BIOINFORMATICS & COMPUTATIONAL GENETICS MSc
PROBABILITY AND STATISTICS

Chapter 3. Statistical Analysis

Statistical analysis involves the informal/formal comparison of hypothetical or predicted behaviour with experimental results. For example, we wish to be able to compare the predicted outcomes of an experiment, and the corresponding probability model, with a data histogram. We will use both *qualitative* and *quantitative* approaches.

3.1 General Framework, Notation and Objectives

Suppose that an experiment or **trial** is to be repeated n times under identical conditions. Let X_i be the random variable corresponding to the outcome of the ith trial, and suppose that each of the n random variables $X_1, ..., X_n$ takes values in sample space $\mathbb{X}$.

Often, assumptions can reasonably be made about the experimental conditions that lead to simplifications of the joint probability model for the random variables. Essentially, the assumption of identical experimental conditions for each of the n trials implies that the random variables corresponding to the trial outcomes are **identically distributed**, that is, in the usual notation, the (marginal) mass/density function of X_i is denoted $f(x)$ dropping the subscript on the function f. Another common assumption is that the random variables $X_1, ..., X_n$ are **independent**. Thus $X_1, ..., X_n$ are usually treated as **i.i.d.** random variables.

In practice, it is commonly assumed that f takes one of the familiar forms (*Binomial*, *Poisson*, *Exponential*, *Normal* etc.). Thus f depends on one or more parameters (θ, λ, (μ, σ) etc.). The role of these parameters could be indicated by re-writing the function $f(x)$ as

$$f(x) \equiv f(x; \theta) \qquad x \in \mathbb{X} \tag{(*)}$$

where θ here is a generic parameter, which may possibly be vector-valued.

It is important here to specify precisely the range of values which this parameter can take; in a Poisson model, we have parameter $\lambda > 0$, and in a Normal model, we have parameters $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$. In the general case represented by (*) above, we have parameter $\theta \in \Theta$ where Θ is some subset of $\mathbb{R}^d$ and $d = 1, 2$, say, is the number of parameters. We refer to Θ as the **parameter space**. In practice, of course, parameter θ is **unknown** during the experiment.

3.1.1 Objectives of a statistical analysis

After the experiment has been carried out, a sample of **observed data** will have been obtained. Suppose that we have observed outcomes $x_1, ..., x_n$ on the n trials (that is, we have observed $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$), termed a **random sample**. This sample can be used to answer qualitative and quantitative questions about the nature of the experiment being carried out. The objectives of a statistical analysis can be summarized as follows. We want to, for example,

- **SUMMARY** : **Describe** and **summarize** the sample $\{x_1, ..., x_n\}$ in such a way that allows a specific probability model to be proposed.
- **INFERENCE** :**Deduce** and **make inference about** the parameter(s) of the probability model θ.
- **TESTING** :**Test** whether θ is "**significantly**" larger/smaller/different from some specified value.
- **GOODNESS OF FIT** :**Test** whether the probability model encapsulated in the mass/density function f, and the other model assumptions are **adequate** to explain the experimental results.

The first objective can be viewed as an **exploratory** data analysis exercise - it is crucially important to understand whether a proposed probability distribution is suitable for modelling the observed data, otherwise the subsequent formal inference procedures (estimation, hypothesis testing, model checking) cannot be used.

3.2 Exploratory Data Analysis

Objectives : To summarize and describe the sample data $\{x_1, ..., x_n\}$, in order to propose a plausible probability model.

The aim of this stage of analysis is first to produce summaries of the data in order to convey general trends or features that are present in the sample. Secondly, in order to propose an appropriate probability model, we seek to **match** features in the observed data to features of one of the conventional (Poisson, Exponential, Normal) probability distributions that may be used in more formal analysis. The four principal features that we need to assess in the data sample are

(1) The **location**, or the "average value" in the sample.
(2) The **mode**, or "most likely" value or interval observed in the sample.
(3) The **scale** or **spread** in the sample.
(4) The **skewness** or **asymmetry** in the sample.

These features of the sample are important because we can relate them **directly** to features of probability distributions.

3.2.1 Numerical Summaries

Sample mean	$\bar{x}$	$= \dfrac{1}{n}\sum_{i=1}^{n} x_i$
Sample variance	S^2	$= \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
(S^2 or s^2 may be used)	s^2	$= \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Sample quantiles		Suppose that the sample has been sorted into ascending order and re-labelled $x_{(1)} < ... < x_{(n)}$ Then the pth quantile, $0 < p < 100$, is given by
	$x^{(p)}$	$= x_{(k)}$ where k is the nearest integer to $pn/100$.
Median	m	$= x^{(50)}$, the 50th quantile
Lower quartile	q_{25}	$= x^{(25)}$, the 25th quantile
Upper quartile	q_{75}	$= x^{(75)}$, the 75th quantile
Inter-quartile range	IQR	$= q_{75} - q_{25}$
Sample minimum	x_{min}	$= x_{(1)}$
Sample maximum	x_{max}	$= x_{(n)}$
Sample range	R	$= x_{(n)} - x_{(1)}$
Sample skewness	κ	$= \dfrac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{S^2}$

Each of these summary statistics can be routinely computed from sample data using a calculator or statistical computer package.

3.2.2 Connection between sample statistics and probability models.

Consider the discrete probability distribution defined on the set of observed sample outcomes $\{x_1, ..., x_n\}$, by placing equal probability $1/n$ on each value, that is, the probability distribution specified by mass function denoted $f_{(n)}$

$$f_{(n)}(x) = \frac{1}{n} \qquad x \in \{x_1, ..., x_n\}.$$

Then the expectation of this probability distribution is given by

$$\mathrm{E}_{f(n)}[X] \quad = \sum_{i=1}^{n} x_i f_{(n)}(x_i) = \sum_{i=1}^{n} x_i \left\{\frac{1}{n}\right\} \quad = \quad \frac{1}{n}\sum_{i=1}^{n} x_i \quad = \quad \bar{x}$$

that is, the sample mean. Similarly, the variance of this probability distribution is equal to sample variance,

$$\mathrm{Var}_{f(n)}[X] = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = S^2.$$

In fact, each of the summary statistics listed above can be viewed as a feature of the probability distribution described by mass function $f_{(n)}$.

Now, consider this probability distribution as n increases to infinity. Then the sample mass function $f_{(n)}$ tends to a function f which can be regarded as the "true" mass/density function, and the sample mean, variance, percentiles etc. tend to the true mean, variance, percentiles of the distribution from which the data are generated.
In practice, of course, n is always finite, and thus the true distribution, true mean etc., cannot be known exactly. Therefore, we approximate the true distribution by an appropriately chosen distribution (Poisson, Exponential, Normal etc.) with parameters chosen to correspond to the observed sample properties.

3.2.3 Graphical Summaries

The most common graphical summary technique is the **histogram**. Typically, the sample space $\mathbb{X}$ is divided into a number of subsets $\mathbb{X}_1, ..., \mathbb{X}_H$, and the frequency with which a data value in the sample is observed to lie in subset $h = 1, ..., H$ is noted. This procedure leads to a set of counts $n_1, ..., n_H$ (where $n_1 + ... + n_H = n$) which are then plotted on a graph as a set of bars, where the hth bar has height n_h and occupies the region of $\mathbb{X}$ corresponding to $\mathbb{X}_h$.

The histogram again aims to approximate the "true" probability distribution generating the data by the observed sample distribution. It illustrates graphically the concepts of location, mode, spread and skewness and general shape features that have been recognized as important features of probability distributions.

3.2.4 Outliers

Each of the summaries described above is used on the presumption that each of the observed data values $x_1, ..., x_n$ is essentially an observation from the same probability distribution, and, because of this, each x_i is treated as equally important. Sometimes, however, for example due to slight variation in experimental conditions, one or two values in the sample may be much larger or much smaller in magnitude than the remainder of the sample. Such observations are termed **outliers** and must be treated with care, as they can distort the impression given by some of the summary statistics. For example, the sample mean and variance are extremely sensitive to the presence of outliers in the sample. Other summary statistics, for example those based on sample percentiles, are less sensitive to outliers. Outliers can usually be identified by inspection of the raw data, or from careful plotting of histograms.

3.3 Estimation

It is often of interest to draw inference from data regarding the parameters of the proposed probability distribution; recall that many aspects of the standard distributions studied are controlled by the distribution parameters.

EXAMPLE Lifetime modelling
Suppose that 20 product lifetime measurements are to be made; beginning at time $x = 0$, each product survives for some random time until failure. This yields a sample of observed failure times $x_1, x_2, ..., x_{20}$. Typically, we would assume a probability model, such as the Exponential, for such data; this distribution has a single parameter λ, which is unknown in practice, and thus requires estimation. Subsequently, we may be interested in prediction for a future product. For example, we may be interested in assessing the probability that an identical product would survive for longer than x^*; this probability is given under the Exponential model by $\mathrm{P}[\ X > x^*\ ] = e^{-\lambda x^*}$, which we may only evaluate if we know, or have an estimate of λ.

It is therefore important to find a simple and yet general technique for parameter estimation.

3.3.1 Maximum Likelihood Estimation

Maximum Likelihood Estimation is a systematic technique for estimating parameters in a probability model from a data. Suppose a sample $x_1, ..., x_n$ has been obtained from a probability model specified by mass or density function $f(x;\theta)$ depending on parameter(s) θ lying in parameter space Θ. The **maximum likelihood estimate** or **m.l.e.** is produced as follows;

STEP 1 Write down the **likelihood function**, $L(\theta)$, where

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

that is, the product of the n mass/density function terms (where the ith term is the mass/density function evaluated at x_i) viewed as a function of θ.

STEP 2 Take the natural log of the likelihood, and collect terms involving θ.

STEP 3 Find the value of $\theta \in \Theta$, $\hat{\theta}$, for which $\log L(\theta)$ is maximized, for example by differentiation. If θ is a single parameter, find $\hat{\theta}$ by solving

$$\frac{d}{d\theta}\ \{\log L(\theta)\} = 0$$

in the parameter space Θ. If θ is vector-valued, say $\theta = (\theta_1, ..., \theta_d)$, then find $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_d)$ by simultaneously solving the d equations given by

$$\frac{\partial}{\partial\theta_j}\ \{\log L(\theta)\} = 0 \quad j = 1, ..., d$$

in parameter space Θ.

Note that, if parameter space Θ is a bounded interval, then the maximum likelihood estimate may lie on the boundary of Θ.

STEP 4 Check that the estimate $\hat{\theta}$ obtained in STEP 3 truly corresponds to a maximum in the (log) likelihood function by inspecting the second derivative of $\log L(\theta)$ with respect to θ. If

$$\frac{d^2}{d\theta^2}\ \{\log L(\theta)\} < 0$$

at $\theta = \hat{\theta}$, then $\hat{\theta}$ is confirmed as the m.l.e. of θ (other techniques may be used to verify that the likelihood is maximized at $\hat{\theta}$).

This procedure is a systematic way of producing parameter estimates from sample data and a probability model; it can be shown that such an approach produces estimates that have good properties. After they have been obtained, the estimates can be used to carry out *prediction* of behaviour for future samples.

EXAMPLE A sample $x_1, ..., x_n$ is modelled by a Poisson distribution with parameter denoted λ; hence

$$f(x;\theta) \equiv f(x;\lambda) = \frac{\lambda^x}{x!}e^{-\lambda} \quad x = 0, 1, 2, ...$$

for some $\lambda > 0$.

STEP 1 Calculate the likelihood function $L(\lambda)$. For $\lambda > 0$,

$$L(\lambda) = \prod_{i=1}^{n} f(x_i;\lambda) = \prod_{i=1}^{n} \left\{ \frac{\lambda^{x_i}}{x_i!}e^{-\lambda} \right\} = \frac{\lambda^{x_1+...+x_n}}{x_1!....x_n!}e^{-n\lambda}$$

STEP 2 Calculate the log-likelihood $\log L(\lambda)$.

$$\log L(\lambda) = \sum_{i=1}^{n} x_i \log -n\lambda - \sum_{i=1}^{n} \log(x_i!)$$

STEP 3 Differentiate $\log L(\lambda)$ with respect to λ, and equate the derivative to zero.

$$\frac{d}{d\lambda}\{\log L(\lambda)\} = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0 \Longrightarrow \hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

Thus the maximum likelihood estimate of λ is $\hat{\lambda} = \bar{x}$

STEP 4 Check that the second derivative of $\log L(\lambda)$ with respect to λ is negative at $\lambda = \hat{\lambda}$.

$$\frac{d^2}{d\lambda^2}\{\log L(\lambda)\} = -\frac{\sum_{i=1}^{n} x_i}{\lambda^2} < 0 \text{ at} \lambda = \hat{\lambda}$$

3.4 Sampling Distributions

Maximum likelihood can be used systematically to produce estimates from sample data.

EXAMPLE : If a sample of data $x_1, ..., x_n$ are believed to have a Normal distribution with parameters μ and σ^2, then the maximum likelihood estimates based on the sample are given by

$$\hat{\mu} = \bar{x} \qquad \hat{\sigma^2} = S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

If five samples of eight observations are collected, however, we might get five different sample means

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	$\bar{x}$
10.4	11.2	9.8	10.2	10.5	8.9	11.0	10.3	10.29
9.7	12.2	10.4	11.1	10.3	10.2	10.4	11.1	10.66
12.1	7.9	8.6	9.6	11.0	11.1	8.8	11.7	10.10
10.0	9.2	11.1	10.8	9.1	12.3	10.3	9.7	10.31
9.2	9.7	10.8	10.3	8.9	10.1	9.7	10.4	9.89

and so the estimate $\hat{\mu}$ of μ is different each time.

We attempt to understand how $\bar{x}$ varies by calculating the **probability distribution** of the corresponding **estimator**, $\bar{X}$.

The estimator $\bar{X}$ is a **random variable**, the value of which is **unknown** *before* the experiment is carried out. As a random variable, $\bar{X}$ has a probability distribution, known as the **sampling distribution**. The form of this distribution can often be calculated, and used to understand how $\bar{x}$ varies. In the case where the sample data have a Normal distribution, the following theorem gives the sampling distributions of the maximum likelihood estimators;

THEOREM

If $X_1, ..., X_n$ are i.i.d. $N(\mu, \sigma^2)$ random variables, then

(1) $\bar{X} \sim N\left(\mu, \sigma^2/n\right)$,

(2) $$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} = \frac{nS^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1},$$

(3) $\bar{X}$ and S^2 are **independent** random variables.

First, we need to derive the probability distribution of the sum of independent Normal random variables. So consider the case of two independent random variables X_1 and X_2 where

$$X_1 \sim N(\mu_1, \sigma_1^2) \qquad X_2 \sim N(\mu_2, \sigma_2^2)$$

It can be shown that the distribution of $Y = X_1 + X_2$ is also Normal, that is, that

$$Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Thus, if $\mu_1 = \mu_2 = \mu$, $\sigma_1 = \sigma_2 = \sigma$,

$$Y = X_1 + X_2 \sim N(2\mu, 2\sigma^2) \ .$$

We therefore deduce that if Y is the sum of n such random variables, then

$$Y = \sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$$

Now, by the properties of the Normal distribution, if $U \sim N(0,1)$ then $V = aU + b \sim N(b, a^2)$, so we have that

$$\bar{X} = \frac{1}{n}\, Y \sim N\left(\mu, \frac{\sigma^2}{n}\right) \ .$$

Thus the estimator $\bar{X}$ has a Normal distribution with parameters μ and σ^2/n.

Proofs of (2) and (3) - omitted.

This theorem tells us how we expect the sample mean and sample variance to behave. In particular, it tells us that

$$\mathrm{E}[\bar{X}] = \mu \qquad \mathrm{E}[S^2] = \frac{n-1}{n}\sigma^2 \qquad \mathrm{E}[s^2] = \sigma^2$$

Interpretation : This theorem tells us how the sample mean and variance will behave if the original random sample is assumed to come from a Normal distribution. For example, if we believe that $X_1, ..., X_{10}$ are i.i.d random variables from a Normal distribution with parameters $\mu = 10.0$ and $\sigma^2 = 25$, then $\bar{X}$ has a Normal distribution with parameters $\mu = 10.0$ and $\sigma^2 = 25/10 = 2.5$.

The result will be used to facilitate formal tests about model parameters. For example, given a sample of experimental, we wish to answer **specific** questions about parameters in a proposed probability model.