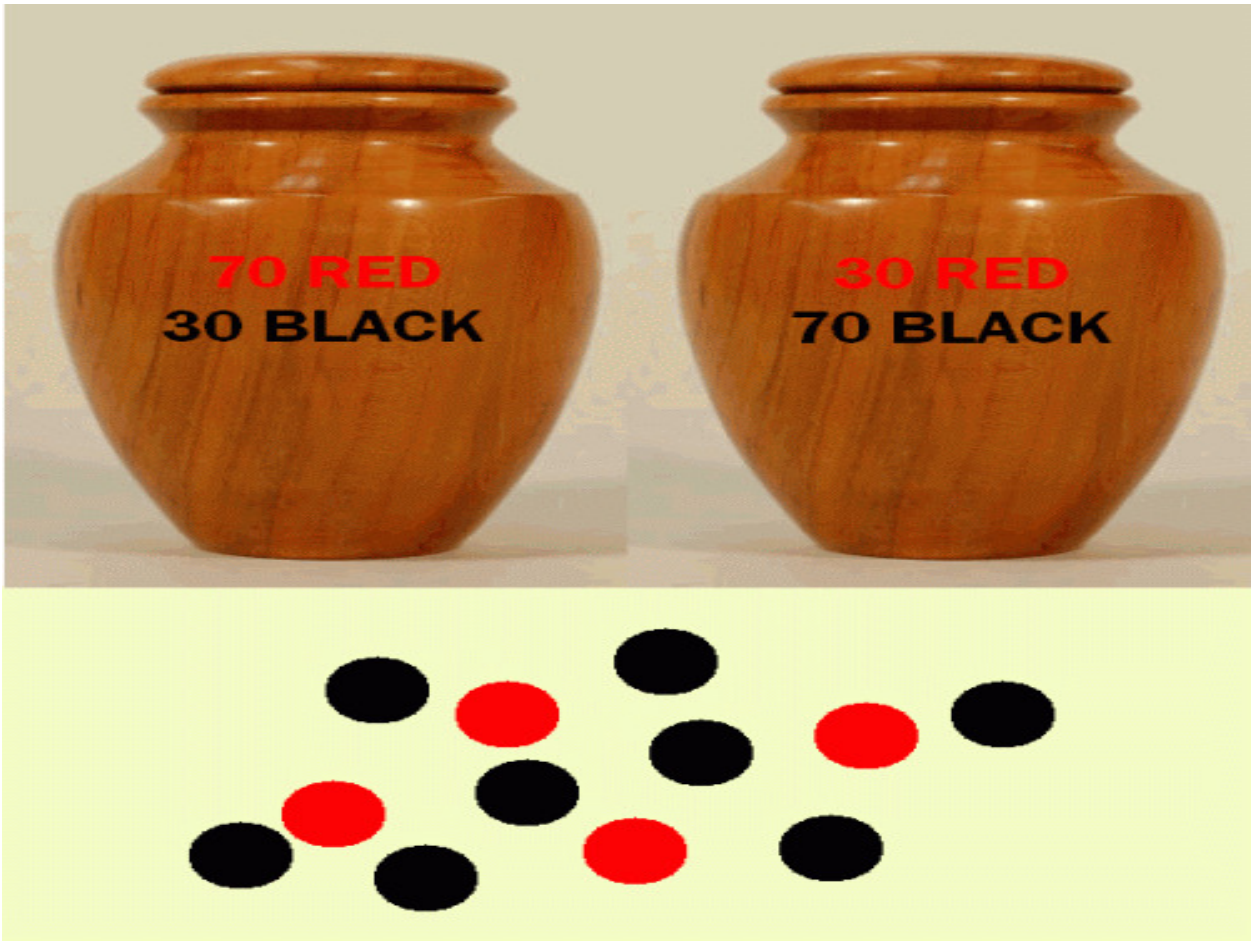
$$p_X(x) = \frac{\binom{k}{x}\binom{m-k}{n-x}}{\binom{m}{n}}$$

$$E[X] = n\frac{k}{m} \qquad \text{Varians } \operatorname{Var}[X] = \frac{nk}{m}\frac{m-k}{m}\frac{m-n}{m-1}$$

KTH VETENSKAP OCH KONST

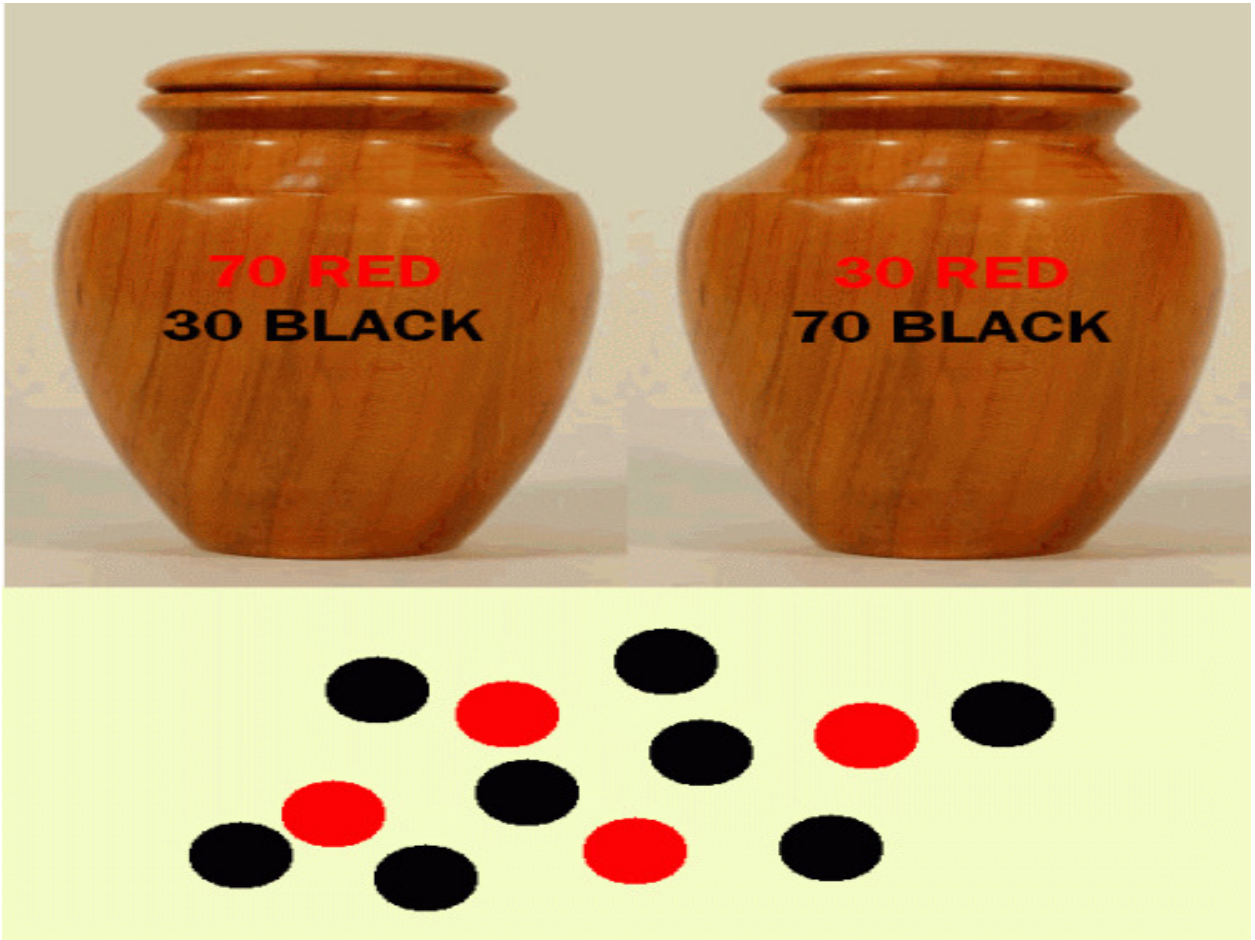
KTH Matematik

You see a couple of urns. The one on the left contains 70 red balls and 30 black. The one on the right contains 30 red and 70 black. While you were not looking, a friend reached into one of these urns and randomly drew out a dozen balls. As you can see, 4 of them were red and 8 were black.




KTH Matematik

## Which urn did the balls come from? Probability model..




KTH Matematik