BIOINFORMATICS & COMPUTATIONAL GENETICS MSc
PROBABILITY AND STATISTICS

SOME MORE EXERCISES

SECTION I : MAXIMUM LIKELIHOOD ESTIMATION

1. Suppose that $x_1, ..., x_n$ is a sample from an exponential distribution with parameter λ. Derive the m.l.e., $\hat{\lambda}$, of λ.

Now consider the function $\tau(\lambda)$ of λ, where τ is defined by

$$\tau(\lambda) = e^{-\lambda} = \mathrm{P}[X > 1]$$

where X is the random variable corresponding to the outcome of a new but identical experiment. Write down the likelihood **in terms of** τ, and thus find the m.l.e. of τ.

2. Find the m.l.e.s of the parameter θ based on a sample $x_1, ..., x_n$ from each of the following probability models;
   (a) $f(x;\theta) = \theta x^{\theta-1}$ $, 0 < x < 1$ for $\theta > 0$.
   (b) $f(x;\theta) = (\theta+1)x^{-\theta-2}$ $, 1 < x$ for $\theta > 0$.
   (c) $f(x;\theta) = \theta^2 x e^{-\theta x}$ $, 0 < x$ for $\theta > 0$.

3. Suppose that $x_1, ..., x_n$ and $y_1, ..., y_m$ are independent samples from Normal distributions with the same mean but different variances, that is,

$$x_1, ..., x_n \quad \text{are modelled using } N(\mu, \sigma_1^2)$$

$$y_1, ..., y_m \quad \text{are modelled using } N(\mu, \sigma_2^2).$$

First, verify that the the m.l.e.s for μ and σ_1^2 **based on** $x_1, ..., x_n$ **only** are given by

$$\begin{aligned} \hat{\mu} &= \bar{x} \\ \hat{\sigma_1^2} &= S^2 \end{aligned}$$

respectively, and then find the m.l.e.s of μ, σ_1^2 and σ_2^2 based on the **pooled** sample comprising both $x_1, ..., x_n$ and $y_1, ..., y_m$.

4. Find the maximum likelihood estimator, $\hat{\lambda}$, of parameter λ derived from a random sample $X_1, ..., X_n$ from a $Poisson(\lambda)$ distribution. By considering the probability distribution of random variable

$$X = \sum_{i=1}^{n} X_i$$

find an approximation to

$$\mathrm{P}[\lambda - \epsilon \leq \hat{\lambda} \leq \lambda + \epsilon] \quad \text{for any } \epsilon > 0$$

for large n using the **Central Limit Theorem** (see attached sheet) that says that

$$Poisson(\lambda) \approx N(\lambda, \lambda) \qquad \lambda \text{ large}$$

SECTION II : HYPOTHESIS TESTING

5. The pH of the catalyst in a chemical reaction is to be measured by a standard titration procedure. In order for the reaction to proceed at the desired rate, the catalyst pH must be very close to 6.4. The following measurements of pH were obtained;

6.492	6.397	6.469	6.343	6.397
6.542	6.467	6.351	6.390	6.446
6.311	6.288	6.396	6.515	6.438

Assuming that the data are well-modelled by a normal distribution with mean μ and variance σ^2, carry out a test of the hypothesis

$$\begin{array}{ll} H_0 & : \mu = 6.4 \\ H_1 & : \mu \neq 6.4 \end{array}$$

at the 5 % significance level, if

(i) σ is **known**, $\sigma = 0.03$,

(ii) σ is **unknown**.

6. Determining small quantities of calcium in the presence of Magnesium is a difficult problem for chemical analysts. Direct precipitation is not feasible, but one proposed procedure involves the use of alcohol as a solvent. The data in the first row below present the results of applying the alcohol method to ten mixtures containing known quantities of CaO (x, in mg). The second row gives, in each instance, the amount of CaO (y, in mg) recovered;

x	4.0	8.0	12.5	16.0	20.0	25.0	31.0	36.0	40.0	40.0
y	3.7	7.8	12.1	15.6	19.8	24.5	31.1	35.5	39.4	39.5

Assuming a linear regression relationship between corresponding random variables X and Y, that is, $\mathrm{E}[Y|X = x] = \alpha + \beta x$, and that the random errors are i.i.d normally distributed, calculate

(a) the m.l.e s of α and β,

(b) the fitted values and residuals from the resulting fit,

(c) the corrected estimate of variance, s^2

(d) the standard errors of the estimates of α and β,

(e) 95% confidence intervals for α and β.

On the basis of the confidence intervals, discuss whether or not there is evidence that $\alpha = 0$, or that $\beta = 1$.

7. The following data are believed to be a random sample from a $N(\mu, \sigma^2)$ distribution

$$249, 254, 243, 268, 253, 269, 287, 241, 273, 306$$
$$303, 280, 260, 256, 278, 344, 304, 283, 310.$$

Test the hypothesis

$$\begin{array}{ll} H_0 & : \sigma = 30.0 \\ H_1 & : \sigma \neq 30.0 \end{array}$$

at the 5 % significance level. Construct a 95 % confidence interval for μ

(i) σ is **known**, $\sigma = 30.0$,

(ii) σ is **unknown**.

and a 95 % confidence interval for σ^2.

9. It can be shown that if U and V are independent Chi-squared random variables with ν_1 and ν_2 degrees of freedom respectively, then the random variable Q defined by

$$Q = \frac{U/\nu_1}{V/\nu_2}$$

has an *F-distribution* with ν_1 and ν_2 degrees of freedom (this distribution forms the basis of the F-test for equal variances). However, statistical tables for the F-distribution only include values in the right tail of the distribution, that is, the value of x which satisfies

$$F_{F_{\nu_1,\nu_2}}(x) = 0.975.$$

where $F_{F_{\nu_1,\nu_2}}(.)$ is the cdf of the F-distribution. Show that

$$F_{F_{\nu_1,\nu_2}}(x) = 0.975 \Longleftrightarrow F_{F_{\nu_2,\nu_1}}(1/x) = 0.025$$

hence values for the left tail of the F-distribution can be calculated directly from the tables.

10. The following data are believed to be random samples from two normal distributions, $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$;

SAMPLE 1	:	$150, 142, 160, 110, 120, 240, 152, 136, 122, 200, 160, 102, 152, 208$
SAMPLE 2	:	$70, 110, 85, 90, 210$

(i) Carry out an appropriate two-sample hypothesis test of $\mu_1 = \mu_2$, if it is assumed that $\sigma_1 = \sigma_2$.

(ii) Is there any evidence that $\sigma_1 = \sigma_2$?

(iii) Construct a 95 % confidence interval for $\mu_1 - \mu_2$.

(iv) Construct a 95 % confidence interval for σ_1/σ_2.

SECTION III : MODEL VALIDATION AND GOODNESS-OF-FIT TESTS

11. Continuous random variable X has a *Weibull* distribution, so that its cdf, F_X, is given by

$$F_X(x) = \mathrm{P}[X \leq x] = 1 - \exp\{-\lambda x^\alpha\} \qquad x \geq 0$$

for parameters $\alpha, \lambda > 0$. Suppose that a sample of data $x_1, ..., x_n$ is available. By considering the solutions $q_1, ..., q_n$ to equations

$$F_X(q_i) = \frac{i}{n+1} \qquad i = 1, ..., n$$

describe how to check that the Weibull model is an appropriate probability model for the sample using a probability plotting approach.

12. A sample of count data was obtained, and a frequency table was formed. The data are summarized below.

Count	Frequency
0	44
1	86
2	56
3	12
4	2

Using the Chi-squared Goodness of fit test, carry out hypothesis tests of whether the following probability models are appropriate for the data:

(i) a uniform distribution on the set $\{0, 1, 2, 3, 4\}$

(ii) a binomial distribution with parameters $n = 4$, $p = 0.25$.

(iii) a binomial distribution with parameter $n = 4$ and p unspecified

(iv) a Poisson distribution with parameter $\lambda = 1.0$

(v) a Poisson distribution with parameter λ unspecified.

THE CENTRAL LIMIT THEOREM

THEOREM

Suppose $X_1, ..., X_n$ are i.i.d. random variables with

$$\mathrm{E}_{f_X}[X_i] = \mu, \qquad \mathrm{Var}_{f_X}[X_i] = \sigma^2$$

Let the random variable Z_n be defined by

$$Z_n = \frac{\displaystyle\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}}$$

Then, as $n \longrightarrow \infty$, $Z_n \longrightarrow Z \sim N(0,1)$ **irrespective** of the distribution of $X_1, ..., X_n$, so probability calculations involving Z_n can be approximated using the unit normal c.d.f, Φ, that is,

$$\mathrm{P}[Z_n \leq z] \approx \Phi(z).$$

The central limit theorem is important in probability theory because it justifies the use of an otherwise obscure pdf, and allows us to make approximate probability statements about other probability distributions. For example, can approximate the Binomial, Poisson and Gamma distributions by the normal, that is

BINOMIAL	$Binomial(n, \theta)$	$\approx N(n\theta, n\theta(1-\theta))$	n, θ large
POISSON	$Poisson(\lambda)$	$\approx N(\lambda, \lambda)$	λ large
GAMMA	$Gamma(\alpha, \beta)$	$\approx N(\alpha/\beta, \alpha/\beta^2)$	α large

These results allows us to approximate some otherwise awkward probabilities.

EXAMPLE : Suppose that $X \sim Binomial(n, \theta)$. To calculate $\mathrm{P}[X \leq x]$ for large n, first define random variables $X_1, ..., X_n$ to be i.i.d. $Bernoulli(\theta)$. Then, by construction,

$$X = \sum_{i=1}^{n} X_i.$$

and $\mathrm{E}_{f_X}[X_i] = \theta = \mu$, $\mathrm{Var}_{f_X}[X_i] = \theta(1-\theta) = \sigma^2$. Hence, for large n,

$$\mathrm{P}[X \leq x] = \mathrm{P}\left[\sum_{i=1}^{n} X_i \leq x\right] = \mathrm{P}\left[\frac{\displaystyle\sum_{i=1}^{n} X_i - n\theta}{\sqrt{n\theta(1-\theta)}} \leq \frac{x - n\theta}{\sqrt{n\theta(1-\theta)}}\right] \approx \Phi\left(\frac{x - n\theta}{\sqrt{n\theta(1-\theta)}}\right).$$

EXAMPLE : Suppose that $X \sim Poisson(\lambda)$. To calculate $\mathrm{P}[X \le x]$ for large n, first define random variables $X_1, ..., X_n$ to be i.i.d. $Poisson(\lambda/n)$. Then, by construction,

$$X = \sum_{i=1}^{n} X_i.$$

and $\mathrm{E}_{f_X}[X_i] = \lambda/n = \mu$, $\mathrm{Var}_{f_X}[X_i] = \lambda/n = \sigma^2$. Hence, for large n,

$$\mathrm{P}[X \le x] \quad = \mathrm{P}\left[\sum_{i=1}^{n} X_i \le x\right] \quad = \mathrm{P}\left[\frac{\sum_{i=1}^{n} X_i - n\lambda/n}{\sqrt{n\lambda/n}} \le \frac{x - n\lambda/n}{\sqrt{n\lambda/n}}\right] \approx \Phi\left(\frac{x-\lambda}{\sqrt{\lambda}}\right).$$

EXAMPLE : Suppose that $X \sim Gamma(n, \lambda)$ (for $n = 1, 2, ...$). To calculate $\mathrm{P}[X \le x]$ for large n, first define random variables $X_1, ..., X_n$ to be i.i.d. $Exponential(\lambda)$. Then, by construction,

$$X = \sum_{i=1}^{n} X_i.$$

and $\mathrm{E}_{f_X}[X_i] = 1/\lambda = \mu$, $\mathrm{Var}_{f_X}[X_i] = 1/\lambda^2 = \sigma^2$. Hence, for large n,

$$\mathrm{P}[X \le x] = \mathrm{P}\left[\sum_{i=1}^{n} X_i \le x\right] \quad = \mathrm{P}\left[\frac{\sum_{i=1}^{n} X_i - n/\lambda}{\sqrt{n/\lambda^2}} \le \frac{x - n/\lambda}{\sqrt{n/\lambda^2}}\right] \approx \Phi\left(\frac{x - n/\lambda}{\sqrt{n/\lambda^2}}\right).$$