Statistical and Numerical Methods for Bioinformatics: Exercises 1

Lecture examples

L1 Using the probability mass function, find the mean and variance of;
(i) The Binomial distribution
(ii) The Geometric distribution
(iii) The Negative Binomial distribution

L2 How would you find the following probabilities using R?
(i) the probability that an observation from a standard normal distribution is less than 0.5,
(ii) the probability that an observation from binomial (10,0.4) distribution is less than 4,
(iii) the probability that a negative binomial random variable (parameters 3 and 0.4) is between 2 and 4 inclusive.

L3 Let $X_1, X_2, \ldots, X_{20}$ denote a random sample of size 20 from the uniform distribution $U(0, 1)$, and let

$$Y = X_1 + X_2 + \ldots + X_{20}$$

Using a suitable approximation, find;
(i) $P(Y \leq 9.1)$
(ii) $P(8.5 \leq Y \leq 11.7)$

Questions by hand

E1 Find the following probabilities;
(i) $P[X = 3]$ if X is Geometric (0.4)
(ii) $P[X > 2]$ if X is Geometric (0.4)
(iii) $P[X < 2]$ if X is Exponential (2)
(Note that the Geometric (p) distribution is the same as the Negative Binomial (1,p)).

E2
(i) A fair 6-sided die is rolled until a 3 is observed. What is the expected number of rolls needed?
(ii) Two fair 6-sided dice are rolled. What is the expected number of rolls needed to observe at least one three on at least one of the two dice?
(iii) What is the smallest number of dice required so that the expected number of rolls needed to observe at least one 3 does not exceed 2?

E3 Let $\bar{X}$ be the mean of a random sample of size 36 from an exponential distribution with mean 3.

(i) Find the mean and variance of $\bar{X}$ and hence state an approximate distribution for $\bar{X}$.
(ii) Find an approximation for $P(2.5 \leq \bar{X} \leq 4)$.

Computer questions

R1 a)Find the probability that an observation from a standard normal distribution is greater than 1.
b)Find the probability that an observation from a Normal distribution with mean 1 and variance 4 lies between 2 and 4.

R2 Find the probability that a binomial (15,0.3) random variable takes values between 3 and 5 inclusive.

R3 A disease has post-operation complication frequency of 20%. A surgeon tests a new procedure on 10 patients with no complications. What would be the probability of operating on 10 patients successfully with the traditional method?

Statistical and Numerical Methods for Bioinformatics: Exercises 2

Lecture examples

L1 Suppose that scores on a standardised test in mathematics taken by students from large and small schools follow the distributions $N(\mu_X, \sigma^2)$ and $N(\mu_Y, \sigma^2)$ respectively, where σ^2 is unknown. Suppose that a random sample of 9 students from large schools gave $\bar{x} = 81.31, s_x^2 = 60.76$ and a random sample of 15 students from small schools gave $\bar{y} = 78.61, s_y^2 = 48.24$. Find a 95% confidence interval for the difference $\mu_X - \mu_Y$.

L2 A machine shop manufactures toggle levers. A lever is flawed if a standard nut cannot be screwed onto the threads. Let p equal the proportion of flawed toggle levers that they manufacture. If there were 24 flawed levers out of a sample of 642 that were selected randomly from the production line,
(i) give a point estimate of p.
(ii) Find an approximate 95% confidence interval for p.

L3 An experiment was conducted to measure people's reaction times to a red light versus a green light. When signaled with either the red or green light, the subject had to hit a switch to turn off the light. When the switch was hit, a clock was turned off and the reaction time in seconds recorded. The following are the reaction times for eight individuals.

Person	Red(X)	Green(Y)	$D = X - Y$
1	0.30	0.43	-0.13
2	0.23	0.32	-0.09
3	0.41	0.58	-0.17
4	0.53	0.46	0.07
5	0.24	0.27	-0.03
6	0.36	0.41	-0.05
7	0.38	0.38	0.00
8	0.51	0.61	-0.10

Find a 95% confidence interval for $\mu_X - \mu_Y$, the mean time of reaction to red light minus the mean time to reaction to green light.

Questions by hand

E1 Assume that the yield per acre of a particular variety of soybeans is $N(\mu, \sigma^2)$. For a random sample of $n = 5$ plots, the yields in bushels per acre were 37.4,48.8,46.9,55.0 and 44.0.
(i) Give a point estimate for μ.
(ii) Find a 90% confidence interval for μ.

E2 The length of life of brand X light bulbs is assumed to be $N(\mu_X, 784)$. The length of life of brand Y light bulbs is assumed to be $N(\mu_Y, 627)$ and independent of that of X. If a random sample of $n = 56$ brand X bulbs yielded a mean of $\bar{x} = 937.4$ hours and a random sample of size $m = 57$ brand Y bulbs yielded a mean of $\bar{y} = 988.9$ hours, find a 90% confidence interval for $\mu_X - \mu_Y$.

E3 A sweet manufacturer selects mints at random from the production line and weighs them. For one week, the day shift weighed $n_1 = 194$ mints, and the night shift weighed $n_2 = 162$ mints. The number of these mints which weighed at most 21 grams was $y_1 = 28$ for the day shift and $y_2 = 11$ for the night shift. Let p_1 and p_2 denote the proportions of mints that weigh at most 21 grams for the day and night shifts, respectively.

(i) Give a point estimate of p_1.
(ii) Give a 95% confidence interval for p_1.
(iii) Give a point estimate of $p_1 - p_2$.
(iv) Give a 95% confidence interval for $p_1 - p_2$.

Computer questions

Each of the questions uses the following data set (data set A).

Data set A - consumption rate (cm^3/hour) of nectar by honey-eating birds
X Sunset 0.9 1.6 1.4 1.2 1.6 1.1 0.8 1.0
Y Sunrise 0.8 1.1 1.2 1.3 1.1 1.0 0.7 0.8

R1. Find a confidence interval for the population mean of the data X, assuming that it is normally distributed with known variance 1.

R2. a) Find a 95% confidence interval for the population mean of data X, assuming it is normally distributed with unknown variance.
b) Similarly find a 95% confidence interval for Y.

R3. Find a 95% confidence interval for the population mean of X minus the population mean of Y, assuming that;
a) there is no relationship between the pairs of data
b) the values in a column can be considered a pair (since they are from the same bird).

Statistical and Numerical Methods for Bioinformatics: Exercises 3

Lecture examples

L1 Suppose that the breaking strength of a type of steel bar has the distribution $n(\mu, 36)$ and that we want to test the null hypothesis $H_0 : \mu = 50$ against the simple alternative hypothesis $H_0 : \mu = 55$. If, for a sample of size 16, the critical region is

$$\bar{X} \geq 53$$

find the significance level of the test α and the probability of a Type II error.

L2 Plot the p.d.f. of $\bar{X}$ from L1 under H_0 and H_1 on the same graph.

L3 A botanist measures the growths of pea stem segments, in millimetres, for $n = 11$ observations of type X

0.8, 1.8, 1.0, 0.1, 0.9, 1.7, 1.0, 1.4, 0.9, 1.2, 0.5

and $m = 13$ observations of type Y

1.0, 0.8, 1.6, 2.6, 1.3, 1.1, 2.4, 1.8, 2.5, 1.4, 1.9, 2.0, 1.2

Find the critical region for testing $H_0 : \mu_X - \mu_Y = 0$ against $H_1 : \mu_X - \mu_Y < 0$ at the 5% level and hence carry out this test.

Questions by hand

E1 Let X equal the forced vital capacity (FVC - the amount of air that a person can force out of their lungs) in litres for a female college student. Assume that the distribution of X is approximately $N(\mu, \sigma^2)$. Suppose it is known that $\mu = 3.4$ for the whole population of students. A volleyball coach claims that the FVC of volleyball players is greater than 3.4 and plans to test the claim using a random sample of size 9.

(i) Define the null hypothesis.
(ii) Define the appropriate alternative hypothesis.
(iii)Define the test statistic.
(iv) Define a critical region for $\alpha = 0.05$. Draw a figure illustrating this critical region.
(v) Calculate the test statistic given the following random sample of FVCs

3.4, 3.6, 3.8, 3.3, 3.4, 3.5, 3.7, 3.6, 3.7

(vi) What is your conclusion?
(vii) What is the approximate p-value of the test?

E2 It was claimed that 75% of dentists recommend a certain design of toothbrush. A consumer group doubted this claim and decided to test $H_0 : p = 0.75$ against $H_1 : p < 0.75$, where p is the proportion of dentists who recommend this design. A survey of 390 dentists found that 273 recommended the design.

(i) Which hypothesis would you accept at the 5% level?
(ii) Which hypothesis would you accept at the 1% level?
(iii) Find the p-value for this test.

E3 Let X and Y denote the weights in grams of male and female gallinules (a type of bird), respectively. Assume that X has distribution $N(\mu_X, \sigma_X^2)$ and Y has distribution $N(\mu_Y, \sigma_Y^2)$.

(i) Given that $n = 16$ observations of X and $m = 13$ observations of Y will be taken, define a test statistic and critical region for testing $H_0 : \mu_X = \mu_Y$ against the alternative hypothesis $H_1 : \mu_X > \mu_Y$ assuming that the variances are equal, and $\alpha = 0.01$.
(ii) If samples are taken so that $\bar{x} = 415.16, s_x^2 = 1356.75, \bar{y} = 347.40, s_y^2 = 692.21$, calculate the test statistic and state the result of the test.
(iii) Assuming that the variances are not equal, use the test proposed by Welch.

Computer questions

Each of these questions is based upon the following data (data set B).

Data set B - lengths of coleoptiles from irradiated and control barley seed.
X Control 28.5 29.0 19.7 24.5 26.1 19.3 23.0 28.9
Y Irradiated 20.8 16.5 17.4 17.3 16.3 16.9 19.8 16.6 16.5 17.2

R1. On the above data set, test the hypothesis that the population mean of X is equal to 23, on the assumption that the variance is unknown, against;
a) a 2-sided alternative
b) the alternative hypothesis that the mean is greater than 23.

R2. Test the hypothesis that the population mean of X and the population mean of Y are the same on the assumption that the variances are equal, against;
a) a 2-sided alternative
b) the alternative hypothesis that the difference in the means is greater than 0.

R3. a) Test the hypothesis that the population mean of X and the population mean of Y are the same without the assumption that the variances are equal (so using Welch's test), against;
a) 2-sided alternative
b) Comment on the solutions to 2 (a) and 3 (a).

Statistical and Numerical Methods for Bioinformatics: Exercises 4

Lecture examples

L1 The following numbers are assessment scores of students going to an American college. Construct an appropriate stem and leaf diagram of this data.

26 , 19 , 22 , 28 , 31 , 29 , 25 , 23 , 20 , 33 , 23 , 26
30 , 27 , 26 , 29 , 20 , 23 , 18 , 24 , 29 , 27 , 32 , 24
25 , 26 , 22 , 29 , 21 , 24 , 20 , 28 , 23 , 26 , 30 , 19
27 , 21 , 32 , 28 , 29 , 23 , 25 , 21 , 28 , 22 , 25 , 24
19 , 24 , 35 , 26 , 25 , 20 , 31 , 27 , 23 , 26 , 30 , 29

L2 At each of two sites in the Lewes Brooks area, fifty sampling units were chosen and the number of molluscs in each unit counted. The data were

Site A: 2, 1, 3, 0, 1, 0, 6, 2, 8, 0, 9, 10, 6, 0, 4, 1, 2, 3, 22, 3, 0, 1, 0, 0, 4,
6, 0, 6, 1, 3, 8, 0, 4, 4, 0, 1, 2, 5, 1, 4, 1, 0, 2, 3, 10, 5, 8, 0, 0, 6,
Site B: 7, 22, 0, 4, 4, 2, 0, 4, 1, 4, 0, 0, 6, 1, 0, 2, 1, 4, 6, 3, 0, 3, 2, 6, 3,
5, 2, 0, 0, 1, 5, 1, 0, 5, 3, 0, 1, 0, 15, 4, 3, 0, 4, 4, 1, 2, 0, 1, 0, 0

(i) Construct histograms for each site.
(ii) Find the five number summary and draw back to back boxplots of the data.
(iii) Is it reasonable to consider the two sites as essentially the same, or are there differences?

L3 The lengths of ten goldfish in centimetres are
5.0, 3.9, 5.2, 5.5, 2.8, 6.1, 6.4, 2.6, 1.7, 4.3

(i) Use the Wilcoxon test to test the hypothesis $H_0 : m = 3.7$ against the alternative $H_1 : m > 3.7$, where m is the median of the distribution of fish sizes.
(ii) What result would you obtain if you used the sign test instead to carry out this test?

Questions by hand

E1 In the casino game roulette, if a player bets one unit on red, the probability of winning is 18/38 and of losing is 20/38 (in American casinos - European ones are more generous!). Suppose that a player begins with five units and let Y be a player's maximum capital, before eventually losing their money. the following data are 100 simulations of this value of Y.

25, 9, 5, 5, 5, 9, 6, 5, 15, 45, 55, 6, 5, 6, 24, 21, 16, 5, 8, 7, 7, 5, 5, 35, 13,
9, 5, 18, 6, 10, 19, 16, 21, 8, 13, 5, 9, 10, 10, 6, 23, 8, 5, 10, 15, 7, 5, 5, 24, 9,
11, 34, 12, 11, 17, 11, 16, 5, 15, 5, 12, 6, 5, 5, 7, 6, 17, 20, 7, 8, 8, 6, 10, 11, 6,
7, 5, 12, 11, 18, 6, 21, 6, 5, 24, 7, 16, 21, 23, 15, 11, 8, 6, 8, 14, 11, 6, 9, 6, 10

(i) Construct an ordered stem and leaf plot.
(ii) Find the five-number summary and draw a boxplot.
(iii) Draw a histogram of the data.

E2 The ages at which the adolescent growth spurt began in a sample of 35 boys and 40 girls who transferred to secondary school were;

Boys: 16.0, 14.9, 14.1, 14.8, 14.4, 14.0, 13.6, 14.6, 16.1, 13.2, 13.2,
14.6, 15.3, 14.4, 14.8, 15.9, 14.7, 14.5, 14.6, 13.5, 15.1, 13.5, 15.0,
15.2, 15.4, 15.9, 13.7, 14.9, 14.1, 15.4, 14.4, 13.8, 15.3, 14.7, 14.8

Girls: 12.2, 13.7, 13.3, 12.3, 12.5, 12.9, 14.1, 11.8, 12.8, 12.9, 11.6, 14.3,
12.3, 11.6, 13.1, 12.6, 11.7, 13.5, 11.9, 11.6, 13.4, 12.4, 12.6, 13.7, 12.1, 13.5,
12.5, 13.4, 13.1, 13.3, 13.5, 14.7, 12.7, 12.7, 12.0, 11.4, 13.5, 12.4, 12.1, 12.1

Make a comparison between boys and girls of the age of onset of the growth spurt using histograms and boxplots, comparing shape, spread and location.

E3 Let X denote the weight in grams of a type of sweet, where m is the median of X. We wish to test $H_0 : m = 5.900$ against $H_1 : m > 5.900$. A random sample of size 25 yielded the following data (in order of increasing size).

5.625, 5.665, 5.697, 5.838, 5.863
5.870, 5.878, 5.884, 5.908, 5.967
6.019, 6.020, 6.029, 6.032, 6.037
6.045, 6.049, 6.050, 6.079, 6.116
6.159, 6.186, 6.199, 6.307, 6.387

(i) Perform the test using the sign test.
(ii) Now perform the test using the Wilcoxon test.
(iii) Thirdly test the same hypotheses using a t-test.
(iv) Write a short comparison of your results.

Computer questions

R1. For the variable X dataset A from sheet 2, find;
the five number summary and hence draw a boxplot of the data,
draw a histogram of the data, and
using a normal Q-Q plot, say whether you think the data may be normally distributed.

R2. Draw back to back boxplots and histograms and comment for;
a) data set A from sheet 2
b) data set B from sheet 3

R3. Perform a Wilcoxon test on each of the data sets X and Y from data set B on sheet 3, to test whether the population mean is 23 in each case, and comment.

Statistical and Numerical Methods for Bioinformatics: Exercises 5

Lecture examples

L1 The following data are 10 pairs of scores in a psychology class, each pair being the score of a preliminary test (X) and the score on the final exam (Y) of a particular student. Fit a regression line to the data, obtaining estimates of all relevant parameters.

X	70	74	72	68	58	54	82	64	80	61
Y	77	94	88	80	71	76	88	80	90	69

L2 The grade point averages of 20 American students at high school and college are written below as pairs (X, Y).

(3.75,3.19) (3.45,3.34) (2.87,2.23) (3.60,3.46) (3.42,2.97)
(4.00,3.79) (2.65,2.55) (3.10,2.50) (3.47,3.15) (2.60,2.26)
(4.00,3.76) (2.30,2.11) (2.47,2.11) (3.36,3.01) (3.60,2.92)
(3.65,3.09) (3.30,3.05) (2.58,2.63) (3.80,3.22) (3.79,3.27)

(i) Find the correlation coefficient of the data.
(ii) Find the regression line.
(iii) Plot the regression line and the data on the same graph.

L3 In the manufacture of commercial wood products it is important to estimate the relationship between the density of a wood product and its stiffness. The following data show the density (X) and stiffness (Y) for 30 particleboards, given in pairs (X, Y).

(9.5,14814) (8.4,17502) (9.8,14007) (11.0,19443) (8.3,7573) (9.9,14191)
(8.6,9714) (6.4,8076) (7.0,5304) (8.2,10728) (17.4,43243) (15.0,25319)
(15.2,28028) (16.4,41792) (16.7,49499) (15.4,25312) (15.0,26222) (14.5,22148)
(14.8,26751) (13.6,18036) (25.6,96305) (23.4,104170) (24.4,72594) (23.3,49512)
(19.5,32207) (21.2,48218) (22.8,70453) (21.7,47661) (19.8,38138) (21.3,53045)

(i) Plot the data.
(ii) Fit a linear regression line to the data, and add the line to the plot.
(iii) Transform the data appropriately, and repeat parts (i) and (ii).

Questions by hand

E1 Let X and Y equal lengths in inches of a foot and a hand, respectively. The following measurements were made on 15 women

X	9.00	8.50	9.25	9.75	9.00	10.00	9.50	9.00
Y	6.50	6.25	7.25	7.00	6.75	7.00	6.50	7.00

X	9.25	9.50	9.25	10.00	10.00	9.75	9.50
Y	7.00	7.00	7.00	7.50	7.25	7.25	7.25

(i) Calculate the least squares regression line for these data, and find an estimate of the variance of the error term.
(ii) Plot the points and the regression line on a graph.

E2 Chemists often use Ion Sensitive Electrodes to measure the ion concentration of aqueous solutions. These measure the migration of charge in millivolts (mV). The following data gives pairs of values, concentration (in particles per million, ppm) versus the reading in mV.

ppm	0	0	0	50	50	50	75	75	75
mV	1.72	1.68	1.74	2.04	2.11	2.17	2.40	2.32	2.33

ppm	100	100	100	150	150	150	200	200	200
mV	2.91	3.00	2.89	4.47	4.57	4.43	6.67	6.66	6.57

(i) Find the correlation coefficient of the data.
(ii) Find the least squares regression line.
(iii) Plot the points and the regression line on the same graph and comment.

E3 The following data represent the height in centimetres (X) and weight (Y) in grams of a type of plant. A sample of ten plants was taken.

X	4.7	6.2	6.4	6.9	7.6	7.8	8.1	8.7	9.2	10.4
Y	2.2	4.6	5.0	6.8	9.2	9.2	10.9	13.6	15.9	22.1

(i) Plot the data.
(ii) Fit a linear regression line to the data, and add the line to the plot.
(iii) Transform the data appropriately, and repeat parts (i) and (ii).

Computer questions

R1. Using the following data (data set C) plot the data and perform a linear regression of Y on X, adding the regression line to your plot.

Data set C - Mercuralis plants
X the fresh weight (g) versus Y height (cm) multiplied by stem diameter (cm)
X 0.42 0.64 0.81 1.23 1.32 1.80 2.24 2.39 2.56 3.13
Y 0.21 0.22 0.23 0.40 0.53 0.56 1.21 1.38 0.86 1.19
X 3.05 3.28 3.72 3.86 3.94 5.76 5.82 7.05 8.01 8.67
Y 1.77 1.78 1.85 1.73 2.42 3.17 3.80 3.50 4.65 5.13

R2. a) Using the following data (data set D) plot the data and perform a linear regression of Y on X, adding the regression line to your plot.
b) Draw a Q-Q plot of the residuals of the fit and comment.
c) Transform the data by taking logarithms of Y, and repeat the above.

Data set D

X	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
Y	10.52	15.70	24.39	43.18	68.09	119.12	154.60	199.02	488.41

R3. Find the correlation coefficient for each of the data sets C and D in questions 1 and 2 and comment.

Statistical and Numerical Methods for Bioinformatics: Exercises 6

Lecture examples

L1 Four groups of three pigs each were fed individually four different feeds for a specified length of time to test the hypothesis $H_0 : \mu_i = \mu_2 = \mu_3 = \mu_4$, where μ_i is the mean weight gain for each of the feeds $i = 1, 2, 3, 4$. Determine whether the null hypothesis is accepted or rejected at the level $\alpha = 0.05$ when the observed weight gains were

$X_1 : 194.11, 182.80, 187.43$
$X_2 : 216.06, 203.50, 216.88$
$X_3 : 178.10, 189.20, 181.33$
$X_4 : 197.11, 202.68, 209.18$

L2 repeat the previous question using calculations from R.

L3 A window manufactured for a car has five studs attached to it. A company that manufactures these windows performs "pull-out" tests to determine the force needed to pull a stud out of the window. Let X_i equal the force required at position i $(i = 1, \ldots, 5)$, and assume that X_i is $N(\mu_i, \sigma^2)$. Test the null hypothesis

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

using seven independent observations at each position with $\alpha = 0.01$. The data values are as follows:

$X_1 : 92, 90, 87, 105, 86, 83, 102$
$X_2 : 100, 108, 98, 110, 114, 97, 94$
$X_3 : 143, 149, 138, 136, 139, 120, 145$
$X_4 : 147, 144, 160, 149, 152, 131, 134$
$X_5 : 142, 155, 119, 134, 133, 146, 152$

Questions by hand

E1 let μ_i be the average yield in bushels per acre of a variety i of corn, $i = 1, 2, 3, 4$. In order to test at the 5% level, the null hypothesis that $H_0 : \mu_i = \mu_2 = \mu_3 = \mu_4$, four test plots for each of the varieties of corn were planted. Given that the yields are those given below, do we accept or reject H_0?

$X_1 : 68.82, 76.99, 74.30, 78.73$
$X_2 : 86.84, 75.69, 77.87, 76.18$
$X_3 : 90.16, 78.84, 80.65, 83.58$
$X_4 : 61.58, 73.51, 74.57, 70.75$

E2 The driver of a diesel-powered car decided to test the quality of three types of diesel fuel, based upon miles per gallon. Test the null hypothesis that the three means are equal using the data below, using the significance level $\alpha = 0.05$ and making the usual assumptions.

Brand A: $38.7, 39.2, 40.1, 38.9$
Brand B: $41.9, 42.3, 41.3$
Brand C: $40.8, 41.2, 39.5, 38.9, 40.3$

E3 If you concluded for either of questions E1 or E2 that there was a difference, explore where the difference lies using pairwise comparisons and explain your conclusion.

Computer questions

Data set E - individual test scores. Nine volunteers were divided into three groups and asked to perform the test associated with that group. The scores were
Test 1 2, 7, 3 Test 2 7, 9, 5 Test 3 9, 12, 9

R1. a) Perform an analysis of variance on the above data (data set E).
b) Do you think that there is any difference between the groups?
c) If so, without performing any tests, which groups do you think differ?

Data set F - box sizes
Different sizes of nails are packaged in one-pound boxes. Let X_i be the weight of a box with nail size $4iC$, $i = 1, 2, 3, 4, 5$, where $4C, 8C, 12C, 16C, 20C$ are the sizes of the sinkers from smallest to largest.

$X_1 : 1.03, 1.04, 1.07, 1.03, 1.08, 1.06, 1.07$
$X_2 : 1.03, 1.10, 1.08, 1.05, 1.06, 1.06, 1.05$
$X_3 : 1.03, 1.08, 1.06, 1.02, 1.04, 1.04, 1.07$
$X_4 : 1.10, 1.10, 1.09, 1.09, 1.06, 1.05, 1.08$
$X_5 : 1.04, 1.06, 1.07, 1.06, 1.05, 1.07, 1.05$

R2. a) Perform an analysis of variance on the above data (data set F).
b) Do you think that there is any difference between the two groups?
c) If so, without performing any tests, which groups do you think differ?

R3. If you concluded that there was a difference between the groups in either of the two questions above, use pairwise comparisons to show whether these difference are significant;
a) without the Bonferroni correction,
b) with the Bonferroni correction.
c) Comment on your answers.

Statistical and Numerical Methods for Bioinformatics: Exercises 7

Lecture examples

L1 Let X denote the number of heads that occur when four coins are tossed at random. Under the assumption that the four coins are independent and the probability of heads for each coin is 0.5, X is Binomial (4,0.5). One hundred repetitions of the experiment resulted in the following numbers for each of the number of heads (in brackets).

7 (0), 18 (1), 40 (2), 31 (3), 4 (4)

Do these results support the above assumptions?

L2 We wish to see if two groups of nurses distribute their time in six different categories about the same way. That is, we wish to test $H_0 : p_{i1} = p_{i2}$ for $i = 1, \ldots, 6$. To carry out this test, nurses were observed at random throughout several days, each observation resulting in a mark in one of the six categories. The data are summarised below.

Category	1	2	3	4	5	6	Total
Group I	95	36	71	21	45	32	300
Group II	53	26	43	18	32	28	200

Perform the test and state your conclusions.

L3 Repeat the above analysis using R.

Questions by hand

E1 In the Michigan Daily Lottery, each weekday a three-digit integer is generated one digit at a time. Let p_i denote the probability of generating digit $i, i = 0, 1, \ldots, 9$. Use the following 50 digits to test $H_0 : p_0 = p_1 = \ldots = p_9 = 0.1$, using $\alpha = 0.05$.

1, 6, 9, 9, 3, 8, 5, 0, 6, 7, 4, 7, 5, 9, 4, 6, 5, 6, 4, 4, 4, 8, 0, 9, 3,
2, 1, 5, 4, 5, 7, 3, 2, 1, 4, 6, 7, 1, 3, 4, 4, 8, 8, 6, 1, 6, 1, 2, 8, 8.

E2 A random sample of 50 women who were tested for cholesterol were classified according to age and cholesterol level and grouped into the following table.

Age/ Chol.	<180	180-210	>210	Total
<50	5	11	9	25
$\geq$ 50	4	3	18	25
Total	9	14	27	50

Test the null hypothesis that age and cholesterol are independent at the 1% and 5% levels.

E3 For the above cholesterol data, where does any difference in the two groups principally lie? Show some numerical explanation for your answer.

Computer questions

R1. For the following data (data set G) test the hypothesis that the data are observations from a binomial distribution with parameters 5 and 0.3.
Hint: you should find the binomial probabilities and then use the chisq.test command in the "stats" library.

Data set G - count data. There are 113 pieces of data in all, with 32 zeros, 34 ones, 21 twos, 15 threes, 7 fours and 4 fives.

R2. a) For the contingency table below (data set H), test whether there is any association between the rows and columns.
b) If there is an association, where do you think it lies?

Data set H - roommates. It is claimed that the roommate of a student at an American university will have a great influence on their grades. To test this 200 random students were selected and classified according to two attributes;
a) the ranking of their roommate from 1 to 5 from difficult to live with/discouraged scholarship to congenial/encouraged scholarship,
b) the student's first year grade point average (GPA).

The data are summarised in the table below

Rank/ GPA	<2.00	2.00-2.69	2.70-3.19	3.20-4.00	Total
1	8	9	10	4	31
2	5	11	15	11	42
3	6	7	20	14	47
4	3	5	22	23	53
5	1	3	11	12	27
Total	23	35	78	64	200

R3. For the following contingency table (data set I), test whether there is any difference between the groups. Do you think that they are independent? Explain your reasoning.

Data set I - male/female breakdown at Aberdeen University (BSc students, 1985)

	Men	Women	Total
Agriculture	107	48	155
Engineering	304	30	334
Forestry	93	7	100
Pure Science	636	635	1271
Total	1140	720	1860

Statistical and Numerical Methods for Bioinformatics: Exercises 8

Lecture examples

L1 Prove that the variance of the estimator a of the intercept term of the simple linear regression α is given by

$$Var[a] = \frac{\sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \sigma^2$$

L2 By the method of least squares, perform a multiple regression

$$Y = \beta_1 x_1 + \beta_2 x_2 + \beta_3$$

using the following 12 data triples X_1, X_2, Y.
(Hint: consider β_3 to be associated with a factor x_3 which always takes the value 1).

(0,0,3) (1,0,7) (2,0,8) (3,0,9) (0,1,6) (1,1,7)
(2,1,10) (3,1,11) (0,2,7) (1,2,6) (2,2,11) (3,2,11)
(0,3,6) (1,3,9) (2,3,12) (3,3,14)

L3 Repeat the above analysis using R and explain the output.

Questions by hand

E1 For the data from question $E1$ from Exercise sheet 5, find a confidence interval for;
(i) the slope of the regression line,
(ii) the mean length of the hand of a woman with a foot of length 9.00 inches.

E2 By the method of least squares, perform a multiple regression

$$Y = \beta_1 x_1 + \beta_2 x_2 + \beta_3$$

using the following 12 data triples X_1, X_2, Y.
(Hint: consider β_3 to be associated with a factor x_3 which always takes the value 1).

(1,1,6) (0,2,3) (3,0,10) (-2,0,-4) (-1,2,0) (0,0,1)
(2,1,8) (-1,-1,-2) (0,-3,-3) (2,1,5) (1,1,1) (-1,0,-2)

You are given that the summary statistics are $\sum x_1^2 = 26, \sum x_1 x_2 = 5, \sum x_1 = 4$,
$\sum x_2^2 = 22, \sum x_2 = 4, n = 12, \sum x_1 y = 75, \sum x_2 y = 37, \sum y = 23$

E3 For the data from $E3$ from exercise sheet 5, explain how you would carry out a polynomial regression of weight Y on height X. For a quadratic model (i.e. including terms up to X^2), find the appropriate summary statistics, and form the equations (but do not solve them).

Computer questions

R1. From the data set C from sheet 5, if the value of X is 7;
a) obtain a 95% confidence interval for the mean of Y at this value,
b) obtain a 95% prediction interval for a randomly chosen value of Y at this value

R2. For the data below (data set J), perform a polynomial regression of Y on X. By considering fits up to different powers of X, find the best fit to the data.

Data set J -
X 10.00 12.00 14.00 16.00 18.00 20.00 22.00 24.00 26.00 28.00
Y 87.3 128.4 156.7 218.3 287.6 384.2 413.5 548.2 602.5 727.9

R3. For the data from data set K perform a multiple regression, finding estimates of the parameters. By considering different models with some of the variables removed, explore which is the best model for the data.

Data set K - this is the Freeny Revenue Data file labelled "freeny" in R library "stats".

Statistical and Numerical Methods for Bioinformatics: Exercises 9

Lecture examples

L1 For the following data triples (X_1, X_2, X_3), find the sample variances, covariances and correlations, and hence the correlation matrix of the data.
What are the adjusted values of the data, represented by (U_1, U_2, U_3)?

(7,4,3), (2,3,3), (6,1,2), (4,5,6),
(9,6,5), (4,2,2), (3,2,4), (8,3,3)

L2 Find the eigenvectors and eigenvalues of the following matrices

$$\begin{vmatrix} 1 & 0.8 \\ 0.8 & 1 \end{vmatrix} \tag{1}$$

$$\begin{vmatrix} 1 & -0.2 \\ -0.2 & 1 \end{vmatrix} \tag{2}$$

L3 (i) For the following correlation matrix, use R to find all of the principal components.
(ii) Using the Kaiser criterion, which components should be included in a sensible reduced set of components?

$$\begin{vmatrix} 1 & 0.8 & 0.3 & 0.3 & -0.2 \\ 0.8 & 1 & 0.1 & 0.4 & 0.1 \\ 0.3 & 0.1 & 1 & 0.9 & 0.5 \\ 0.3 & 0.4 & 0.9 & 1 & 0.4 \\ -0.2 & 0.1 & 0.5 & 0.4 & 1 \end{vmatrix} \tag{3}$$

Questions by hand

E1 Find the eigenvalues and eigenvectors of the following matrices.

$$\begin{vmatrix} 1 & 0.727 \\ 0.727 & 1 \end{vmatrix} \tag{4}$$

$$\begin{vmatrix} 1 & -0.924 \\ -0.924 & 1 \end{vmatrix} \tag{5}$$

E2 Consider the pairs of data from Exercises 5, $E1$, the hand and foot data. relabel the data (X_1, X_2), where X_1 are the hand measurements, and X_2 the foot measurements.
(i) Find the sample variance for X_1 and X_2 and the sample correlation between X_1 and X_2
(ii) find the largest eigenvalue from the data and its associated eigenvector, and hence the principal component.

E3 For each of the cases L3 and R1, plot the size of the eigenvalue against the ordering of its relative size (so the third largest eigenvalue is plotted against 3). Determine where the curve flattens out into a straight line in each case; this is the number of principal components that we should take. [This graphical procedure is known as the scree test].

Computer questions

R1. For the following correlation matrix, find the principal components and use the Kaiser criterion to decide which components should be included.

$$\begin{vmatrix} 1 & 0.714 & 0.454 & 0.673 & 0.215 \\ 0.714 & 1 & 0.277 & 0.878 & 0.310 \\ 0.454 & 0.277 & 1 & 0.664 & 0.501 \\ 0.673 & 0.878 & 0.664 & 1 & 0.226 \\ 0.215 & 0.310 & 0.501 & 0.226 & 1 \end{vmatrix} \tag{6}$$

R2. For the following data triples, find the correlation matrix, and hence find the principal components.

Data set M -
(6.71,3.12,4.15), (2.34,3.45,2.87), (6.12,1.45,2.31), (4.46,5.04,5.99),
(9.67,6.22,4.79),(3.78,2.20,2.39),(3.07,2.73,4.02),(7.86,3.51,3.18)

R3. For the following data set (Data set N), find the principal components and their eigenvalues, and hence decide how many components are required.

Data set N - this is the "sim10var" data file in R library "pcurve".

Statistical and Numerical Methods for Bioinformatics: Exercises 10

Lecture examples

L1 Let $X_1, \ldots, X_n$ be a random sample from the exponential distribution with p.d.f

$$f(x;\theta) = \frac{1}{\theta} exp\left(-\frac{x}{\theta}\right) \qquad 0 < x < \infty$$

where $\theta \in \Omega = \{\theta : 0 < \theta < \infty\}$. Find the maximum likelihood estimator of θ.

L2 Let $X_1, \ldots, X_n$ be a random sample of size n from the distribution with p.d.f.

$$f(x;\theta) = \theta x^{\theta-1} \qquad 0 < x < 1, 0 < \theta < \infty$$

(i) Sketch the graph of this p.d.f. for $\theta = 0.25, 1, 4$.
(ii) Use the method of moments to find an estimate of θ.

L3 Assume that the weight X in ounces of "10 pound" bag of sugar is $N(\mu, 5)$. Test the hypothesis $H_0 : \mu = 162$ against $H_1 : \mu \neq 162$ using the likelihood ratio test, when $n = 20$ and $\bar{x} = 161.1$.

(i) Do we accept H_0 if $\alpha = 0.10$?
(ii) Do we accept H_0 if $\alpha = 0.05$?
(iii) What is the p-value of the test?

Questions by hand

E1 Let $X_1, \ldots, X_n$ be a random sample of size n from the geometric distribution with success parameter p, i.e.

$$f(x) = p(1-p)^{x-1} \qquad x = 1, 2, 3, \ldots$$

(i) Use the method of moments to find a point estimate of p.
(ii) Explain intuitively why this estimate makes sense.
(iii) Find a point estimate of p, given the following data
3, 34, 7, 4, 19, 2, 1, 19, 43, 2, 22, 4, 19, 11, 7, 1, 2, 21, 15, 16

E2 Let $X_1, \ldots, X_n$ be a random sample of size n from the distribution with p.d.f.

$$f(x;\theta) = \theta x^{\theta-1} \quad 0 < x < 1, 0 < \theta < \infty$$

(i) Sketch the graph of this p.d.f. for $\theta = 0.5, 1, 2$.
(ii) Show that the maximum likelihood estimator of θ is given by

$$\hat{\theta} = -\frac{n}{ln(\prod_{i=1}^{n} X_i)}$$

(iii) For each of the following three sets of observations from this distribution, calculate the values of the maximum likelihood estimate and the methods of moments estimate for θ.
a) 0.0256, 0.3051, 0.0278, 0.8971, 0.0739, 0.3191, 0.7379, 0.3671, 0.9763, 0.0102
b) 0.9960, 0.3125, 0.4374, 0.7464, 0.8278, 0.9518, 0.9924, 0.7112, 0.2228, 0.8609
c) 0.4698, 0.3675, 0.5991, 0.9513, 0.6049, 0.9917, 0.1551, 0.0710, 0.2110, 0.2154

E3 Let $X_1, X_2, \ldots, X_n$ be a random sample from an exponential distribution with mean θ. Show that the likelihood ratio test of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ has a critical region of the form

$$\sum_{i=1}^{n} x_i \leq c_1 or \sum_{i=1}^{n} x_i \geq c_2$$

How would you modify this to allow the use of χ^2 tables?

Statistical and Numerical Methods for Bioinformatics: Exercises 11

E1 Cars pass a certain tree on a quiet country road at a rate of one every two minutes.
(i) What is the probability that exactly five cars pass the tree in ten minutes?
(ii) What is the probability that exactly five cars pass the tree in ten minutes given that in the first five of these minutes one car passes the tree?
(iii) What is the probability that all of the periods 10.00-10.02, 10.02-10.04, 10.04-10.06, 10.06-10.08, 10.08-10.10 contain at least one event of a car passing the tree?
(iv) Would a Poisson process be a good model for a busy road?

E2 Ten unemployed former students join a job club. As soon as one gets a job, they leave. Suppose that each is equally likely to get a job and receives offers (which are always accepted) at rate 3 per year.
a) What is the probability that all the students have got a job after 2 years ?
b) What is the probability that the last student to get a job gets it in the third year ?
c) How long must we wait before the probability that all the students have accepted a job is greater than 0.5 ?

E3 Consider a simple birth and death process with birth rate 2 and death rate 1, starting with two individuals. What is the probability that;
(i) the population becomes extinct ?
(ii) the population reaches 5 before it becomes extinct ?
(iii) the population becomes extinct, given that it reaches 5 ?
(iv) the population reaches 5, given that it becomes extinct ?
(v) the population is extinct at time 2 ?
(vi) What is the expected number of offspring for a given individual ?

R1 If the number of messages arriving at a telephone exchange occurs according to a Poisson process of rate 3 per minute, find the probability that
(i) no messages arrive in a minute
(ii) greater than two messages arrive in a minute
(iii) the 3rd message arrives in the second minute.

R2 A population starting with 16 individuals following a death process of rate 3 per year has been running for six months. Find the probability that:
(i) the population size is 3.
(ii) the population size is no more than 3.

R3 A population following a birth process of rate 0.5 per day starts with 3 individuals. Find the probability that
(i) there are still 3 individuals after one day (i.e. there have been no births)
(ii) there are more than 5 individuals after one day
(iii) there are between 100 and 200 individuals after ten days.

Statistical and Numerical Methods for Bioinformatics: Exercises 12

E1 Find the stationary distribution of the Markov chain with transition matrix

$$\begin{vmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{vmatrix} \tag{7}$$

E2 In the Markov chain defined in E1, find the probability that the occupied state at time 3 is E_1, given that the initial distribution (at time 0) is (0.8, 0.2).

E3 Consider the random walk described by the following matrix (similar to that in the notes, except that when either state 0 or n is reached, the next step is forced to be 1 from 0, or $n-1$ from n.

	0	1	2	3	...	$n-3$	$n-2$	$n-1$	n
0	0	1	0	0	...	0	0	0	0
1	q	0	p	0	...	0	0	0	0
2	0	q	0	p	...	0	0	0	0
...	...	...	...	...	...	...	...	...	...
$n-2$	0	0	0	0	...	q	0	p	0
$n-1$	0	0	0	0	...	0	q	0	p
n	0	0	0	0	...	0	0	1	0

(8)

a) Show that this Markov chain is periodic (you only need to show that at least one state can only be reached at given regularly spaced times).
b) Construct equations for the stationary distribution ϕ and solve them for the case where $p = 0.5$

R1 Simulate a continuous time Markov chain with the following transition ($\mathbf{Q}$) matrix, up to time 30, obtaining a series of states and the times of each transition. Repeat the process twice and compare the results.

$$\begin{vmatrix} -0.2 & 0.1 & 0.1 \\ 0.5 & -0.6 & 0.1 \\ 0.1 & 0.2 & -0.3 \end{vmatrix} \tag{9}$$

R2 A continuous time Markov chain has all possible transitions between the states $i, j = 1, 2, 3$ $i \neq j$. Observations were made at the following times

Times - 0, 0.1, 0.3, 0.7, 0.8, 0.9, 1.2, 1.4, 1.6, 1.7, 2.1, 2.2

The following sequence gives the observed states

States - 1, 1, 1, 2, 2, 1, 1, 1, 3, 3, 3, 1

Obtain estimates of the transition rates between states and the mean time spent in any given state in one visit.

R3 Consider the aneurism data in the data file *aneur*. 838 male patients over 65 were followed, with transitions occurring in sequence between grades of aortic aneurisms Aneurism free, Mild aneurism, Medium aneurism, Severe aneurism.

Obtain estimates of the transition rates between the states.

General Repeat the simulation from R1. Using the data that you obtain estimate the transition PROBABILITIES between the states. By comparing the real transition rates to the formulae in the notes, find the true transition probabilities, and comment.

Statistical and Numerical Methods for Bioinformatics: Exercises 13

E1 Suppose that we wish to cover a sequence of length 10000 bases with N pieces each of length 500 bases, randomly chosen.
(i) What is the expected proportion of coverage for the sequence if N=50?
(ii) What value would N have to be to give 99% coverage?

E2 A sequence contains 100000 bases, and the frequency of a in the sequence is 0.3. What is the probability that the maximum number of a's in a row in the whole sequence is at least 15?

E3 Using the simple scoring system
number of matches-number of mismatches-number of gaps
find the best alignment of $x = cttgac$ in $y = cagtatcgtac$
(i) where gaps are not allowed (simply state the highest score obtained),
(ii) where gaps are allowed (it will take a long time to do this exhaustively; try to find good alignments by eye).

R1 Investigate the data *woodmouse* which is a set of 15 sequences of woodmouse DNA data, using the R help system. Describe the data in more detail.

R2 Find the base frequencies for each of a, g, c and t for the woodmouse data.
Use this to test the hypothesis of equal frequencies within the data (Hint: if you can find the number of observations from the data file you can use the frequencies given to calculate the total number of each base observed).
What conclusions do you reach?
Are there any problems with this testing idea? If so, how would you modify your test?

R3 Generate the PAM1 matrix in R, and hence find the PAM6 matrix using matrix multiplication.

General What is the probability that a given amino acid A does not occur at a given location at each of 6 successive PAM time units if all transitions between amino acids are equally likely? How would you work this out if they were not?

Statistical and Numerical Methods for Bioinformatics: Exercises 14

E1 Define a Hidden Markov model Λ with the following parameters:
Three states S_1, S_2, S_3, alphabet $A = \{, 1, 2, 3\}$, transition matrix P given by

$$\begin{vmatrix} 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{vmatrix} \tag{10}$$

$\pi = (1, 0, 0)$
$b_1(1) = 0.5, b_1(2) = 0.5, b_1(3) = 0$
$b_2(1) = 0.5, b_2(2) = 0, b_2(3) = 0.5$
$b_3(1) = 0, b_3(2) = 0.5, b_3(3) = 0.5$

What are all possible state sequences for the observed sequences O, and what is $P(O|\Lambda)$, for the following sequences?
(i) $O = 1, 2, 3$
(ii) $O = 1, 3, 1$

E2 A Hidden Markov model is defined as follows:
There are two states S_1, S_2, alphabet $A = \{1, 2\}$, transition matrix P given by

$$\begin{vmatrix} 0 & 1 \\ 0.5 & 0.5 \end{vmatrix} \tag{11}$$

$\pi = (0.6, 0.4)$
$b_1(1) = 0.5, b_1(2) = 0.5$
$b_2(1) = 1, b_2(2) = 0$

If the observations $1, 1$ occur, what is the most plausible underlying sequence of states?

E3 Consider the five amino acid sequences $WRCCTGC$, $WCCGGCC$, $WCGCC$, $WCCCGCC$, $WCCGC$. Suppose that their respective paths through a protein model HMM of length 8 are

$$\begin{array}{cccccccccc} m_0 & m_1 & i_1 & m_2 & m_3 & m_4 & m_5 & d_6 & m_7 & m_8 \\ m_0 & m_1 & m_2 & m_3 & m_4 & m_5 & m_6 & m_7 & m_8 & \\ m_0 & m_1 & m_2 & d_3 & d_4 & m_5 & m_6 & m_7 & m_8 & \\ m_0 & m_1 & m_2 & m_3 & m_4 & m_5 & m_6 & m_7 & m_8 & \\ m_0 & m_1 & m_2 & m_3 & d_4 & m_5 & d_6 & m_7 & m_8 & \end{array} \tag{12}$$

Give the alignment of the sequences that these paths determine.

R1 Generate 100 random observations from a gamma distribution with parameters 2 and 1, and then perform a one sample bootstrap on the data set, taking 1000 replicates. Find a 95% confidence interval for the true mean and provide a histogram of the sample means.

R2 Repeat R1 using the exponential distribution with parameter 100.

R3 Generate two data samples using random numbers with Normal distributions with
a) mean 1, variance 1
b) mean 0, variance 1
Carry out a 2-sample bootstrap for the difference of the means of the distributions, giving a 95% confidence interval for this difference and plot a histogram of the sample mean differences.

General Generate a new random sample as in R3 a). Plot the empirical distribution function for this sample and the true distribution function for the underlying distribution. Repeat this for the distribution in R2. Comment.

Statistical and Numerical Methods for Bioinformatics: Exercises 15

E1 Using the Metropolis Hastings algorithm, find a 3-state Markov chain with a stationary distribution (0.2,0.5,0.3).

E2 Suppose that we have a two dimensional vector $\mathbf{Y}$, where each element can take values 1 or 0 only, so there are four possible vectors only
(0,0), (0,1), (1,0) and (1,1).
Suppose that these vectors occur in relative proportions (0.2,0.4,0.45,0.05).

Using the Gibbs sampling method, construct the appropriate Markov chain with this stationary distribution.

E3 For the multiple sequence alignment problem from the Gibbs sampling example, show that if we can consider $q_{ij}(s)$ and $q^*_{ij}(s)$ as approximately equal, the relative entropy between $q^*_{ij}(s)$ and p_j

$$\sum_{i=1}^{W}\sum_{j=1}^{20} q^*_{ij}(s) log\left(\frac{q^*_{ij}(s)}{p_j}\right)$$

has an approximate linear relationship with the logarithm of the probability of being in state s

$$C\prod_{i=1}^{W}\prod_{j=1}^{20}\left(\frac{q^*_{ij}(s)}{p_j}\right)^{c_{ij}(s)}$$

[Note that X and Y have a linear relationship if for some constants a and b, $Y = aX + b$].

The three R-based questions all concern the routine *MCMCmetrop1R* in the MCMCpack library. Visit the help system and read about this routine. In particular the first example is about logistic regression which serves all three questions. This example is divided into three main components; the function to optimise, generation of the data (these are simulated) and the application of the routine with some related commands.

R1 Copy the central part of the example (from the line beginning x1 to that beginning yvector) and run this in R. Obtain plots of the data to get a feel for how y depends upon the x-values (noting that y are Bernoulli variables, taking values only 0 or 1) .

R2 Now copy first the example function (logitfun to }) and then run it, with related commands, (using post.samp to summary(post.samp)). Give estimates of the three variables, and briefly explain the other output.

Note that the function being optimised is

$$\sum (ylog(p) + (1-y)log(1-p))$$

where

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}}$$

R3 Repeat R2, but this time choose a different example function. You can do this by entering each line in turn and making a single change to the form of 'p' (but make sure that this always lies between 0 and 1, as p is a probability).

General Construct a continuous-time Markov chain whose transition RATES correspond to the transition probabilities that you found in E1. Hence simulate a sequence of transitions of the process from E1 How else could you perform such a simulation without working out these rates?.

Statistical and Numerical Methods for Bioinformatics: Exercises 16

E1 Find the proportion of each zygote A_iA_j under Hardy-Weinberg equilibrium for the following sets of alleles and frequencies;
(i) A_1 occurs in frequency 0.7 and A_2 in frequency 0.3.
(ii) A_1 occurs in frequency 0.45, A_2 in frequency 0.33 and A_3 in frequency 0.22.

E2 The zygotes A_1A_1, A_1A_2 and A_2A_2 occur in a sample with the following frequencies:
Case (i) 37, 126, 132
Case(ii) 72, 75, 68
For each of the above, perform a chisquare test for Hardy-Weinberg equilibrium.

E3 Repeat question E2, but this time perform Fisher's test in each case.
Compare your results and comment.

R1 Suppose that a population has two alleles A and B at a given locus, and that the numbers of each genotype are as follows
AA 23, AB 55, BB 58
Test for Hardy-Weinberg equilibrium using
(i) a χ^2 test,
(ii) Fisher's exact test.

R2 Suppose that a population has three alleles A, B and C at a given locus, and that the numbers of each genotype are as follows
AA 17, AB 32, BB 71, AC 44, BC 63, CC 49
Test for Hardy-Weinberg equilibrium using a χ^2 test.

R3 Estimate the level of disequilibrium, and find a 95% confidence interval for this level, for
(i) the data from question R1
(ii) the same data, but with the number of AAs being 53 instead of 23
(iii) the same data, but with the number of AAs now 8 and the number of BBs 38 (instead of 58).

General Simulate a new generation of individuals descended from those in R1, assuming that the population follows Hardy-Weinberg.

Statistical and Numerical Methods for Bioinformatics: Exercises 17

E1 Obtain the distances between each of the nodes in the following tree and verify that these distances satisfy the distance properties (i)-(iii) from the notes.

E2 For the five species a, b, c, d and e with distances given in the following table, reconstruct the tree using the algorithm in 17.3, starting with two species and adding species sequentially.

	a	b	c	d	e
a	0	2	8	8	8
b		0	8	8	8
c			0	4	4
d				0	2
e					0

E3 Use the UPGMA algorithm to construct the tree between the five species with distances given in the following table.

	a	b	c	d	e
a	0	9	8	7	8
b		0	3	6	7
c			0	5	6
d				0	3
e					0

R1 Consider the phylogeny data set *bird.orders*. Use *R* to show that the tree is ultrametric and binary and obtain a plot of the tree.

R2 Generate a random tree with
a) 7 species
b) 25 species.
Plot the tree for each case.
In each case convert the tree to an ultrametric one, and obtain plots of the new trees.

R3 The data file *bird.orders* and the random trees above are data sets of the class *phylo*. By looking at the plots of the trees from questions R1 and R2, and the information given when you type the name of the data file, explain the information in a 'phylo' data set.

General Repeat R2 with a tree with 5 species, find the distance matrix between the species, and then use one of the methods from the notes to reconstruct the tree from the distance matrix.

Statistical and Numerical Methods for Bioinformatics: Exercises 18

E1 Verify that the stationary distribution for the Kimura model is (0.25, 0.25, 0.25, 0.25)

E2 The Jukes-Cantor model is a special case of the Kimura model (so that if you choose $\alpha = \beta$ in the Kimura model you get the Jukes-Cantor) and also of the Felsenstein model. Is either the Kimura model or the Felsenstein model a special case of the other?

E3 Discuss the example in section 18.4 of the notes. Some groupings were consistent across all methods, others were not. By examining the data, explain why particular groupings would be consistent and others not. in particular why is the location of 'human' so different?

R1 For each of the ultrametric random trees and the *bird.order* data from the previous question sheet, obtain the matrix of distances associated with the tree.

R2 Use a Kimura evolutionary model to obtain a distance matrix for the data *woodmouse*. Repeat this with a Jukes-Cantor model.

R3 Use a Kimura evolutionary model to obtain the distance matrix for the following three DNA strings, which are subsequences of the first three sequences of the woodmouse data from the previous question.
ATCAGTCACT
ATCAACCACT
ATCAATCACT

General Draw the tree relating to the sequences from R3 (by hand). Comment on any problems that you have, and why they might be occurring. Repeat R3 with different sequences, and compare your results.

Statistical and Numerical Methods for Bioinformatics: Exercises 19

E1 A genetic algorithm has six individuals in each generation. The fitness of the current six individuals are 0.7,1.2,1.3,1.5,1.6 and 1.7 respectively.
(i) What is the probability that the first individual in the mating population is a copy of the fittest individual?
(ii) What is the probability that exactly two out of the six individuals in the mating population are copies of the fittest individual?

E2 Two individuals from the mating population have genomes (1,1,2,1,1) and (1,2,1,1,2). List
(i) all of the possible genomes under the random allocation model,
(ii) all of the possible pairs of genomes under the single crossover model.
(iii) For each of parts (i) and (ii), what is the probability that a genome will contain a sequence of (at least) four 1s?

E3 The values of the first allele of two mating individuals, from a continuous set, are 1.342 and 1.765 respectively. Assuming that random allocation occurs in recombination, and that mutations occur with probability $r = 0.15$ according to a normal distribution with a mean of the current allele and variance 1:
(i) What is the probability that, if the chosen value from recombination is 1.342, the new allele will have value less than 1?
(ii) What is the probability that the new allele has value less than 1, if the chosen value of the allele after recombination is unknown?