# References

F. Biagini, M. Campanino. Elementi di Probabilità e Statistica. Springer.

B. de Finetti. Theory of Probability: A Critical Introductory Treatment. Wiley.

W. J. Ewens, G. R. Grant. Statistical, Methods in Bioinformatics. Springer.

## Probability Theory

The goal of Probability Theory is to quantify our degree of uncertainness.

Probability Theory and Statistics, that is based on Probability Theory and uses its language, have applications to almost all Sciences and humain activities.

The basic object of study of Probability Theory are *random objects* and in particular *random numbers*.

A random number is a well defined number, whose value however is not necessarily known. For example the result of a specific experiment or observation.

We will denote random numbers with capital letters.

Given a random number $X$ we denote by $I(X)$ the set of its *possible values*.

A randon number $X$ is said to be *upper bounded* if

$$\sup I(X) < \infty$$

, *lower bounded* if

$$\inf I(X) > -\infty$$

and bounded if both inequalities are true.

Similarly for two random numbers $X, Y$ we will denote by $I(X, Y)$ the *set of the possible pairs* of $(X, Y)$ and in general if $X_1, X_2, \ldots, X_n$ are $n$ random numbers, then $I(X_1, X_2, \ldots, X_n)$ is the *set of possible $n$-tuples* $(X_1, X_2, \ldots, X_n)$.

A particular case of random numbers are *events*.

Events are random numbers whose set of values is contained in the set $\{0, 1\}$. We will often use for events the letters $E, F, G, H, A, B, C, \ldots$. If an event $E$ takes the value 1, then we say that $E$ takes place; if it takes the value 0 then we say that it does not take place.

In order to define events we will use the following notation. A proposition that can be true or false inserted between brackets will denote a number equal to 1 if the proposition is true and equal to 0 if the proposition is false. For example if the random number $X$ represents the result obtained by thowing a die with set of possible values $I(X) = \{1, 2, 3, 4, 5, 6\}$. Then we can define the event $E =$( X is even). This notation will be used in general and not just for defining events.

## Operations on random numbers and events.

On random numbers and events we can perform operations and obtiain other random numbers. In particular we can perform arithmetical operations. We will also consider the following operations that are in particular important in the case of events. Given $x, y$ real numbers, we define

$$x \wedge y = \min(x, y)$$

$$x \vee y = \max(x, y)$$

$$\tilde{x} = 1 - x$$

.

It is easy to see that they satisfy the following properties:

$$x \wedge y = y \wedge x$$

$$x \vee y = y \vee x$$

$$x \wedge (y \wedge z) = (x \wedge y) \wedge z$$

$$(x \vee y) \vee z = x \vee (y \vee z)$$

$$x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$$

$$x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z)$$

$$\widetilde{(x \wedge y)} = \tilde{x} \vee \tilde{y}$$

$$\widetilde{(x \vee y)} = \tilde{x} \wedge \tilde{y}$$

$$\tilde{\tilde{x}} = x$$

.

Similar relations hold when the operations involve $n$ numbers instead of just two. It is not necessary in this case to insert parentheses as the operations satisfy the associative properties.

Let $E$ $F$ be two events.

$E \vee F$ is called the *logic sum* (or *union* in the set theoretic interpretation) of the events $E$ and $F$. It is the event. It the event that takes place if and only if at least one of the events $E$ of $F$ takes place. Thefore it corresponds to the preposition "or".

$E \wedge F$ is called the *logic product* (or *intersection* in the set theoretic interpretation) of the events $E$ and $F$. As the events take only the values 0 and 1, $E \wedge F$ is equal to the orsinary product $EF$. It is the event. It the event that takes place if and only if both events $E$

of $F$ take place. Thefore it corresponds to the preposition "and".

$\tilde{E}$ is called the *complemetary* event of the event $E$. It is the event that takes place if and only if the event $E$ does not take place.

The properties that we have stated at the beginning can be applied to the logical operations on events. As we have remarked, the logical product can simply expressed by means of the ordinary arithmetic product. This is not true for the logical sum, as if two events both take place their logical sum is equal to 1, whereas their arithmetical sum is equal to 2 and therefore it is no an event but just a random number. If however two events are uncompatible (i. e. they cannot both take place at the same time), then their logical and arithmetical sum are equal.

Let us show a simple application of these relations.

$$1 - (E \vee F) = \widetilde{(E \vee F)} = \tilde{E} \wedge \tilde{F}$$

$$= \tilde{E}\tilde{F} = (1 - E)(1 - F) = 1 - E - F + EF.$$

It follows that

$$E \vee F = E + F - EF.$$

Similarly it can be shown that

$$E \vee F \vee G = E + F + G - EF - FG - EG + EFG.$$

**Conditional probability and expectation.**

Given a random number $X$ and an event $H$, the conditional expectation of $X$ given the event $H$ is the expectation that we assign to $X$ given that in addition to the present information we also know that the event $H$ has taken place. We denote it by $\mathbb{P}(X|H)$.

It is easy to see that by coherence priciple it must satisfy the following properties that are analogues of those satisfied by ordinary expectation:

$$\inf I(X|H) \leq \mathbb{P}(X|H) \leq \sup I(X|H)$$

where $I(X|H)$ denotes the set of the values that $X$ can assume whent the event $H$ takes place. In general $I(X|H) \subset I(X)$. Therefore in general $\inf I(X|) \leq \inf I(X|H)$ and $\sup I(X|H) \leq \sup I(X)$.

$$\mathbb{P}(X + Y|H) = \mathbb{P}(X|H) + \mathbb{P}(Y|H)$$

$$\mathbb{P}(aX) = a\mathbb{P}(X).$$

Principle of composite expectation.

It can be shown that coherence priciple also implies that conditional expectation satisfies the following relation, that is called *principle of composite expectation*:

$$\mathbb{P}(XH) = \mathbb{P}(X|H)\mathbb{P}(H).$$

When $\mathbb{P}(H) > 0$ we have

$$\mathbb{P}(X|H) = \frac{\mathbb{P}(XH)}{\mathbb{P}(H)}$$

that sometimes is taken as a definition of conditional expectation.

If $E$ and $H$ are events then $\mathbb{P}(E|H)$ is called the conditional probability of $E$ given $H$.

If $\mathbb{P}(H) > 0$ then

$$\mathbb{P}(E|H) = \frac{\mathbb{P}(EH)}{\mathbb{P}(H)}.$$

That is the conditional probability of $E$ given $H$ is the probability that both $E$ and $H$ take place divided by the probability of $H$.

It follows from the properties of conditional expectation to see that if $H \subset E$, then $\mathbb{P}(E|H) = 1$.

On the other side if $E \subset H$, then $EH = E$ and therefore

$$\mathbb{P}(E|H) = \frac{\mathbb{P}(E)}{\mathbb{P}(H)}.$$

**Positive correlation, negative correlation and non-correlation.**

The event $E$ is *positively correlated* with the event $H$ if $\mathbb{P}(E|H) > \mathbb{P}(E)$.

The event $E$ is *negatively correlated* with the event $H$ if $\mathbb{P}(E|H) < \mathbb{P}(E)$.

The event $E$ is *non-correlated* with the event $H$ if $\mathbb{P}(E|H) = \mathbb{P}(E)$.

It is easy to see that if $E$ is positively correlated with $H$, then $\tilde{E}$ is negatively correlated with $H$.

If $E$ is negatively correlated with $H$, then $\tilde{E}$ is positively correlated with $H$.

If $E$ is non-correlated with $H$, then also $\tilde{E}$ is non-correlated with $H$.

Assume that $\mathbb{P}(E) > 0$ and $\mathbb{P}(H) > 0$, then we can formulate the previous properties in a symmetric way.

Therefore we can say that $E$ and $H$ are *positively correlated* if $\mathbb{P}(EH) > \mathbb{P}(E)\mathbb{P}(H)$.

$E$ and $H$ are *negatively correlated* if $\mathbb{P}(EH) < \mathbb{P}(E)\mathbb{P}(H)$.

$E$ and $H$ are *non-correlated* or *stochastically independent* if $\mathbb{P}(EH) = \mathbb{P}(E)\mathbb{P}(H)$.

Independence of $n$ events.

We say that $n$ events $E_1, E_2, \ldots, E_n$ are *stochastically independent* if for every subset

$$\{i_1, i_2, \ldots, i_k\} \subset \{1, 2, \ldots, n\}$$

we have

$$\mathbb{P}(E_{i_1} E_{i_2} \ldots E_{i_k}) = \mathbb{P}(E_{i_1})\mathbb{P}(E_{i_2}) \ldots \mathbb{P}(E_{i_k}).$$

This is equivalent to the following condition

$$\mathbb{P}(E_1^* E_2^* \ldots E_n^*) = \mathbb{P}(E_1^*)\mathbb{P}(E_2^*) \ldots \mathbb{P}(E_n^*),$$

for every choice of $E_i^*$ between $E_i$ and $\tilde{E}_i$ for $i = 1, 2, \ldots, n$.

**Bernoulli scheme.**

A sequence of events $E_1, E_2, \ldots$ stochastically independent and equiprobable (i. e. such that for every $i$ $\mathbb{P}(E_i) = p$ for some $p$ is called *Bernoulli scheme*. It represents a sequence of *trials* or experiments each one performed in the same conditions, cannot influence each other and can have two results called respectively *success* and *failure*.

**Distributions related to Bernoulli scheme.**

The number $S_n$ of successes in the first $n$ trials of a Bernoulli scheme of parameter $p$ is said to have *binomial distribution with parameters $n$ and $p$*. It follows from the properties of expectation, variance and covariance that the expectation of $S_n$ is $np$ as the sum of $n$ random numbers with expectation $p$ and that its variance is $np(1-p)$ as the sum of $n$ non-correlated random numbers each one with variance $p(1-p)$.

The number of the trial where the first success is obtained in a Bernoulli scheme with parameter $p$ is said to have *geometrical distribution with parameter $p$*. It can be shown that this distibution has expectation $\frac{1}{p}$ and variance $\frac{1-p}{p^2}$.

## Limits of these distribution

## Poisson distribution

Assume $p$ is very small and $n$ is very large in such a way that $np = \lambda$.

Then the binomial distribution with these parameters can be approximated by *Poisson distribution with parameter* $\lambda$ a distribution with possible values $0, 1, 2, \ldots$ and with

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

. The expectation and the variance of $X$ are both equal to $\lambda$

## Exponential distribution

A number $Y = X/n$ where $X$ has geometrical distribution with parameter $p$ has approximatively a continuous *exponential distribution with parameter* $\lambda$ with parameter $\lambda$, possible values all positive numbers and $\mathbb{P}(Y \leq x) = 0$ for $x < 0$ and $\mathbb{P}(Y \leq x) = 1 - e^{-\lambda x}$ for $x > 0$. The expectation of this distribution is $\frac{1}{\lambda}$ and the variance is $\frac{1}{\lambda^2}$. This is a an *absolutely continuous* distribution with a *probability densitry* $p(x)$ given by 0 for $x < 0$ and $\lambda e^{-\lambda x}$. The probability that a number with this distribution belongs to the interval $[a, b]$ with $a < b$ is given by

$$\int_a^b p(x)dx.$$

## Standard Normal or Gaussian distribution

Let $X$ be a random number with binomial distribution with parameters $n$ and $p$. We can perform on it the operation of standardization and obtan the random number $Y$ given by

$$Y = \frac{X - np}{\sqrt{np(1-p)}}.$$

$Y$ is obtained from $X$ by means of a linear transformation in such a way that its expectation is 0 and its variance is 1. It can be shown that when $n$ is large $Y$ has approximatively a *standard normal* or *standard Gaussian* distribution, an absolutely continuous distribution with probability density $n(x)$ given by

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

.

The probability that a number with this distribution belongs to the interval $[a, b]$ with $a < b$ is given by

$$\int_a^b n(x)dx = N(b) - N(a),$$

where

$$N(x) = \int_{-\infty}^{x} n(y)dy.$$

$N(x)$ gives the probability that a number with standard normal distribution is less than or equal to $x$. $N(x)$ cannot be expressed in terms of elementary function but its values for non-negative $x$ can be found in every text book of elementary statistics. The values for negative $x$ can be obtained by means of the relation

$$N(x) = 1 - N(-x).$$

**General normal or gaussian distribution.**

The general *normal* or *Gaussian* distribution is absolutely continuous distribution with two parameters $m$ and $\sigma$. Its probability density is given by

$$\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-m)^2}{2\sigma^2}}.$$

If $Y$ has normal distribution with parameters $m$ and $\sigma$, then $X = \frac{Y-m}{\sigma}$ has standard normal distribution. Therefore we can use the tables of $N(x)$ to compute the probability that $Y$ belongs to the interval $[a, b]$ with $a < b$

$$\mathbb{P}(a \leq Y \leq b) = \int_a^b \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

is given by

$$\mathbb{P}(a \leq Y \leq b) = N(\frac{b-m}{\sigma}) - N(\frac{a-m}{\sigma}).$$

It follows from the properties of expectation and variance that the expectation and variance of $Y$ can be obtained from those of $X$. The expectation of $Y$ is equal to $m$ and its variance is equal to $\sigma^2$.

## Markov chains

Let $S$ be a finite or denumerable set.

Then a *homogeneous Markov chain* is a sequence of random numbers

$$X_0, X_1, X_2, \ldots$$

with *state space* $S$ is defined by giving the *initial distribution* $\rho_s$ for $s \in S$ and the *matrix of transition probabilities* $\Pi = (p_{s,s'})$ for $s, s' \in S$,

by saying that

$$\mathbb{P}(X_0 = s_0, X_1 = s_1, \ldots, X_n = s_n) =$$

$$\rho_{s_0} p_{s_0,s_1} \cdots p_{s_{n-1},s_n}$$

or (Markov property)

$$\mathbb{P}(X_n = s_n | X_0 = s_0, X_1 = s_1, \ldots, X_{n-1} = s_{n-1}) =$$

$$p_{s_{n-1},s_n}.$$

They satisfy

$$0 \leq \rho \leq 1 \qquad \sum_{s \in S} \rho_s = 1.$$

$$\mathbf{\Pi} = \begin{pmatrix} p_{11} & p_{12} & \cdots \\ p_{21} & p_{22} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

$$0 \leq p_{ss'} \leq 1 \qquad \sum_{s' \in S} p_{ss'} = 1 \quad \forall s \in S.$$

**Examples of random walks.**

Random walk on $\mathbb{Z}$.

The state spase $S$ is $\mathbb{Z}$. For some $p$, $0 < p < 1$ $p_{s,s+1} = p$, $p_{s,s-1} = 1 - p$.

Random walk on the interval $[a, b] \subset \mathbb{Z}$ with absorbing boundary conditions.

**Urn models.**

## Transition probabilities in more steps.

$$p_{s,s'}^{(n)} \triangleq \mathbb{P}(X_{k+n} = s' | X_k = s) = (\Pi^k)_{s,s'}$$

**Ergodic theorem for Markov chains.**

If $S$ is finite and the Markov chain is *irreducible* and aperiodical then for every states $s, s' \in S$ the following limit exists

$$\lim_{n\to\infty} p_{s,s'}^{(n)} = \mu_{s'}$$

and does not depend on $s$.

Moreover $\mu_s$ is the unique solution of the following system of equations

$$\sum_s \mu_s = 1$$

$$\sum_s \mu_s p_{s,s'} = \mu_{s'}.$$

This implies that $\mu_s$ is the (unique) *invariant* (or *stationary*) distribution for the Markov chains. This means that if we put the initial distribution of the Markov chain equal to $\mu_s$, then the distribution of $X_n$ for all $n \geq 0$ is $\mu_n$.

Moreover $\mu_s$ is the percentage of time that the Markov chain is in state $s$ for every initial distribution.

## Hypothesis Testing: Examples

Unknown mean $\mu$ of a normal distribution with known variance $\sigma^2$. 1.645 is the value $\overline{x}$ such that the normal standard distribution has a probability 0.05 to be larger than $\overline{x}$. Let

$$Z = \frac{(\overline{X} - \mu_o)\sqrt{n}}{\sigma}.$$

If $Z \geq \overline{x}$, then the hypothesis $\mu = \mu_0$ is rejected.

Case in which $\sigma^2$ is unknown.

In this case a *one-sample t-test* is used. Here one estimates the unknown variance $\sigma^2$ by

$$s^2 \triangleq \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}.$$

The test statistic is then

$$t = \frac{(\overline{x} - \mu_0)\sqrt{n}}{s}.$$

The test is then similar to that when the variace is known. Only in this case one uses instead of standard normal distribution the *Student t-distribution* with $\nu = n - 1$ degrees of freedom with probability density

$$f(t) \triangleq \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})(1+\frac{t^2}{\nu})^{(\nu+1)/2}}$$

for $-\infty < t < \infty$.

A protein coding gene is a segment of the DNA that codes for a particular protein (or proteins). In any given cell type at any given time, this may or may not be needed. Each cell will generate the proteins it needs, which will usually be some small subset of all possible proteins. If a protein is generated in a cell, we says, we say that the gene coding for this protein is *expressed* in that cell type. When this happens we say that the gene is *differentially expressed* between the two cell types. There are several techniques for measuring the level of gene expression in a cell type. All of these methods are subject to both biological and experimental variability. Therefore, one cannot simply measure the level of expression once in each cell type to test for differential expression. Instead, one must repeat each experiment several times and perform a statistical test of the hypothesis that they are expressed at the same or different levels.

Suppose that the mean expression levels of a given gene in two cell types, for example normal and tumor (cancerous) cells, are to be compared. In statistical terms, this comparison can be framed as the test of the equality of two unknon means. For the moment we assume that the (unknown) variance of expression level in normal cells is identical to that in tumor cells. To test for equality of the two means, we plan to measure the expression levels of $m$ cells of one type and compare these with the expression levels of $n$ cells of another type.

Suppose that, before the experiment, the measurements $X_{11}, X_{12}, \ldots, X_{1m}$ from the first cell type are thought as $m$ $NID(\mu_1, \sigma^2)$ random variables, and the measurements $X_{21}, X_{22}, \ldots, X_{2n}$ from the second cell type are thought as $n$ $NID(\mu_2, \sigma^2)$ random variables. The null hypotheses states that $\mu_1 = \mu_2$. We assume for the moment that the alternative hypothesis leaves both $\mu_1$ and $\mu_2$ unspecified, so that our test is two-sided.

Let $\nu_1 = m - 1$, $\nu_2 = n - 1$ , $\nu = \nu_1 + \nu_2$, $\nu s^2 = \nu_1 s_1^2 + \nu_2^2$, $\delta = \mu_1 - \mu_2$ then

$$t = \frac{(\delta - (\overline{x_1} - \overline{x_2}))}{s(1/n_1 + 1/n_2)^{\frac{1}{2}}}$$

has Student's t-distribution with $\nu$ degrees of freedom.

Case with different variances

Let $\delta = \mu_1 - \mu_2$.

$$d = \frac{(\delta - (\overline{x_1} - \overline{x_2}))}{(s_1^2/n_1 + s_2^2/n_2)^{\frac{1}{2}}}$$

has *Behren's distribution* with $\nu_1$ and $\nu_2$ degrees of freedom

$$d = t_1 \cos(\tilde{\omega}) - t_2 \sin(\tilde{\omega}).$$

and angle $\tilde{\omega}$ given by

$$\tan(\tilde{\omega}) = (s_2/\sqrt{n_2})/(s_1/\sqrt{n_1}).$$

Testing for the parameters in a multinomial distribution

$$X^2 = \sum_{i=1}^{k} \frac{(Y_i - np_i)^2}{np_i}$$

has approximately $\chi^2$ ( *chi square* ) distribution with $k-1$ degreed of freedom.

## Maximum likelihood for Markov chains

The probability of a sequence $s_1, s_2, \ldots, s_n$, given that the initial state is $s_0$ is given by

$$p_{s_0,s_1} p_{s_1,s_2} \cdots p_{s_{n-1},s_n}.$$

This can be written as

$$\prod_{s,s'} p_{s,s'}^{n_{s,s'}},$$

where $n_{s,s'}$ denotes the number of transitions from $s$ to $s'$. The maximum likelihood point is obtained for

$$\overline{p}_{s,s'} = \frac{n_{s,s'}}{n_{s,\cdot}}.$$

**Association tests.**

Tests of associations are used typically when observations are categorized into a certain number of "row" categories and into a certain number of "column" categories. The test evaluate the association between row and column categories. We start with the case of two rows and two columns.

**Examples:**

Suppose that $N$ laboratory mice, $n$ of which are males and $N - n$ females are irradiated. We wish to test whether a certain mutation is more likely to arise in male mice than in females. After the radiation it is found that, in all, there are $m$ new mutant mice in the joint sample of males and females, and thust $N-m$ mice which are non-mutant. Conditional on the event that the total number of mutant mice is $m$, the number $Y$ of mutant males has the hypergeometric distribution if there is no association between gender and the propensity to be a mutant.

Suppose that the probability that a mouse of either gender is a mutant is $p$. The number of mutant male mice has a binomial distribution with parameters $p$ and $n$ and the number of mutant female mice has a binomial distribution with parameters $p$ and $N-n$. The total number of mutant mice has a binomial distributions with parameters $p$ and $N$.

Let $A_1$ be the event that $y$ male mice are mutant and $A_2$ be the event that, in all, $m$ mice are mutants. Using the formula of conditional probability we can derive the hypergeometric distribution for the number of mutant male mice, given the total number $m$ of mutant mice. The event $A_1A_2$ is the event thet $y$ males and $m-y$ females are mutants. This allows to compute the probability of $A_1A_2$.

We want to test the hypothesis that there is no association between gender and propensity to be a mutant, the alternative hypothesis of interest being that males are more likely to be mutants than are females.

Suppose that $n$ male mice and $N - n$ female mice are irradiated, and that in all a total of $m$ mutant mice is observed and thus a total of $N - m$ non-mutants. These four totals are taken as given.

To illustrate the calculations, suppose that $n = 8$, that $N = 20$ and that $m = 9$. Of the males, $y = 6$ are mutants. These data may be arranged in the form of a two-by-two contingency table, as shown below.

| | mutant | non-mutant | total |
|---|---|---|---|
| male | 6 | 2 | 8 |
| female | 3 | 9 | 12 |
| total | 9 | 11 | 20 |

What is the probability of observing a value 6 or larger in the upper left-hand cell in the table, assuming that the null hypothesis is true and that the four marginal totals $8, 12, 9$ and $11$ are given. The hypergeometric formula shows that this probability is

$$\frac{\binom{8}{6}\binom{12}{3}}{\binom{20}{9}} + \frac{\binom{8}{7}\binom{12}{2}}{\binom{20}{9}} + \frac{\binom{8}{8}\binom{12}{1}}{\binom{20}{9}}$$

## Tables of arbitrary size

Assume that categorization count data arise in the form of a two-way table with an arbitrary number $r$ of rows and an arbitrary number $c$ of columns.

| | | column | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | $\cdots$ | c | Total |
| row | 1 | $Y_{11}$ | $Y_{12}$ | $Y_{13}$ | $\cdots$ | $Y_{1c}$ | $y_{1\cdot}$ |
| | 2 | $Y_{21}$ | $Y_{22}$ | $Y_{23}$ | $\cdots$ | $Y_{2c}$ | $y_{2\cdot}$ |
| | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| | r | $Y_{r1}$ | $Y_{r2}$ | $Y_{r3}$ | $\cdots$ | $Y_{rc}$ | $y_{r\cdot}$ |
| | Total | $y_{\cdot 1}$ | $y_{\cdot 1}$ | $y_{\cdot 1}$ | $\cdots$ | $y_{\cdot c}$ | $y$ |

It is assumed that the $y$ observations leading to the counts $\{Y_{jk}\}$ are independent of each other. This fact is often overlooked in the application of chi-square procedures in bioinformatics.

Given the row and column totals, it can be shown that if the two categories are independent in the population then $Y_{jk}$ is a random number with expectation value $E_{jk}$ given by

$$E_{jk} = \frac{y_{j\cdot} y_{\cdot k}}{y}$$

## Association tests

$$\sum_{jk} \frac{(Y_{jk} - E_{jk})^2}{E_{jk}}$$

has approximately $\chi^2$ distribution with

$$\nu = (r-1)(c-1)$$

degrees of freedom.

## The Analysis of one DNA sequence

many overlapping small piecec each of the order of 500 bases (nucleotides)

assembling fragments into one long "contig"

"shotgun sequencing"

*n-times coverage* or *nX* coverage if the length of the original sequence is $G$ the total length of the fragments sequenced is $nG$.

probabilistic issues

$N$ fragments each of length $L$

The coverage $a$ is given by

$$a = NL/G.$$

The number $Y$ of fragments whose left-hand end is located within an interval of length $L$ to the left of randomly-chosen point therefore has a Poisson distribution with mean $a$, so that the probability that at least one fragment arises in this interval is $1 - \Pr(Y = 0) = 1 - e^{-a}$.

What is the means proportion of the genome covered by contigs?

What is the mean number of contigs?

What is the mean contig size?

| a | 2 | 4 | 6 |
|---|---|---|---|
| Mpgc | .864665 | .981684 | .997521 |

| a | 8 | 10 | 12 |
|---|---|---|---|
| Mpgc | .999665 | .999955 | .999994 |

The mean number of contigs is the number $N$ of fragments multiplied by the probability that a fragment is the rightmost member of a contig.

Mean number of contig= $Ne^{-a} = Ne^{-NL/G}$.

| a | .5 | .75 | 1 | 1.5 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|---|
| Mnc | .60.7 | 70.8 | 73.6 | 66.9 | 54.1 | 29.9 | 14.7 |

| a | 5 | 6 | 7 |
|---|---|---|---|
| Mnc | 6.7 | 3.0 | 1.3 |