CHAPTER-V

SOME NEW ADVANCED TECHNIQUES FOR BIOSTATISTICS
CHAPTER - V
SOME NEW ADVANCED TECHNIQUES FOR BIOSTATISTICS

5.1 INTRODUCTION

The word statistical science is sometimes used for statistical theory and its applications to the natural and social sciences and to science based technology. Statistical Consulting gives the provision of statistical suggestions and / or services to those who request it, applies the statistical techniques to problems in other disciplines. Biometric consulting is the application of statistical expertise in the biological, medicine and public Health - Biostatistical consultant plays a vital role in the various organizations such as Medical institutions, Pharmaceutical Companies, Governmental agencies etc. since, most new statistical methodology arises from realistic problems, being a consulting Biostatistician is a way to learn about new projects requiring advances in statistical methodology. In the present research work the various new advanced techniques for Biostatistics have been developed by using statistical inference.

5.2 TREND TESTS FOR BIOSTATISTICS

In the comparison of counts or proportions across different populations, it is often important to consider the intrinsic ordering of the populations with respect to some particular characteristic. For instance, one may be interested in assessing whether the proportion of women reporting Insomnia increases with age group or whether the number of accidents is increasing over calendar periods. This type of comparison can be accomplished through the application "Trend Test". Trend Test
arise generally within a wide variety of biostatistical applications, such as Bioassays, epidemiologic studies and evaluations of environmental exposures etc. in which a
Dose-Response relationship may be considered. The characteristic of the
population may be measured on a continuous scale, such as an assigned treatment
level, or on an ordinal scale (ordered categorical data), such as age group or initial
severity of a health condition.

Consider \( Y_i \) be a random variable representing the count of interest for the
\( i^{th} \) population; \( X_i \) be quantitative (continuous or ordinal) covariate for the
\( i^{th} \) population; and \( w_i \) be a known design variable for the \( i^{th} \) population (often
relates to the sample or population size).

Now, \( R_i = \frac{Y_i}{W_i} \) represents a rate of a certain event.

The form of data for a Trend Test may be given by

<table>
<thead>
<tr>
<th>Population</th>
<th>Population covariate</th>
<th>Weight</th>
<th>Observed count</th>
<th>Rate</th>
<th>Expected count</th>
</tr>
</thead>
<tbody>
<tr>
<td>( i )</td>
<td>( X_i )</td>
<td>( W_i )</td>
<td>( Y_i )</td>
<td>( R_i )</td>
<td>( E(Y_i) )</td>
</tr>
<tr>
<td>1</td>
<td>( X_1 )</td>
<td>( W_1 )</td>
<td>( Y_1 )</td>
<td>( R_1 )</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>( X_2 )</td>
<td>( W_2 )</td>
<td>( Y_2 )</td>
<td>( R_2 )</td>
<td></td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( K )</td>
<td>( X_K )</td>
<td>( W_K )</td>
<td>( Y_K )</td>
<td>( R_K )</td>
<td>( W_K )</td>
</tr>
</tbody>
</table>

The expected count relates to the covariate through a continuous function \( f \)
(\( x_i \)) may be written as

278
\[ E(Y_i) = W_i f(X_i) \]  \hspace{1cm} (5.2.1)

One may state the thus null hypothesis as, there is no difference in expected counts due to differences in \( X_i \), so that \( H_0 \) may be written as

\[ H_0 : f(X_i) = f(X_j), \ \forall \ i \neq j = 1, 2, \ldots k \]

And \( H_1 : f(X_i) \neq f(X_j), \ \forall \ i \neq j \).

The one sided alternatives may be written as,

\[ H_{11} : f(X_i) < f(x_j), \ \forall \ X_i < X_j, \ \text{an increasing trend alternative}, \]

\[ H_{12} : f(X_i) > f(x_j), \ \forall \ X_i < X_j, \ \text{a decreasing trend alternative}. \]

Consider \( k \) independent random samples drawn from each of the \( i = 1, 2, \ldots, k \) populations.

The function \( f(x) \) may be considered as either linear function of \( x \) say

\[ f(x) = \alpha + \beta, \ \text{or a monotone (increasing or decreasing) continuous function of} \]

\[ \alpha + \beta x \] as \( f(x) = g(\alpha + \beta x) \)  \hspace{1cm} (5.2.2)

For instance \( g(x) = 1 - \exp[-(\alpha + \beta x)] \)  \hspace{1cm} (5.2.3)

Generally, the inverse function \( g^{-1}[f(x)] \) is known as the 'Link function' to be modeled as the linear function \( (\alpha + \beta x) \). For example, the link functions for the Normal, Logistic and extreme value models are respectively given by probit, logit and complementary log – log link functions.
By choosing an appropriate model, the null hypothesis may be stated as,

\[ H_0 : \beta = 0 \] against

\[ H_{i_1} : \beta > 0, \text{ an increasing trend or } H_{i_2} : \beta < 0, \text{ decreasing trend} \]

For the trend test, any of discrete probability distribution may be assumed for the count random variables \( y_i \).

5.3 TREND TEST FOR BINOMIAL COUNTS AND RATES

Suppose that, \( Y_i \sim \text{Binomial distribution and } W_i = n_i, \text{ sample size for the } i^{th} \) population.

Also let \( f(x_i) = p_i = g(\alpha + \beta x_i) \) \hspace{1cm} (5.3.1)

One may have, \( E[Y_i] = n_i, p_i = n_i g(\alpha + \beta X_i) \) \hspace{1cm} (5.3.2)

The \( H_0 \) may be stated as

\[ H_0 : p_1 = p_2 = \ldots = p_k \quad \sim \quad H_{i_1} : p_1 < p_2 < \ldots < p_k \]

and \( H_{i_2} : p_1 > p_2 > \ldots > p_k \).

To test the null hypothesis, first one may obtain the maximum likelihood estimator for \( \beta \) as follows:

Consider the likelihood function for binomial distribution as

\[ L(\alpha, \beta) = \prod_{i=1}^{k} \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i} \]

\[ = \prod_{i=1}^{k} \binom{n_i}{y_i} (\alpha + \beta x_i)^{y_i} [1 - (\alpha + \beta x_i)]^{n_i - y_i} \] \hspace{1cm} (5.3.3)
Here, $g(x)$ is the identity function which is linear. The maximum likelihood (ML) estimators for $\alpha$ and $\beta$ may be obtained by solving the following score equations:

$$S(\hat{\alpha}, \hat{\beta}) = \sum_{i=1}^{k} \frac{1}{y_i - n_i \hat{p}_i} = 0$$  \hspace{1cm} (5.3.4)

Where $\hat{p}_i = g(\hat{\alpha} + \hat{\beta} x_i)$ \hspace{1cm} (5.3.5)

The ML estimator for $\beta$ is solve by

$$\hat{\beta} = \frac{\sum_{i=1}^{k} x_i (y_i - n_i \hat{p})}{\sum_{i=1}^{k} n_i (x_i - \bar{x})^2}$$ \hspace{1cm} (5.3.6)

Where $\hat{p} = \frac{\sum y_i}{\sum n_i}$ and $\bar{x} = \frac{\sum x_i n_i}{\sum n_i}$.

Remarks: (i) For the logistic regression model, $p_i$ may be written,

$$p_i = \frac{\exp \{ \alpha + \beta x_i \}}{1 + \exp \{ \alpha + \beta x_i \}}$$ \hspace{1cm} (5.3.7)

In this case, the score equations may be solved by using some methods in the numerical analysis such as the Newton-Raphson or Fisher scoring Algorithm or iterative technique, to obtain the ML estimation $\hat{\alpha}$ and $\hat{\beta}$.

(ii) the link function $g^{-1}f(x)$ may be considered as a second degree polynomial $(\alpha + \beta x + \gamma x^2)$ and one can obtain the ML estimations $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$.
The score test statistic to test $H_0 : \beta = 0$ is given by:

$$S'(\alpha, \beta) \Gamma^{-1}(\alpha, \beta, \beta) S(\alpha, \beta)$$  \hspace{1cm} (5.3.8)

Where, $\beta_0 = 0$ and $\Gamma^{-1}(\alpha, \beta_0)$ is the inverse of the information matrix, evaluated the null hypothesis, one may express,

$$I(\alpha, \beta) = \begin{bmatrix} I_{\alpha'} & I_{\alpha \beta} \\ I_{\alpha \beta} & I_{\beta'} \end{bmatrix}$$

$$= - \begin{bmatrix} \frac{\partial^2 \log L(\alpha, \beta)}{\partial \alpha^2} & \frac{\partial^2 \log L(\alpha, \beta)}{\partial \alpha \partial \beta} \\ \frac{\partial^2 \log L(\alpha, \beta)}{\partial \beta \partial \alpha} & \frac{\partial^2 \log L(\alpha, \beta)}{\partial \beta^2} \end{bmatrix}$$

$$= \sum_{i=1}^k n_i p_i (1 - p_i) \begin{bmatrix} 1 \\ x_i \\ x_i^2 \end{bmatrix}$$  \hspace{1cm} (5.3.9)

$$= \sum_{i=1}^k \text{Var}(y_i) X_i X_i'$$  \hspace{1cm} (5.3.10)

Where $x_i' = [1 \ x_i]$.

$$I^{-1}(\alpha, \beta) = (I_{\beta'} - I_{\alpha \beta} I_{\alpha}' I_{\alpha \beta})$$

or $I^{-1}(\alpha, \beta) = p (1 - p) \sum_{i=1}^k n_i (x_i - \bar{x})^2$  \hspace{1cm} (5.3.11)

Now, the score test statistic is given by

$$\frac{S^2(\alpha, \beta)}{I^{-1}(\alpha, \beta)}$$  \hspace{1cm} (5.3.12)
\[
= \frac{\sum x_i (y_i - \bar{n}\hat{p})}{\hat{p}(1 - \hat{p}) \sum n_i (x_i - \bar{x})^2}
\]  \hspace{1cm} (5.3.13)

In the matrix form, the score test statistic is given by

\[
Z^2_{\text{score}} = X' [Y - E] [X' \Sigma X]^{-1}
\]  \hspace{1cm} (5.3.14)

Where, \(X = [(x_1 - \bar{x}), \ldots (x_k - \bar{x})]\),

\[Y = [y_1, y_2, \ldots, y_k], \quad E = [n, \hat{p}, \ldots, n_k \hat{p}]
\]

And \(V\) is the diagonal matrix with elements \(n, \hat{p}, (1 - \hat{p})\) on the diagonal.

Here, \(Z^2_{\text{score}}\) follows asymptotically the \(\chi^2\) distribution with one degree of freedom.

The trend test may be frequently used in the analysis of animal bioassay with reference to tumor incidence experiments, in which the animals are randomized to various exposure or dose levels of a drug, chemical or other stimulus and the proportion exhibiting the response of interest is observed. A typical form of bioassay data with binomial counts for lung tumors in female mice exposed 1,2, dichloroethane, for the application of trend test is given by:

<table>
<thead>
<tr>
<th>Dose (mg/kg)x1</th>
<th>Number exposed (w_i)</th>
<th>Number with tumor (y_i)</th>
<th>Percentage with tumor ((y/w) 100)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x_0)</td>
<td>(w_0)</td>
<td>(y_0)</td>
<td>(p_0)</td>
</tr>
<tr>
<td>(x_1)</td>
<td>(w_1)</td>
<td>(y_1)</td>
<td>(p_1)</td>
</tr>
<tr>
<td>(x_2)</td>
<td>(w_2)</td>
<td>(y_2)</td>
<td>(p_2)</td>
</tr>
</tbody>
</table>

283
5.4 TREND TEST FOR POISSON COUNTS AND RATES:

Consider \( Y_1, Y_2, \ldots, Y_k \) be independent poisson random variables and \( x_i \) be an ordered covariate assume that \( E[Y_i] = W_i \cdot f(x_i) \) \hspace{1cm} (5.4.1)

Also consider \( f(x_i) = \lambda_i \), the mean of poisson variable \( y_i \).

One may consider the weights \( W_i \) arising from one of two situations: either,

(i) \( Y_i \) may be the number of rare events during an internal of length \( W_i \),

where \( \lambda_i \) is the event rate per unit time

Or (ii) \( Y_i \) may be the sum of \( W_i \) independent Poisson random variables, i.e.,

\[
Y_i = \sum_{j=1}^{W_i} Y_i
\] \hspace{1cm} (5.4.2)

Where \( Y_{i1}, Y_{i2}, \ldots, Y_{im} \) are identically distributed with mean \( \lambda_i \).

For instance, incidence of AIDS or cancer cases per calendar year; number of injuries or accidents over a set time period, number of bacteria per unit volume of suspension; or number of tumors observed in \( W_i \) animals exposed to dose \( x_i \) in an animal bioassay.

One may test for an increasing or decreasing trend in the means \( \lambda_i = E[y_i]/w_i \) with increasing levels of \( x_i \).

The relationship between \( \lambda_i \) and \( x_i \) may be specified as

\[
\lambda_i = g(\alpha + \beta x_i)
\] \hspace{1cm} (5.4.3)

Frequently under poison regression, one may specify
\[ g(x_i) = e^{\lambda_i} \text{ or } \log(\lambda_i) = \alpha + \beta x_i \quad (5.4.4) \]

For the more general specification given by (5.4.3), the likelihood function may be written as

\[ L(\alpha, \beta) = c(Y_1, Y_2, \ldots, Y_n) \prod_{i=1}^{n} \exp\left\{ -w_i g(\alpha + \beta x_i) \right\} \left\{ g(\alpha + \beta x_i) \right\}^{X_i} \quad (5.4.5) \]

Where, \( c \) is a constant independent of \( \alpha \) and \( \beta \). The ML estimators \( \hat{\alpha} \) and \( \hat{\beta} \) can be obtained by solving score equations,

\[ \frac{\partial \log L(\alpha, \beta)}{\partial \alpha} = 0 \]

and \( \frac{\partial \log L(\alpha, \beta)}{\partial \alpha} = 0 \) simultaneously.

As in the case of Binomial counts, iterative numerical analysis methods may be used to obtain the ML estimators \( \hat{\alpha} \) and \( \hat{\beta} \).

The score test statistic for testing \( H_0 : \beta = 0 \) is given by

\[ Z^2_{\text{score}} = \frac{\sum_{i=1}^{n} x_i (Y_i - W_i \bar{Y})}{\bar{Y} \sum_{i=1}^{n} w_i (X_i - \bar{X})^2} \quad (5.4.6) \]

Where, \( \bar{Y} = \sum_{i=1}^{n} Y_i / \sum_{i=1}^{n} w_i \).

\( Z^2 \) Poisson follows asymptotically \( \chi^2 \) distribution with one degree of freedom.
A typical form to data with poisson counts for new cases of melanoma and lung, stomach, reported between a time interval for six age groups, along with the person - years of employment in each age group. In these cases, the variance of poisson counts appears to be inflated relative to the mean.

<table>
<thead>
<tr>
<th>Age group mid point $x_i$</th>
<th>Number of observed melanoma cases $y_i$</th>
<th>Person - years of exposure $w_i$</th>
<th>Observed rate per 100000 person - years $(y/w_i) \times 10^5$</th>
<th>Predicted rate per 100000 person - years $(\lambda_i \times 10^5)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>$y_1$</td>
<td>$w_1$</td>
<td>$r_1$</td>
<td>$(\lambda_1 \times 10^5)$</td>
</tr>
<tr>
<td>$x_2$</td>
<td>$y_2$</td>
<td>$w_2$</td>
<td>$r_2$</td>
<td>$(\lambda_2 \times 10^5)$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$x_6$</td>
<td>$y_6$</td>
<td>$w_6$</td>
<td>$r_6$</td>
<td>$(\lambda_6 \times 10^5)$</td>
</tr>
</tbody>
</table>

Here the predict rate is based on the fitting of the model, given in (5.4.3)

5.5 TESTS FOR THE EQUALITY BETWEEN THE POISSON COUNTS

5.5.1 Test for comparing the equality of Two Poisson counts

Suppose that $Y_1$ and $Y_2$ be the two poisson counts taken over time periods $W_1$ and $W_2$ respectively. The two average frequencies or rates are given by $R_1 = (y_1/w_1)$ and $R_2 = (y_2/w_2)$. To test for the equal rates, the test statistic is given by
\[ t^2 = \frac{\left[ R_1 - R_2 \right]^2}{\frac{R_1}{W_1} + \frac{R_2}{W_2}} - \chi_i^2 \quad (5.5.1) \]

For large number of counts, the normal approximation is given by

\[ Z = \frac{R_1 - R_2}{\sqrt{\frac{R_1}{W_1} + \frac{R_2}{W_2}}} \sim N(0,1) \quad (5.5.2) \]

5.5.2 Test For Equality Of More Than Two Poisson Counts

(i) Equal Timings for Poisson Counts

Consider \( Y_i \) be the \( i \)th count and the same times to obtain the counts are all the same.

To test the null hypothesis, \( H_0 : Y_1 = Y_2 = \ldots = Y_k = Y \) (say) the test statistic is given by

\[ \chi^2 = \sum_{i=1}^{k} \frac{(Y_i - \bar{Y})^2}{\bar{Y}} - \chi_{k-1}^2 \quad (5.5.3) \]

Where \( \bar{Y} = \frac{1}{k} \sum_{i=1}^{k} Y_i \) \quad (5.5.4)

(ii) Unequal Timings for Poisson Counts

Suppose that the time to obtain the \( i \)th count \( y_i \) be \( w_i \), \( i = 1, 2, \ldots, k \). define,

\[ \frac{\sum_{i=1}^{k} Y_i}{\sum_{i=1}^{k} w_i} \quad (5.5.5) \]

To test for the equality between the \( k \) poisson counts, the test statistic is given by
\[ t = \sum_{i=1}^{k} \left( Y_i - \bar{W}_i \bar{R}_i \right)^2 - \chi^2_4, \]  
(5.5.6)

5.6 LOGISTIC DISTRIBUTION

A random variable \( X \) is said to have Logistic distribution with mean \( \mu \) and variance \( \sigma^2 \), if it has a cumulative distribution function,

\[
F(x, \mu, \sigma) = 1 + \exp \left( -\frac{\pi (x - \mu)}{\sigma \sqrt{3}} \right) 
- \infty < x < \infty, \quad (5.6.1)
\]

\[-\infty < \mu < \infty \]
\[\sigma > 0\]

and probability density function related to its distribution function,

\[
f(x, \mu, \sigma) = \frac{n}{\sigma \sqrt{3}} F(x, \mu, \sigma) [1 - F(x, \mu, \sigma)] \quad (5.6.2)
\]

Alternatively, these functions may be expressed as

\[
F(x, \mu, \sigma) = \frac{1}{2} \left( 1 + \tanh \frac{\pi (x - \mu)}{2(\sigma \sqrt{3})} \right) \quad (5.6.3)
\]

\[
\text{and } F(x, \mu, \sigma) = \frac{n}{4 \sigma \sqrt{3}} \text{sech}^2 \left( \frac{\pi (x - \mu)}{2 (\sigma \sqrt{3})} \right) \quad (5.6.4)
\]

Properties of Logistic Distribution

1. The density \( f(x, \mu, \sigma) \) is bell shaped and symmetrical with heavier tails than a normal density with the same mean and variance.
2. The canonical form of the logistic distribution, which corresponds to the random variable $z$, with mean $\mu = 0$ and variance $\sigma^2 = \frac{\pi^2}{3}$ and has cumulative distribution function and probability density functions.

\[
G(z) = \frac{1}{1 + e^{-z}} \quad \text{(5.6.5)}
\]

and $g(z) = G(z)(1 - G(z)) \quad \text{(5.6.6)}$

Equation (5.6.6) and therefore equation (5.6.2) characterizes the logistic distribution and it is equivalent to the linearity of the transformation (known as logit)

\[
\log \left[ \frac{G(z)}{1 - G(z)} \right] = z \quad \text{(5.6.7)}
\]

3. The most popular application of the logistic distributions is the logit in the content of modeling Quantal response data and performing Logistic regression.

4. The distribution function of the standardized random variable $z/(\pi \sqrt{3})$ is very close to the standard normal distribution.

5. The sum of independent logistic random variables is not a logistic random variable.

6. The characteristic function of $z$ is given by

\[
\phi_z(t) = \Gamma (1 - it) \Gamma (1 + it) = \prod_{i=1}^{n} \left(1 - \frac{t^2}{i^2} \right)^{-1} \quad \text{(5.6.8)}
\]
7. The absolute moments are given by

$$E \left| z \right|^k = 2 \Gamma (k + 1) \left[ 1 - \frac{1}{2^{k+1}} \xi(k) \right]$$

(5.6.9)

8. The Logistic distribution can be obtained from a mixture of the extreme value distributions and the exponential distribution. The Logistic distribution is infinitely divisible.

9. If a random variable $Y$ is uniformly distributed on $[0,1]$ then the logit transformation of $Y$, say $\log \left[ \frac{y}{1-y} \right]$ has the logistic distribution function $G$.

5.7 SPECIFICATION OF LOGISTIC REGRESSION MODEL

Generally linear regression model is used to approximate the relationship between a continuous response variable and a set of predictor variables. However the response variable is often categorical rather than continuous variable, for such cases, linear regression is not appropriate, but the biostatistician can turn to an analogous method, Logistic regression, which is similar to linear regression in many ways.

Logistic regression refers to techniques for describing the relationship between a categorical response variable and a set of predictor variables. In other words the goal of a logistic regression analysis is to find the best fitting and most parsimonious, yet biologically reasonable, model to describe the relationship between a response variable and a set of predictor or independent variables.
Generally the response variable in the Logistic regression model is categorical and usually Binary or dichotomous variable.

There are two main reasons for choosing the Logistic regression model by Biostatistician. These are:

(i) In the mathematical point of view, it is an extremely flexible and easily used function;

(ii) It trends itself to a biologically meaningful interpretation.

Suppose that $\pi(x) = E[y/x]$ be the conditional mean of $y$ given $x$. The logistic regression model is given by

$$
\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}
$$

(5.7.1)

The logit transformation $g(x)$ in terms of $\pi(x)$ is given by

$$
g(x) = \ln \left( \frac{\pi(x)}{1 - \pi(x)} \right) = \beta_0 + \beta_1 x
$$

(5.7.2)

The logit $g(x)$ has many of the desirable properties of linear regression model. The logit $g(x)$ is linear in its parameters, may be continuous and may range from $-\infty$ to $+\infty$ depending on the range of $x$.

In the linear regression analysis, one may assume, that an observation of the outcome variable may be expressed as

$$
y = E\{y/x\} + \epsilon
$$

(5.7.3)
Such that the error variable $\epsilon$ follows a normal distribution with mean zero and some constant variance. That is constant across the levels of the independent variable. This implies that the conditional distribution of the dependent variable ($y$) given $x$ is normal with mean $E[y|x]$ and constant variance.

In the case of Logistic regression model, one may express the value of the response variable given $x$ as

$$Y = \pi(x) + \epsilon$$

(5.7.4)

Here, $\epsilon$ may assume one of are two possible values. If $y = 1$ then $\epsilon = 1 - \pi(x)$ with probability $\pi(x)$ and if $y = 0$ then $\epsilon = -\pi(x)$ with probability $1 - \pi(x)$. i.e., $\epsilon$ has a distribution with mean zero and variance equal to $\pi(x)[1 - \pi(x)]$.

Thus, the conditional distribution of the response variable follows a Binomial distribution with probability given by the conditional mean $\pi(x)$.

A simple example of logistic regression is as follows: suppose that medical researchers are interested in exploring the relationship between patient age ($x$) and the presence (1) or absence (0) of a particular disease ($y$). For a number of patients, generally, this relationship shows logistic regression.
5.8 Estimating the Logistic Regression Model

Consider \((x_i, y_i), i = 1, 2, \ldots, n\) be a sample of \(n\) independent pairs of observations on two variables \(X\) and \(Y\), where \(y_i\) denotes the value of a dichotomous response variable and \(x_i\) denotes the value of the independent variable for the \(i^{th}\) subject.

Assume that \(Y\) has been coded as 0 or 1 representing the absence or presence of the characteristic respectively.

Write the logistic regression model as

\[
\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{5.8.1}
\]

To estimate the logistic regression model, the maximum likelihood method of estimation may be used to obtain optimum estimators for \(\beta_0\) and \(\beta_1\).

Suppose that \(Y\) is coded as 0 or 1, then one may have

\[
\pi(x) = P(y = 1|x) \text{ and } 1 - \pi(x) = P(y = 0|x)
\]

For the pairs \((x_i, y_i)\), where \(y_i = 1\) and \(y_i = 0\), the likelihood function may be expressed as

\[
L(\beta_0, \beta_1) = \prod_{i=1}^{n} \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \tag{5.8.2}
\]

\[
\Rightarrow \ln L(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i \ln[\pi(x_i)] + (1-y_i) \ln[1-\pi(x_i)]] \tag{5.8.3}
\]
Under the method of maximum likelihood estimation, one may obtain the maximum likelihood equations as
\[
\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \quad (5.8.4)
\]
and \[
\sum_{i=1}^{n} x_i [y_i - \pi(x_i)] = 0 \quad (5.8.5)
\]

Since, the maximum likelihood equations are nonlinear in parameters, iterative methods may be used to obtain the maximum likelihood estimators \( \hat{\beta}_c \) and \( \hat{\beta}_l \) for \( \beta_c \) and \( \beta_l \) respectively. In particular, the solutions to the equations (5.8.4) and (5.8.5) may be obtained by using a generalized weighted least squares estimation.

By substituting \( \hat{\beta}_c \) and \( \hat{\beta}_l \), the maximum likelihood estimate of \( \pi(x_i) \) is obtained as \( \hat{\pi}(x_i) \).

Hence, \( \hat{\pi}(x_i) \) gives the fitted or predicted value for \( y_i \).

Since, \( P(y = 1 / x_i) = \pi(x_i) \)

From (5.8.4), one may obtain
\[
\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i) \quad (5.8.6)
\]

\( \Rightarrow \) Sum of the observed values of \( y_i \) = sum of the predicted values of \( y_i \).
5.9 ASSESSING THE SIGNIFICANCE OF EFFECT OF INDEPENDENT VARIABLE ON RESPONSE VARIABLE

For the purposes of assessing the significance of an independent variable, one may state the null hypothesis as \( H_0 : \beta_i = 0 \)

To test the \( H_0 \), one may use the following two test statistics:

(i) Likelihood Ratio Test Statistic:

\[
G = -2 \ln \left[ \frac{\text{Likelihood without predictor variable}}{\text{Likelihood with the predictor variable}} \right] \tag{5.9.1}
\]

Under \( H_0 \), \( G \) follows \( \chi^2 \) distribution with one degree of freedom.

(ii) Wald Test Statistic:

\[
W = \frac{\beta_i}{\text{SE}(\hat{\beta})} \tag{5.9.2}
\]

Under \( H_0 : \beta_i = 0 \), follows a standard normal distribution.

Both the test statistics Likelihood Ratio and Wald test statistics require the computation of the maximum likelihood estimate for \( \beta_i \). For a single predictor, \( \hat{\beta} \) may be easily obtained. However, for large data sets with many variables, the iterative computational methods may be used to obtain the maximum likelihood estimators for the parameters of the logistic regression model.

For the specific case of a single predictor variable, when the variable, is not in the model, the maximum likelihood estimate of \( \beta_0 \) in \( \ln \left( \frac{n_i}{n_o} \right) \),
Where \( n_1 = \Sigma y_i \) and \( n_0 = \Sigma (1 - y_i) \), the predicted value is constant \( \hat{\beta}_1 \).

Now, the value of \( G \) is given by

\[
G = -2 \ln \left( \frac{\left( \frac{n_1}{n} \right)^{n_1} \left( \frac{n_0}{n} \right)^{n_0}}{\prod_i \hat{\pi}_i^{n_1} (1 - \hat{\pi}_i)^{(1-n_i)}} \right)
\]

or \( G = 2 \left[ \sum_{i=1}^{n} [y_i \ln (\hat{\pi}_i) + (1 - y_i) \ln(1 - \hat{\pi}_i)] - [n_1 \ln(n_1) + n_0 \ln(n_0) - n \ln(n)] \right] \)

Under \( H_0 = \beta_1 = 0 \), the Likelihood Ratio test statistic \( G \) follows a \( \chi^2 \) distribution with one degree of freedom.

### 5.10 SPECIFICATION OF MULTIPLE LOGISTIC REGRESSION MODEL

In the multiple logistic regression model, more than one predictor variable is used to classify the binary response variable. Consider a set of \( k \) predictor variables, which is denoted by a vector \( X^i = (x_1, x_2, \ldots, x_k) \) and assuming that each of these predictor variables is at least interval scaled.

Suppose that the conditional probability that the response is present to denoted by \( P(y=1/X) = \pi(x) \). \hspace{1cm} (5.10.1)

Now the logit of the Multiple logistic Regression model is specified as

\[
g(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \hspace{1cm} (5.10.2)
\]

Here, the multiple logistic regression model is given by

\[
\pi(X) = \frac{\exp[r(X)]}{1 + \exp[r(X)]} \hspace{1cm} (5.10.3)
\]
When some of the predictor variables are discrete, nominal scaled variables such as Race, Sex, Treatment, group, and so forth, then it is in appropriate to include them in the model as if they were interval scaled. In this case, a set of Design variables or dummy variables should be used. Most Logistic regression software will provide the design variables and some programs have a choice of several different techniques. In general, if a nominal scaled variable has $p$ possible values then $(p-1)$ design variables will be needed. Suppose that, the $i^{th}$ predictor $x_i$ has $p_i$ levels $(p_i - 1)$ design variables may be denoted by $D_{p_i}$ and their coefficients are denoted by $\beta_q$, $q = 1, 2, \ldots, (p_i - 1)$.

Now the Logit for a model with $k$ predictors and the $j^{th}$ predictor being discrete is given by

$$g(X) = \beta_0 + \beta_1 X_1 + \ldots + \sum_{p_i-1}^{k} \beta_{p_i} X_{p_i}$$

(5.10.4)

5.11 ESTIMATING THE MULTIPLE LOGISTIC REGRESSION MODEL

Consider a sample of $n$ independent pairs of observations $(X_i, Y_i)$, $i = 1, 2, \ldots, n$. As in the case of Univariate Logistic Regression model, the estimation of Multiple Logistic Regression model requires to obtain the estimators of the parametric vector $\beta^i = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$. The maximum likelihood method of estimation is used as in the univariate case. The likelihood function is nearly identical to that given in the univariate situation, with the only change being that $\pi(x)$ is defined as
\[ \pi(X) = \frac{\exp[g(x)]}{1 + \exp[g(x)]} \quad (5.11.1) \]

Under the maximum likelihood estimation, by maximizing differentiating
the log likelihood function with respect to the \((k+1)\) parameters, one may obtain
the following \((k+1)\) maximum likelihood equations:

\[ \sum_{i=1}^{n} [Y_i - \pi(X_i)] = 0 \quad (5.11.2) \]

and

\[ \sum_{i=1}^{n} X_{ij} [Y_i - \pi(X_i)] = 0, \quad j = 1, 2, \ldots, k \quad (5.11.3) \]

Suppose that \(\hat{\beta}\) be the solution vector to these maximum likelihood
equations. Then the estimated multiple logistic regression model is given by
\(\hat{\pi}(x_i)\), which is computed using \(\hat{\beta}\) and \(X_i\).

5.12 ASSESSMENT OF THE SIGNIFICANCE OF THE EFFECTS OF
PREDICTOR VARIABLES IN THE MULTIPLE LOGISTIC
REGRESSION MODEL

After estimating the model, the most scrutiny is the process of assessment of the
multiple logistic regression model. Under the process of assessment one may test
for the significance of the predictor variables in the multiple logistic regression
model. As in the invariable situation, the Likelihood Ratio test statistic \(G\) may be
used to assess the significance of the predictor variables. The only difference is
that the estimated values \( \hat{\pi} \), under the model are based on the vector of \((k+1)\) parameters \( \hat{\beta} \).

Under \( H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \), the test statistic \( G \) follows the \( \chi^2 \) distribution with \( k \) degrees of freedom.

i.e., \( G = -2 \ln \left[ \frac{\text{Likelihood without the predictors}}{\text{Likelihood with the predictors}} \right] \) \hspace{1cm} (5.12.1)

The Wald test statistics for testing the assessment of the multiple logistic regression model are given by

\[
W_j = \frac{\hat{\beta}_j}{S \Sigma (\hat{\beta}_j)} \sim N (0,1) \hspace{0.5cm} j = 1, 2, \ldots, k \hspace{1cm} (5.12.2)
\]

To provide an example of estimating a multiple logistic regression model, consider the data for the low birth weight study containing the predictors Age, weight of the mother at her last menstrual period, Race and Number of physician visits during the first trimester of the pregnancy. The main goal of this study is to identify the risk factors associated with giving birth to a low birth weight baby.

5.13 ODDS RATIO: INTER RELATION OF COEFFICIENTS IN LOGISTIC REGRESSION MODEL.

After fitting a model the emphasis shifts from the computation and assessment of significance of estimated coefficients to interpretation of their values. Generally, interpretation involves two issues: (i) determining the functional relationship between the response variable and the predictor variable
(ii) appropriately defining the unit of change for the predictor variable. In the Logistic regression model, the estimate of the slope coefficient of predictor represents the change in the logit for a change of one unit in the predictor variable \(X\), i.e.,

\[
\beta_j = g(x_j + 1) - g(x_j)
\]  

(5.13.1)

This implies that \(\beta_j\) gives the change in the log odds for an increase of '1' until \(X_j\). Most often the value of '1' will not be biologically very interesting. Proper interpretation of the coefficient is a logistic regression model depends on being able to place meaning on the difference between the logits.

Consider the predictor \(x\) is coded as either 0 or 1. This indicates that \(x\) be a Dichotomous predictor variable. Under this model, there are two values of \(\pi (x)\) and equivalently two values of \(1 - \pi (x)\). These values may be displayed in the following (2x2) table.

Values of the Logistic Regression model when the predictor variable is

Dichotomous variable

<table>
<thead>
<tr>
<th>Response Variable (Y)</th>
<th>Predictor Variable (X)</th>
<th>(x = 1)</th>
<th>(x = 0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(y = 1)</td>
<td>(\pi (1) = \frac{\exp (\beta_0 + \beta_1)}{1 + \exp (\beta_0 + \beta_1)})</td>
<td>(\pi (0) = \frac{\exp (\beta_0)}{1 + \exp (\beta_0)})</td>
<td></td>
</tr>
<tr>
<td>(y = 0)</td>
<td>(1 - \pi (1) = \frac{1}{1 + \exp (\beta_0 + \beta_1)})</td>
<td>(1 - \pi (0) = \frac{1}{1 + \exp (\beta_0)})</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1.0</td>
<td>1.0</td>
<td></td>
</tr>
</tbody>
</table>

300
The odds ratio is defined as the ratio of the odds for \( x = 1 \) to the odds for \( x = 0 \) and is given by

\[
\text{OR} = \psi = \left[ \frac{\pi(1)/[1 - \pi(1)]}{\pi(0)/[1 - \pi(0)]} \right]
\]

(5.13.2)

The log Odds Ratio or log odds is given by

\[
\ln(\psi) = \ln \left[ \frac{\pi(1)/[1 - \pi(1)]}{\pi(0)/[1 - \pi(0)]} \right] = g(1) - g(0).
\]

(5.13.3)

Thus log odds gives the logit differences one may have,

\[
g(1) = \ln \left( \frac{\pi(1)}{1 - \pi(1)} \right)
\]

(5.13.4)

and \( g(0) = \ln \left( \frac{\pi(0)}{1 - \pi(0)} \right) \)

(5.13.5)

By substituting the expressions for \( \pi(0) \) and \( \pi(1) \) in the above logits, the Odds Ratio is given by

\[
\psi = \exp(\beta_1)
\]

Also the log odds is given by,

\[
\ln(\psi) = \ln \exp(\beta_1) = \beta_1.
\]

Remark: The log odds for a change of \( c \) units in \( x \) is obtained from the logit difference \( g(x+c)-g(x) = c\beta_1 \) and the associated Odds Ratio is obtained by exponentiating this difference,

\[
\psi(c) = \psi(x + c, x) = \exp(c\beta_1)
\]

(5.13.6)
An estimate of the log odds may be obtained by substituting the maximum likelihood estimate $\hat{\beta}_1$ or $\beta$.

5.14. SPECIFICATION OF POISSON REGRESSION MODEL

For response variables that have count or frequencies as outcomes, it is often reasonable to assume an underlying Poisson distribution. The impact of predictor variables on the means of these response variables can be described by same regression function known as Poisson Regression model. Poisson regression models are widely applicable class of models particularly useful in Biostatistics, emerged in the late 1970’s.

As an example, consider a randomly chosen household members from a random sample of a town, were asked to note which stressful events had occurred within the last one and half year (18 months) and to report the month of occurrence of these events.

Distribution by months prior to interview of stressful events reported from subjects:

<table>
<thead>
<tr>
<th>Month</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>...</th>
<th>...</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Recalls (y)</td>
<td>$Y_1$</td>
<td>$Y_2$</td>
<td>$Y_3$</td>
<td>...</td>
<td>...</td>
<td>$Y_{18}$</td>
</tr>
</tbody>
</table>

To specify a Poisson regression model, assume that (i) the number of recalls is a random variable $Y$ follows Poisson distribution with mean $\mu$ and (ii) $\mu$ is same function of $X$, the number of months before interview.
To specify the basic version of a Poisson regression model, suppose that $Y_1, Y_2, \ldots, Y_n$ be response variables on which one may have observations $y_1, y_2, \ldots, y_n$, assumed to be independently distributed Poisson variates with means $\mu_1, \mu_2, \ldots, \mu_n$

$$i.e. \ f(y_i/\mu_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \quad (5.14.1)$$

The systematic component of the model is specified by some regression function $\eta$, depending on regression parameters $\beta_1, \beta_2, \ldots, \beta_k$, with each component relating values $X_{i1}, X_{i2}, \ldots, X_{ik}$ of predictor variables to respective means.

$$i.e., \mu_i = \eta_i(\beta) = \eta_i(x_{i1}, x_{i2}, \ldots, x_{ik}; \beta_1, \beta_2, \beta_k) \quad (5.14.2)$$

Assuming that $E(y_i) = \mu_i$ and there exists a function $g(\mu_i) = \eta_i$, then $\eta$ is a linear predictor.

or $\log(\mu_i) = \sum_{j=1}^{k} x_{ij} \beta_j \quad i = 1,2,\ldots,n \quad (5.14.3)$

In this situation, the function $g(.)$ is called the link function and the model specified in this manner is an instance of a Generalized Linear Model.

For $\eta(\beta) = \exp \left( \sum_{j=1}^{k} x_{ij} \beta_j \right) \quad (5.14.4)$
One may have the log linear model
\[ \log \mu_i = \sum_{j=1}^{k} x_{ij} \beta_j \]  
(5.14.5)

Sometimes, one may specify the Poisson regression model with link function as
\[ Y_i = E(Y_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n \]  
(5.14.6)

\[ f(y_i) = \frac{\varepsilon_i \beta^\eta}{y_i}, \quad y_i = 0, 1, 2, \ldots, \mu > 0 \]  
(5.14.7)

Assuming that \( E(y_i) = \mu_i \) and link function \( g(\mu_i) = \eta_i \), a linear predictor,

or \( g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} \)  
(5.14.8)

or \( g(\mu_i) = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \ldots + \beta_k z_{ik} \)  
(5.14.9)

Where \( z_{i1} = x_{i1}, z_{i2} = x_{i2}, \ldots, z_{ik} = x_{ik} \)

For Identity link function, one may write
\[ E(y_i) = \mu_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \ldots + \beta_k z_{ik} \]  
(5.14.10)

For the model specified by (5.14.1), (5.14.2) and (5.14.3), the log likelihood function may be written as
\[ \ln \gamma(\beta) = \sum_{i=1}^{n} \{y_i \log \{\eta_i(\beta)\} - \eta_i(\beta) - \log(y_i!)} \]  
(5.14.11)

If the model is log linear then one may write
\[ \ln \gamma(\beta) = \sum_{j=1}^{k} \left( \sum_{i=1}^{n} x_{ij} y_i \right) \beta_j - \sum_{i=1}^{n} \exp \left( \sum_{j=1}^{k} x_{ij} \beta_j \right) - \sum_{i=1}^{n} \log(y_i!) \]  
(5.14.12)
The maximum likelihood method of estimation may be often used to obtain estimates for the unknown parameters of the Poisson regression model.

Likelihood equations for the maximum likelihood method of estimation are given

\[(i) \sum_{i=1}^{n} \frac{\partial}{\partial \beta_j} \eta_i(\beta) \frac{1}{n(\beta)} [y_i - \eta_i(\beta)] = 0 \quad \text{for} \quad (5.14.11)\]

\[(ii) \sum_{i=1}^{n} x_{ij} y_i - \exp \left( \sum_{i=1}^{n} x_{ij} \beta_j \right) = 0 \quad \text{for} \quad (5.14.12)\]

The solutions of these likelihood equations can be obtained by Fishers Scoring Algorithm method or an Iterative Reweighted Least Squares Method, where in each step of the iterative algorithm, a weighed Least squares problem is to be solved.

As an example, consider the data from a Radioimmunoassay, a widely used technique to measure the quantity of a given biological substance by identifying the amount of a radioactive labeled antibody from a reagent by subsamples of increasing concentration. Here, the response variable is the amount of radioactive material remaining measured in counts per minute. If these are not large, an underlying Poisson distribution for the counts may be assumed. For counts \(y_1, y_2, \ldots, y_n\) and concentrations \(x_1, x_2, \ldots, x_n\), the Poisson regression model may be specified to describe the relationship between mean counts and concentrations.
5.15 GOODNESS OF FIT FOR FITTING OF THE POISSON REGRESSION MODEL

A test statistic capable of measuring the amount of support given by the data to a particular value of the parameter compared to its maximum likelihood estimate is the 'Deviance statistic' which is given by

\[ D_y(\beta) = -2 \ln \frac{L_y(\beta)}{L_y(\hat{\beta})} \]

\[ = -2 \left[ \ln L_y(\beta) - \ln L_y(\hat{\beta}) \right] \]

\[ = -2 \sum_{i=1}^{n} y_i \ln \left( \frac{\eta_i(\beta)}{\eta_i(\hat{\beta})} \right) - \left[ \eta_i(\beta) - \eta_i(\hat{\beta}) \right] \]

(5.15.1)

Here, the Deviance provides a measure of distance between the model described by \( \beta \) and the model characterized by the most likely estimate \( \hat{\beta} \). Assuming that \( \beta \) to be the true parametric vector than \( D_y(\beta) \) follows asymptotically \( \chi^2 \) distribution with \( k \) degrees of freedom, where \( k \) is the number of parameters in \( \beta \).

To test for the goodness of fit, one may write the test statistic as

\[ D_y(\beta) = 2 \sum \left[ y_i \ln \left( \frac{Y_i}{\eta_i(\hat{\beta})} \right) - [Y_i - \eta_i(\hat{\beta})] \right] \]

(5.15.2)

For large sample distribution of \( D_y(\beta) \) the Deviance itself can be approximated by
\[ \chi^2 = \sum \left( \frac{(y_i - \eta_i(\hat{\beta}))^2}{\eta_i(\hat{\beta})} \right) \]  

(5.15.3)

Which is known as the Pearson's goodness of fit statistic.

Remark: The likelihood Ratio Test and Wald test may also be used to test for performing significance of fitting of Poisson regression model.

5.16 TEST FOR COMPARING SEQUENTIAL CONTINGENCIES ACROSS K GROUPS IN A THREE WAY (2X2XK) MANIFOLD CONTINGENCY TABLE USING LOG ODDS RATIO:

Consider a 2x2xk manifold contingency table for three attributes A, B, C with two, two and k levels respectively.

<table>
<thead>
<tr>
<th></th>
<th>C₁</th>
<th>C₂</th>
<th>……</th>
<th>Cₖ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B₁</td>
<td>B₂</td>
<td>……</td>
<td>B₁</td>
</tr>
<tr>
<td>A₁</td>
<td>O₁₁</td>
<td>O₁₂</td>
<td>O₁₁₂</td>
<td>O₁₂₂</td>
</tr>
<tr>
<td>A₂</td>
<td>O₂₁</td>
<td>O₂₂</td>
<td>O₂₁₂</td>
<td>O₂₂₂</td>
</tr>
</tbody>
</table>

One may wish to test the significance of the differences in sequential connections or contingencies across k groups. A fundamental distinction may be made between the measures of association in contingency tables which are either
sensitive or insensitive to the marginal (row) totals. Logit transformation provides 
a measure to the marginal total. The Logit is defined by

\[
\text{Logit}(P) = \ln \left( \frac{P}{1-P} \right)
\]  
(5.16.1)

Now, one may define log odds ratio statistic for each 2x2 contingency table 
of k levels of attribute C as

\[
Q_i = \log \left[ \frac{O_{1i} O_{2i}}{O_{1i} O_{2i}} \right], \quad i = 1, 2, 3, \ldots, k
\]  
(5.16.2)

In order to test whether \(Q_i\)'s different across k groups, one may use the \(\chi^2\) test statistic as

\[
\chi^2 = \sum_{i=1}^{k} \left( \frac{(Q_i - \bar{Q})^2}{Q} \right) - \chi^2_{k-1}
\]  
(5.16.3)

Where \(\bar{Q} = \sum_{i=1}^{k} Q_i / k\)

If the \(\chi^2\) statistic be significant of a given level of significance, one may 
test the difference between any two groups contingencies by using the Z-test 
statistic, which is given by

\[
Z = \frac{Q_i - Q_j}{\sqrt{\frac{1}{O_{11} O_{21}} + \frac{1}{O_{1j} O_{2j}}}}, \quad i \neq j = 1, 2, \ldots, k
\]  
(5.16.4)

Here, \(Z\) follows \(N(0,1)\)
As an example, consider, Husband's (H) antecedent behaviour taking one of two values: H=1 for negative effect and H=0 for positive effect; and similarly, Spouse's (W) consequent behavior taking one of two values: W=1 for negative effect and W=0 for positive effect; in k different types of couples in a place. Here, one may apply the proposed test to the following table:

Table: Distribution of k Groups of couples according to their Antecedent behaviour

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>......</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Husband</td>
<td>Spouse (W)</td>
<td>Spouse (W)</td>
<td>Spouse (W)</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.17. A NEW TREND TEST BASED ON REVERSE ARRANGEMENT

Consider a sequence of n observed values of a random variable \( x_1, x_2, \ldots, x_n \). The \( i^{th} \) reverse arrangement \( R_i \) is the number of times that \( x_i > x_j \) where \( x_i \neq x_j \).

The total number of reverse arrangements is given by

\[
R = \sum_{j=1}^{n} R_j
\]  
(5.17.1)

Where \( k \) is the number of distinct observations in seq.

The number \( R \) may range between 0 and \( \frac{k(k+1)}{2} \)

309
Define mean and variance of \( R \) as:

\[
\text{Mean} = \mu_R = E[R] = \frac{\sum R_i}{K} \quad (5.17.2)
\]

and variance:

\[
\sigma_R^2 = \frac{\sum (R_i - \mu_R)^2}{K} \quad (5.17.3)
\]

This test can be used for evaluating whether the given sequence of ordered data is derived from independent observations of the same random variable by detecting whether a significant trend underlines the observations. It is a nonparametric test, making no assumptions about the distribution of the observations in the data and about a model for the possible trend.

If an increasing or a decreasing trend underlines the data, one may expect \( R \) to be respectively greater or lower than \( \mu_R \). The distribution of \( R \) is derived by Kendall and Stuart (1967) and critical values for \( K \) between 10 and 100 is tabulated by Bendat and Piersol (1986). However, the tendency to normality is extremely rapid. When \( K \geq 14 \), the test statistic is given by

\[
\frac{R - \mu_R}{\sigma_R} \sim N(0,1) \quad (5.17.4)
\]

Here, \( Z \) approximately follows a standard normal distribution and it can be used to reject the null hypothesis with little loss of accuracy.

As an example, consider a series of Diastolic Blood pressure values measured daily in a patient during one mother monitoring period. The proposed test may be applied to test the null hypothesis of independence at a given level of signi...
5.18 A NEW TEST FOR COMPARING SEQUENTIAL CONTINGENCIES ACROSS n GROUPS OF THREE WAY MANIFOLD CLASSIFICATION USING COEFFICIENTS OF CONTINGENCY.

Consider three attributes A, B and C with levels p, q and k respectively.

Write the three way pxqxk manifold contingency table as

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>..........</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>B</td>
<td>..........</td>
<td>B</td>
</tr>
<tr>
<td>1 2</td>
<td>.......... q</td>
<td>1 2</td>
<td>.......... q</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Denote the observed \((i, j, k)^{th}\) cell frequency as \(O_{ijk}\) where, \(i = 1, 2, \ldots, p ; j = 1, 2, \ldots, q; k = 1, 2, \ldots, r\).

First obtain the \(\chi^2\) statistic values for each of the \(r\) groups of \((pxq)\) contingency tables by using \(\chi^2\) test for the independence of two attributes A and B.

\[
\chi_i^2 = \sum_i \sum_j \frac{(O_{ijk} - E_{ijk})^2}{E_{ijk}}, \quad k = 1, 2, \ldots, r
\]

Where \(E_{ijk}\)'s are the expected cell frequencies correspond to the observed cell frequencies \(O_{ijk}\)'s. Find the coefficients of contingency for the \(r\) groups as...
\[ C_k = \sqrt{\frac{\chi^2(k)}{\chi^2 + N_k}} \quad k = 1, 2, \ldots, r \]

Where \( N_k \) is the total frequency of \((p \times q)\) contingency table in the \( k^{th} \) group.

To test the significance of the difference in sequential contingencies (connections) across \( r \) groups, one may use the test statistic as

\[ \chi^2 = \sum_{k=1}^{r} \left( \frac{c_k - \bar{c}}{\bar{c}} \right)^2 \quad - \chi^2 (r - 1) \]

Where \( \bar{c} = \sum_{k=1}^{r} c_k \)

Remark: The Analysis of variance technique for Three way classified data with single observation per cell can be applied to \( p \times q \times r \) manifold contingency table after taking Angular Transformation to the cell frequencies.

5.19. MEASURES OF DIAGNOSTIC ACCURACY IN BIOSTATISTICS.

The field of diagnostic medicine is complex in Nature. This is due to the fact that the process of medical diagnosis is dynamic and it is difficult to formulate straightforward scientific questions are enable to simple study designs. In interpreting the result of an individual test the doctor must consider the context in which it is applied. Many diagnostic tests, especially Radiologic and psychometric tasks, are evaluated subjectively, leading typically to test results that are classified in ordinal categories which are defined verbally. One may find two different kinds of errors of diagnosis, False Positives and False Negatives. Generally, Medical researchers have been satisfied with evaluating tests on the basis of measures of diagnostic accuracy.
Consider a diagnostic test result (outcome) denoted by $x$ and let $D$ be a Binary indicator of the True disease status, where,

$$D = 1 \Rightarrow \text{Presence of disease}$$

$$D = 0 \Rightarrow \text{Absence of disease}$$

Let $F_x(x) = P[X \leq x / D = 1]$ be the Distribution of the test result in diseased cases; and let $G_x(x) = P[X \leq x / D = 0]$ be the corresponding distribution in control subjects, i.e., patients suspected of having the disease who are candidates for testing in the relevant medical context.

Most of the measures of diagnostic accuracy are based on a binary classification of the test result. Suppose that the classification is $x$,

i.e. the test is positive if $X > x$

and the test is negative if $X \leq x$.

validity of the test used to be measured in terms of sensitivity, specificity and predictivities.

Sensitivity of the test is defined as the proportion of diseased patients who are classified as diseased; specificity is the corresponding proportion of control patients who are classified as normal.

i.e.,

Sensitivity to the test = $1 - F_x(x)$ \hspace{1cm} (5.19.1)

Specificity of the test = $G_x(x)$ \hspace{1cm} (5.19.2)
There are other measures diagnostic accuracy which are given by

\[ \text{False Negative Ratio} = \frac{1}{1 - \text{Sensitivity}} = F_X(x) \]  \hspace{1cm} (5.19.3)

\[ \text{False positive Ratio} = 1 - \text{Specificity} = 1 - G_X(x) \]  \hspace{1cm} (5.19.4)

\[
\text{Positive predictive Value} = P[D = 1/X \leq x] = \frac{\pi(1-F_X(x))}{\pi[1-F_X(x)] + (1-\pi)[1-G_X(x)]}
\]  \hspace{1cm} (5.19.5)

Where, \( \pi = p(D = 1) \) is the prior probability of disease = prevalence of disease in the population under study.

\[
\text{Negative predictive value} P[D = 0/X \leq x] = \frac{(1-\pi)G_X(x)}{(1-\pi)G_X(x) + \pi F_X(x)}
\]  \hspace{1cm} (5.19.6)

Accuracy means the overall relative frequency of correct diagnosis in the study. If \( A(x) \) denotes the accuracy, then

\[ A(x) = \pi[1-F_X(x)] + (1-\pi)G_X(x) \]  \hspace{1cm} (5.19.7)

There are two likelihood ratios corresponding to a Negative and Positive test result.

\[ \text{Likelihood Ratio corresponding to Negative test result } \text{(LR)}^- = \frac{F_X(x)}{G_X(x)} \]  \hspace{1cm} (5.19.8)

\[ \text{Likelihood ratio corresponding to positive test result } \text{(LR)}^+ = \frac{[1-F_X(x)]}{[1-G_X(x)]} \]  \hspace{1cm} (5.19.9)
All the aforementioned measures are limited by the fact that they correspond to a specific, and possibly arbitrary, classification point $x$. Changing the classification point will either increase the sensitivity at the expense of the specificity or vice versa, with corresponding effects on the error rates and the predictive values.

It is a problem when diagnostic tests are being compared, or when the same test is used in different studies with different classification points, since the classifications for the two tests are unlikely to be calibrated in practice.

As a remedy, Receiver Operating Characteristic (ROC) curve analysis has become a preferred method for evaluating and comparing tests.

The ROC curve is a plot of $F_X(x)$ vs $G_X(x)$.

ROC plot lies along the $45^\circ$ line $\Rightarrow$ The test is random and hence uninformative. The higher, ROC curve lies above the $45^\circ$ line $\Rightarrow$ The test is more accurate. Thus, the area under the ROC curve is often used as a measure of accuracy that does not require a specific classification point. The area of the ROC curve is given by

$$\text{Area} = A = 1 - \int_0^1 F_X(x) \, dG_X(x)$$

(5.19.10)

The area $A$ can be interpreted as the probability that a randomly chosen diseased subject has a test result that is greater than that of a randomly chosen control subject.
5.20. RECENT DEVELOPMENTS ON DIAGNOSTIC TESTS IN BIOSTATISTICS.

There has been a substantial recent increase in research activity in the biostatistician literature on methods pertaining to the evaluation of diagnostic tests.

Diagnostic tests are an important part of medical decision making. In daily clinical practice, many tests are performed to obtain diagnoses.

A perfect test is positive in all patients with the disease and negative in all patients who do not have the disease. Usually this test is called the Gold standard test or Reference test. After applying the gold standard test, one can get information that which patients have disease and which patients are free of disease. However most tests are imperfect. Sensitivity and specificity are the two measures to assess the performance of a diagnostic test.

Problems may arise when the sensitivity and the specificity of the Reference test are unknown. The 2x2 Table of test \( r \). Reference test contains too little information to estimate all unknown parameters, even if the prevalence of the disease is known.

The relationship between the results of a test and Gold Standard test is given in the following 2x2 table.
<table>
<thead>
<tr>
<th>Results of tests for disease under study</th>
<th>Results of Gold Standard Test</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Disease Present</td>
<td>Disease Absent</td>
</tr>
<tr>
<td>Positive</td>
<td>True Positive</td>
<td>False Positive</td>
</tr>
<tr>
<td>Negative</td>
<td>False Negative</td>
<td>True Negative</td>
</tr>
<tr>
<td>Total</td>
<td>Diseased Patients</td>
<td>Non Diseased Patients</td>
</tr>
</tbody>
</table>

Suppose that \( D \) represents disease status \( D = 1 \) if diseased and \( D = 0 \) if not.

\[ T_i = \text{The result of the } i^{th} \text{ test, } i = 1, 2, \]

\[ T_i = 1 \text{ if the test is positive and } T_i = 0 \text{ if the rest is negative} \]

\[ S_{ei} = \text{Sensitivity of } i^{th} \text{ test } = P(T_i = 1 / D = 1) \]

\[ S_{pi} = \text{Specificity of } i^{th} \text{ test } = P(T_i = 0 / D = 0) \]

Let there be \( G \) subpopulations or groups

With prevalences \( \pi_g = P(D = 1 / \text{group } g) \), \( g = 1, 2, \ldots, G \)

Under conditional independence of \( T_1 \) and \( T_2 \) given \( D \), The probabilities in the \( T_1 \) \( X \) \( T_2 \) contingency table in group ‘\( g \)’ are given by

\[
p[T_1 = t_1, T_2 = t_2 / \text{group } g] = \pi_g \left[ S_{ei}^g (1 - S_{pi}^g)^{t_2 +} \right] \left[ S_{ei}^g (1 - S_{pi}^g)^{t_1 -} \right] \\
+ \pi_g \left[ (1 - S_{pi}^g)^{t_2 +} S_{pi}^g \right] \left[ (1 - S_{pi}^g)^{t_1 -} S_{ei}^g \right]
\]  \( (5.20.1) \)

Suppose that one may have a population of patients in which proposition \( P \), truly have a particular disease and the remainder, \((1 - p)\), do not have disease.
In other words, the background prevalence of disease is \( p \). If the diagnostic test was performed as each member of the population, then the distribution of test results that would occur in patients with and without disease can be presented in the following table:

<table>
<thead>
<tr>
<th>Diagnostic Test Result</th>
<th>Disease Status</th>
<th>Predictive value (Post-test probability)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Present</td>
<td>Absent</td>
</tr>
<tr>
<td>Positive</td>
<td>( p(1-\beta) )</td>
<td>( (1-p)\alpha )</td>
</tr>
<tr>
<td></td>
<td>( \frac{p(1-\beta)}{p(1-\beta)+(1-p)\alpha} )</td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td>( p\beta )</td>
<td>( (1-p)(1-\alpha) )</td>
</tr>
<tr>
<td></td>
<td>( \frac{p\beta}{p\beta+1-p(1-\alpha)} )</td>
<td></td>
</tr>
</tbody>
</table>

Here, sensitivity = \( P \) (Positive test / disease present) = \( 1-\beta \) = positive predictivity

Specificity = \( P \) (Negative test / disease not present) = \( 1 - \alpha \) = Negative Predictivity

\( \alpha \) and \( \beta \) are analogous to the Type I and Type II errors in the context of Hypothesis Testing.

The likelihood Ratios are given by,

Likelihood Ratio for a positive test result (LR)\(^+\) = \( \frac{1-\beta}{\alpha} = \frac{\text{Sensitivity}}{1-\text{specificity}} \)

Likelihood Ratio for a Negative test result (LR)\(^-\) = \( \frac{1-\text{Sensitivity}}{\text{specificity}} \)
Odds are an alternative way of expressing the likelihood of an event.

\[
\text{Odds} = \frac{\text{Probability}}{1 - \text{Probability}} \quad \text{or} \quad \text{probability} = \frac{\text{Odds}}{1 + \text{Odds}}
\]

One may have,

\[
\text{Post-test odds for Positive Test} = \left[ \frac{P}{1-P} \right] \left[ \frac{1-\beta}{\alpha} \right] = [\text{Pre-test odds}] [(LR)^+]
\]

\[
\text{Post test odds for Negative Test} = \left[ \frac{P}{1-P} \right] \left[ \frac{\beta}{1-\alpha} \right] = [\text{pre-test odds}] [(LR)]^{-1}
\]

Likelihood Ratio is a useful simple measure of the diagnostic information conveyed by a diagnostic test. Compared with Sensitivity and Specificity, Likelihood Ratio offers a more interpretable measure of the impact of the test result on the probability of disease and simplification in the computation.