Chapter 2. Review of basic Statistical methods

1 Distribution, conditional distribution and moments

We consider two kinds of random variables: discrete and continuous random variables.

- For discrete random variable X, the possible values of X is finite or countable infinite $x_1, x_2, \ldots, x_n$, X will have a probability function
$$p(x_i) = P(X = x_i)$$
with the properties
$$p(x_i) > 0$$
$$\sum_{i=1}^{\infty} p(x_i) = 1$$

- For continuous random variable X, there is a non-negative function $f(x)$ called density function with the property, for any real number a and b
$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

- The moments of a random variable X include the expectation $E(X)$, variance $Var(X) = E[X - E(X)]^2$ etc., where
$$E(X) = \sum_{i=1}^{\infty} x_i p(x_i) \text{ for discrete r.v.} X$$
$$E(X) = \int_a^b x f(x)dx \text{ for continuous r.v.} X$$
$$Var(X) = \sum_{i=1}^{\infty} [x_i - E(X)]^2 p(x_i) \text{ for discrete } X$$
$$Var(X) = \int_a^b [x - E(X)]^2 f(x)dx \text{ for continuous } X$$

- For more than one random variables, the probability function for discrete random variable and density functions are defined in the similar way.

- For two discrete random variables (X, Y), suppose that the possible values for X are $x_1, x_2, \ldots,$ and the possible values for Y are $y_1, y_2, \ldots$ The conditional probability function
$$p(x_i|Y = y) = \frac{P(X = x_i, Y = y)}{P(Y = y)}$$
For two continuous random variables (X, Y), let $f(x, y)$ is the join density of (X, Y), the conditional density of X given $Y = y$ is
$$f(x|y) = \frac{f(x, y)}{f_Y(y)},$$
where $f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx$ is the marginal density of X.

- The conditional moments such as expectation of X and variance of X given $Y = y$ is similar to the formula given above, but replace the probability or density by conditional probability or conditional density given $Y = y$.

1.1 Some special distribution

1. Binomial distribution: $B(n, p)$

   If a random variable X with possible values 0, 1, 2, . . . , n and the probability function

$$P(X = k) = C_n^k p^k (1-p)^{n-k},$$

   we say that X has a binomial distribution denoted by X ˜ $B(n, p)$ and

$$E(X) = np; Var(X) = np(1-p).$$

   - Background: If a experiment only have two outcomes "success" and "failure", let $p = P(\text{success})$, and we do the experiment n times independently. Let X denote the number of success among the n experiments, then X ˜ $B(n, p)$.
   - Example: Consider a biallelic marker with codominant alleles A and a, and n individuals. Let p denote the frequency of allele A. Let X denote the number of allele A among the n individuals, then X ˜ $B(2n, p)$.

2. Multinomial distribution (generalization of Binomial distribution)

   - The experiment have m possible outcomes. Let $p = (p_1, ..., p_m)$ with p_i being the probability of ith outcome ($\sum_{i=1}^m p_i = 1$). We do the experiment n times independently. Let X_i denote the number of ith outcomes among the n experiments and $X = (X_1, ..., X_m)$,then we say X has a multinomial distribution denoted by

$$X \sim M_m(n, p).$$

     Let $N = (n_1, ..., n_m)$ with $\sum_{i=1}^m n_i = n$,then

$$P(X = N) = P(X_1 = n_1, ..., X_m = n_m)$$

$$= \frac{n!}{n_1! \cdots n_m!} p_1^{n_1} \cdots p_m^{n_m}.$$

     and

$$E(X_i) = np_i; Var(X_i) = np_i(1-p_i).$$

     Note: $M_2(n, p) = B(n, p_1)$
   - Example: Consider a marker with m codominant alleles $A_1, ..., A_m$ and n individuals. Let p_i denote the frequency of allele A_i. Let X_i denote the number of allele A_i among the n individuals, then

$$X = (X_1, ..., X_m) \sim M_m(2n, p),$$

     where $p = (p_1, ..., p_m)$. So, $E(X_i) = 2np_i$.

3. Some continuous random variables: such as Normal, χ^2,Gamma etc.

2 Likelihood, Maximum Likelihood Estimator and EM Algorithm

1. Likelihood

When we have collected data, the probability

$$P(data|parameters)$$

is called likelihood. The statistical inference is usually based on the likelihood. Mathematically, the data set we collected is also called sample denoted by $X_1, X_2, ..., X_n$. In most of the cases, the sampled individuals are iid (independent identical distributed) and have the same distribution with a random variable X called population. If the sampled individuals are independent, the likelihood will be

$$L(\theta) = P(X_1, ..., X_n|\theta) = \prod_{i=1}^{n} P(X_i|\theta)$$

$$= \prod_{i=1}^{n} p(X_i|\theta) \text{ for discrete r.v}$$

$$= \prod_{i=1}^{n} f(X_i|\theta) \text{ for continuous r.v}$$

2. The maximum estimator $\hat{\theta}(X_1, X_2, ..., X_n)$ of parameter θ is the maximizer of the likelihood function

$$L(\hat{\theta}) = \max_{\theta} L(\theta)$$

- Example: Consider a marker with three codominant alleles A, B, C. We sample n individuals and get the data

1. - $n_{AA}, n_{BB}, n_{CC}, n_{AB}, n_{AC}$ and n_{BC} the number of individuals with Genotype AA, BB, CC, AB, AC and BC, respectively.

   Let p_A, p_B and p_C denote the population frequencies of alleles A, B and C,respectively. Find the MLE of p_A, p_B and p_C.Solution: Note the multinomial distribution of $n_{AA}, n_{BB}, n_{CC}, n_{AB}, n_{AC}$ and n_{BC}, we can get the likelihood function of the data up to a constant, denoting $\theta = (p_A, p_B, p_C)$,

   $$L(\theta) = (p_A^2)^{n_{AA}}(p_B^2)^{n_{BB}}(p_C^2)^{n_{CC}}(2p_Ap_B)^{n_{AB}}(2p_Ap_C)^{n_{AC}}(2p_Bp_C)^{n_{BC}}.$$

   The log-likelihood is

   $$\begin{aligned} logL(\theta) &= 2n_{AA}\log p_A + 2n_{BB}\log p_B + 2n_{CC}\log p_C + n_{AB}(\log p_A + \log p_B) \\ &\quad + n_{AC}(\log p_A + \log p_C) + nBC(\log pB + \log pC) \\ &= (2n_{AA} + n_{AB} + n_{AC})\log p_A + (2n_{BB} + n_{AB} + n_{BC})\log p_B + (2n_{CC} + n_{AC} + n_{BC})\log p_C \end{aligned}$$

   So, the MLE of p_A, p_B and p_C are given by

   $$\hat{p}_A = \frac{2n_{AA} + n_{AB} + n_{AC}}{2n}$$

   $$\hat{p}_B = \frac{2n_{BB} + n_{AB} + n_{BC}}{2n}$$

   $$\hat{p}_A = \frac{2n_{CC} + n_{AC} + n_{BC}}{2n}$$

   - If the marker in last example is ABO locus, there will be some difficulties to estimate the parameters. For n sampled individuals, the data are n_A, n_B, n_{AB} and n_O for phenotype A, B, AB and O. How to get the MLE of allele frequencies of alleles A, B and O at this case? The data we get is not complete as given in last example. The complete data should be $n_{AA}, n_{BB}, n_{OO}, n_{AB}, n_{AO}$ and n_{BO}. In this case we can use EM algorithm.

3. EM algorithm

- Let X denote the unobserved complete data ($n_{AA}, n_{BB}, n_{OO}, n_{AB}, n_{AO}$ and n_{BO} in the example). Let Y denote the observed incomplete data(n_A, n_B, n_{AB} and n_O). Some function $t(X) = Y$ collapses X onto Y. The general idea is to choose X so that the maximum likelihood become trivial for the complete data.

  The complete data are assume to have a probability density (or likelihood) $f(X|\theta)$ that is a function of θ as well as X. The EM-algorithm have two steps:

  (a) E-step: we calculate the conditional expectation

  $$G(\theta|\theta_m) = E(\log f(X|\theta)|Y, \theta_m)$$

  where θ_m is the current estimated value of θ.

  (b) M-step: we maximize

  $$G(\theta|\theta_m)$$

  with respect to θ. This yield a new estimator of θ denoted by θ_{m+1}.
  We repeat the two steps until convergence occur. The convergence can be θ_m or the likelihood of the observed data $f(Y|\theta)$.

Consider the allele frequencies of alleles A, B and O at ABO locus.

The log-likelihood of the complete data

$$\log f(X|\theta) = (2n_{AA} + n_{AB} + n_{AC}) \log p_A + (2n_{BB} + n_{AB} + n_{BC}) \log p_B + (2n_{CC} + n_{AC} + n_{BC}) \log p_C$$

1. - E-step: $\theta = (\theta_1, \theta_2, \theta_3) = (p_A, p_B, p_O)$

$$\begin{aligned} G(\theta|\theta_m) = &(2E(n_{AA}|Y, \theta_m) + E(n_{AB}|Y, \theta_m) + E(n_{AO}|Y, \theta_m)) \log p_A \\ &+(2E(n_{BB}|Y, \theta_m) + E(n_{AB}|Y, \theta_m) + E(n_{BO}|Y, \theta_m)) \log p_B \\ &+(2E(n_{OO}|Y, \theta_m) + E(n_{AO}|Y, \theta_m) + E(n_{BO}|Y, \theta_m)) \log p_C. \end{aligned}$$

  We know that $E(n_{OO}|Y, \theta_m) = n_{OO}$, $E(n_{AB}|Y, \theta_m) = n_{AB}$. The conditional expectations we needed to calculate are $E(n_{AA}|Y, \theta_m), E(n_{AO}|Y, \theta_m), E(n_{BB}|Y, \theta_m)$ and $E(n_{BO}|Y, \theta_m)$. Intuitively, within Blood type A (n_A individuals the genotype may be A/A and A/O), the proportion of the individuals with genotype A/A is

  $$\frac{p_{AA}}{p_{AA} + p_{AO}} = \frac{p_{mA}^2}{p_{mA}^2 + 2p_{mA}p_{mO}}$$

  So, the expected number of individuals with genotype A/A is

  $$n_A \frac{p_{mA}^2}{p_{mA}^2 + 2p_{mA}p_{mO}}$$

  Mathematically, to calculate $E(n_{AA}|Y, \theta_m)$, we need to know the conditional distribution of n_{AA} given Y , that is, given $n_A = n_{AA} + n_{AO}$. Given n_A, n_{AA} can be considered as a random variable with a binomial distribution $B(n_A, p^*)$ where p^* is the probability of an individual with genotype AA given this individual having blood type A. So,

  $$p^* = P(AA|\text{blood type } A)$$

$$= \frac{P(AA)}{P(AA) + P(AO)}$$
$$= \frac{p_{mA}^2}{p_{mA}^2 + 2p_{mA}p_{mO}}.$$

It follows that

$$E(n_{AA}|Y, \theta_m) = E(n_{AA}|n_A) = n_A p^* = n_A \frac{p_{mA}^2}{p_{mA}^2 + 2p_{mA}p_{mO}}.$$

Similar to the argument above, we can calculate

$$E(n_{AO}|Y, \theta_m), E(n_{BB}|Y, \theta_m) \text{ and } E(n_{BO}|Y, \theta_m).$$

- M-step: with complete data X, we know that the MLE of the parameters are

$$\hat{p}_A = \frac{2E(n_{AA}|Y, \theta_m) + E(n_{AB}|Y, \theta_m) + E(n_{AO}|Y, \theta_m)}{2n}$$
$$\hat{p}_B = \frac{2E(n_{BB}|Y, \theta_m) + E(n_{AB}|Y, \theta_m) + E(n_{BO}|Y, \theta_m)}{2n}$$
$$\hat{p}_O = \frac{2E(n_{OO}|Y, \theta_m) + E(n_{BO}|Y, \theta_m) + E(n_{AO}|Y, \theta_m)}{2n}$$

3 Parametric Test Statistics

1. General set of Hypothesis Testing

   Suppose the sample $Y = (Y_1, ...Y_n)'$ has joint density (or probability) function $f(y|\theta)$ where $\theta \in \Omega \subseteq R^d$ (here means that $\theta = (\theta_1, ..., \theta_d)'$ is a d dimensional vector). Given $\Omega_0 \subset \Omega$ (strictly), it is desired to test $H_0 : \theta \in \Omega_0$ versus $H_A : \theta \in \Omega - \Omega_0$.

   - The general procedure of testing problem
     (a) Given a test statistics $T(Y)$ which has no relation with parameters θ
     (b) decide the rejection region, for example, $T(Y) > C$
     (c) For a given significant value α, calculate the value of C or calculate the p-value of the test

     $$\text{p-value} = P(T(Y) > T(y))$$

     where y is the observed value of sample Y. This step need to know the distribution of $T(Y)$.
   - In summary, a testing problem need to know (1) test statistics (2) the form of rejection region and (3) the distribution of the test statistic.

2. Three methods to construct the test

   (a) Likelihood ratio (Neyman and Pearson 1928). Let θ_Ω and θ_{Ω_0} are the MLE of the parameter θ under Ω and Ω_0, respectively, and let $L(\theta)$ denote the likelihood function of the sample Y. The likelihood ratio test statistic is

   $$G^2(Y) = -2\log(L(\theta_{\Omega_0}) - L(\theta_\Omega)).$$

   As $n \to \infty, G^2$ approximately has a χ^2 distribution with degrees of freedom $d - s$ where d and s are the dimensions of Ω and Ω_0, respectively. For a given significant value α, the likelihood ratio test rejects H_0 if and only if $G^2 > \chi^2_{\alpha,d-s}$. The p-value of the test is $P(\chi^2_{\alpha,d-s} > G^2(y))$.

(b) Score test (Rao 1947)

The ith score of Y is $S_i(\theta) = \frac{\partial \log L(\theta))}{\partial \theta_i}$ for $i = 1, ...d$. The vector $S(\theta) = (S_1(\theta), ..., S_d(\theta))$ satisfies $S(\theta_\Omega) = 0$. This observation suggests rejecting H_0 when $S(\theta_{\Omega_0})$ is far from the origin. The score test statistic is

$$X^2(Y) = (S(\theta_{\Omega_0}))' I^{-1}(\theta_{\Omega_0})(S(\theta_{\Omega_0})).$$

Here, $I(\theta)$ is the $d \times d$ information matrix with (i, j)th element

$$(I(\theta))_{ij} = -E_\theta[\frac{\partial^2 \log L(\theta)}{\partial \theta_i \partial \theta_j}]$$

As $n \to \infty, X^2$ approximately has a χ^2 distribution with degrees of freedom $d - s$. For a given significant value α, the score test rejects H_0 if and only if $X^2 > \chi^2_{\alpha,d-s}$. The p-value of the test is $P(\chi^2_{\alpha,d-s} > X^2(y))$.

(c) Wald (1943) test: Wald test is to design for testing the null hypothesis $\Omega_0 = \{\omega \in \Omega : h_1(\theta) = 0, ..., h_r(\theta) = 0\}$ such as $\theta_1 + \theta_3 = 0$, $\theta_2 + \theta_4 = 0$. Let $h(\theta) = (h_1(\theta), ..., h_r(\theta))'$. The idea of the test is to reject $H_0 : h(\theta) = 0$ when the vector $h(\theta_\Omega)$ is far from the null value of zero. The Wald test statistic is $W = (h(\theta_\Omega))'[(\nabla h(\theta_\Omega))' I^{-1}(\theta_\Omega) \nabla h(\theta_\Omega)]^{-1} h(\theta_\Omega)$ where $\nabla h(\theta_\Omega)$ is the $d \times r$ matrix with (i, j)th element $(\nabla h(\theta_\Omega))_{ij} = \frac{\partial h_j(\theta)}{\partial \theta_i}$.This test statistic also approximately has a χ^2 distribution with degree freedom r (note $s = d - r, r = d - s$)

Example 1 Let Y ˜ $M_m(n, p)$ with unknown parameter p belonging to the $m - 1$ dimensional simplex $\Omega = \{(p_1, ..., p_d) : p_i \geq 0, \sum_{i=1}^d p_i = 1\}$. The null hypothesis is $H_0 : p = p^0$ *i.e.* $\Omega_0 = \{p^0\}$. Construct likelihood ratio test, score test, and Wald test.

Example 2 Using the test you constructed to test the following problem: We have sampled 100 individuals. Each individual has genotype at a marker with three codominant alleles A_1, A_2,and A_3. Among the 100 individuals, we observed 60 allele A_1, 80 allele A_2 and 60 allele A_3. Let p_1, p_{12},and p_3 are the allele frequencies of A_1, A_2,and A_3, respectively. Test the null hypothesis $H_0 : p_1 = p_2 = p_3 = 1/3$.

- Likelihood ratio test:

1. (a) log-Likelihood function up to a constant

$$\log L(p) = \sum_{i=1}^m Y_i \log p_i$$

The MLE of under Ω is $\hat{p}_i = \frac{Y_i}{n}$, $i = 1, ..., m$. The MLE under Ω_0 is $p = p^0$. So, the likelihood ratio test statistic

$$G^2(Y) = -2(\log L(p^0) - \log L(\hat{p}))$$

$$= -2 \sum_{i=1}^m (Y_i \log p_i^0 - Y_i(\log Y_i - \log n))$$

$$= 2 \sum_{i=1}^m Y_i \log \frac{Y_i}{np_i^0}$$

which has an asymptotic distribution χ^2_{m-1}. For the specific test problem, $m = 3, n = 200, Y_1 = 60, Y_2 = 80$, and $Y_3 = 60$. So, $G^2(Y) = 2(60 \log(\frac{9}{10}) + 80 \log(\frac{6}{5}) + 60 \log(\frac{9}{10})) = 3.88$.

$$\text{p-value} = P(\chi^2_2 > 3.88) = 0.1433.$$

- Score test: The log-likelihood is

  1. (a)

$$\log L(p) = \sum_{i=1}^{m-1} Y_i \log p_i + Y_m \log(1 - p_1 - ... - p_{m-1}).$$

So,

$$S_j(p) = \frac{\partial \log L(p)}{\partial p_j} = \frac{Y_j}{p_j} - \frac{Y_m}{p_m}, 1 \le j \le m-1$$

and

$$\frac{\partial^2 \log L(p)}{\partial p_k \partial p_j} = \begin{cases} -\frac{Y_i}{p_j^2} - \frac{Y_m}{p_m^2}, & if\ k = j, \\ -\frac{Y_m}{p_m^2}, & if\ k \neq j, \end{cases}$$

Calculation gives

$$-Ep(\frac{\partial^2 \log L}{\partial p_k \partial p_j}) = \begin{cases} \frac{n}{p_j} + \frac{n}{p_m}, & if\ k = j, \\ \frac{n}{p_m}, & if\ k \neq j, \end{cases}$$

and thus the information matrix is

$$I(p) = n[diag(\frac{1}{p_1}, ..., \frac{1}{p_{m-1}}) + \frac{1}{p_m} \mathbf{1}_{m-1} \mathbf{1}'_{m-1}$$

where $\mathbf{1}_{m-1} = (1, 1, ...1)'$ a $m-1$ dimensional vector with elements 1. The MLE under H_0 is $p^0 = (p_1^0, ..., p_{m-1}^0)'$, the inverse is

$$I^{-1}(p^0) = \frac{1}{n}[diag(p_1^0, ..., p_{m-1}^0) - p^0(p^0)'].$$

The score test statistic

$$\begin{aligned} X^2 = & \ S(p^0)' I^{-1}(p^0) S(p^0) \\ = & \ \tfrac{1}{n}(\tfrac{Y_1}{p_1} - \tfrac{Y_m}{p_m}, ..., \tfrac{Y_{m-1}}{p_{m-1}} - \tfrac{Y_m}{p_m}) \\ & \times [diag(p_1^0, ..., p_{m-1}^0) - (p_1^0, ..., p_{m-1}^0)(p_1^0, ..., p_{m-1}^0)'] \\ & \times (\tfrac{Y_1}{p_1} - \tfrac{Y_m}{p_m}, ..., \tfrac{Y_{m-1}}{p_{m-1}} - \tfrac{Y_m}{p_m})' \\ = & \ \sum_{i=1}^{m} \tfrac{[Y_i - np_i^0]^2}{np_i^0} \sim \chi^2_{m-1} \end{aligned}$$

which is standard Pearson's Chi-square test

$$X^2 = \sum_{i=1}^{m} \frac{[Y_i - EY_i]^2}{EY_i}$$

For the specific test problem:

$$X^2 = \frac{[60 - 200/3]^2}{200/3} + \frac{[80 - 200/3]^2}{200/3} + \frac{[60 - 200/3]^2}{200/3} = 4$$

$$\text{p-value} = P(\chi_2^2 > 4) = 0.135.$$

- Wald test: Let $h_i(p) = p_i - p_i^0, i = 1, ..., m-1$. we can construct the Wald test (similar to the argument above)

$$W = \sum_{i=1}^{m} \frac{[Y_i - np_i^0]^2}{Y_i} \sim \chi^2_{m-1}$$

  For the specific problem,

$$W = \frac{[60 - 200/3]^2}{60} + \frac{[80 - 200/3]^2}{80} + \frac{[60 - 200/3]^2}{60} = 3.7$$

$$\text{p-value} = P(\chi^2_2 > 3.7) = 0.157.$$