Probability and Statistics

Kristel Van Steen, PhD[2]

Montefiore Institute - Systems and Modeling

GIGA - Bioinformatics

ULg

kristel.vansteen@ulg.ac.be

CHAPTER 3: SOME IMPORTANT DISTRIBUTIONS

1 Discrete case

1.1 Bernoulli trials

Binomial distribution – sums of binomial random variables

Hypergeometric distribution

Geometric distribution

Memoryless distributions

Negative binomial distribution

1.2 Multinomial distribution

1.3 Poisson distribution

Sums of Poisson random variables

1.4 Summary

2 Continuous case

2.1 Uniform distribution

2.2 Normal distribution

Probability tabulations

Multivariate normality

Sums of normal random variables

2.3 Lognormal distribution

Probability tabulations

2.4 Gamma and related distributions

Exponential distribution

Chi-squared distribution

2.5 Where discrete and continuous distributions meet

2.6 Summary

1 Discrete case

- This part deals with some distributions of random variables that are important as models of scientific discrete phenomena.
- An understanding for the situations in which these random variables arise enables us to choose an appropriate distribution for a scientific phenomenon under consideration.
- Hence, in alignment with what we discussed in Chapter 1, we will dwell upon "induction": choosing a model on the basis of factual understanding of the physical phenomenon under investigation
    - induction is reasoning from detailed facts to general principles and
    - deduction is reasoning from the general to the particular

2.7 Bernoulli trials and binomial distributions

- Suppose X represents a random variable representing the number of successes S in a sequence of n Bernoulli trials, regardless of the order in which they occur.
- Then X is a discrete random variable
- What is the probability mass function of X? $P_X(k) = ?$
- Answer: Compute the total number of possible arrangements of outcomes of the n Bernoulli trials that satisfy the property. In particular, count the number of ways that k letters S can be placed in n boxes:
    - n choices for first S
    - n-1 choices for second S
    - …
    - N-(k-1) choices for kth S

Divide by the number of ways k S letters can be arranged in k boxes: $k!$

The number of ways k successes can happen in n trials is therefore:

$$\frac{n(n-1)\cdots(n-k+1)}{k!} = \frac{n!}{k!(n-k)!},$$

and the probability associated with each is $p^k q^{n-k}$:

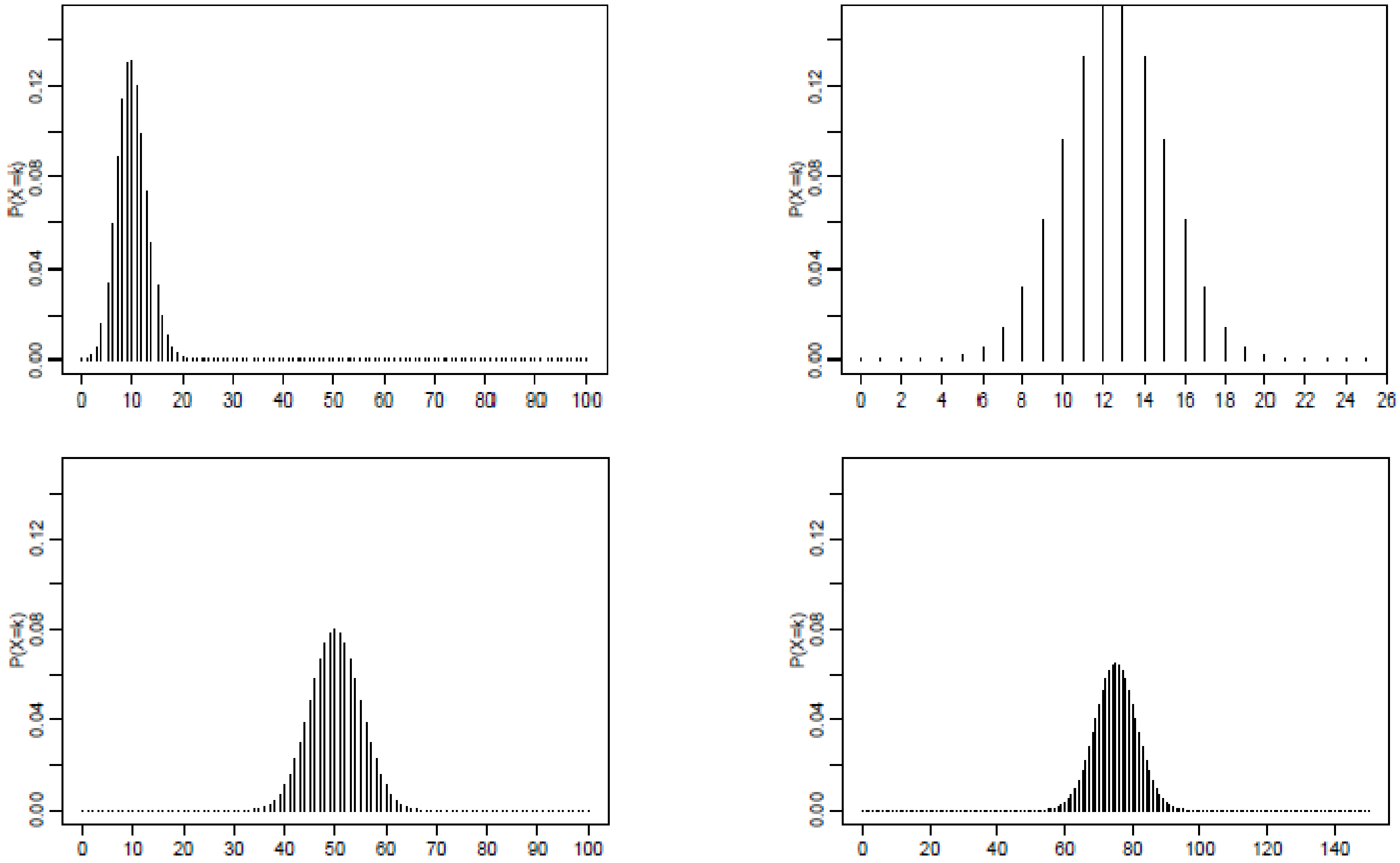
$$p_X(k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, 2, \ldots, n,$$

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

the binomial coefficient in the binomial theorem

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

- Binomial probabilities P(X = x) as a function of x for various choices of n and π. On the left, n=100 and π=0.1,0.5. On the right, π=0.5 and n=25,150

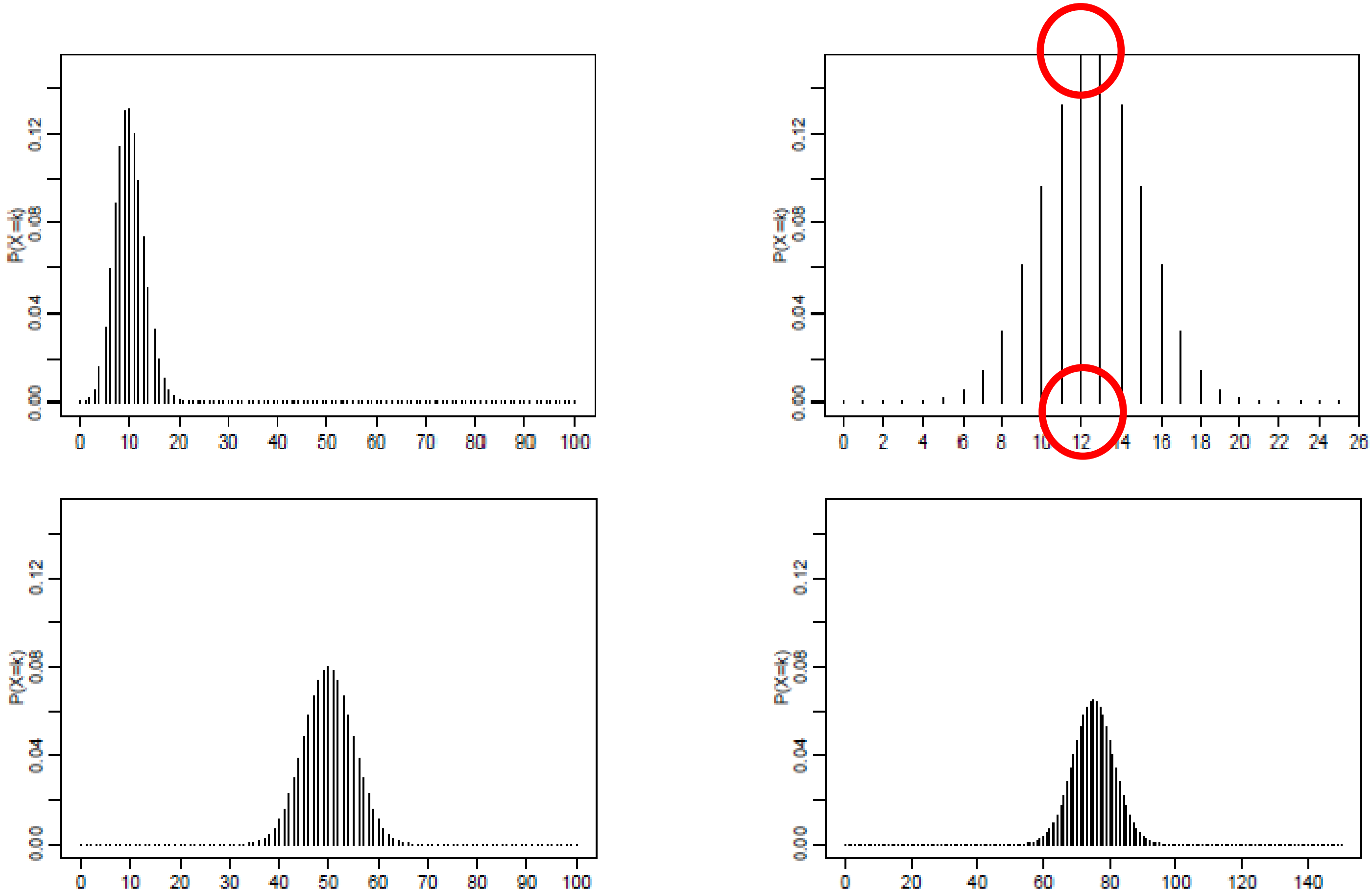


- More insight into the behavior of $P_X(k)$ can be gained by taking the ratio:

$$\frac{P_X(k)}{P_X(k-1)} = \frac{(n-k+1)p}{kq} = 1 + \frac{(n+1)p-k}{kq}$$

- Hence,
  - o $P_X(k)$ is greater than $P_X(k-1)$ when $k < (n+1)p$ and is smaller when $k > (n+1)p.$
  - o If we define an integer $k^\star$ as $(n+1)p - 1 < k^\star \leq (n+1)p,$ the value of $P_X(k)$ increases monotonically and attains its max value at $k = k^\star$, then decreases monotonically
  - o If $(n+1)p$ happens to be an integer, the max value takes place at both $P_X(k^\star - 1)$ and $P_X(k^\star)$
  - o The integer $k^\star$ is a *mode* of this distribution and often referred to as the "most probable number of successes"

- Binomial probabilities P(X = x) as a function of x for various choices of n and π. On the left, n=100 and π=0.1,0.5. On the right, π=0.5 and n=25,150



Example

- What is the probability distribution of the number of times a given pattern occurs in a random DNA sequence $L_1, \ldots, L_n$?
    - o New sequence $X_1, \ldots, X_n$:

      $$X_i=1 \text{ if } L_i=A \text{ and } X_i=0 \text{ else}$$
    - o The number of times N that A appears is the sum

      $$N=X_1+\ldots+X_n$$
    - o The prob distr of each of the X_i:

      $$P(X_i=1) = P(L_i=A)=p_A$$

      $$P(X_i=0) = P(L_i=C \text{ or } G \text{ or } T) = 1 - p_A$$
- What is a “typical” value of N?

Example

- What is the probability distribution of the number of times a given pattern occurs in a random DNA sequence $L_1, \ldots, L_n$?
  - o New sequence $X_1, \ldots, X_n$:

    $$X_i=1 \text{ if } L_i=A \text{ and } X_i=0 \text{ else}$$

  - o The number of times N that A appears is the sum

    $$N=X_1+\ldots+X_n$$

  - o The prob distr of each of the X_i:

    $$P(X_i=1) = P(L_i=A)=p_A$$

    $$P(X_i=0) = P(L_i=C \text{ or } G \text{ or } T) = 1 - p_A$$

- What is a “typical” value of N?
  - o **Depends on how the individual X_i (for different *i*) are interrelated**

Exact computation via closed form of relevant distribution

- The formula for the binomial probability mass function is :

$$P(N = j) = \binom{n}{j} p^j (1-p)^{n-j}, \text{ j = 0,1, …,n}$$

and therefore

$$P(N \geq 300) = \sum_{j=300}^{1000} \binom{1000}{j} (1/4)^j (1-1/4)^{1000-j}$$

$$= 0.000193590321949658 41$$

Approximate via Stirling's formula

- Factorials start off reasonably small, but by 10!, we are already in the millions, and it doesn't take long until factorials "explode". Unfortunately there is no shortcut formula for n!, you have to do all of the multiplication.
- On the other hand, there is a famous approximate formula, named after the Scottish mathematician James Stirling (1692-1770), that gives a pretty accurate idea about the size of n!:

$$\text{Stirling's formula } n! \approx \sqrt{2\pi n}\,(\frac{n}{e})^n$$

- n factorial involves nothing more sophisticated than ordinary multiplication of whole numbers, which Stirling's formula relates to an expression involving square roots, π (the area of a unit circle), and e (the base of the natural logarithm).
- What are the consequences of using this approximation?

$$\begin{array}{lllll} 1! = 1 & 2! = 2 & 3! = 6 & 4! = 24 & 5! = 120 \\ 6! = 720 & 7! = 5040 & 8! = 40320 & 9! = 362880 & 10! = 3628800 \end{array}$$

$$\begin{array}{lllll} 1! \approx 0.92 & 2! \approx 1.92 & 3! \approx 5.84 & 4! \approx 23.51 & 5! \approx 118.02 \\ 6! \approx 710.08 & 7! \approx 4980.39 & 8! \approx 39902.39 & 9! \approx 359536.87 & 10! \approx 3598695.62 \end{array}$$

- In fact the approximation $1! \approx 0.92$ is accurate to 0.08, while $10! \approx 3598695.62$ is only accurate to about 30,000. *[compute the difference between the exact and approximated values]*
- You can see that the larger n gets, the better the approximation proportionally. The proportional error for 1! Is (1!-0.92)/1! – 0.0800 while for 10! It is (10! -3598695.62)/10! = 0.0083, ten times smaller.
- This is the correct way to understand Stirling's formula:
  as n gets large, the proportional error

$$[n! - \sqrt{2\pi n}\,(n/e)^n]/n!$$

goes to zero.

Approximate via Central Limit Theory

- The central limit theorem offers a 3rd way to compute probabilities for a binomial distribution
- It applies to sums or averages of iid random variables
- Assuming that $X_1, \ldots, X_n$ are iid random variables with mean μ and variance σ^2, then we know that for the sample average

$$\bar{X}_n = \frac{1}{n}(X_1 + \ldots + X_n),$$

$$\mathrm{E}\bar{X}_n = \mu \text{ and } \mathrm{Var}\,\bar{X}_n = \frac{\sigma^2}{n}$$

- Hence,

$$E\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right) = 0, Var\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right) = 1$$

Approximate via Central Limit Theory

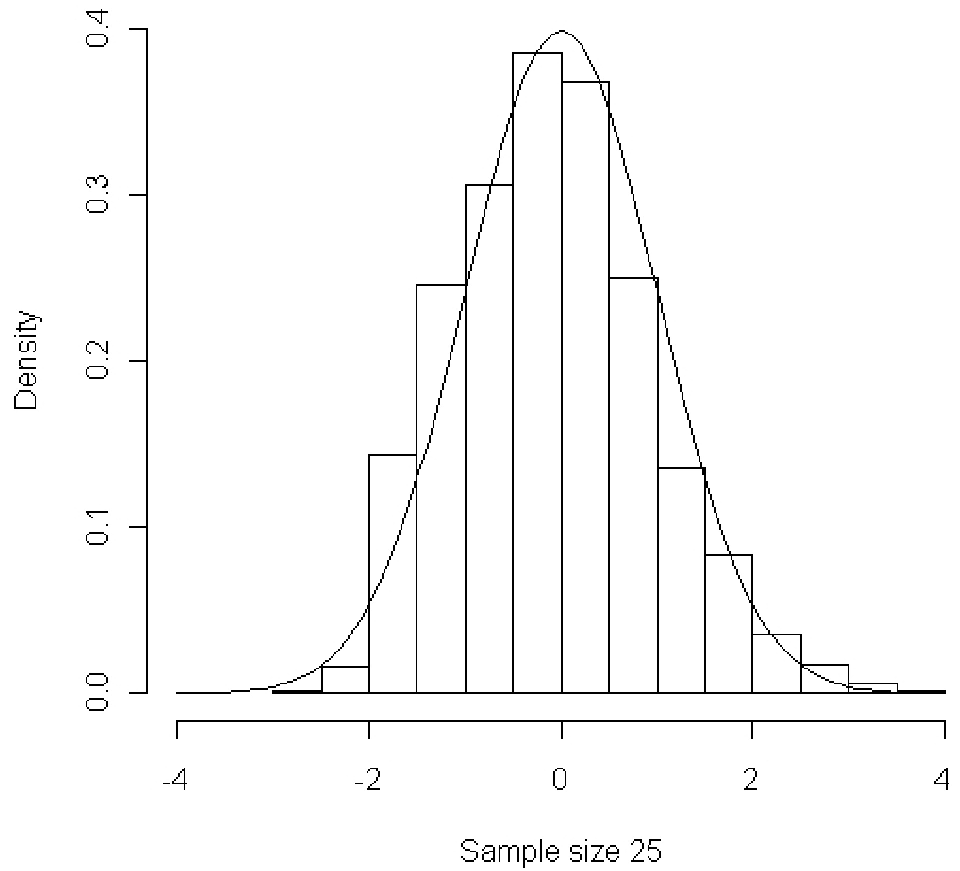
- The central limit theorem states that if the sample size n is large enough,

$$P\left(a \leq \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq b\right) \approx \phi(b) - \phi(a),$$

with $\phi(.)$ the standard normal distribution defined as

$$\phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \phi(x)dx$$

Approximate via Central Limit Theory



Approximate via Central Limit Theory

- Estimating the quantity $P(N \geq 300)$ when N has a binomial distribution with parameters n=1000 and p=0.25,

$$E(N) = n\mu = 1000 \times 0.25 = 250,$$

$$sd(N) = \sqrt{n}\,\sigma = \sqrt{1000 \times \frac{1}{4} \times \frac{3}{4}} \approx 13.693$$

$$P(N \geq 300) = P\left(\frac{N - 250}{13.693} > \frac{300 - 250}{13.693}\right)$$

$$\approx P(Z > 3.651501) = 0.0001303560$$

- Now consider all estimates of $P(N \geq 300)$ and you will see that all of these compare really well …

Approximate via Poisson distribution

- When n gets large, the computation of mass probabilities may become cumbersome:
    - Use Stirling's formula (see before)
    - Use the central limit theorem (see before)
    - Use Poisson's approximation to the binomial distribution (see later)

Sum of binomial distributed random variables

Problem: let X_1 and X_2 be two independent random variables, both having binomial distributions with parameters (n_1, p) and (n_2, p), respectively, and let $Y = X_1 + X_2$. Determine the distribution of random variable Y.

Answer: the characteristic functions of X_1 and X_2 are,

$$\phi_{X_1}(t) = (pe^{jt} + q)^{n_1}, \phi_{X_2}(t) = (pe^{jt} + q)^{n_2}.$$

the characteristic function of Y is simply the product of $\phi_{X_1}(t)$ and $\phi_{X_2}(t)$. Thus,

$$\begin{aligned}\phi_Y(t) &= \phi_{X_1}(t)\phi_{X_2}(t)\\ &= (pe^{jt} + q)^{n_1+n_2}.\end{aligned}$$

By inspection, it is the characteristic function corresponding to a binomial distribution with parameters $(n_1 + n_2, p)$. Hence, we have

$$p_Y(k) = \binom{n_1 + n_2}{k} p^k q^{n_1+n_2-k}, \quad k = 0, 1, \ldots, n_1 + n_2.$$

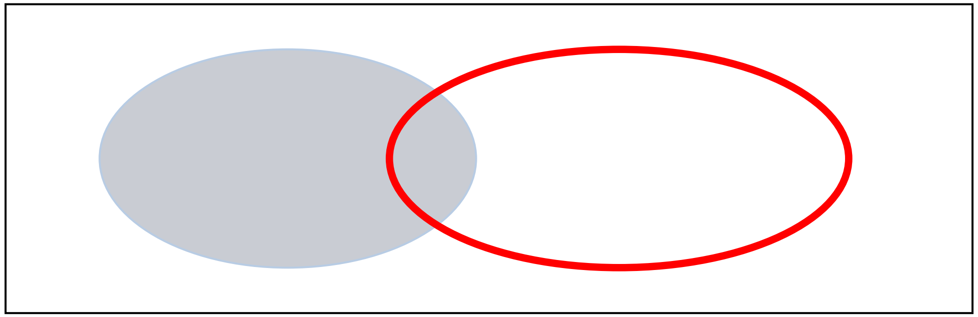
Recall:

- The characteristic function approach is particularly useful in analysis of linear combinations of independent random variables
- The characteristic function provides an alternative way for describing a random variable; it completely determines behavior and properties of the probability distribution of the random variable X
- If a random variable admits a density function, then the characteristic function is its dual, in the sense that each of them is a Fourier transform of the other.
- If a random variable has a moment-generating function, then the domain of the characteristic function can be extended to the complex plane, and $\phi_X(-it) = M_X(t)$
- The characteristic function of a distribution **always** exists, even when the probability density function or moment-generating function do not.

The conditional probability mass function of a binomial random variable X, conditional on a given sum m for X+Y (Y an independent from X binomial random variable)

$X \sim Bin(n_1, p)$ and $Y \sim Bin(n_2, p)$,

$$X + Y = m, 0 \leq m \leq n_1 + n_2$$



a) Box = total possibilities (n_1+n_2)
b) Blue = those having the property (n_1)
c) Red = selection (k out of m selected have the property)

<u>Solution</u>:

For $k \leq min(n_1, m)$,

$$
\begin{aligned}
P(X=k|X+Y=m) &= \frac{P(X=k\cap X+Y=m)}{P(X+Y=m)} \\
&= \frac{P(X=k\cap Y=m-k)}{P(X+Y=m)} = \frac{P(X=k)P(Y=m-k)}{P(X+Y=m)} \\
&= \frac{\binom{n_1}{k}p^k(1-p)^{n_1-k}\binom{n_2}{m-k}p^{m-k}(1-p)^{n_2-m+k}}{\binom{n_1+n_2}{m}p^m(1-p)^{n_1+n_2-m}} \\
&= \binom{n_1}{k}\binom{n_2}{m-k}\Big/\binom{n_1+n_2}{m}, \quad k=0,1,\ldots,\min(n_1,m),
\end{aligned}
$$

having used the result that X+Y is binomially distributed with parameters (n_1+n_2, p)

- This distribution is known as the **hypergeometric distribution**.

Example: over-representation of terms

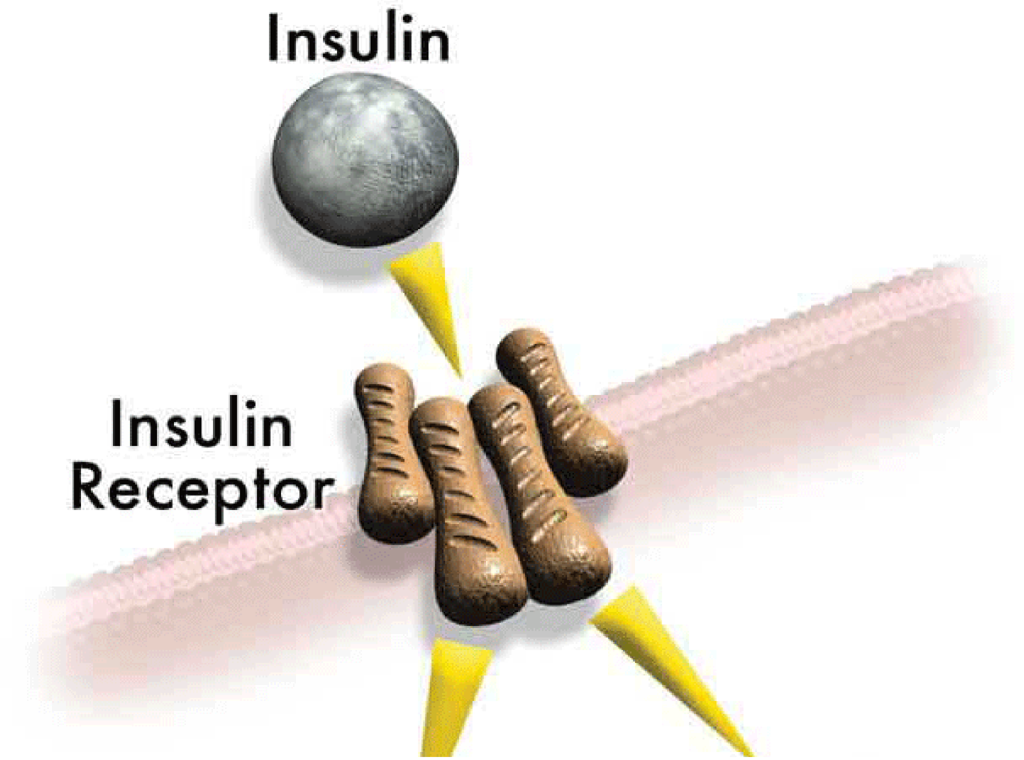
- Gene Ontology (GO) is a collection of controlled vocabularies describing the biology of a gene product in any organism
- There are 3 independent sets of vocabularies, or so-called “ontologies”:

    - **Molecular Function** (MF)
    - **Cellular Component** (CC)
    - **Biological Process** (BP)

- Question: In a given list of genes of interest (eg. Differentially Expressed), is there a Gene Ontology term that is more represented than what it would be expected by chance only?

Molecular function

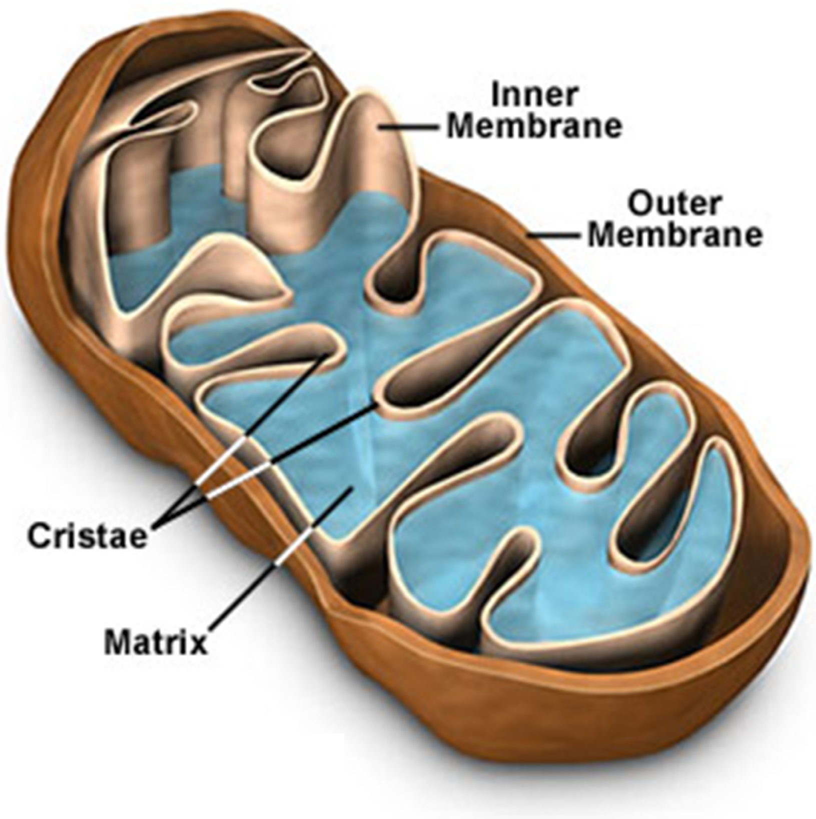
- *… activities or jobs of a gene product*



(e.g., insulin binding or receptor activity)

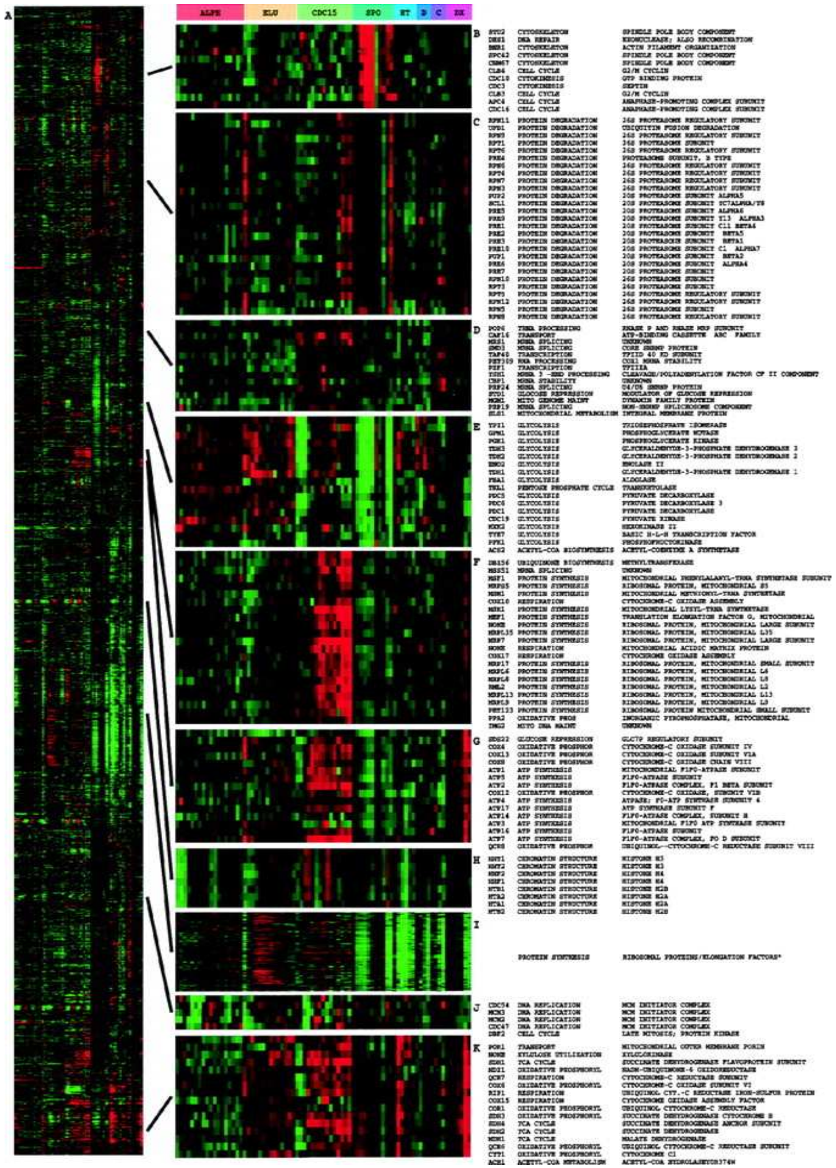
Cellular component

- … where a gene product acts

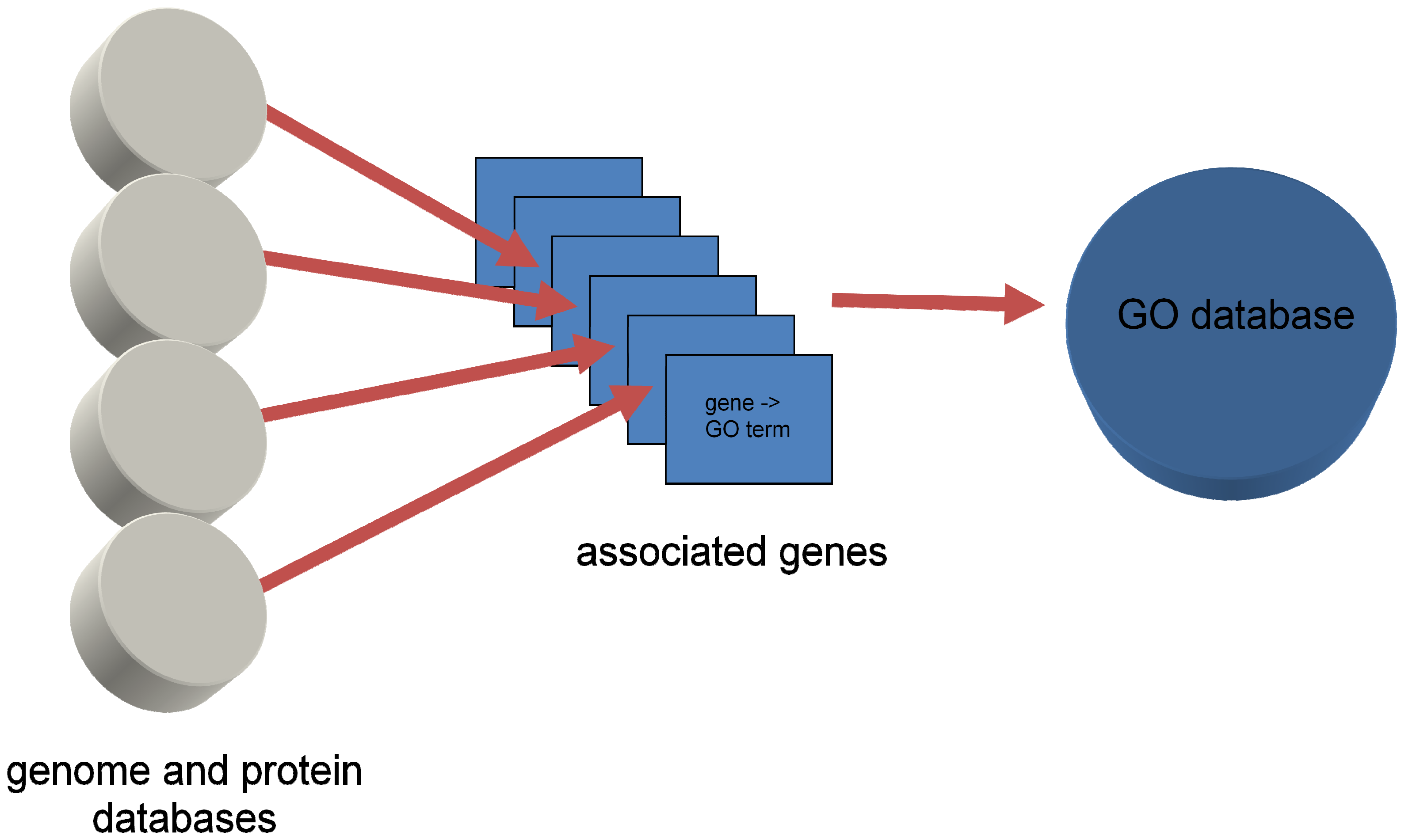


Biological processes

- A set of gene product functions make up a biological process, such as in courtship behavior



Gene ontology analysis makes life easier for the researcher: it allows making inferences across large numbers of genes without researching each one individually

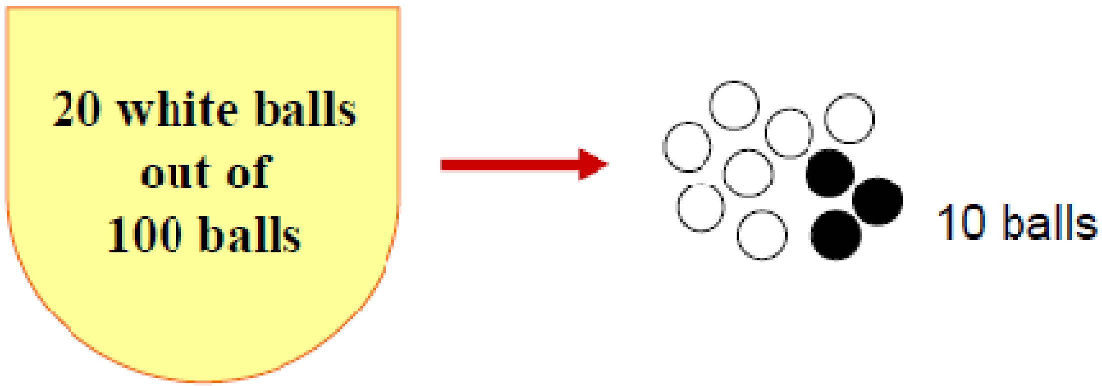

gene ->
GO term
GO database
associated genes
genome and protein
databases

- Solution:

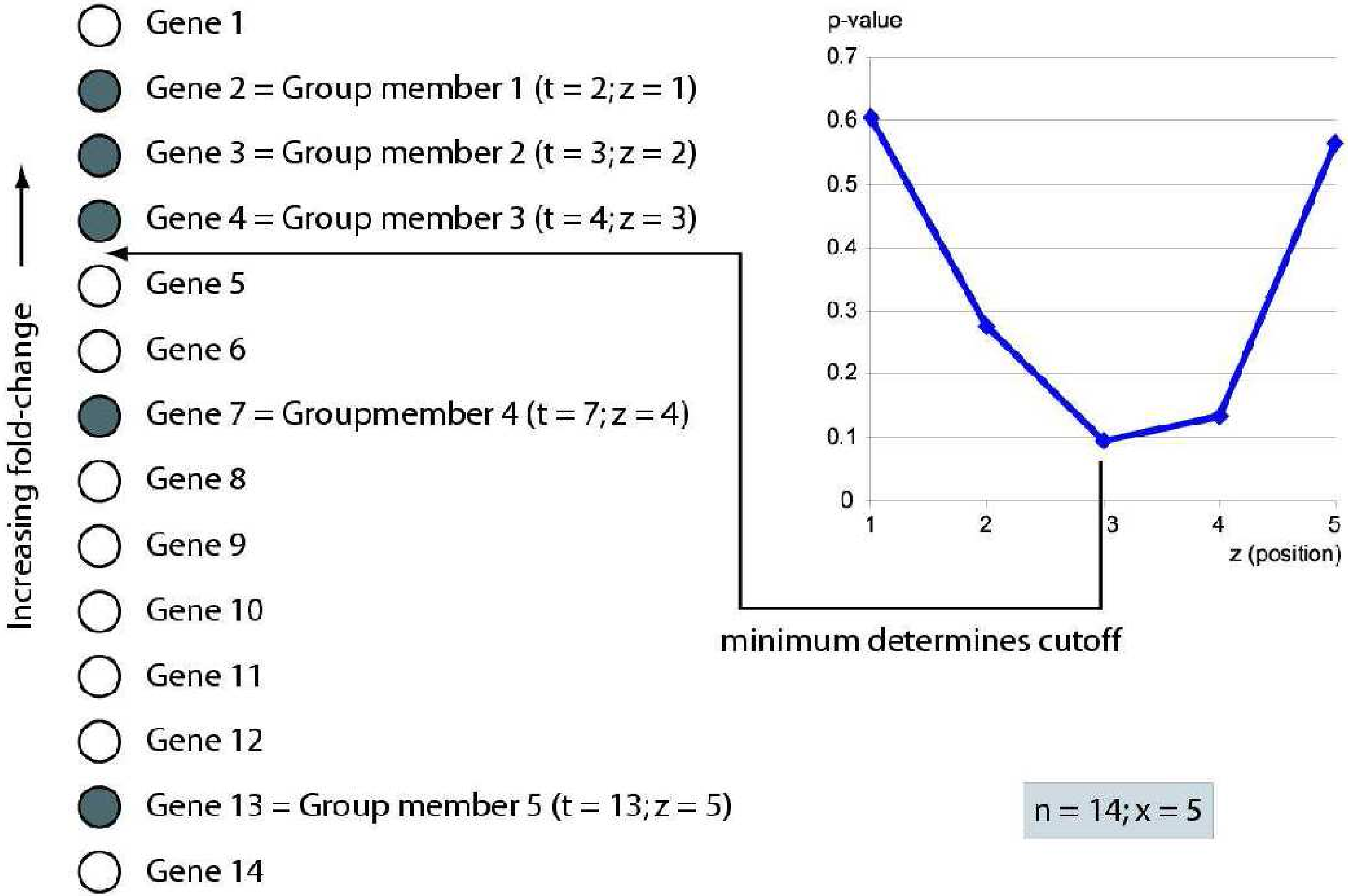
  - Most GO tools work in a similar way:
    - input a gene list and a subset of ‘interesting’ genes
    - tool shows which GO categories have most interesting genes associated with them i.e. which categories are ‘enriched’ for interesting genes
    - tool provides a *statistical measure* to determine whether enrichment is significant … and here the geometric distribution comes around

- This can be seen in the following way:

  The hypergeometric distribution naturally arises from sampling from a fixed population of balls .



  Here, a typical problem of interest is to to calculate the probability for drawing 7 or more white balls out of 10 balls given the distribution of balls in the urn → hypergeometric test → p-value (see later).
- Now the "property" is not the color of a ball, but whether a gene can be linked to a GO term or group of interest.


Gene 1
Gene 2 = Group member 1 (t = 2; z = 1)
Gene 3 = Group member 2 (t = 3; z = 2)
Gene 4 = Group member 3 (t = 4; z = 3)
Gene 5
Gene 6
Gene 7 = Groupmember 4 (t = 7; z = 4)
Gene 8
Gene 9
Gene 10
Gene 11
Gene 12
Gene 13 = Group member 5 (t = 13; z = 5)
Gene 14
Increasing fold-change
p-value
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
z (position)
minimum determines cutoff
n = 14; x = 5