Statistics for Bioinformatics

Theoretical distributions of probability

Jacques van Helden
Jacques.van.Helden@ulb.ac.be
Jacques.van-Helden@univ-amu.fr

Combinatorial analysis

Problem - oligomers

- How many oligomers contain exactly a single occurrence of each monomer, for oligonucleotides and oligopeptides, respectively ?

Permutations within a set - the factorial

- How many distinct permutations can be made from a set of x elements ?
  - $x = 2$ 2
  - $x = 3$ $3*2 = 6$
  - $x = 4$ $4*3*2 = 24$
  - *any x* $x*(x-1)...1 = x!$
- The **factorial** ***x!*** represents the number of possible permutations between x objects.
- Solution to the problem of oligomers
  - There are 4!=24 distinct oligonucleotides with a single occurrence of each nucleotide (A, C, G, T)
  - There are 20!=2.4*10^{18} distinct oligopeptides with a single occurrence of each amino acid.

Problem - Selection of a subset of elements

- A genome contains n=6000 genes.
- We select a series of genes in the following way :
  - Once a gene has been selected once, it cannot be selected anymore (**no replacement**)
  - We are not interested in the order of the selection: if A and B were selected, we do not consider whether A came out in first or in second position.
- How many possibilities do we have to select
  - 1 gene ?
  - 2 genes ?
  - 3 genes ?
  - x genes ?

Selection of a subset of elements

Selection	Possible outcomes		Possible orderings		Distinct outcomes
	calculation	value	calculation	value	(orderless)
1	6000	6.00E+03	1	1	6.00E+03
2	6000*5999	3.54E+07	2*1	2	1.77E+07
3	6000*5999*5998	2.16E+11	3*2*1	6	3.60E+10
...	...	...	...	...	...
10	6000*5999*...*5991	6.00E+37	10*9*...*1	3628800	1.65E+31

- Number of possible outcomes
  - n size of the set
  - x size of the subset

$$n(n-1)(n-2)...(n-x+1) = \frac{n!}{(n-x)!}$$

- Possible permutations among the elements of a subset

$$x!$$

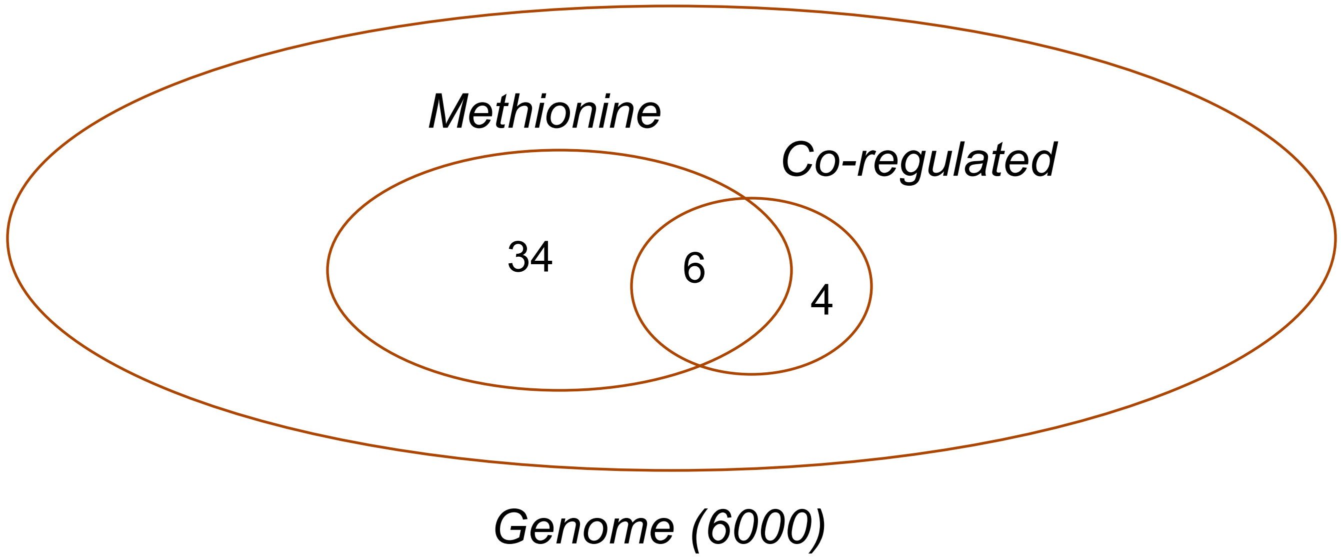
- Number of distinct selections (orderless).
- The coefficient C_x^n represents the number of distinct choices of x elements among n. For this reason, it is called "***Choose x among n***". It is also called ***binomial coefficient*** (we will see later why).

$$C_n^x = \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Set comparisons

Problem - selection within a set with classes

- A given organism has 6,000 genes, 40 of which are involved in methionine metabolism.
- A set of 10 genes were reported as co-regulated in a microarray experiment. Among them, 6 are related to methionine metabolism.
- How significant is this observation ? More precisely, what would be the probability to observe such a correspondence by chance alone ?



Selection within a set with classes

- Let us define
  - g = 6000 number of genes
  - m = 40 genes involved in methionine metabolism
  - n = 5960 genes not involved in methionine metabolism
  - k = 10 number of genes in the cluster
  - x = 6 number of methionine genes in the cluster

- We calculate the number of possibilities for the following selections
  - **C1**: 10 distinct genes among 6,000
  - **C2**: 6 distinct genes among the 40 involved in methionine
  - **C3**: 4 genes among the 5960 which are not involved in methionine
  - **C4**: 6 methionine and 4 non-methionine genes

$$C1 = C_{m+n}^{k} = \frac{6000!}{10!5990!} = 1.65e^{31}$$

$$C2 = C_{m}^{x} = \frac{40!}{6!34!} = 3.8e^{6}$$

$$C3 = C_{n}^{k-x} = \frac{5960!}{4!5956!} = 5.2e^{13}$$

$$C4 = C_{m}^{x} C_{n}^{k-x} = 2.0e^{20}$$

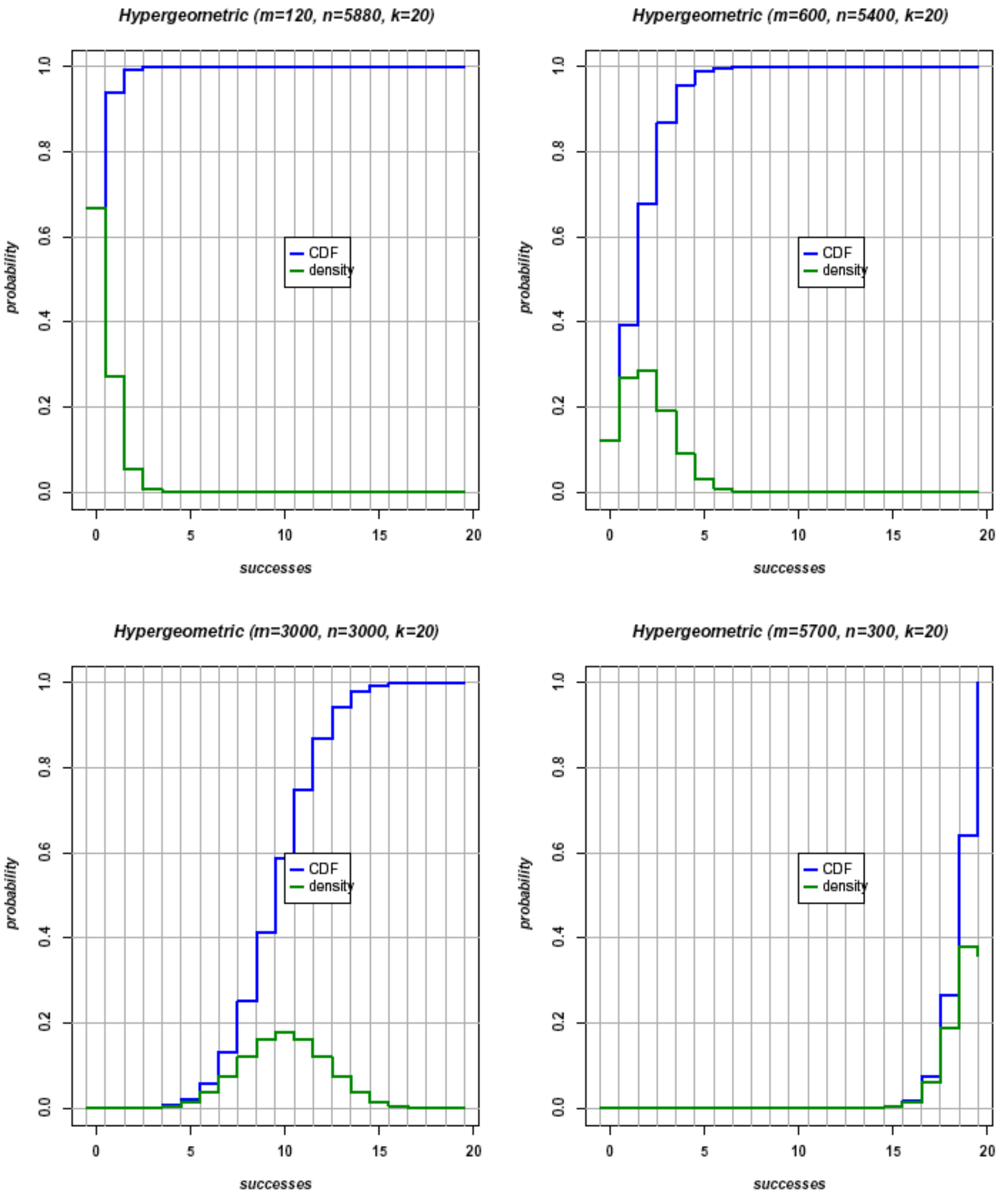
- Probability to have exactly 6 methionine genes within a selection of 10

$$P(X = 6) = \frac{C_{m}^{x} C_{n}^{k-x}}{C_{m+n}^{k}} = 1.219e^{-11}$$

- Probability to have at least 6 methionine genes within a selection of 10

$$P(X \geq 6) = \sum_{i=x}^{k} \frac{C_{m}^{i} C_{n}^{k-i}}{C_{m+n}^{k}} = 1.222e^{-11}$$

The hypergeometric distribution



- The hypergeometric distribution represents the probability to observe x successes in a sampling without replacement
  - ❑ *m* *number of marked elements in the set*
  - ❑ *n* *number of non-marked elements in the set*
  - ❑ *k* *sample size*
  - ❑ *x* *number of marked elements in the sample*

$$P(X = x) = \frac{C_m^x C_n^{k-x}}{C_{m+n}^k}$$

- The shape of the distribution depends on the ratio between *m* and *n*
  - ❑ $m << n$ *i-shaped*
  - ❑ $m \sim n$ *bell-shaped*
  - ❑ $m >> n$ *j-shaped*
- ❑ The distribution is bounded on both sides (0 ≤ x ≤ k).
- ❑ Statistical parameters

$$\min(x) = 0$$
$$\max(x) = \min(k,m)$$
$$\mu = km/(m+n)$$
$$\sigma^2 =$$

Statistics Applied to Bioinformatics

Bernoulli Schemas

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

Bernoulli trial

- A Bernoulli trial is an experiment whose outcome is random and can lead to either of two possible outcomes, called ***success*** and ***failure***, respectively.
- Examples :
  - Selection of a random nucleotide. Success if the nucleotide is a G.
  - Looking at a position from an alignment of two sequences. Success if this position corresponds to a match.
  - Selection of one gene from the yeast genome; success if the gene belongs to a specific functional class (e.g. Methionine biosynthesis).

Bernoulli schema

- A Bernoulli schema is a succession of n trials, each of which can lead or not to the realization of an event A.
  - Trials must be independent from each other
  - The probability of success is constant during the n trials
    - p is the probability of success at each trial
    - $q = 1 - p$ is the probability of failure at each trial
- Examples :
  - generation of a random sequence of length n; event X is the addition of a purine

Extreme cases - all successes or all failures

- What is the probability to observe n successes during the n trials ?

  - We can apply the joint probability for stochastically independent events :

$$\boxed{P(A_1, A_2, \ldots, A_n) = P(A_1)P(A_2)\ldots P(A_n)}$$

  - And since the probability of success is constant during the trials

$$\boxed{P(A_1, A_2, \ldots, A_n) = P(A)^n = p^n}$$

- What is the probability to observe *n* failures during the *n* trials ?

$$\boxed{P(\neg A_1, \neg A_2, \ldots, \neg A_n) = P(\neg A)^n = (1 - p)^n = q^n}$$

Problems - series of successes/failures

- In a random gapless alignment of two DNA sequences, what is the probability to observe a succession of exactly 10 matches at a given position ?

```
ATTAGTACCGTAGTAA
||||||||||-|--||
ATTAGTACCGCACAAA
```

- In a random sequence with equiprobable nucleotides, what is the probability to observe the first G at the 30th position ?

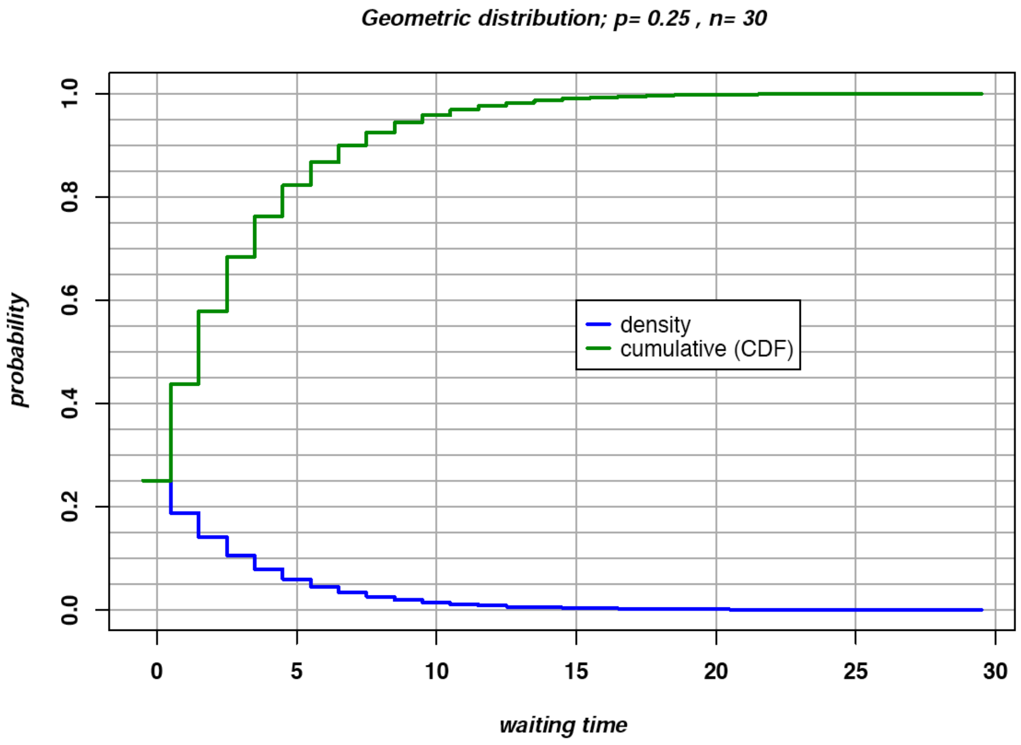
```
123456789012345678901234567890
ATTACTCTTACTCTCATCTATCTTTCATCG
```

Series of successes/failures

- In a random gapless alignment of two DNA sequences, what is the probability to observe a succession of exactly 10 matches at a given position ?
  - $P(match) = p = 0.25$
  - $P(10\ matches) = p^{10} = 9.54e^{-7}$
  - $P(mismatch) = 1 - p = 0.75$
  - $P(10\ matches\ and\ 1\ mismatch) = p^{10}(1 - p) = 7.15e^{-7}$

- In a random sequence with equiprobable nucleotides, what is the probability to observe the first G at position 30 ?
  - $P(G) = p = 0.25$
  - $P(not\ G) = 1 - p = 0.75$
  - $P(no\ G\ between\ positions\ 1\ and\ 29) = (1 - p)^{29} = 2.38e^{-4}$
  - $P(first\ G\ at\ position\ 30) = (1 - p)^{29}p = 5.95e^{-5}$

The geometric distribution



- The geometric distribution is used to calculate the probability to observe
  - x consecutive successes followed by a failure

$$P(X = x) = p^x (1 - p)$$

  - x consecutive failures followed by a success

$$P(X = x) = (1 - p)^x p$$

Defined succession of successes and failures

- What is the probability to first observe s consecutive successes, followed by n-s consecutive failures ?

$$\boxed{P(A_1, A_2, \ldots, A_s) = p^s}$$

$$\boxed{P(\neg A_{s+1}, \neg A_{s+2}, \ldots, \neg A_n) = q^{n-s}}$$

$$\boxed{P(A_1, A_2, \ldots, A_s, \neg A_{s+1}, \ldots, \neg A_n) = p^n q^{n-s}}$$

Permutations of successes and failures

- How many ways are there to permute s successes and n-s failures ?

- The number of permutations of x **distinct** objects is given by the factorial

$$\boxed{\begin{array}{l} 0! = 1 \qquad\qquad 1! = 1 \\ x! = x(x-1)! \\ x! = x(x-1)(x-2)\ldots2 \quad when \quad x \quad is \quad l\arg e \end{array}}$$

- However
  - The s successes are not distinct from each other
  - The n-s failures are not distinct from each other

- The number of permutations of s objects of one type and n-s objects of the other type is given by the *binomial coefficient*

$$\boxed{C_n^s = \binom{n}{s} = \frac{n!}{s!(n-s)!}}$$

The binomial distribution (Bernoulli distribution)

- What is the probability to observe x successes during the n trials (irrespective of the particular order of succession) ?

$$\boxed{P(X=x)=\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}=C_n^x p^x(1-p)^{n-x}}$$

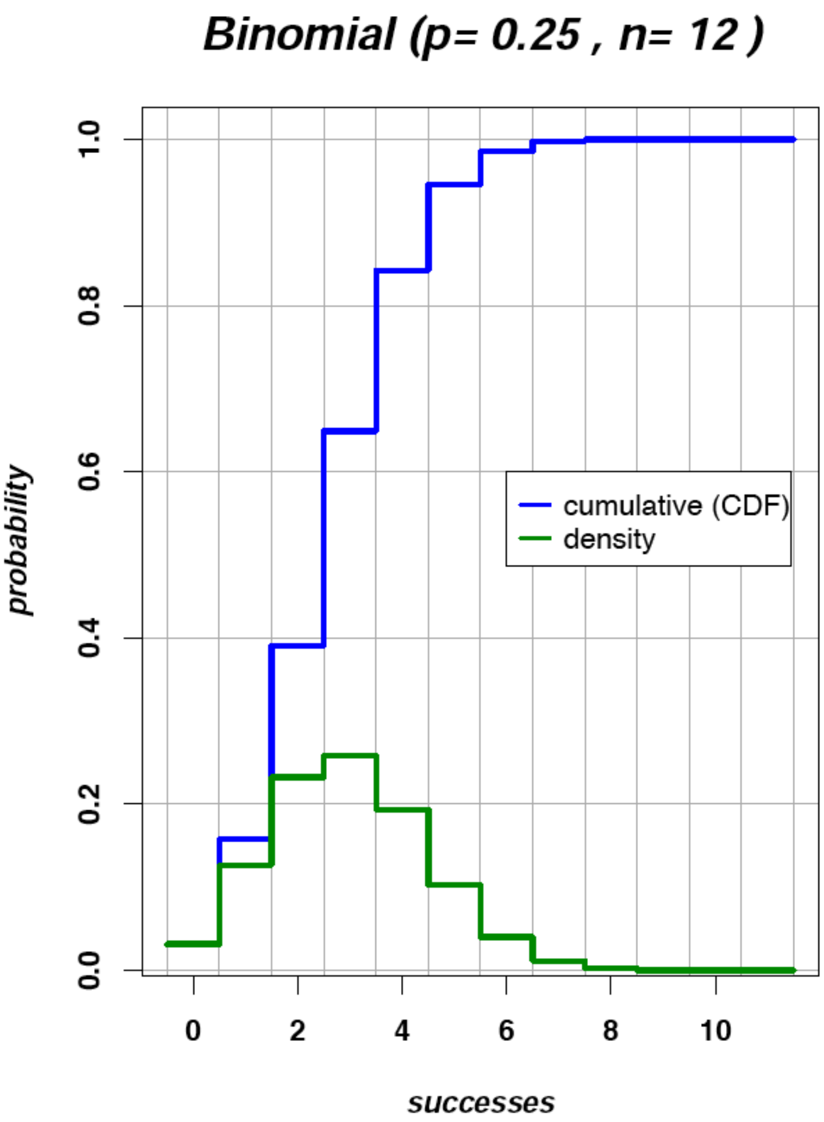
- This is the **binomial probability**.
- In this formula, the term C_n^x (*choose* x *among* n) is called the **binomial coefficient**.

- What is the probability to observe **up to** x successes during the n trials (irrespective of the particular order of succession) ?

$$\boxed{P(X\leq x)=\sum_{i=0}^{x}\frac{n!}{i!(n-i)!}p^i(1-p)^{n-i}}$$

- This is the **binomial cumulative distribution function** (**CDF**).

The binomial distribution



- The binomial distribution represents the probability to observe x successes in a Bernoulli trial (such as a sampling with replacement).
- Parameters
  - *p* *the probability of success at each trial*
  - *n* *number of trials*
  - *x* *number of successes in the sample*
- Values (X axis) are
  - always positive
  - comprised between *0* and *n*
- Probabilities (Y axis) are comprised between 0 and 1
- In R
  - dbinom(x,n,p) ## Density function
  - pbinom(x,n,p) ## CDF, left tail, inclusive
  - pbinom(x,n,p,lower.tail=F) ## CDF, right tail, exclusive
  - pbinom(x-1,n,p,lower.tail=F) ## CDF, right tail, inclusive

$$\min(s) = 0$$

$$\max(s) = n$$

$$\mu = np$$

$$\sigma^2 = np(1-p) = npq$$

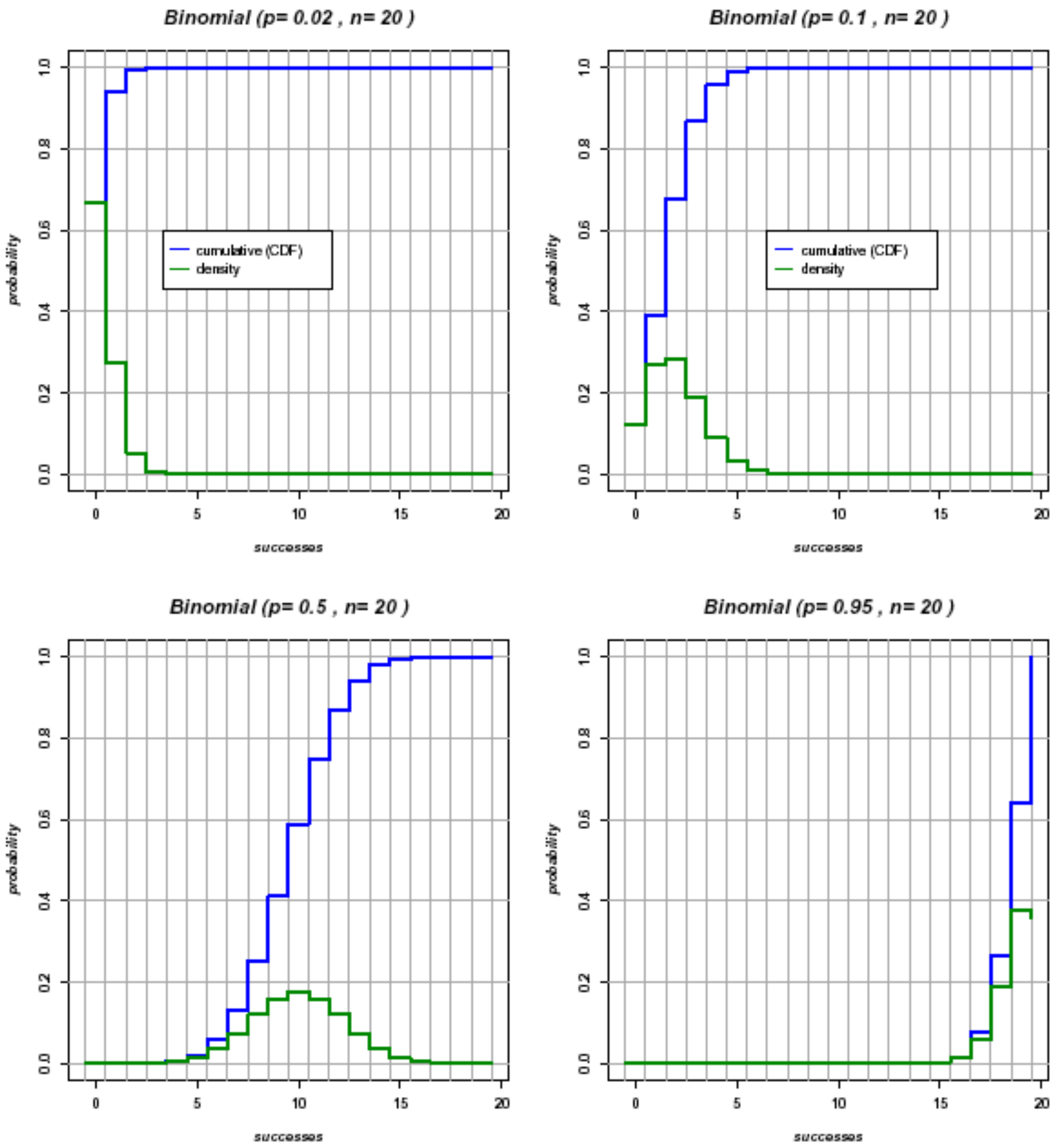
Binomial : efficient computation

- The binomial probability can be computed efficiently by using a recursive formula.
- This drastically reduces the computation time.

$$P(X = 0) = (1 - p)^n$$

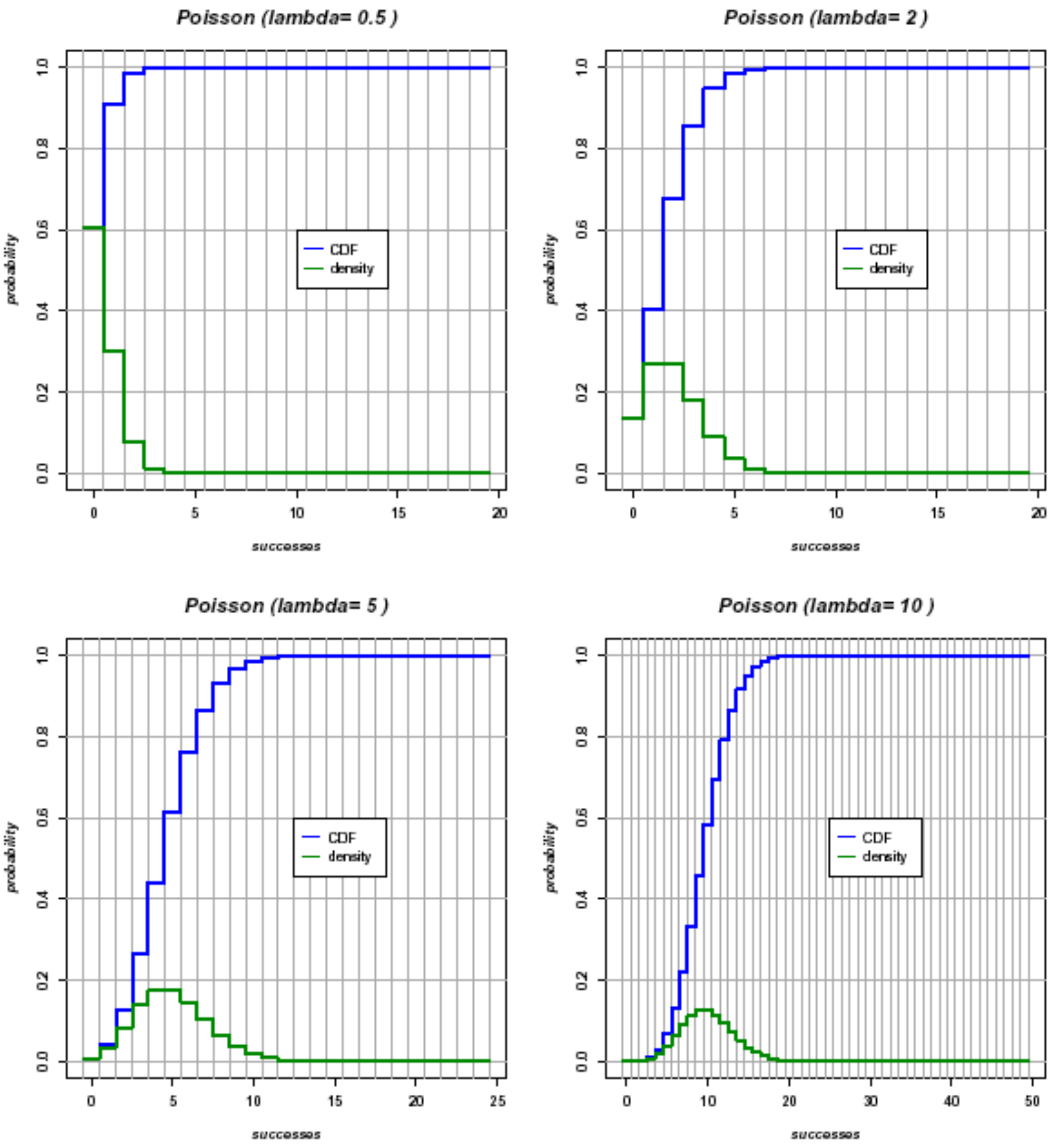
$$P(X = x + 1) = P(X = x)\frac{p(n - x)}{(1 - p)(x + 1)}$$

Binomial - effect of p (probability of success)



- The curve can take different shapes
    - i-shaped (small p)
    - bell-shaped (intermediate p)
    - j-shaped (high p)
- The curve is asymmetric, except when p=0.5
- The curve is bounded on both sides (0 ≤ s ≤ n)

Poisson distribution



$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$\min(X) = 0 \qquad \max(X) = \infty$$
$$\mu = \lambda \quad \sigma^2 = \lambda$$

- The Poisson distribution is characterized by a single parameter, λ, which is the mean of the distribution.
- The Poisson distribution can be used as an approximation of the binomial when
  - $n \to \infty$
  - $p \to 0$
  - $\lambda = p*n$ is small (e.g. < 5)
- The curve is bounded on the left (min=0).

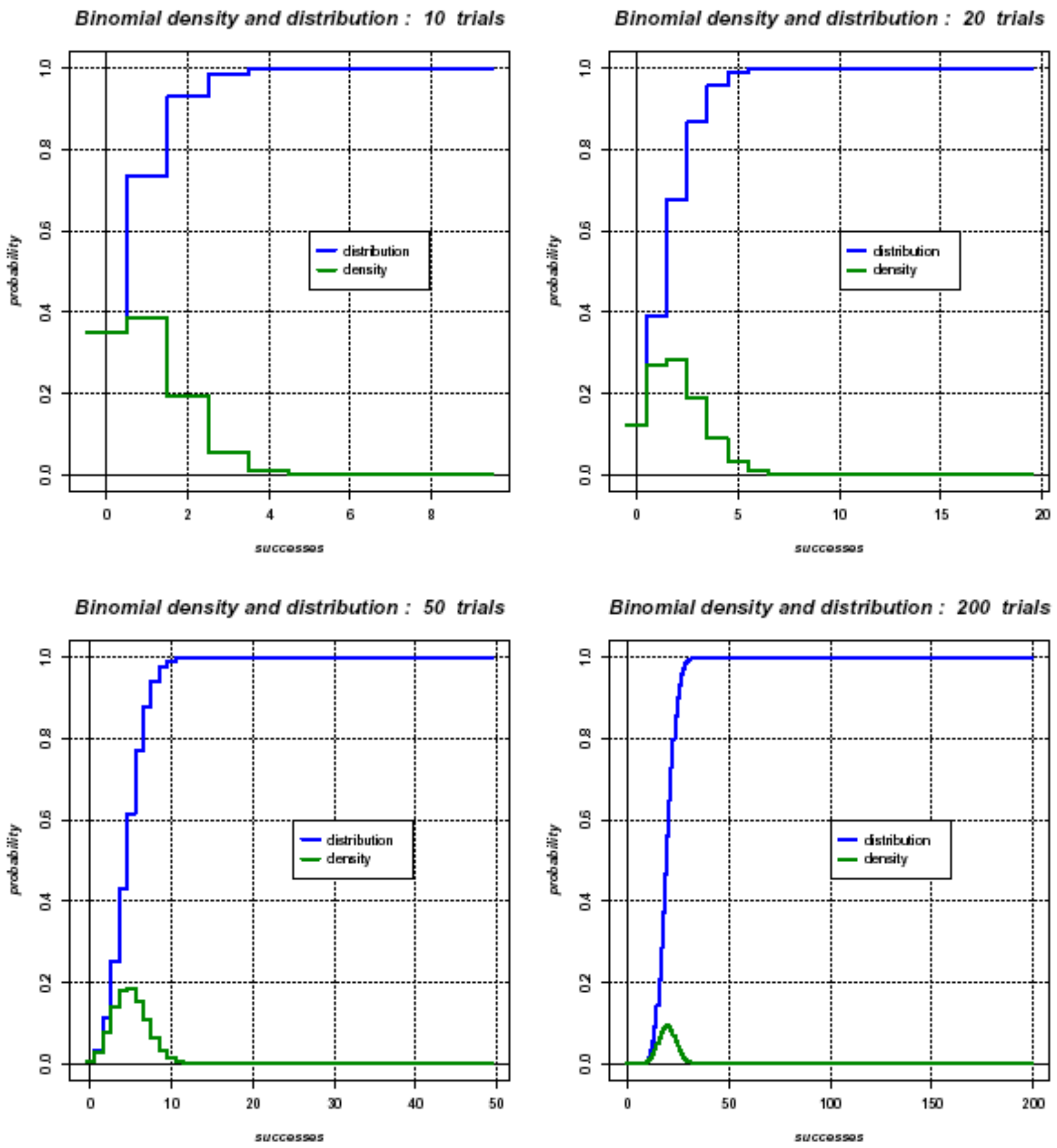
Poisson - efficient computation

- The Poisson probability can be calculated efficiently with a recursive formula

$$P(X=0)=e^{-E_W}$$

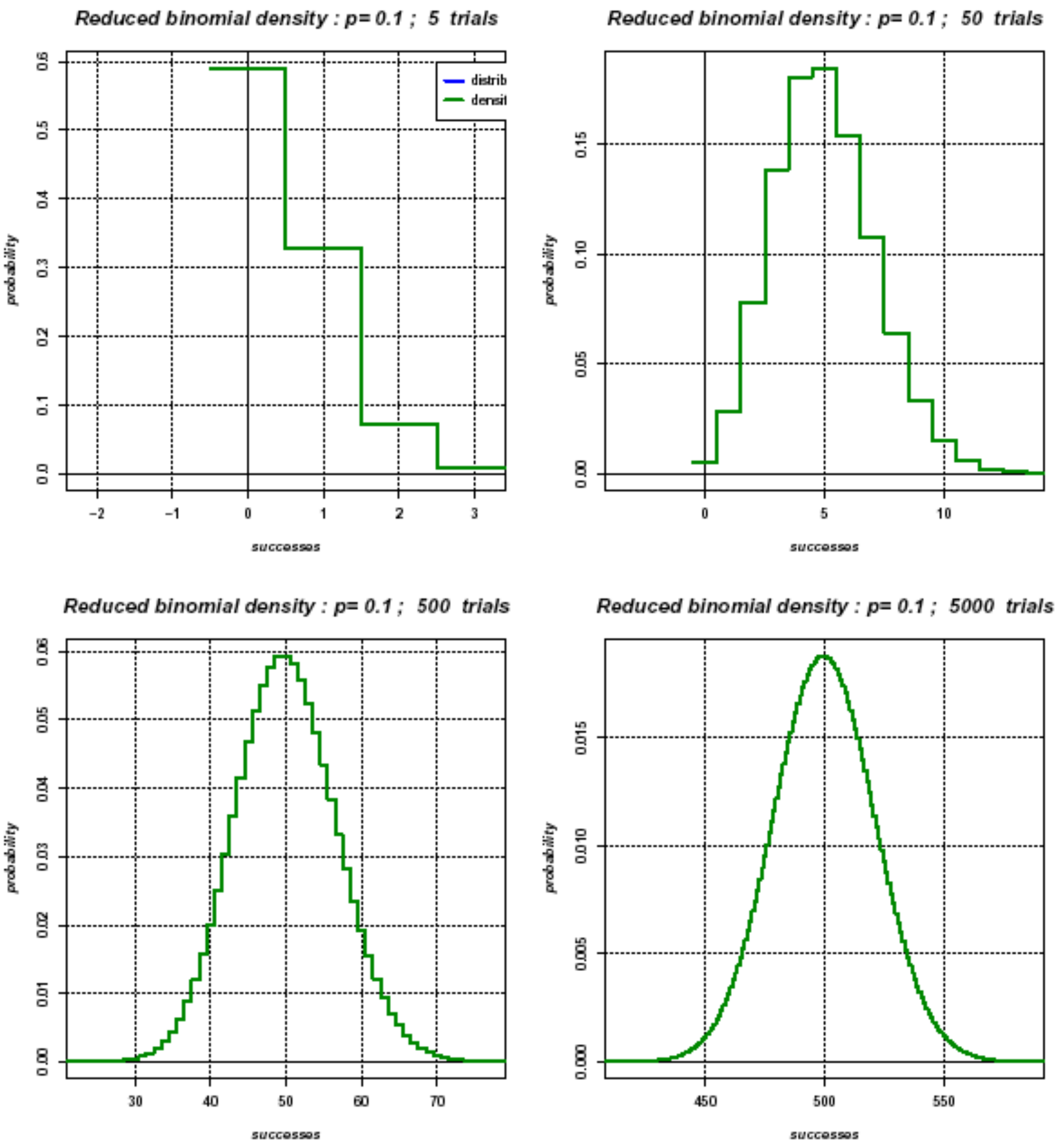
$$P(X=C_w+1)=P(X=C_w)\frac{E_w}{(z+1)}$$

Binomial - effect of n (number of trials)



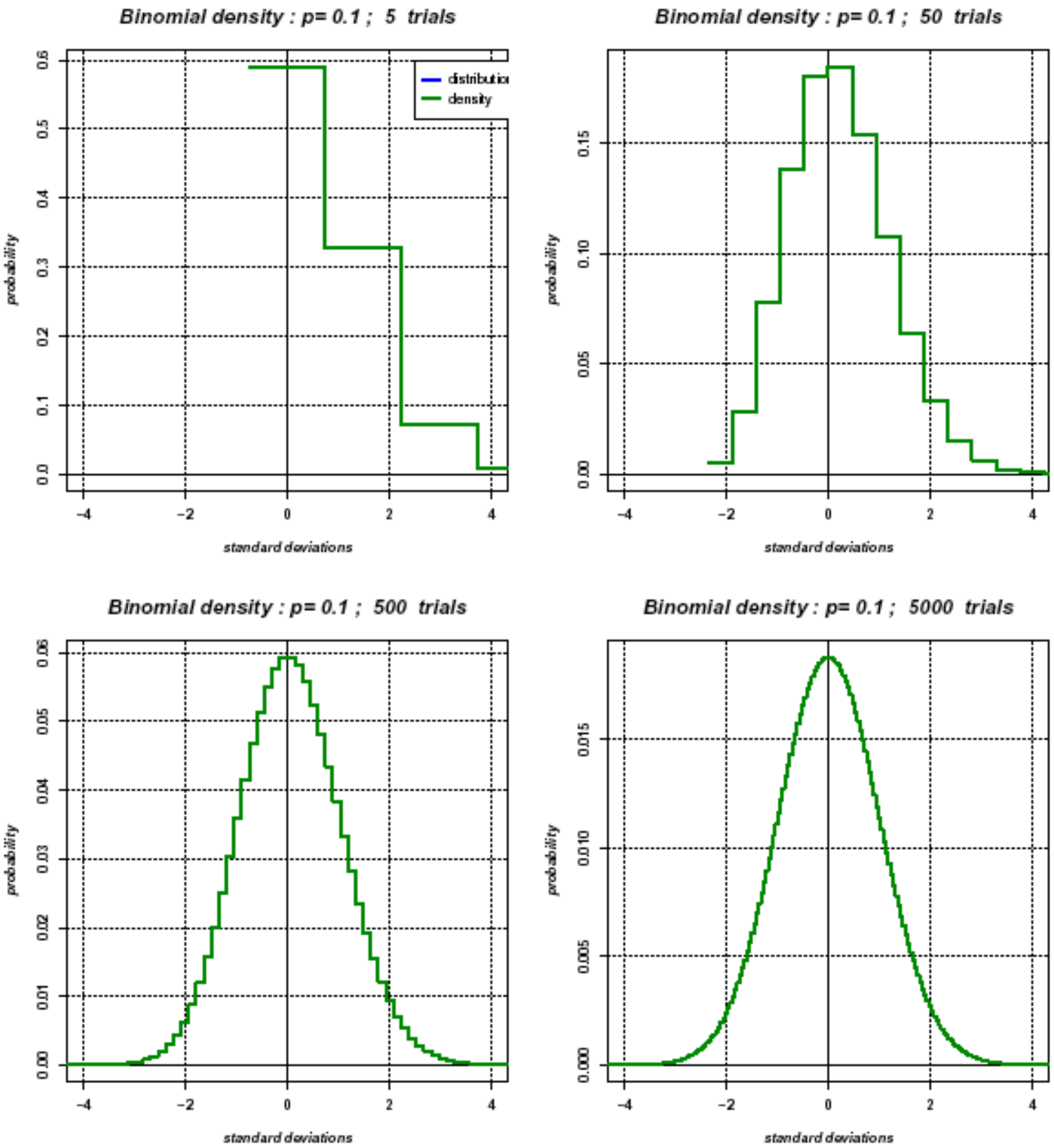
- When the number of trials increases
  - The number of distinct values for s increases
  - The probability of each value decreases
  - The binomial tends towards a bell-shaped curve

Binomial - effect of n (number of trials)



- On this figure, the density is displayed around the mean of the binomial ($\mu = np$).
- When n increases :
  - The number of distinct values for s increases.
  - The probability of each value decreases.
  - The binomial tends towards a bell-shaped curve.
- When $n \rightarrow \infty$
  - The binomial tends towards a continuous density function

Reduced binomial distribution -> Normal



- Starting from a binomial distribution, let n -> *Inf*
- Let us replace x by the *reduced variable* U

$$U = \frac{x-\mu}{\sigma} = \frac{x-np}{\sqrt{np(1-p)}} = \frac{x-np}{\sqrt{npq}}$$

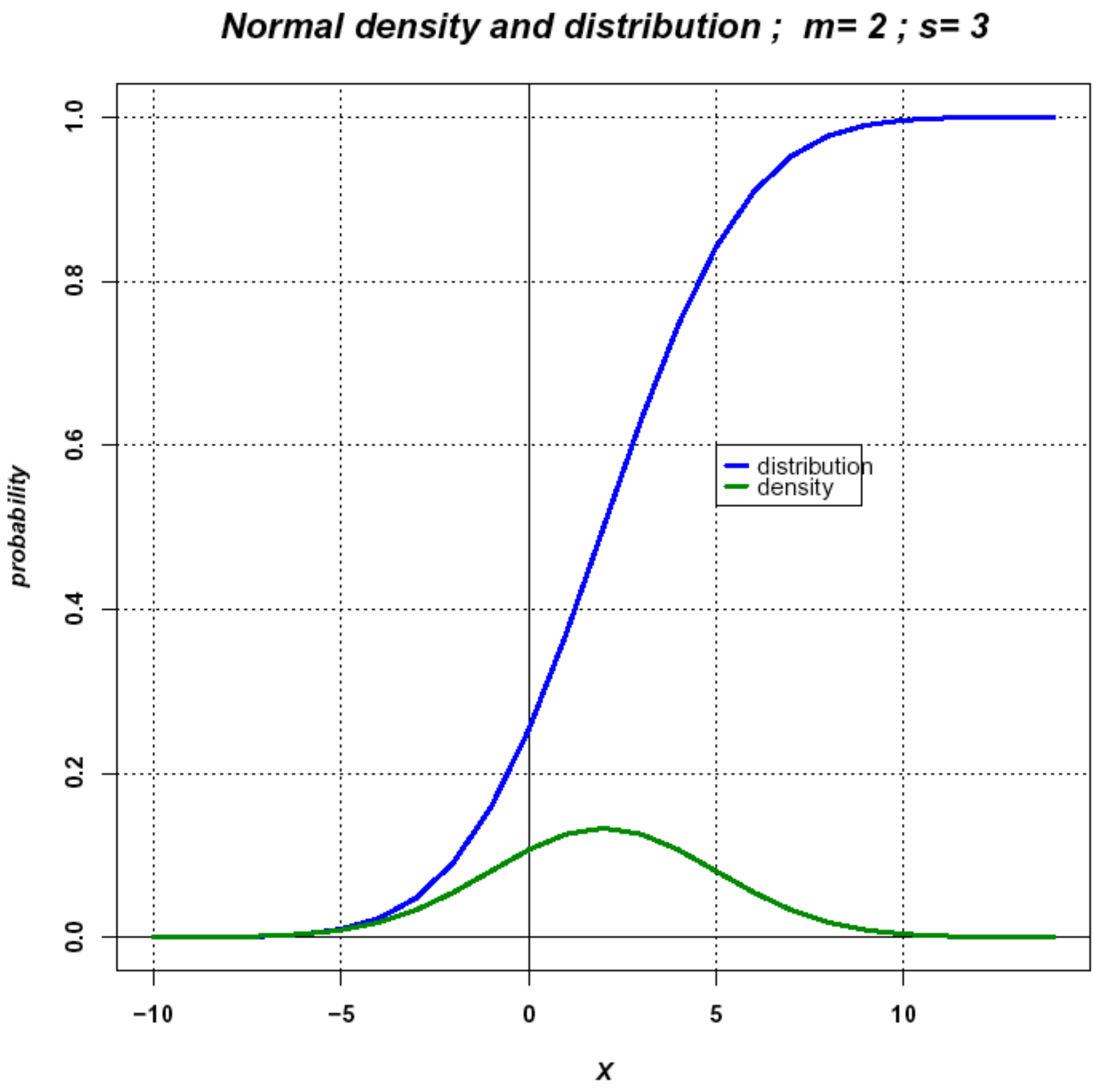
- When $n \to \infty$, the binomial tends towards the ***standard normal density function***

$$f_N(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$$

- The ***cumulative density function*** (**CDF**) is obtained by integrating the density function

$$F_N(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-u^2/2} du$$

Normal distribution



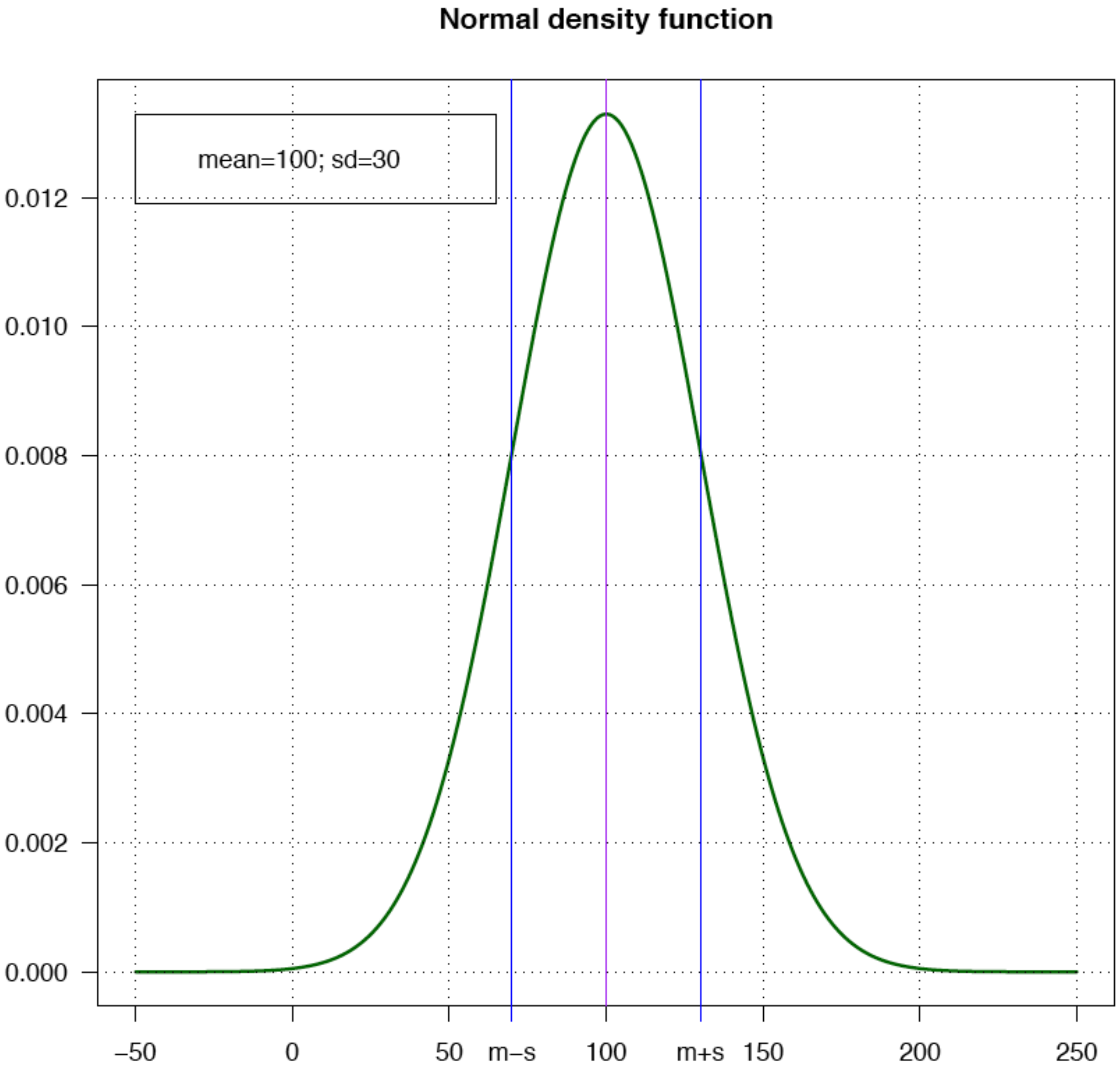
- A normal distribution with mean μ and a variance σ^2 is defined by the density function

$$f_N(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- The distribution function is obtained by integrating the density function from -∞ to x

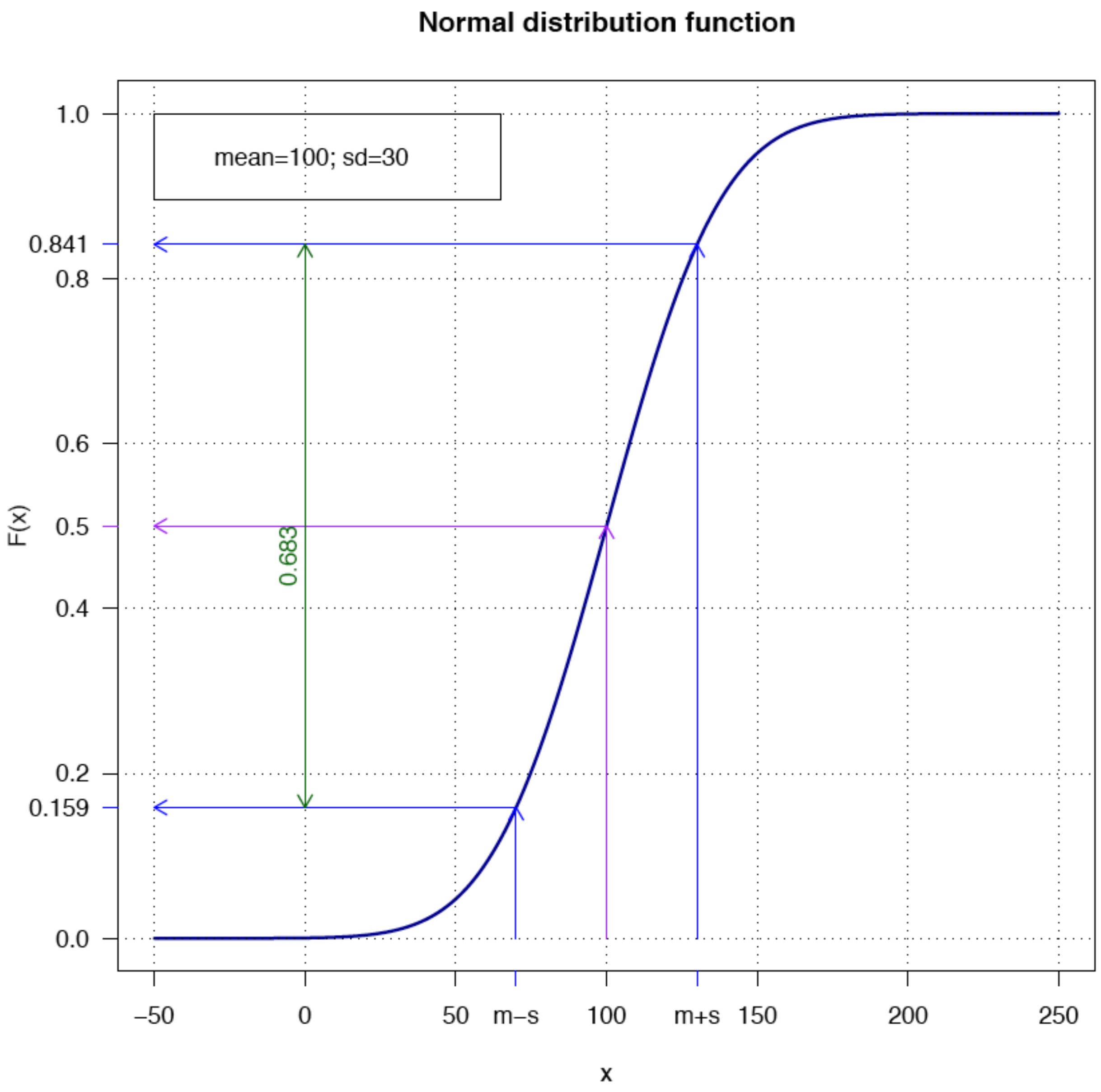
$$F_N(x) = \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

The density function



- For continuous probability distributions, the density represents the limit of the probability per interval, when the range of this interval tends towards 0.
- The normal density function is continuous.
- It is defined from -∞ to +∞
- In R, the normal density function is
  - dnorm(x,m,s)

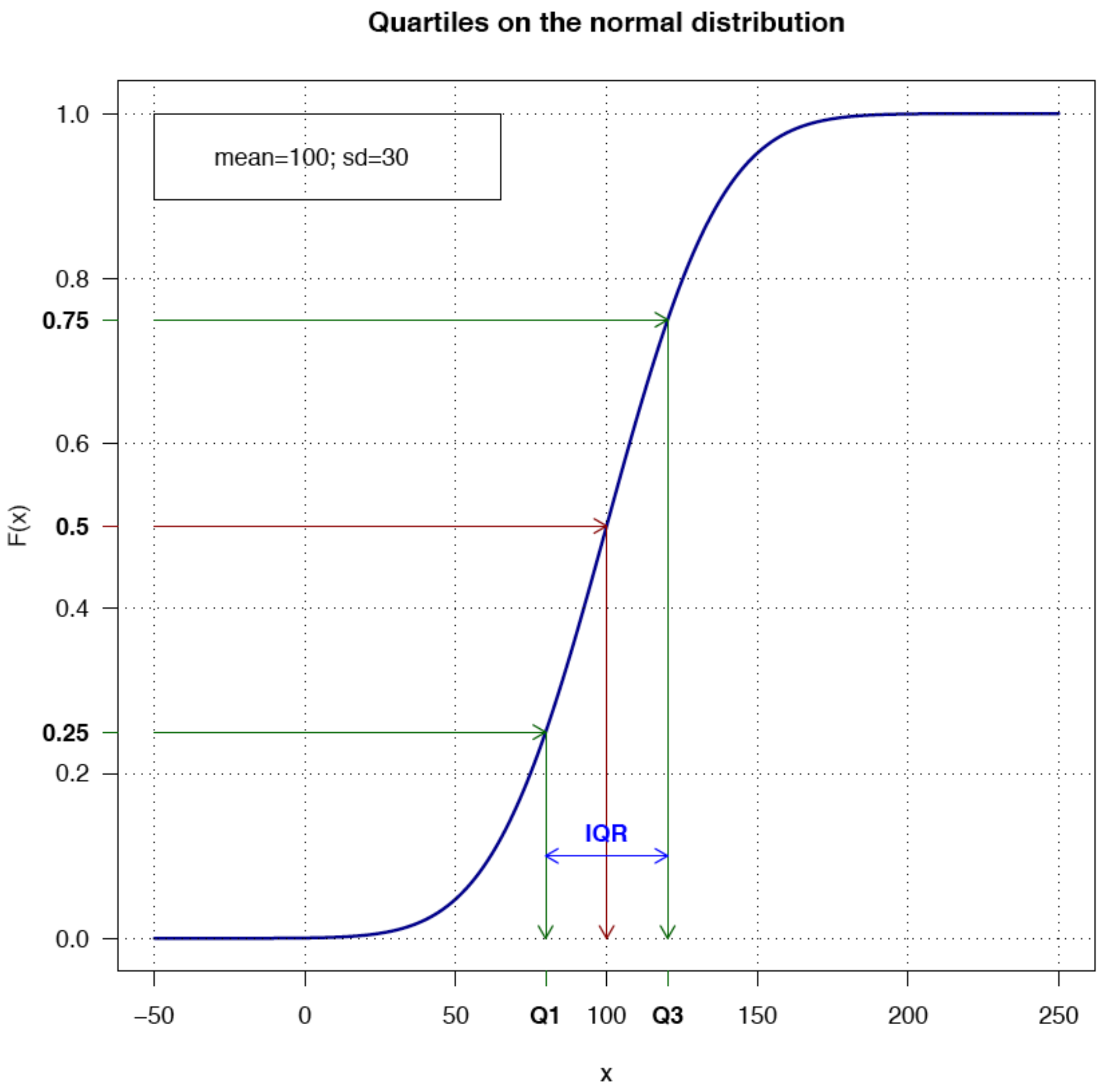
The distribution function



- The distribution function $F(x)$ allows to easily calculate the probability of an interval.
- $F(x)$ gives the probability to observe a value smaller than x.
- The probability to observe a value $x_1 \leq x \leq x_2$, is the difference $F(x_2)$-$F(x_1)$
- In R, the normal distribution function is
  - pnorm(x,m,s)

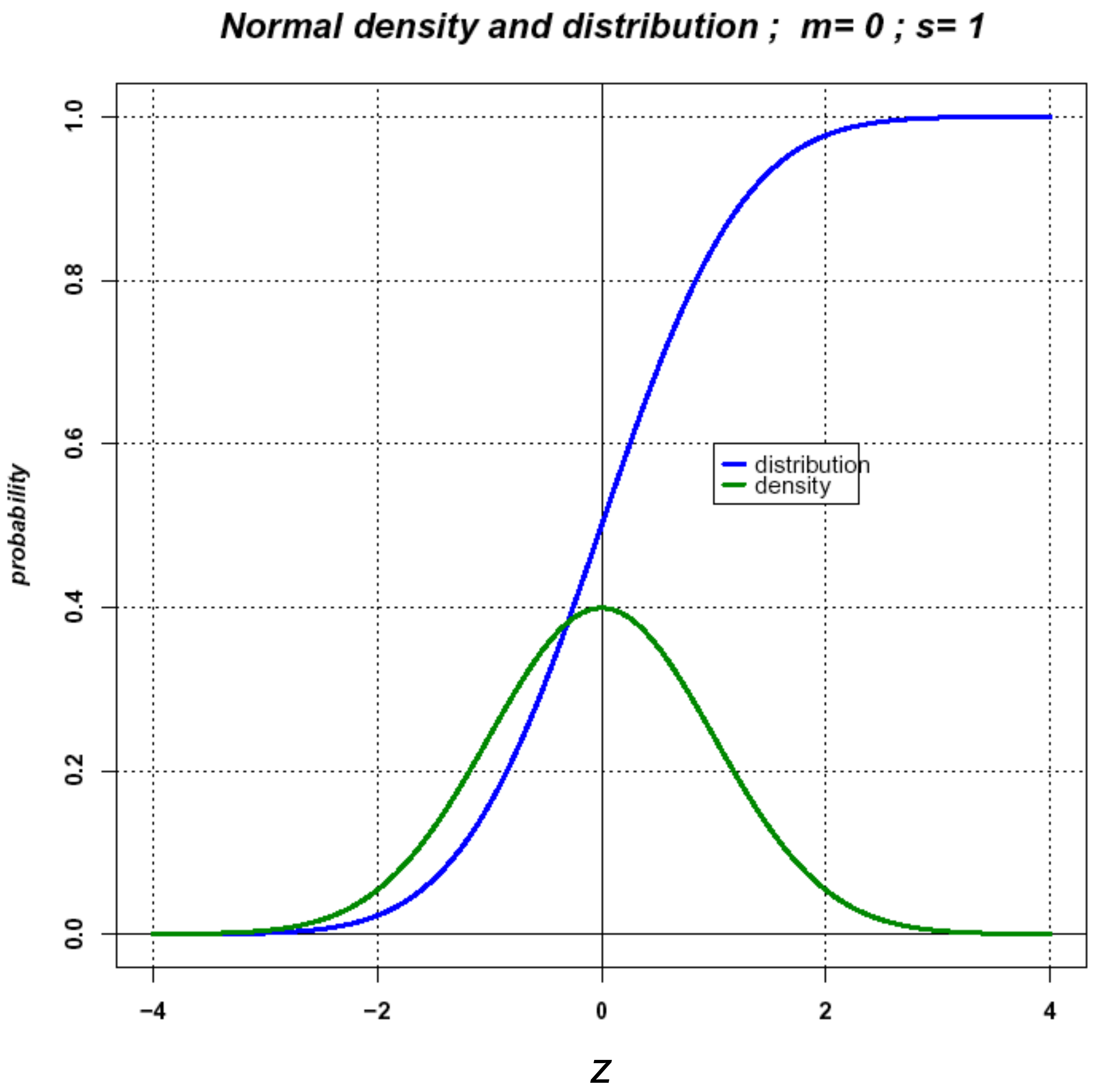
$$
\begin{aligned}
P(x_1 \leq X \leq x_2) &= \int_{x_1}^{x_2} f(x)dx \\
&= \int_{-\infty}^{x_2} f(x)dx - \int_{-\infty}^{x_1} f(x)dx \\
&= F(x_2) - F(x_1)
\end{aligned}
$$

Quartiles on a distribution function



- The first quartile *Q1* is the *x* which leaves 25% of the observations on its left. It is thus the *x* value such that
  - *F(Q1)=0.25*.
- The third quartile *Q3* is the *x* which leaves 75% of the observations on its left. It is thus the *x* value such that
  - *F(Q3)=0.75*.
- The inter-quartile range *IQR* is the difference between the third and the first quartiles.
  - *IQR=Q3-Q1*

Standard normal distribution



- The standard normal is obtained by the transformation

$$z = \left(\frac{x - \mu}{\sigma} \right)$$

- This distribution has
  - mean $\mu = 0$
  - variance $\sigma^2 = 1$

$$f_N(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

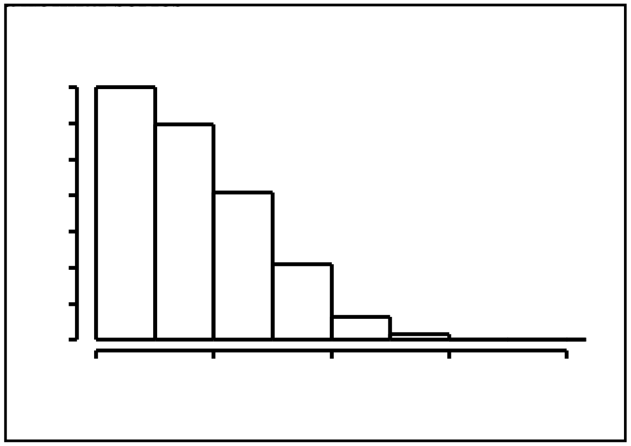
$$F_N(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}z^2} dx$$

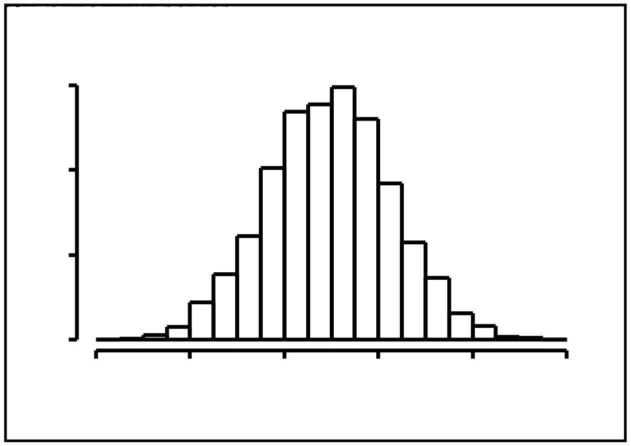
Standard normal distribution - some landmarks

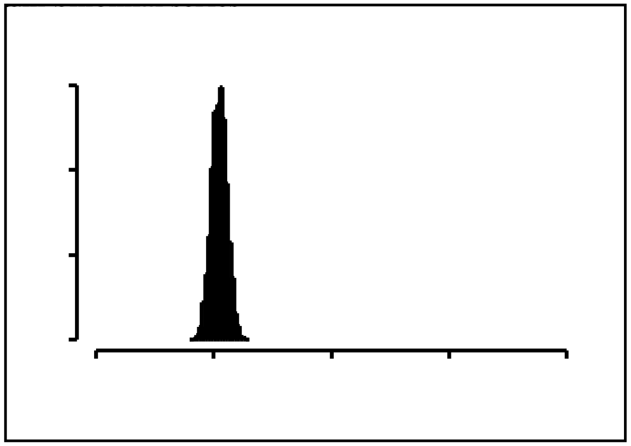
- Parameters of the reduced normal distribution
  - $m = 0$ the **standard** normal distribution is centered around 0
  - $\sigma^2 = 1$ the **standard** normal distribution has a unit variance
  - $\beta_3 = 0$ the normal distribution is symmetric
  - $\gamma_2 = 0$ the normal distribution is mesokurtic
- Some landmarks
  - $P(-\sigma < u < \sigma) = 68.3\%$
  - $P(-2\sigma < u < 2\sigma) = 95.4\%$
  - $P(-3\sigma < u < 3\sigma) = 99.7\%$

Central limit theorem

- Laplace-Liapounoff theorem
  - Any sum of n independent random variables $X_1, X_2, \ldots X_n$ is asymptotically normal
- This naturally extends to the mean of n independent variables, since the mean is the sum divided by a constant.
- Mean of a series of binomial variables
  - Let us take a set of 100 random binomial variables, each with a small mean (e.g. $n*p = 2.1$).
  - Each individual variable is far from normal : it is strongly asymmetric and has an inferior boundary at 0 (there can be no negative values).
  - The sum of these variables however fits a normal distribution.







The chi-squared (χ^2)distribution

- If we have N standard normal random variables
  - $X_1,\dots X_N$
- The variable

$$S_n = \sum_{i=1}^{n} X_i^2$$

has a chi^2_n distribution with n degrees of freedom

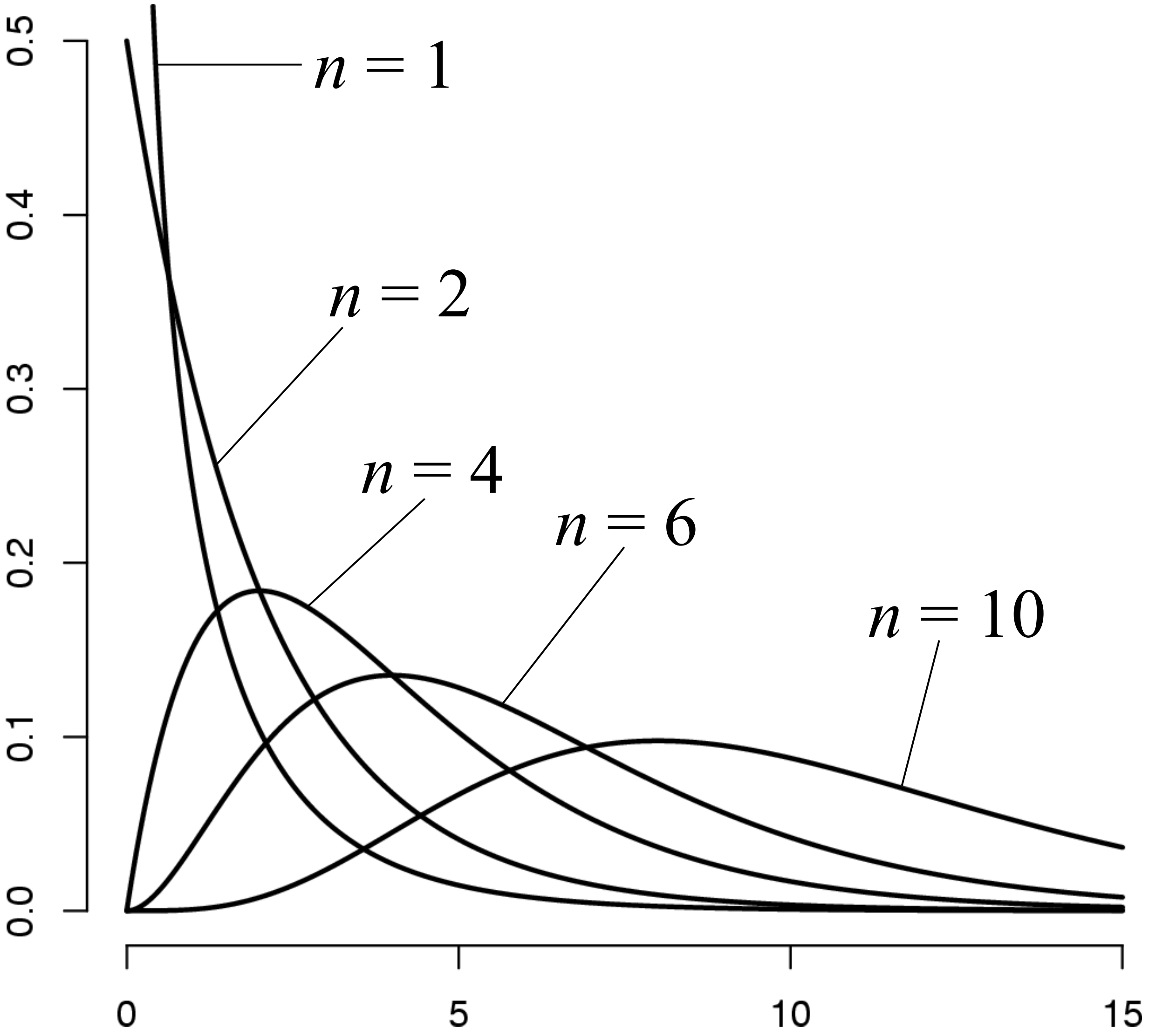
- Density
- Expectation
- Variance

$$f_n(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}$$

$$E[S_n] = n$$

$$V[S_n] = 2n$$

Shapes of χ^2 distributions



slide from Lorenz Wernisch

Student (t) *distribution*

- Z ~ N(0,1) independent of U ~ χ_n^2
- then

$$S = Z \Big/ \sqrt{U \big/ n}$$

has a *t* distribution with n degrees of freedom

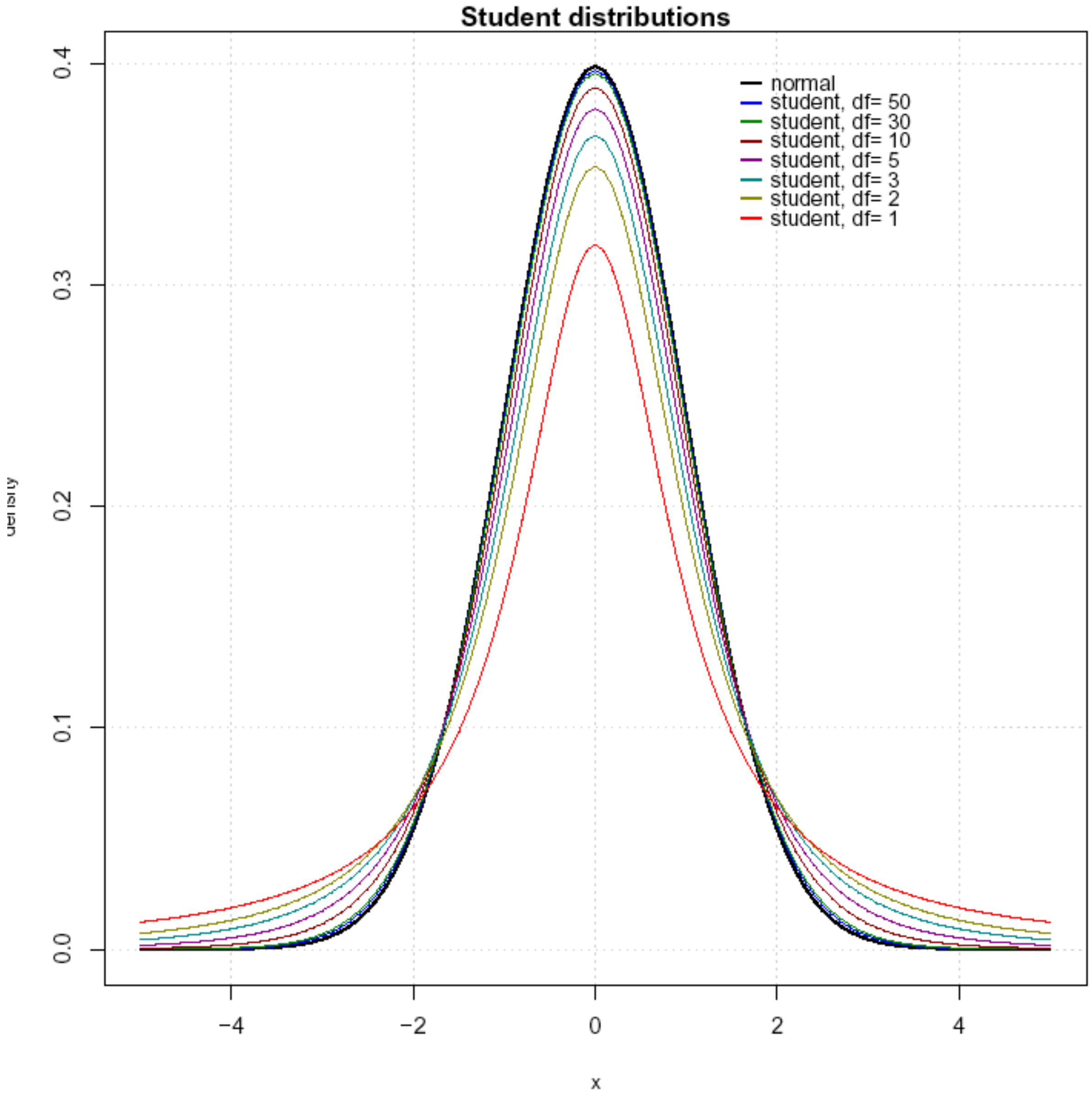
- density

$$f_n(x) = \frac{\Gamma\left(n + 1\big/2\right)}{\sqrt{n\pi}\ \Gamma(n/2)}\left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}$$

```
pt(x,n)
dt(x,n)
rt(num,x,n)
```

slide from Lorenz Wernisch

Shape of Student t distributions



- There is a family of Student distributions, defined by a degree of freedom (n).
- Platykurtic. The degree of kurtosis (flatness) decreases with the degrees of freedom.
- Approaches the normal N(0,1) distribution for large n (n > 30)

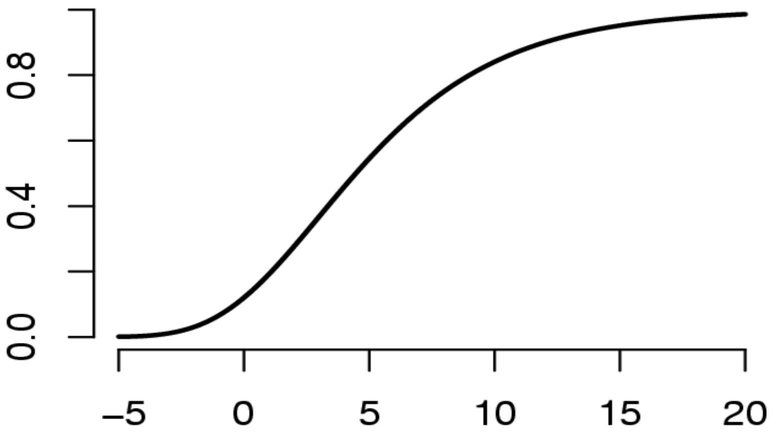
Extreme value distribution

- Cumulative distribution CDF

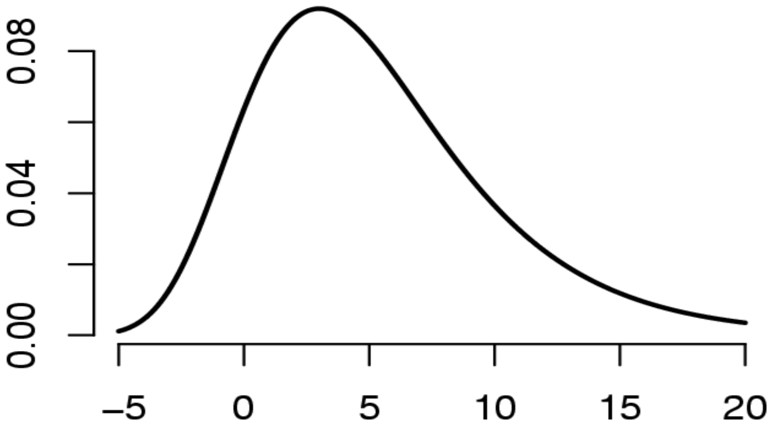
$$\Pr[X < x] = \exp\left(-e^{-(x-\mu)/\sigma}\right)$$

- Probability density PDF

- $$f_{EV}(x;\mu,\sigma) = \frac{1}{\sigma} e^{-(x-\mu)/\sigma} \exp\left(-e^{-(x-\mu)/\sigma}\right)$$

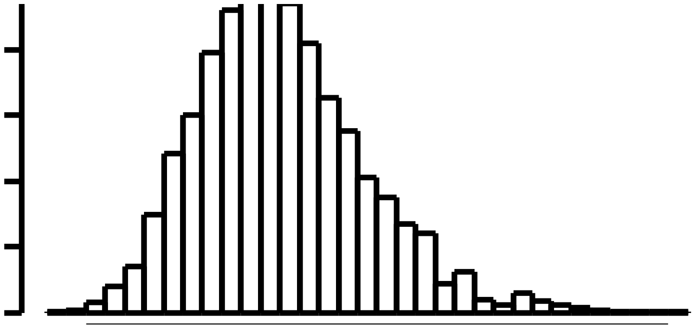


Extreme Value
$\mu = 3, \sigma = 4$



slide from Lorenz Wernisch

Extreme value distributions - random example



- Generate 100 random numbers
  - with standard normal random generator ($m=0, \sigma=1$)
- Take the maximum
- Repeat 1000 times
- The distribution of maxima is
  - Asymmetrical (right-skewed)
  - Bell-shaped
  - Centered around 2.5
  - Less dispersed than the normal populations from which it originated.
- Note that this is different from the central limit theorem :
  - Extreme value distributions are obtained by taking the min or the max of several variables.
  - The central limit theorem applies to the sum or mean of several variables.

Extreme value distribution - applications

- The extreme value distribution has a particular importance in bioinformatics, for its role in BLAST
  - Aligning two sequences consists in searching the alignment with maximum score
  - Aligning a sequence against a whole database amounts to get, for each database entry, the maximum alignment score
  - BLAST scores have thus an extreme value distribution
  - (more details in the course on sequence analysis)

Other distributions not (yet) covered here

- Compound Poisson
- Snedecor (F)
- Beta function
- Gamma function

Exercises - theoretical distributions

Exercises - theoretical distributions

- In which cases is it appropriate to apply a hypergeometric or a binomial distribution, respectively ?
- Does the hypergeometric distribution correspond to a Bernoulli schema ?
- What are the relationships between binomial, Poisson and normal distributions ?

Exercise - Word occurrences in a sequence

- A sequence of length 10,000 has the following residue frequencies
  - F(A) = F(T) = 0.325
  - F(C) = F(G) = 0.175
- What is the probability to observe the word GATAAG at a given position of a sequence (assuming a Bernoulli model).
- What would be the probability to observe, in the whole sequence
  - 0 occurrences
  - at least one occurrence
  - exactly one occurrence
  - exactly 15 occurrences
  - at least 15 occurrences
  - less than 15 occurrences

Exercise - substitutions of a word

- A sequence is generated with equiprobable nucleotides. What is the probability to observe the word GATAAG or a single-base substitution of it, at the first position ?
- Same question with at most 3 substitutions.