

M.Sc. BIOINFORMATICS

CAS34 BIODATA ANALYSIS
REVIEW EXERCISES + PROB. BASIS

SOLUTIONS

Q.1 (i) Let $X =$ r.v. (R.A. Status)
Let $Y =$ r.v. (No./month)
 \therefore Joint Univariate / Marginal Dist: measure
(Across the Columns, X only)
 $\therefore P\{X = \text{None}\} = 77/398 = 0.1935$

(ii) Conditional X given Y
 $\therefore P\{X = \text{Maintained} | Y > 1\}$
 $= \frac{P\{X = \text{Maintained} \cap Y > 1\}}{P\{Y > 1\}}$
 $= \frac{67/398}{141/398} = \frac{67}{141}$

Note:
Sample Space reduced
 $= 0.4752$

(iii) Joint or Bivariate X and Y
 $\therefore P\{X = \text{Not Maint.} \cap Y = 0\}$
 $= 17/398$
 $= 0.0427$

Q.1 contd.

(iv) Either X or Y Addition but may not be Mutually Exclusive
 $\therefore P\{X = \text{None or } Y = 1\} = P\{X = \text{None}\} + P\{Y = 1\}$
 $- P\{X = \text{None and } Y = 1\}$

$= \frac{77}{398} + \frac{177}{398} - \frac{37}{398}$

$= 0.5452$

Q.2 Exercise on Conditional & Total Probability

Let event S = Exhibits Symptoms
 $\therefore D =$ Disease Present
Given $P\{S\} = 0.2$ (and $P\{\bar{S}\} = 1 - P\{S\}$)

$P\{D\} = 0.23$
 $P\{D \cap S\} = 0.18$

Want $P\{D | \bar{S}\}$ i.e. P{Disease Present given no symptoms}

Cond: $P\{D | \bar{S}\} = \frac{P\{D \cap \bar{S}\}}{P\{\bar{S}\}}$

and $P\{D \cap S\} = P\{D | S\} \cdot P\{S\}$
 \therefore Tot. Prob. $P\{D\} = P\{D | S\} \cdot P\{S\} + P\{D | \bar{S}\} \cdot P\{\bar{S}\}$
all possibilities

- 3 -

Q2 contd.

$$\begin{aligned} \therefore P\{D \leq 3\} &= \frac{P\{D\} - P\{D \leq 5\} \cdot P\{S\}}{P\{S\}} \\ &= \frac{P\{D\} \cdot P\{D \leq 5\}}{P\{S\}} \\ &= \frac{0.23 \cdot 0.18}{0.8} = \underline{\underline{0.0625}} \end{aligned}$$

Q3 Situation is Binomial Either/or
i.e. binary data - Colour Blind/Not C.B.
Binomial parameters n, p = 25, 0.1 here

Calculation options:

- First principles (lot of effort)
- Tables
- Approximation

To illustrate what is involved.

(i) Want $P\{X \leq 5\}$ i.e. Prob. Five or fewer are C.B.

p.d.f. Binomial (B(n,p))

$$P\{X=x\} = \binom{n}{x} p^x q^{n-x}$$

$\left\{ \begin{array}{l} \text{or } n C x p^x q^{n-x} \\ \text{C = No. of ways (combination)} \end{array} \right\}$

- 4 -

Could \therefore calculate: \square First principles

$$\begin{aligned} P\{X=0\} &= \binom{25}{0} (0.1)^0 (0.9)^{25} = 0.0718 \\ P\{X=1\} &= \binom{25}{1} (0.1)^1 (0.9)^{24} = 0.1994 \\ P\{X=2\} &= \binom{25}{2} (0.1)^2 (0.9)^{23} = 0.2659 \\ P\{X=3\} &= \binom{25}{3} (0.1)^3 (0.9)^{22} = 0.2265 \\ P\{X=4\} &= \binom{25}{4} (0.1)^4 (0.9)^{21} = 0.1384 \\ P\{X=5\} &= \binom{25}{5} (0.1)^5 (0.9)^{20} = 0.0646 \end{aligned}$$

\therefore Sum of these = 0.9666 $P\{\text{Five or fewer are C.B.}\}$ (A)

OR a Tables: Want n = 25, p = 0.1

W, Y & S tables are cumulative, i.e.

$$\text{give } \sum_{x=r}^n \binom{n}{x} p^x (1-p)^{n-x} = P\{X \geq r\}$$

$$\therefore P\{X \leq 5\} = 1 - P\{X \geq 6\} \text{ complementary probability}$$

but not actually tabulated for n = 25, p = 0.1

in this set of tables. (Note other tables

may include slightly different ranges

of n, p). However W, Y & S summarise what

to do next p.6 - Approximate

(Note: Interpolation may be possible within

our given distribution but

o.g. n = 20, p = 0.1; n = 50, p = 0.1

are not same binomial distributions)

Approximate 1:

Poisson? Strictly large n , $p < 0.1$

$\lambda = \text{Poisson parameter} = np = 2.5$

(Again, could calculate from form of p.d.f. but have Tables for $\lambda = 2.5$)

w/ 1.5 Tables again cumulative sense, so

$$P\{X \leq 5\} = 1 - P\{X \geq 6\} \text{ for } \lambda = 2.5$$

$$= 1 - 0.04202$$

$$= \underline{\underline{0.95798}} \text{ c.f. soln. from (A)}$$

(B)

Approximate 2:

Normal w/ 1.5 p. 6 suggests $\mu = np$

$$\sigma = \sqrt{npq}$$

$$= 1.118034$$

$$(q = 1 - p)$$

Form of Normal Transform, p. 59

$$U = \frac{X - \mu}{\sigma}$$

$$\therefore P\{X \leq 5\} = P\left\{U \leq \frac{5 - 2.5}{1.118034}\right\}$$

$$= P\{U \leq 1.666\}$$

Tabulated Values in Standardized Normal Table

are $P\{U > 1.66\}$ and $P\{U > 1.67\}$

giving a value of 0.04813

interpolating between the two

So, finally

$$P\{X \leq 5\} = P\{U \leq 1.666\}$$

$$= 1 - P\{U > 1.666\}$$

$$= 1 - 0.04813$$

$$= \underline{\underline{0.95187}} \text{ c.f. (A)}$$

(C)

Note: That skill not exact because

approximating a discrete by a continuous distribution and strictly should take the "continuity correction" into account.

However for most purposes $P\{X \leq 5\} = 0.96$ is adequate!

(ii) want prob. at least 6 or c.d.

$$\therefore P\{X \geq 6\}$$

Note this is just the complement of

previous part

$$P\{X \geq 6\} = 1 - P\{X \leq 5\}$$

$$= \underline{\underline{0.0334}}$$

or more approximately

$$\underline{\underline{0.04}}$$

-7-

(iii) Between 6 and 9 colour-blind options for calculation as before, but clearly, $P\{6 \leq X \leq 9\} = P\{X \leq 9\} - P\{X \leq 5\}$.
 Though we do not have exact table, it is obvious that $P\{X \leq 9\} \approx 1$ and we have already obtained $P\{X \leq 5\}$.
 $\therefore P\{6 \leq X \leq 9\} \approx 1 - 0.9666$
 $= 0.0334$

$$(iv) P\{X=2,3 \text{ or } 4\} = P\{2 \leq X \leq 4\}$$

↑
 So, if did the initial calculation from first principles, have this already
 $= P\{X=2\} + P\{X=3\} + P\{X=4\}$
 $= 0.2659 + 0.2265 + 0.1384$
 $= 0.6308$ (Prob. 2, 3 or 4 are C.B.)

Alternatively, could apply one or other of the approximations as before

-8-

Q.4 Poisson, with $\lambda = 2$
 Straightforward use of Tables, bearing in mind that these are cumulative

$$(i) P\{X \leq 1\} = 1 - P\{X \geq 2\} \quad \text{where } P\{X \geq 2\} \text{ tabulated}$$

$$= 1 - 0.59399$$

$$= 0.406$$

$$(ii) P\{X=3\} = P\{X \geq 3\} - P\{X \geq 4\}$$

$$= 0.32332 - 0.14288$$

$$= 0.180$$

$$(iii) P\{X \geq 5\} = 0.0527 \quad \text{Directly from tables}$$

Q.5 Poisson, with $\lambda = 13$
 Options again as for Q.4, since table for $\lambda = 13$ not given in w.Y.S. + Additional options.

Note: a) Crude interpolation is possible here \therefore only one parameter, but answers will be approximate.

-9-

Illustration: for part (iii)

$$P\{X \leq 12\} = 1 - P\{X \geq 13\} \quad \{\text{value } \geq \text{is from for } \omega, y \leq \text{table}\}$$

$$\therefore \text{for } \lambda = 12: 1 - P\{X \geq 13\} = 1 - 0.42403 = 0.57597$$

$$\text{for } \lambda = 13: 1 - P\{X \geq 13\} = 1 - 0.64154 = 0.35846$$

\therefore Interpolation $\rightarrow 0.467$
(See more accurately below)

b) More usefully: simple recursion formula applies to calculation of Poisson s.t.

$$P\{X+1\} = \frac{\lambda P\{X\}}{X+1}$$

$$\therefore P\{1\} = \frac{\lambda P\{0\}}{1} \text{ etc.}$$

So, whichever method, choose to use:

(i) $P\{X=10\}$ for Poisson, $\lambda=13$
e.g. $P\{X=10\} = e^{-13} \frac{13^{10}}{10!} = 0.086$

(ii) $P\{X \geq 8\}$:
 $P\{0\} = e^{-13} \frac{13^0}{0!} = e^{-13} = 0.00002263$

-10-

$$P\{1\} = \frac{13 P\{0\}}{1}$$

$$P\{2\} = \frac{13 P\{1\}}{2} = \frac{13^2 P\{0\}}{2 \cdot 1}$$

$$\vdots P\{7\} = \frac{13^7 P\{0\}}{7!}$$

$$\therefore P\{X \geq 8\} = 1 - P\{X \leq 7\}$$

$$= 1 - e^{-13} \left[1 + 13 + \frac{13^2}{2} + \frac{13^3}{6} + \frac{13^4}{24} + \frac{13^5}{120} + \frac{13^6}{720} + \frac{13^7}{5040} \right]$$

$$= 1 - 0.05409$$

$$= \underline{\underline{0.946}}$$

(iii) $P\{X \leq 12\}$ - easy to continue from previously

$$= e^{-13} \frac{13^8}{8!} \left[1 + \frac{13}{9} + \frac{13^2}{10 \cdot 9} + \frac{13^3}{11 \cdot 10 \cdot 9} + \frac{13^4}{12 \cdot 11 \cdot 10 \cdot 9} \right] + [0.05409]$$

$$= \underline{\underline{0.463}}$$

(iv) $P\{9 \leq X \leq 15\} = P\{X \leq 15\} - P\{X \leq 8\}$
Again, most of work is done, so substituting
 $= \underline{\underline{0.664}}$

(v) $P\{X \leq 7\}$ from part (ii) directly
 $= 0.05409$

Q.6 M.g.f. Binomial

$$M_x(t) = E\{e^{tx}\} \\ = \sum_x e^{tx} f(x) \\ = \sum_{x=0}^{\infty} e^{tx} \binom{n}{x} p^x (1-p)^{n-x} \\ = [1-p+pe^t]^n \text{ from Notes}$$

$$M_1 = E\{X\} = \left. \frac{d}{dt} M_x(t) \right|_{t=0} \\ = np e^t [1-p+pe^t]^{n-1} \Big|_{t=0} \\ = np$$

$$M_2 = E\{X^2\} = \left. \frac{d^2}{dt^2} M_x(t) \right|_{t=0} \\ = \left\{ n(n-1)pe^t [1-p+pe^t]^{n-2} \right. \\ \left. + np e^t [1-p+pe^t]^{n-1} \right\} \Big|_{t=0} \\ = n(n-1)p^2 + np$$

$$\therefore \text{Var}\{X\} = E\{X^2\} - [E\{X\}]^2$$

$$= n(n-1)p^2 + np - (np)^2$$

$$= np(1-p) \quad [q=1-p]$$

Q.7 Set-up for migration described in class

$$\text{i.e. } p_n = \left[1 - \sum_{i=1}^k m_i \right] p_0 + \sum_{i=1}^k (m_i p_i) \\ = p_0 + \sum_{i=1}^k [m_i (p_i - p_0)]$$

So that change in allelic frequency

$$\Delta p = p_n - p_0 \\ = \sum_{i=1}^k [m_i (p_i - p_0)]$$

Now (i) Some idea

Start with allelic frequency p_0 but add (or subtract) mutation to or from this original proportion p_0 in native originally.

$\therefore p_n = p_0(1-u) + (1-p_0)v$
 diminished proportion not originally this type now by mutation from original matching to

$\therefore \Delta p = \text{allelic frequency change}$

$$= p_n - p_0 \\ = [v(1-p_0) - u p_0]$$

Clearly, both mutation and migration could occur together.

(ii) For $\Delta p = 0$ Equilibrium

$$\text{i.e. } \frac{p_n}{1-p_n} = \frac{v}{u}$$

$$\text{or } p_n = \frac{v}{u+v}$$

(Note: In general $u \gg v$)

• General point on selection

Form for genotypic frequencies for genotypes AA, Aa, aa given in class

$$p_{AA} = p_A^2$$

$$p_{Aa} = 2p_A p_a$$

$$p_{aa} = p_a^2$$

and only 2 possible (i.e. $p_A = 1 - p_a$)

Interested in A, so numerator from 1st two

$$\therefore \text{Form } p_A = \frac{f_1 p_A^2 + f_2 p_A (1-p_A)}{f_1 p_A^2 + 2f_2 p_A (1-p_A) + f_3 (1-p_A)^2}$$

Fraction of allele A in Total genotypic freq.

$$\begin{aligned} \text{Q. 6 } \mu &= p^2 a + 2pqd - q^2 a \\ &= ((p-q)a + 2pqd) \end{aligned}$$

for this Model
(see Notes)

(i) Deviation of genotypic value of AA from population mean i.e.

$$\begin{aligned} a - \mu &= a - [(p-q)a + 2pqd] \\ &= 2q(a-pd) \end{aligned} \quad \text{--- I}$$

So for Aa similarly:

$$\begin{aligned} d - \mu &= d - [(p-q)a + 2pqd] \\ &= a(q-p) + d(1-2pq) \end{aligned} \quad \text{--- II}$$

and for aa

$$\begin{aligned} -a - \mu &= -a - [(p-q)a + 2pqd] \\ &= -2p(a+qd) \end{aligned} \quad \text{--- III}$$

(ii) A gamete containing allele A results in progeny with genotypes AA and Aa with frequency p and q respectively. Similarly for gamete allele a \rightarrow Aa, aa with p and q respectively.

Mean values of genotypes produced for the two types of gametes thus:

$$\begin{aligned} A &: pa + qd \\ a &: pd - qa \end{aligned}$$

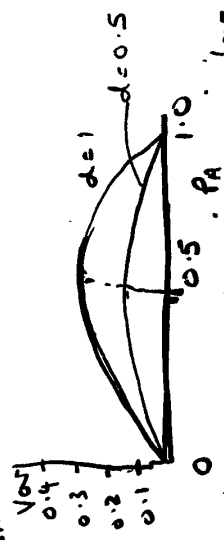
Dominance Variance = Variance of Dominance Deviations

$$\sigma_D^2 = p^2(-2q^2d)^2 + 2pq(2pqd)^2 + q^2(-2p^2d)^2 = 4p^2q^2d^2$$

\therefore clearly $\sigma_A^2 + \sigma_D^2 = \sigma_G^2$

(vi) Could pick any values and substitute. Clearly just fix probability (frequency) of allele, since $p_a = 1 - p_A$

Dominance Variance - easiest to see



Additive Genotypic Variance - similar to above (in shape) for $d=0$ (Maximum again $p_A=0.5$), becomes more right-skewed as d increases.

Similarly, (more pronounced) for Total Variance in ref. to variance values (as might expect), since σ_D^2 and σ_A^2 add to give Total σ_G^2