Statistics and Probability in Bioinformatics

It is a MUST. It is not ENOUGH.

A very brief introduction.

Statistics	Probability
Data $\{y_i\}$	Prob distribution
Estimate:	**Calculate**:
Average	Expectation
Deviation	Variance
	Moments, $\cdots$
Various Estimators	Any $f(Y)$
Testing hypotheses	
"Useful but no theory"	"Theory of no use"

The last row is an exaggeration by ZHENG Wei-mou.

K. Lange, *Mathematical and Statistical Methods for Genetic Analysis*, Springer, 1997.

W. J. Ewens, G. R. Grant, *Statistical Methods in Bioinformatics*, Springer, 2001.

Statistical methods alone are not strong enough to amplify the difference between a DNA sequence and a random sequence or the difference between two DNA sequences. Need more "deterministic" approaches.

This is a working programme, not a summary of accomplished research.

Discrete Random Systems

Sampling **space** consists of (finite or infinite) discrete points.

1. Coin tossing: $\{\text{Head}, \text{Tail}\}$, $\{0, 1\}$, $\{-1, 1\}$
2. Dice tossing (a cube): $\{1, 2, 3, 4, 5, 6\}$
3. A Nucleotide Die (a tetrahedron): $\{a, c, g, t\}$
4. An Amino Acid Die: $\{A, C, \cdots, W, Y\}$

Get used to think in terms of **spaces** even when the latter contain a finite number of points.

Plato Polyhedron: 4, 6, 8, 12, 20.

Random Variables

It is a good practice to use two symbols, e.g., Y and y:

Y — name of a random variable, an abstraction, may be defined in words.

y — a value that Y takes at an **observation**, at a **realization**, or at a **sampling**.

When y takes discrete values Y is a discrete random variable.

The collection of all possible $\{y\}$ — sampling space, may be finite or infinite.

Probability Fuction:

$$P_Y(y) = Prob(Y = y)$$

— the probability that Y takes value y.

Frequency, Probability, Energy, and Temperature

Frequency of a, c, g, t in a DNA sequence:

$$N_a\, N_c\, N_g\, N_t$$

Normalization:

$$N_a + N_c + N_g + N_t = N$$

Divide by N to get probability of nucleotides:

$$p_a + p_c + p_g + p_t = 1$$

A useful trick: introduction of "energy" and "temperature":

$$p_a \to e^{-\frac{E_a}{T}}$$

Different "energies" but the same "temperature" (in energy unit or write kT in degrees Kelvin).

Temperature as a useful parameter

Two (or three) limits:

1. $T \to 0$ singles out the lowest energy state ("ground state").
2. $T \to \infty$: energy difference indifferent. Essence of simulated annealing
3. Might consider $T \to -\infty$: picking up the highest energy state.

Probability Notations

$P_Y(y)$

Prob$(Y = y)$

$P_Y(y; \theta)$ — θ stands for one or more parameters, written explicitly only when parameters are emphasized.

Normalization of probability:

$$\sum_{\{y\}} P_Y(y) = 1$$

Expection value = mean:

$$\mu \equiv E(Y) = \sum_{\{y\}} y P_Y(y)$$

This is a theoretical calculation. μ is determined by parameter(s) of the distribution.

Average:

$$\overline{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

$\overline{y}$ is a random variable. It is obtained from experiments.

mean $\neq$ average

$\overline{y}$ may be used as an estimator for μ.

Expectation of Y^2

$$E(Y^2) = \sum_{\{y\}} y^2 P_Y(y)$$

$E(Y^2)$ contains contribution from $E(Y)^2 = \mu^2$. In order to highlight the real "nonlinear", "self-correlation", define **Variance**:

$$\sigma^2 \equiv Var(Y) = E((Y-\mu)^2) = E(Y^2) - \mu^2$$

It is different from the **Average of the Squares**:

$$\overline{y^2} = \frac{1}{N} \sum_{i=1}^{N} y_i^2$$

$\overline{y^2}$ contains contribution of $(\overline{y})^2$.

$\overline{y^2} - (\overline{y})^2$ may be used as an estimator for σ^2.

A Trick to calculate Mean and Variance

Starting from the normalization equality

$$\sum_{\{y\}} P_Y(y;\theta) = 1$$

Taking derivatives on both sides:

$$\frac{d}{d\theta}\left(\sum_{\{y\}} P_Y(y;\theta)\right) = 0$$

Solve the above equation to get μ.

Taking derivative again:

$$\frac{d^2}{d\theta^2}\left(\sum_{\{y\}} P_Y(y;\theta)\right) = 0$$

From the above two Eqs. $\Rightarrow \sigma^2$.

Estimators: Unbiased and Biased

Estimating μ by $\overline{y}$ and σ^2 by $\overline{y^2} - \overline{y}^2$ are the simplest particular cases of estimating probabilistic characteristics by statistical quantities (Estimators).

In general, if

$$E(\text{Prob Characteristic} - \text{Estimator}) = 0$$

it is an **unbiased** estimator. Otherwise, it is a **biased** estimator.

The art of constructing unbiased estimators (look for Russians):

V. G. Voinov, M. S. Nikulin, *Unbiased Estimators and Their Applications*, Springer, 1993.

vol. 1 *Univariate Case*

vol. 2 *Multivariate Case*

Example of Estimators

The sample average $\overline{y}$ is an **unbiased** estimator of μ.

The sample deviation $\overline{y^2} - \overline{y}^2$ is a **biased** estimator of variance σ^2, because

$$E(\overline{y^2} - \overline{y}^2) = \frac{n}{n-1}\sigma^2$$

for Gaussian IID random variables.

Therefore, the **unbiased** estimator for σ^2 is

$$\frac{n-1}{n}(\overline{y^2} - \overline{y}^2)$$

Unbiased estimators are not necessarily better than biased estomators. There are cases when unbiased estimators simply do not exist.

Examples of Discrete Distributions

1. Bernulli test
2. Binomial distribution
3. Poisson distribution
4. Geometric distribution
5. Uniform distribution

Single Bernulli test (1)

Probability of success: p

Probability of failure: $1 - p$

Random variable Y:
$y = \begin{cases} 1, & \text{with } p \\ 0, & \text{with } (1-p) \end{cases}$

Distribution:
$P_Y(y) = p^y(1-p)^{1-y}, \;\; y = 0, 1$

Expectation: $\mu = p$

Variance: $\sigma^2 = p(1-p)$

Single Bernulli test (2)

Probability of success: p

Probability of failure: $1-p$

Define the random variable Z in a different way:

$$z = \begin{cases} 1, & \text{with } p \\ -1, & \text{with } (1-p) \end{cases}$$

Distribution:

$P_Z(z) = p^{\frac{1+z}{2}}(1-p)^{\frac{1-z}{2}}, \;\; z = 1, -1$

Home work:

Expectation: $\mu = 2p - 1$

Variance: $\sigma^2 = 4p(1-p)$

The Law of Big Numbers

Perform n Bernulli tests with success probability p. Denote the number of successes by S_n, then
$\lim_{n\to\infty} \frac{S_n}{n} = p.$

In probability theory notations:
$\lim_{n\to\infty} (|\frac{S_n}{n} - p| \leq \epsilon) = 0, \ \forall \epsilon > 0$

Roughly speaking: when n is big,
frequency $\to$ probability.

Binomial Distribution (1)

Many single Bernulli tests each with probability of success p

N independent tests

Random variable Y = number of successes in N tests: $y = 0, 1, \cdots, N$

Probability of $Y = y$:

$$B_Y(y; N, p) = C_N^y p^y (1-p)^{N-y}, \quad y = 0, 1, 2, \cdots, N$$

Symmetric only when $p = 0.5$.

1. Number of tests N fixed beforehand.
2. Independent tests.
3. Same p for all tests.

Binomial Distribution (2)

How to remember it:

$$(a+b)^N = \sum_{y=0}^{N} C_N^y a^y b^{N-y}$$

Let $a = p$, $b = 1 - p$ to get

$$1 = \sum_{y=0}^{N} C_N^y p^y (1-p)^{N-y}$$

That is:

$$\sum_{y=0}^{N} B(y; N, p) = 1$$

$$B_Y(y; N, p) = C_N^y p^y (1-p)^{N-y}, \quad y = 0, 1, 2, \cdots, N$$

Home work:

1. Expectation: $\mu = Np$.
2. Variance: $\sigma^2 = Np(1-p)$.

A limit of Binomial Distribution at

$N \to \infty$

$p \to 0$

$Np = \lambda$ finite:

$B_Y(y; N, p) \to \frac{e^{-\lambda}\lambda^y}{y!}$ (Poisson distribution)

Poisson Distribution

S. D. Poisson (1837)

Distribution of rare (in time or space) events:

$$P_Y(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$$

It is a **one-parameter** distribution. Almost a symmetric peak when $\lambda > 5$.

1. Number of α particle decays in a time interval.
2. Number of deaths due to horse running-mad in Prussian army.
3. Many instances of distribution of K-tuples in DNA sequences.
4. The percentage represented at a certain coverage X in a sequencing project.
5. Lander-Waterman curve for number of contigs versus coverage.

Poisson Distribution

How to remember it?

Decomposition of unit:
$1 = e^{-\lambda} e^{\lambda}$

Insert the series expansion

$$e^{\lambda} = \sum_{n=0}^{\infty} \frac{\lambda^n}{n!}$$

to get

$$1 = \sum_{n=0}^{\infty} \frac{e^{-\lambda} \lambda^n}{n!}$$

$$P_Y(y; \lambda) = \frac{e^{-\lambda} \lambda^y}{y!}$$

We have

$$\sum_{n=0}^{\infty} P_Y(y; n, \lambda) = 1$$

The percentage represented at a certain coverage X

G — the genome size

L — the read size (usually $L \approx 500bp$)

Probability that a designated nt is in a read $\propto \frac{L}{G}$

Probability that a designated nt is **not** in a read $\propto 1 - \frac{L}{G}$

Probability that a designated nt is **not** in N copies of reads $\propto (1 - \frac{L}{G})^N$

Let $\frac{L}{G} = \frac{NL}{NG} = \frac{X}{N}$, as $NL = XG$

X is called **coverage**.

The above probability $\propto (1 - \frac{X}{N})^N \rightarrow e^{-X}$

$Prob$(a designated nt is represented in the reads) $= 1 - e^{-X}$

(Clarke and Carbon, 1976)

The percentage represented at coverage X

G — the genome size

L — the read size (usually $L \approx 500bp$)

N — number of sequenced reads

$X = \frac{N \times L}{G}$ — coverage

The percentage that a designated nt is represented in the reads:

Clarke-Carbon formula (1976):

$$f = 100 \times (1 - e^{-X})$$

X	1	2	3	4	5	6
f	63	86.5	95	98	99.4	99.75

X	7	8	9	10
f	99.91	99.97	99.99	99.995

What is e^{-X}

Suppose that the probability that y copies of the designated nt are present in the reads is given by Poisson distribution

$$P_Y(y;\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

The probability that $y = 0$ is

$$P_Y(0;\lambda) = e^{-\lambda}$$

Therefore, $\lambda = X$ in our problem.

Consequently, the probability that $y = 1$:

$$P_Y(1;X) = Xe^{-X}$$

This is an over-simplified version of the Lander-Waterman curve.

Maximum at $X = 1$ as seen from

$$\frac{dP_Y(1;X)}{dX} = e^{-X}(1 - X) = 0$$

A more realistic discussion must consider overlaps and assembling of reads into **contigs**.

We need **geometric distribution** to proceed.

Geometric Distribution

Single Bernulli test with success probability p, probability of failure $q = 1 - p$

Random variable $Y = y$ if y consecutive successes followed by a failure.

Probability distribution:
$P_Y(y) = p^y(1-p), \quad y = 0, 1, 2, \cdots,$

Applications:

1. Number of Q (Gln, Glutamine) runs in SWISS-PROT.
2. Probability of single-letter runs in a Markov Model.

Biased Geometric Distribution

$y_{\text{min}} = c$, i.e.,
$y = c, c+1, c+2, \cdots, c+k, \cdots$

Probability distribution:
$P_Y(k+c) = p^k(1-p), \quad k = 0, 1, 2, \cdots,$

Expressed via probability of failure $q - 1 - p$:
$P_Y(k+c) = (1-q)^k q$

Applications:
Intron length distribution:
Minimal intron length: 50-90 bp depending on species.

Lander-Waterman Curve (1)

G — Haploid genome size

L — Length of a read in bp

N — Number of reads sequences

X — Coverage of the genome: $N \times L = X \times G$

T — Minimal overlap to assemble two reads into a contig, $\theta = T/L$, $\sigma = 1 - \theta$

Probability of encountering a read: $\alpha = \frac{N}{G}$

Probability of encountering an isolated read: $\alpha(1-\alpha)^{L-T}$

Probability of encountering two overlapping reads: $\alpha^2(1-\alpha)^{L-T}$

"Stopping propability":

$$(1-\alpha)^{L-T} = (1-\frac{N}{G})^{L\sigma} \rightarrow e^{-X\sigma}$$

Lander-Waterman Curve (2)

"Stopping propability":

$$(1-\alpha)^{L-T} = (1-\frac{N}{G})^{L\sigma} \rightarrow e^{-X\sigma}$$

Compare to the Poisson distribution at $y = 0$: $\lambda = X\sigma$

Poisson distribution for $y = 1$ gives the essense of Lander-Waterman:

$$X\sigma e^{-X\sigma}$$

# of contigs = # of exits from a read: $\alpha e^{-X\sigma}$

# of contigs at coverage X:

$$G \times \alpha e^{-X\sigma} = \frac{XG}{L} e^{-X\sigma}$$

Return to Clarke-Carbon at $\theta = 0$, i.e., $\sigma = 1$

Physical Mapping vs. Sequencing

	Physical Mapping By Fingerprinting	Sequencing By WGS
G	Haploid genome size	
L: Length of	Clone	Read
N: # of	Fingerprinted clones	Sequenced reads
$X = \frac{LN}{G}$	Coverage	Coverage
	Islands	Contigs
	# of islands	# of contigs
T	Minimal overlap for extension $\theta = T/L$ $\sigma = 1 - \theta$	
$\alpha = N/G$		
$p = L/N$		

Continuous Distributions

Examples:

1. Normal distribution $N(\mu, \sigma^2)$
2. Exponential distribution
3. Extreme value distribution
4. Gamma distribution

Probability Density Function (PDF): $\rho(x)$, may not exist.

Distribution Function $F(x)$ always exists:

$$F(x) = \int_{-\infty}^{x} \rho(y)dy$$

Normal Distribution

Probability density for continuous random variable X with mean μ and variance σ^2:

$N(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

In terms of normalized variable $z = \frac{x-\mu}{\sigma}$:

$N(0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$

Tabulated is the distribution function:

$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-z^2/2} dz, \ \Phi(-x) = -\Phi(x)$

Main application:

$Prob(-1 < z < 1) = 2\Phi(1) = 0.683,$

$Prob(-2 < z < 2) = 2\Phi(2) = 0.954,$

$Prob(-3 < z < 3) = 2\Phi(3) = 0.997,$

$Prob(-4 < z < 4) = 2\Phi(4) = 0.9999$

Extreme Value Distributioni (EVD)

EVD is behind the scoring scheme of BLAST
("Basic Local Alignment Search Tool"
by Altschul et al., 1990)

Gapped-BLAST

PSI-BLAST

Note: EVD is a particular case of **order statistics**.

Central Limiting Theorems

iid random variables $\{x_i\}$ with finite μ and σ^2. Consider random variables:

$$S_n = \sum_{i=1}^{n} x_i,\ \overline{X} = \frac{S_n}{n}$$

$$E(S_n) = n\mu,\ \sigma^2(S_n) = n\sigma^2$$

Then
$\frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma} \sim N(0, 1)$

The sum of a great number of **iid** random variables tend to obey normal distribution.

May be relaxed to dependent case.

Chebeshev Inequality

For any distribution with finite mean μ and finite variance σ^2:

$$Prob(|X - \mu| \leq d) \leq \frac{\sigma^2}{d}$$

Extra notes:
1. Two distinguished students of Chebeshev: Markov and Laypunov.
2. Chebeshev polynomials as best finite approximants in fitting any function. Finite Taylor's expansion being the worst.

Moments:

Given N samples of a random variable $\{x_i\}$:

1st moment: $\mu_1 = E(X) \Leftarrow \frac{1}{N}\sum_{i=1}^{N} x_i$

2nd momemnt: $\mu_2 = E(X^2) \Leftarrow \frac{1}{N}\sum_{i=1}^{N} x_i^2$

k-th moment: $\mu_k = E(X^k) \Leftarrow \frac{1}{N}\sum_{i=1}^{N} x_i^k$

How to calculate them all? Calculate the expectation of a convenient function of the random variable X, for example, e^{tX}, where i is the imaginary unit.

Moment Generating Function (mgf):

$$M(t) = E(e^{tX}) = \sum_{j=0}^{\infty} \frac{t^j E(X^j)}{j!} = \sum_{i=0}^{\infty} \frac{t^j \mu_j}{j!}$$

$$\mu_j = \frac{d^j}{dt^j} M(t)|_{t=0}$$

Cummulants:

Given N samples of a random variable $\{x_i\}$. Recall the **average**, **variance**, $\cdots$ of a random variable:

$$c_1 \equiv \mu \Leftarrow \frac{1}{N}\sum_{i=1}^{N} x_i, \text{ (1st cummulant)}$$

$$c_2 \equiv \sigma^2 \Leftarrow \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2, \text{ (2nd cummulant)}$$

$$c_3 = \mu_3 - 3c_1c_2 + 2c_1^3, \text{ (3rd cummulant)}$$

Key point: highlight the contribution of the highest order nonlinear terms by subtracting combinations of lower ones. How to calculate them all? Define a **Cummulant Generating Function (cgf):**

$$C(t) = \sum_{j=0}^{\infty} \frac{t^j c_j}{j!}$$

It is a matter of **Exponentiation** of the **mgf**:

$$M(t) = e^{C(t)} \quad \text{or} \quad C(t) = \ln M(t)$$

On Exponentiation

	Exponentiation $\Rightarrow$	
Statistics	Frequency p_i (probability)	"Energy" $e^{-\frac{E_i}{T}}$
Probability Theory	Moments	Cummulants
Graph Theory	Number of graphs	Number of connected graphs
Field Theory	Wick's Theorem	
Complex Analysis	Unit circle	Origin
Continuous Group Theory	Lie groups	Lie algebras

Essense of **Statistics**
and **Statistical Physics**

Maximal uncertainty of input data, observation, predicates, $\cdots$

Minimal uncertainty of results, conclusion, inference, $\cdots$

Maximal Likelihood $\Leftrightarrow$ Minimal Entropy

Bridge between "microscopic" and "macroscopic" descriptions: from huge data to few characteristics (thermodynamic quantities, political decisions, $\cdots$)

Generalized Averages of **Renyi** (1)

$$\tilde{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$\tilde{x} = \left(\frac{1}{N} \sum_{i=1}^{N} x_i^2 \right)^{\frac{1}{2}}$$

$$\tilde{x} = \left(\frac{1}{N} \sum_{i=1}^{N} x_i^3 \right)^{\frac{1}{3}}$$

$$\tilde{x} = \left(\frac{1}{N} \sum_{i=1}^{N} x_i^k \right)^{\frac{1}{k}}$$

$$\tilde{x} = \Phi^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \Phi(x_i) \right)$$

Generalized Averages of **Renyi** (2)

Now take

$$\Phi(y) = e^{-\frac{y}{kT}}$$

and solve it for y to get Φ^{-1}:

$$y = -kT \ln \Phi(y), \quad \Phi^{-1}(.) = -kT \ln(.)$$

Express "macroscopic probability" in the same way as microscopic ones:

$$e^{-\frac{F(T)}{kT}} = \frac{1}{N} \sum_{\{j\}} e^{-\frac{E_j}{kT}}$$

Just denote the summation over all possible states by $Z(T)$ (the **partition function**), we get

$$F(T) = -kT \ln Z(T)$$

Statistical physics is nothing but doing Renyi average of the microscopic world to get macroscopic description.

Renyi's Theorem

There are only two choices of $\Phi(y)$ that allow for additivity of independent events:

1. $\Phi(y) = y$ — linear function.
2. $\Phi(y) = e^{\lambda y}$ — exponential function.

Statistical Physics in a Nutshell

Trilogy for equilibrium states:

1. Spectrum: j-th state with energy E_j
   Probability of that state: $\propto e^{-\frac{E_j}{kT}}$.

2. Normalization of probabilities $\rightarrow$ Partition Function:

$$\sum_j e^{-\frac{E_j}{kT}} = Z(T), \quad P(E_j) = \frac{e^{-\frac{E_j}{kT}}}{Z(T)}$$

3. Relation with thermodynamics via Free Energy and its derivatives:

$$F(T) = -kT \ln Z(T)$$

$$S = -\frac{\partial F(T)}{\partial T}, \quad p = -\frac{\partial F(T,V)}{\partial V}$$

Bayesian Statistics

Joint Probability $P(A, B)$ of two events A and B

Conditional Probability $P(A|B)$ — the probability of A conditioned on that of B. From the trivial relation

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A),$$

we get the

Thomas Bayer's Formula (1764):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

This "innocent" formula becomes much more meaningful if we interpret A as **Model** and B as **Data**:

$$P(\text{Model}|\text{Data}) = \frac{P(\text{Data}|\text{Model})P(Model)}{P(\text{Data})}$$

Posteriori $\leftarrow$ Likelihood + Priori

Information and Probability

Given a set of N possible outcomes with equal probability $p = 1/N$ for each, the **Information** I gained by learning that one outcome has realized (Hartley, 1928)

$$I = \log N = -\log p$$

When $\log_2$ is used, the unit information is called a **bit**. When natural logarithm ln is used it is called a **nat**.

Shannon (1948) extended Hartley's definition to a set of outcomes with different probabilities $\{p_i\}$:

$$I = -\sum_{i=1}^{N} p_i \log p_i$$

When $p_i = p$ for all i, Shannon reduces to Hartley.

Why taking logarithm? Additivity for **independent** events.

Both papers appeared in *Bell System Technical Journal*

Distance between Probability Distributions

Given two discrete distributions on the same set of events: $P = \{p_i\}|_{i=1}^{N}$ and $Q = \{q_i\}|_{i=1}^{N}$, how to define a **distance** between the two?

One possible definition: the Kullback-Leibler distance

$$D(P, Q) = \sum_i p_i \log \frac{p_i}{q_i}$$

Symmetrization: $\frac{1}{2}(D(P, Q) + D(Q, P))$

Another possible definition:

$$D(P, Q) = \sum_i \frac{2(p_i - q_i)^2}{p_i + q_i}$$

Positivity. Symmetry. Concavity.