BIOINFORMATICS & COMPUTATIONAL GENETICS MSc
PROBABILITY AND STATISTICS

MORE SOLUTIONS

1. Maximum likelihood for λ;

STEP 1 Likelihood function $L(\lambda)$, for $\lambda > 0$,

$$L(\lambda) = \prod_{i=1}^{n} f(x_i;\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n \exp\left\{-\lambda \sum_{i=1}^{n} x_i\right\}$$

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

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

STEP 3 Derivative $\log L(\lambda)$ with respect to λ equated to zero.

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

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

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

Now reparameterize in terms of $\tau = e^{-\lambda} \Longleftrightarrow \lambda = -\log\tau$.

STEP 1 Likelihood function $L(\tau)$, for $\tau > 0$,

$$L(\tau) = \prod_{i=1}^{n} f(x_i;\tau) = \prod_{i=1}^{n} [-\log\tau] e^{-[-\log\tau]x_i} = [-\log\tau]^n \exp\left\{-[-\log\tau] \sum_{i=1}^{n} x_i\right\}$$

STEP 2 Log-likelihood $\log L(\tau)$.

$$\log L(\tau) = n\log[-\log\tau] + \log\tau \sum_{i=1}^{n} x_i$$

STEP 3 Derivative $\log L(\tau)$ with respect to τ equated to zero.

$$\frac{d}{d\tau}\{\log L(\tau)\} = -\frac{n}{\tau\log\tau} + \frac{1}{\tau}\sum_{i=1}^{n} x_i = 0 \Longrightarrow \hat{\tau} = e^{-1/\bar{x}}.$$

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

Hence if $\tau = \tau(\lambda)$ then $\hat{\tau} = \tau(\hat{\lambda})$

2. Can use ML procedure to show that m.l.e.s of θ are

(a) $\dfrac{n}{-\sum_{i=1}^{n} \log x_i}$ (b) $\dfrac{n}{\sum_{i=1}^{n} \log x_i} - 1$ (c) $\dfrac{2n}{\sum_{i=1}^{n} x_i}$

3. Mles based on $x_1, ..., x_n$ obtained using standard ML techniques; can simplify likelihood by noting that

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$$

which gives $\hat{\mu} = \bar{x}$ directly.

Mles based on $x_1, ..., x_n$ and $y_1, ..., y_m$ are obtained from the likelihood $L(\mu, \sigma_1, \sigma_2)$ where

$$L(\mu, \sigma_1, \sigma_2) = \prod_{i=1}^{n} f(x_i; \mu, \sigma_1) \prod_{i=1}^{m} f(y_i; \mu, \sigma_2)$$

so that

$$\hat{\mu} = \frac{\displaystyle\sum_{i=1}^{n} x_i + \sum_{i=1}^{m} y_i}{n + m} \qquad \hat{\sigma_1^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \qquad \hat{\sigma_2^2} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{\mu})^2$$

4. Can show $\hat{\lambda} = \bar{x}$ (done explicitly in lecture notes); so the ML estimator is $\bar{X}$.

Now, if $X_1, ..., X_n \sim Poisson(\lambda)$ i.i.d., then can show that $Y = \displaystyle\sum_{i=1}^{n} X_i \sim Poisson(n\lambda)$, so the sampling distribution of the estimator is the distribution of random variable $T = \dfrac{1}{n}Y$, and hence

$$\mathrm{E}_{f_T}[T] = \frac{1}{n}\mathrm{E}_{f_Y}[Y] = \frac{1}{n}.n\lambda = \lambda \qquad \mathrm{Var}_{f_T}[T] = \frac{1}{n^2}\ \mathrm{Var}_{f_Y}[Y] = \frac{1}{n^2}.n\lambda = \frac{\lambda}{n}$$

as the expectation and variance of a $Poisson(n\lambda)$ random variable are both equal to $n\lambda$.

Now for $\epsilon > 0$,

$$\mathrm{P}[\lambda - \epsilon \leq T \leq \lambda + \epsilon] \;= \mathrm{P}\left[\lambda - \epsilon \;\leq\; \frac{1}{n}\sum_{i=1}^{n} X_i \;\leq\; \lambda + \epsilon\right] = \mathrm{P}\left[-\epsilon\sqrt{\frac{n}{\lambda}} \;\leq\; \frac{\displaystyle\sum_{i=1}^{n} X_i - n\lambda}{\sqrt{n\lambda}} \;\leq\; \epsilon\sqrt{\frac{n}{\lambda}}\right]$$

$$\approx \Phi\left(\epsilon\sqrt{\frac{n}{\lambda}}\right) - \left(-\epsilon\sqrt{\frac{n}{\lambda}}\right) \longrightarrow 1 \quad \text{as } n \longrightarrow \infty.$$

by the Central Limit Theorem, so the probability distribution of the estimator T becomes more and more concentrated around λ as n gets larger and hence the estimator can be regarded as an good estimator of the parameter of interest.

5. Summary statistics : $\bar{x} = 6.4161$, $s^2 = (0.0742)^2$. Fix significance level at $\alpha = 0.05$.

(a) σ known; carry out a Z-test: Test statistic z given by

$$z = \frac{\bar{x} - c}{\sigma/\sqrt{n}} = \frac{6.4161 - 6.4}{0.03/\sqrt{15}} = 2.079 \sim N(0,1) \text{ if } H_0 \text{ is TRUE}$$

Critical Value : 1.96, so test statistic is more extreme than the critical value. Hence $\Longrightarrow$ REJECT H_0.

(b) σ unknown; carry out a T-test: Test statistic t given by

$$t = \frac{\bar{x} - c}{s/\sqrt{n}} = \frac{6.4161 - 6.4}{0.0742/\sqrt{15}} = 0.840 \sim t_{n-1} \text{ if } H_0 \text{ is TRUE}$$

where t_{n-1} is the Student-t distribution with $n - 1 = 14$ degrees of freedom.

Critical Value : 2.1448, so test statistic is not more extreme than the critical value. Hence $\Longrightarrow$ CANNOT REJECT H_0.

6. Summary statistics are

$$S_x = \sum_{i=1}^{n} x_i = 232.5 \quad S_y = \sum_{i=1}^{n} y_i = 229.0 \quad S_{xx} = \sum_{i=1}^{n} x_i^2 = 6974.25 \quad S_{xy} = \sum_{i=1}^{n} x_i y_i = 6884.65$$

(a) Estimates: $\hat{\alpha} = 0.9948$, $\hat{\beta} = -0.2281$.

(b) Fitted values: $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$. Residuals: $e_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta} x_i$.

(c) $s^2 = \dfrac{1}{n-2} \displaystyle\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = (0.2067)^2$

(d) Standard errors: $s.e.(\alpha) = 0.1378$, $s.e.(\beta) = 0.0052$.

(e) 95 % Confidence Intervals: $t_{n-2}(0.975) = 2.3060$

$$\text{Interval for } \alpha \quad : \ (-0.6059 : 0.0297)$$

$$\text{Interval for } \beta \quad : \ (0.9829 : 1.0068)$$

95 % CI for α includes 0, so some evidence that $\alpha = 0$.

95 % CI for β includes 1, so some evidence that $\beta = 1$.

7. Summary statistics : $\bar{x} = 276.89$, $s^2 = (27.08)^2$. Fix significance level at $\alpha = 0.05$.

(a) Test statistic q given by

$$q = \frac{(n-1)s^2}{\sigma^2} = \frac{18\ 733.4}{(30.0)^2} = 14.668 \sim \chi^2_{18} \text{ if } H_0 \text{ is TRUE}$$

Critical Values : $\chi^2_{18}(0.025) = 8.231$, $\chi^2_{18}(0.975) = 31.526$. Test statistic is not in critical region

$$\Longrightarrow \text{CANNOT REJECT } H_0.$$

(i) 95 % CI for μ : $\bar{x} \pm \dfrac{\sigma}{\sqrt{n}} 1.96 = 276.89 \pm 13.49$.

(ii) 95 % CI for μ : $\bar{x} \pm \dfrac{s}{\sqrt{n}} t_{18}(0.975) = 276.89 \pm 13.05$.

95 % CI for σ: from above, if H_0 is TRUE

$$\mathrm{P}[8.231 \leq (n-1)s^2/\sigma^2 \leq 31.526] = 0.95 \Longleftrightarrow \mathrm{P}[(n-1)s^2/31.526 \leq \sigma^2 \leq (n-1)s^2/8.231] = 0.95$$

so a 95 % CI for σ^2 is given by

$$\mathrm{P}[418.74 \leq \sigma^2 \leq 1603.839] = 0.95 \Longleftrightarrow \mathrm{P}[20.46 \leq \sigma \leq 40.05] = 0.95.$$

9. $Q \sim F_{\nu_1,\nu_2}$. Then $\mathrm{P}[Q \leq x] = 0.975 \Longleftrightarrow \mathrm{P}[1/Q \geq 1/x] = 0.975 \Longleftrightarrow \mathrm{P}[1/Q < 1/x] = 0.025$

But $Q^* = 1/Q = \dfrac{V/\nu_2}{U/\nu_1} \sim F_{\nu_2,\nu_1}$. $\mathrm{P}[Q \leq x] = 0.975 \Longleftrightarrow \mathrm{P}[Q^* \leq 1/x] = 0.025$.

10. Summary statistics: $n_1 = 14$, $\bar{x_1} = 153.857$, $s_1^2 = (39.04)^2$, $n_2 = 5$, $\bar{x_2} = 113.000$, $s_2^2 = (56.08)^2$.

(i) Pooled estimate of variance, $s_P^2 = ((n_1-1)s_1^2 + (n_2-1)s_2^2)/(n_1+n_2-2) = (43.65)^2$.

Test statistic t given by

$$t = \frac{x_1 - x_2}{s_P\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} = \frac{153.857 - 113.000}{43.65 \times 0.521} = 1.797 \sim t_{n_1+n_2-2} \text{ if } H_0 \text{ is TRUE}$$

Critical Value : $n_1 + n_2 - 2 = 17$, so $t_{17}(0.975) = 2.1098$. Test statistic is not in critical region

$$\Longrightarrow \text{CANNOT REJECT } H_0.$$

(ii) Test statistic q given by

$$q = \frac{s_1^2}{s_2^2} = \frac{(39.04)^2}{(56.08)^2} = 0.4846 \sim F_{n_1-1,n_2-1} \text{ if } H_0 \text{ is TRUE}$$

Critical Values : $F_{13,4}(0.025) = 1/(3.996) = 0.2502$, $F_{13,4}(0.975) = 8.715$. Test statistic is not in critical region

$$\Longrightarrow \text{CANNOT REJECT } H_0.$$

(iii) 95 % CI for $\mu_1 - \mu_2$:

$$(x_1 - x_2) \pm s_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\, t_{n_1+n_2-2}(0.975) = (153.857 - 113.000) \pm 43.65 \times 0.521 \times 2.1098$$

so the interval is (-7.123:88.837).

(iv) 95 % CI for σ_1/σ_2: from (ii) above, if H_0 is TRUE

$$\mathrm{P}\left[0.2502 \le \frac{(s_1^2/\sigma_1^2)}{(s_2^2/\sigma_2^2)} \le 8.755\right] = 0.95 \Longleftrightarrow \mathrm{P}\left[\frac{s_1^2}{s_2^2\ 8.755} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2\ 0.2502}\right] = 0.95$$

as test statistic q has an F-distribution. Thus a 95 % CI for σ_1/σ_2 is given by

$$\mathrm{P}\left[0.055 \le \frac{\sigma_1^2}{\sigma_2^2} \le 1.937\right] = 0.95 \Longleftrightarrow \mathrm{P}\left[0.235 \le \frac{\sigma_1}{\sigma_2} \le 1.392\right] = 0.95.$$

11. For probability plot, need to evaluate $q_1, ..q_n$, where

$$F_X(q_i) = \frac{i}{n+1} \Longleftrightarrow q_i^\alpha = -\frac{1}{\lambda}\log\left(1 - \frac{i}{n+1}\right)$$

For a probability plot, need to equate q_i to x_i (in sorted sample), that is, if the model is correct $x_i \approx q_i$, or equivalently

$$\log x_i \approx \log q_i = -\frac{1}{\alpha}\log\lambda + \frac{1}{\alpha}\log\left(-\log\left(1 - \frac{i}{n+1}\right)\right)$$

so that a plot of $\log x_i$ against $\log\left(-\log\left(1 - \dfrac{i}{n+1}\right)\right)$ should be a straight line with slope $1/\alpha$ and intercept $-\log\lambda/\alpha$.

12. Data mean is $242/200 = 1.21$. For each model need to calculate expected frequencies $E_i = 200p_i$ where p_i is defined by the model being tested, for $i = 1, ..., k$, where $k = 5$ categories. For (iii), use ML estimate of p, which is $\bar{x}/4 = 1.21/4 = 0.3025$; for (v), use ML estimate of λ, which is $\bar{x} = 1.21$. Relevant critical values of the test are only available at the $\alpha = 0.01$ significance level from the tables provided; $\chi^2_3(0.99) = 11.345$, $\chi^2_4(0.99) = 13.277$ (could use $\alpha = 0.05$ if you have the figures available).

For model (i), $p_i = 0.2$ for each category, degrees of freedom is $k - 0 - 1 = 4$.

For model (ii), p_i given by $Binomial(4, 0.25)$ formula, degrees of freedom is $k - 0 - 1 = 4$.

For model (iii), p_i given by $Binomial(4, 0.3025)$ formula, degrees of freedom is $k - 1 - 1 = 3$.

For model (iv), p_i given by $Poisson(1)$ formula, degrees of freedom is $k - 0 - 1 = 4$.

For model (v), p_i given by $Poisson(1.21)$ formula, degrees of freedom is $k - 1 - 1 = 3$.

Could proceed by pooling last two categories, reducing k to 4.