Discrete Multivariate Analysis - 579 - Exam 2

Alex Towell (atowell@siue.edu)

Contents

Problem 1

A genetic model predicts the ratio of green seedlings to yellow seedlings to be $3 : 1$. Consider data from a sample of $n = 1100$ seedlings, where $n_1 = 860$ is the number of green and $n_2 = 240$ is the number of yellow.

Part (a)

Compute the statistic X^2 for testing the proposed model.

The observed data models the binomial distribution

$$n_g \sim \text{BIN}(n, \pi_g)$$

where n_g and $n_y = n - n_g$ are respectively the number of green and yellow seedlings and π_g and $\pi_y = 1 - \pi_g$ are respectively the probability of a green and yellow seedling.

Under the null model, the odds of green seedlings to yellow seedlings is $3 : 1$, thus

$$H_0 : \pi_{g0} = 0.75 \;\; (\pi_{y0} = 0.25).$$

We perform a chi-squared test for goodness of fit whose test statistic is given by

$$X^2 = \frac{(n_g - \hat{m}_g)}{\hat{m}_g} + \frac{(n_y - \hat{m}_y)^2}{\hat{m}_y}$$

where $\hat{m}_g = n\pi_{g0}$ and $\hat{m}_y = \pi_{y0}$.

The parameter space of the unconstrained binomial model is given by

$$\vec{\pi}(\pi_g) = (\pi_g, 1 - \pi_g)',$$

which has a dimension of $\dim \vec{\pi} = 1$.

The *null model*, which is the specified binomial $\pi_{g0} = 0.75$, has the parameter space $\vec{\pi}_0 = (0.75, 0.25)'$, which is just a *point* with a dimension of $\dim \vec{\pi}_0 = 0$.

1

Thus, under the null model, X^2 is distributed χ^2 with

$$\text{df} = \dim\vec{\pi} - \dim\vec{\pi}_0 = 1 - 0 = 1$$

degrees of freedom, denoted by $\chi^2(1)$.

ANSWER

We compute the observed X^2 statistic with:

```
obs <- c(ng=860,ny=240)
pi.0 <- c(pi.g0=.75,pi.y0=.25)
x2_test <- multinomial_goodness_of_fit(obs, pi.0)$X2
x2_test
```

```
## $stat
## [1] 5.939394
##
## $p
## [1] 0.01480611
##
## $df
## [1] 1
```

We see that X^2 realizes the value $X_0^2 = 5.939$.

Part (b)

Determine the upper 10th percentile for the reference distribution.

Assuming $X^2 \sim \chi^2(1)$, the $(1-\alpha) \times 100\%$ percentile, denoted by $\chi_\alpha^2(1)$, is given by solving the equation

$$\Pr[X^2 \geq \chi_\alpha^2(1)] = \alpha$$

for χ_α^2.

We compute the 10-th percentile with the following R code:

```
alpha <- 0.10
q <- round(qchisq(p=alpha,df=x2_test$df,lower.tail=FALSE),3)
```

ANSWER

We see that the 10-th percentile is $\chi_{0.1}^2(1) = 2.706$.

Part (c)

Provide an interpretation of your result, stated in the context of the problem.

ANSWER

We say that any observed statistic X_0^2 with df $= 1$ greater than $\chi_{0.1}^2(1) = 2.706$ is not compatible with the null model at significance level $\alpha = 0.1$.

Since the observed statistic $X_0^2 = 5.939 > 2.706$, the null model, which is the genetic theory where the ratio of green to yellow is $3 : 1$, is not compatible with the data.

Part (d)

What are the shortcomings of hypothesis testing as a measure of evidence?

ANSWER

Hypothesis testing, as a dichotomous measure of evidence, does not provide as much information as a more quantitative evidence measure. For instance, it does not provide information about effect size.

Part (e)

Compute a 90% confidence interval for π_1.

The MLE of π_g is given by $\hat{\pi}_g = \frac{n_g}{n}$. A $(1 - \alpha) \times 100\%$ confidence interval for π_g is given by

$$\hat{\pi}_g \pm z_{\alpha/2} \hat{\sigma}_{\hat{\pi}_g}.$$

where $\hat{\sigma}_{\hat{\pi}_g} = \sqrt{\frac{\hat{\pi}_g(1-\hat{\pi}_g)}{n}}$.

ANSWER

Letting $\alpha = 0.10$, the MLE and 90% confidence interval for π_g is computed with the following R code:

```
n <- sum(obs)
mle <- obs[1]/n

alpha <- 0.1
s2 <- mle*(1-mle)/n
q <- qnorm(1-alpha/2)   # should be around 1.645
conf <- c(mle - q * sqrt(s2),
          mle + q * sqrt(s2))
```

We see that $\hat{\pi}_g = 0.782$ and a 90% confidence interval is $(0.761, 0.802)$.

Part (f)

Provide an interpretation of your result, stated in the context of the problem.

ANSWER

Based on the observed data, we estimate that the probability π_g of a green strain is between 0.761 and 0.802.

As expected, the null model specifies a value for π_g (0.75) that is not contained in the computed confidence interval.

Part (g)

Compute the statistic G^2 for testing the proposed model.

The same reasoning as in part (a), except we use a different statistic called the likelihood ratio statistic,

$$G^2 = -2 \log \Lambda$$

where

$$\Lambda = \frac{l(\pi_{g0})}{l(\hat{\pi}_g)}.$$

where $\hat{\pi}_g$ is the maximum likelihood estimate n_g/n.

This may be rewritten as

$$G^2 = 2\sum_{j=1}^{c} \log\left(\frac{n_j}{n\pi_{j0}}\right).$$

Under the null model, G^2 is distributed χ^2 with 1 degree of freedom.

ANSWER

We compute the observed G_2 statistic with the following R code:

```
G2_test <- multinomial_goodness_of_fit(obs,pi.0)$G2
G2_test
```

```
## $stat
## [1] 6.120841
##
## $p
## [1] 0.01335972
##
## $df
## [1] 1
```

We see that G^2 realizes the value 6.121.

Part (h)

Compute the p-value for the proposed model.

ANSWER

The reference distribution is the chi-squared distribution with 1 degrees of freedom.

The p-value given 1 degree of freedom is defined as

$$p\text{-value} = \mathrm{P}(G^2 > \chi^2(1)),$$

which was computed to be 0.013. (For comparison, the p-value for the X^2 test was 0.015.)

Part (i)

Provide an interpretation of your result, stated in the context of the problem.

ANSWER

According to the Fisher scale for interpreting p-values, a p-value of 0.01 is considered *strong* evidence against the null model.

Since we obtained a p-value of slightly larger, the data provides *strong* evidence against the genetic theory positing that the ratio of green to yellow seedlings is $3:1$.

Part (j)

What are the shortcomings of using a p-value as a measure of evidence?

ANSWER

The p-value overstates evidence against the null model by comparing the null model to the alternative model that is best supported by the data.

Problem 2

Consider the results from a study on the relationship between dose level and the severity of side effects. (Larger y values correspond to a more severe response.)

Data:

	$y = 0$	$y = 1$	$y = 2$	$y = 3$	$y = 4$
low dose	6	9	6	3	1
medium dose	2	5	6	6	6
high dose	1	4	6	8	8

Preliminary analysis

We model the cross-sectional data as being an observation of a 3×5 random matrix where the row levels, denoted by X, correspond to dosage level and the column levels, denoted by Y, correspond to severity of side effects. The elements (cells) of the random matrix are jointly sampled from the multinomial distribution

$$\{n_{ij}\} \sim \mathrm{MULT}(n, \{\pi_{ij}\}).$$

where n_{ij} denotes the (i, j)-th element of the random matrix, π_{ij} denotes the parameter that specifies the joint probability $\Pr(X = i, Y = j)$, and n denotes the number of trials.

The full model has a parameter space of dimension $3 \times 5 - 1 = 14$.

Part (a)

> Compute the statistic X^2 and the p-value for testing the model with independence between dose level and response severity. Provide an interpretation, stated in the context of the problem.

We consider a simpler model than the full model where X and Y are independent, i.e., $\pi_{ij0} = \Pr(X = i) \Pr(Y = j) = \pi_{i+}\pi_{+j}$. Since we are only free to specify π_{i+} and π_{+j}, the parameter space of the independence model is of dimension $(3 - 1) + (5 - 1) = 6$.

To decide whether the observed data is compatible with X and Y being independent, we perform the hypothesis test

$$H_0 : \{\pi_{ij}\} = \{\pi_{ij0}\}.$$

An estimator of π_{i+} is $\hat{\pi}_{i+} = n_{i+}/n$ and an estimator of $\hat{\pi}_{+j}$ is n_{+j}/n, therefore an estimator of π_{ij0} is given by

$$\hat{\pi}_{ij0} = \hat{\pi}_{i+}\hat{\pi}_{+j} = \frac{n_{i+}n_{+j}}{n^2}.$$

We perform a chi-square test for independence. The test statistic is given by

$$X^2 = \sum_{i=1}^{4} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}$$

where $\hat{m}_{ij} = n\hat{\pi}_{ij0}$. Asymptotically, X^2 is distributed χ^2 with $14 - 6 = 8$ degrees of freedom.

ANSWER

We compute the observed value of X^2 and its p-value with the following R code:

```
obs <- matrix(c(6,9,6,3,1,
                2,5,6,6,6,
                1,4,6,8,8), nrow=3, byrow=TRUE)
colnames(obs) <- c("y=0","y=1","y=2","y=3","y=4")
rownames(obs) <- c("low dose","medimum dose","high dose")
ind_test <- multinomial_goodness_of_fit_independence_test(obs)$X2
ind_test
```

```
## $stat
## [1] 14.35678
##
## $p
## [1] 0.07292708
##
## $df
## [1] 8
```

The test statistic X^2 realizes the value 14.357 whose p-value is 0.073. We consider this p-value to be moderate evidence against the null model. Stated differently, the observed data is moderately incompatible with the independence model.

Part (b)

Now compute the statistic Z and the p-value for testing the independence model using the correlation estimate $\hat{\gamma}$. Provide an interpretation, stated in the context of the problem.

Under the assumption of a monotonic association between X and Y, if X and Y are independent then their γ correlation is zero. To decide whether the observed data is compatible with X and Y being independent, we perform the hypothesis test

$$H_0 : \gamma = 0.$$

The test statistic is given by

$$Z^* = \frac{\hat{\gamma} - 0}{\hat{\sigma}(\hat{\gamma})},$$

which is normally distributed $\mathcal{N}(0,1)$ under the null model.

ANSWER

We compute Z^* and p-value with the following R code:

```
gamma_test <- independence_test_monotonic_association(obs)
```

The observed Z^* is 4.402 and its p-value is less than 0.001, which we consider to be very strong evidence against the independence model.

That is, the observed data provides very strong evidence against the independence model under the assumption of a monotonic association.

Part (c)

Explain why a monotonic association model is better than a general association model in this example.

ANSWER

Estimates for simpler models have smaller variances than for estimates from more complicated models.

Problem 3

Consider the results from a prospective study on the relationship between sex and hair color.

```
obs <- matrix(c(75,16,9,120,64,16),nrow=2,byrow=TRUE)
dimnames(obs) <- list(sex=c("male","female"),
                      hc.level=c("dark","blonde","red"))
print(obs)
```

```
##         hc.level
## sex      dark blonde red
##   male     75     16   9
##   female  120     64  16
```

Part (a)

State the likelihood function for data from a product multinomial sample. State the equation for the maximum likelihood estimate $\hat{\pi}_{j|i}^{(F)}$ under the full model. State the equation for the maximum likelihood estimate $\hat{\pi}_{j|i}^{(I)}$ under the independence model.

ANSWER

In a prospective study, each row of the table represents an independent multinomial. Thus, the data model is given by

$$\{n_{ij}\} \sim \text{PROD\_MULT}(\{n_{i+}\}, \{\pi_{j|i}\}).$$

We are given a table of $I = 2$ rows and $J = 3$ columns, and thus we have I independent multinomials. Therefore, the likelihood function is given by

$$l(\{\pi_{j|i}\}) = \prod_{i=1}^{I} \frac{n_{i+}!}{\prod_{j=1}^{J} n_{ij}!} \prod_{j=1}^{J} \pi_{j|i}^{n_{ij}}.$$

Observe that

$$l(\{\pi_{j|i}\}) \propto \prod_{i=1}^{I} \prod_{j=1}^{J} \pi_{j|i}^{n_{ij}}.$$

Under the full model, the maximum likelihood of $\pi_{j|i}$ is given by

$$\hat{\pi}_{j|i}^{(F)} = \frac{n_{ij}}{n_{i+}}$$

and under the independence model it is given by

$$\hat{\pi}_{j|i}^{(I)} = \frac{n_{+j}}{n}$$

for $i \in \{1, 2\}$ and $j \in \{1, 2, 3\}$.

Part (b)

Compute the table of expected counts $\hat{m}$ under the independence model.

The expected count $\hat{m}_{ij}$ under the independence model is given by

$$\hat{m}_{ij} = \frac{n_{i+}n_{+j}}{n} = n_{i+}\hat{\pi}_{j|i}^{(I)}.$$

ANSWER

We use the R function chisq.test to compute the expected cell counts $\{\hat{m}_{ij}\}$:

```
m.hat <- chisq.test(obs,correct = FALSE)$expected
print(round(m.hat,3))
```

```
##         hc.level
## sex       dark blonde    red
##   male      65 26.667  8.333
##   female   130 53.333 16.667
```

Part (c)

Compute the statistic G^2 and the p-value for testing the independence model. Provide an interpretation of your result, stated in the context of the problem.

In the full model, the paramter space is given by

$$\vec{\pi}^{(F)}(\theta_1,\theta_2,\theta_3,\theta_4) = \begin{pmatrix} \pi_{1|1} = \theta_1 & \pi_{2|1} = \theta_2 & \pi_{3|1} = 1 - \theta_1 - \theta_2 \\ \pi_{1|2} = \theta_3 & \pi_{2|2} = \theta_4 & \pi_{3|2} = 1 - \theta_3 - \theta_4 \end{pmatrix}$$

which is of dimension $I(J-1) = 4$ and under the independence model, the parameter space is

$$\vec{\pi}^{(I)}(\theta_1,\theta_2) = \begin{pmatrix} \pi_{1|1} = \theta_1 & \pi_{2|1} = \theta_2 & \pi_{3|1} = 1 - \theta_1 - \theta_2 \\ \pi_{1|2} = \theta_1 & \pi_{2|2} = \theta_2 & \pi_{3|2} = 1 - \theta_1 - \theta_2 \end{pmatrix}$$

which is of dimension $J - 1 = 2$. Thus, under the independence model, G^2 is distributed χ^2 with df $= I(J-1) - (J-1) = (J-1)(I-1) = 2$

ANSWER

We compute G^2 from the observed counts $\{n_{ij}\}$ and the expected counts $\{\hat{m}_{ij}\}$:

```
I <- dim(obs)[1]
J <- dim(obs)[2]
G2 <- 2*sum(obs*log(obs/m.hat))
df <- (I-1)*(J-1)
p.value <- pchisq(G2,df,lower.tail = FALSE)

print(round(G2,3))
```

```
## [1] 9.325
```

```
print(round(p.value,3))
```

```
## [1] 0.009
```

We obtain a p-value around 0.01, which we consider to be strong evidence against the independence model (the data supports the general association model). That is to say, the conditional distribution

$$\Pr(\text{hair color} \,|\, \text{sex} = \text{male})$$

and
$$\Pr(\text{hair color} \mid \text{sex} = \text{female}).$$

are different.

Part (d)

> Provide an interpretation of the association, stated in the context of the problem, by comparing the observed counts with the expected counts.

ANSWER

We display a table comparing expected and observed such that if the observed is less than the expected, display -1, and if observed is greater than expected, display a 1:

```
print(sign(obs-m.hat))
```

```
##           hc.level
## sex       dark blonde red
##    male      1    -1   1
##    female   -1     1  -1
```

The observed data is compatible with a model where males are more likely to have dark or red hair (and less likely to have blonde hair) and females are more likely to have blonde hair (and less likely to have dark or red hair).

Problem 4

According to the Hardy-Weinberg model, the number of flies resulting from a crossing of genetic traits are in categories 1, 2, 3 with probabilities

$$\pi_{1o}(\theta) = (1-\theta)^2$$
$$\pi_{2o}(\theta) = 2\theta(1-\theta)$$
$$\pi_{3o}(\theta) = \theta^2$$

Part (a)

> Show that the maximum likelihood estimator of θ from data $(n_1, n_2, n_3) \sim \text{MULT}(n, \pi_1, \pi_2, \pi_3)$ is
> $$\hat{\theta} = \frac{n_2 + 2n_3}{2n}.$$

We assume the data model is given by

$$(n_1, n_2, n_3) \sim \text{MULT}(n, \pi_1, \pi_2, \pi_3)$$

where n_j denotes the number of flies in category j and π_j denotes the probability a fly is in category j.

Under the specified model, the likelihood function is given by

$$\text{L}(\theta) \propto \prod_{j=1}^{3} \pi_{jo}(\theta)^{n_j}$$

9

and the kernel of the log-likelihood function is given by

$$
\begin{aligned}
l(\theta) &= \sum_{j=1}^{3} n_j \log \pi_{jo}(\theta) \\
&= n_1 \log \pi_{1o}(\theta) + n_2 \log \pi_{2o}(\theta) + n_3 \log \pi_{3o}(\theta) \\
&= n_1 \log(1-\theta)^2 + n_2 \log 2\theta(1-\theta) + n_3 \log \theta^2 \\
&= 2n_1 \log(1-\theta) + n_2 \log \theta + n_2 \log(1-\theta) + 2n_3 \log \theta + \text{const} \\
&= (2n_1 + n_2) \log(1-\theta) + (n_2 + 2n_3) \log \theta + \text{const}.
\end{aligned}
$$

The derivative of the log-likelihood is thus given by

$$
\frac{dl}{d\theta} = -\frac{2n_1 + n_2}{1-\theta} + \frac{n_2 + 2n_3}{\theta}.
$$

The ML estimate of θ is given by

$$
\left. \frac{dl}{d\theta} \right|_{\hat{\theta}} = 0.
$$

Solving for $\hat{\theta}$, we see that

$$
\frac{2n_1 + n_2}{1-\hat{\theta}} = \frac{n_2 + 2n_3}{\hat{\theta}}.
$$

Next, we take the reciprocal of both sides,

$$
\frac{1-\hat{\theta}}{2n_1 + n_2} = \frac{\hat{\theta}}{n_2 + 2n_3}.
$$

Expanding the LHS, we get

$$
\frac{1}{2n_1 + n_2} - \frac{\hat{\theta}}{2n_1 + n_2} = \frac{\hat{\theta}}{n_2 + 2n_3}
$$

which may be rewritten as

$$
\hat{\theta} \left(\frac{1}{2n_1 + n_2} + \frac{1}{n_2 + 2n_3} \right) = \frac{1}{2n_1 + n_2}.
$$

Dividing both sides by the part in parentheses, we get

$$
\hat{\theta} = \frac{1}{(2n_1 + n_2)\left(\frac{1}{2n_1+n_2} + \frac{1}{n_2+2n_3} \right)},
$$

which simplifies to

$$
\hat{\theta} = \frac{1}{1 + \frac{2n_1+n_2}{n_2+2n_3}} = \frac{1}{\frac{2n_3+2n_1+2n_2}{n_2+2n_3}}.
$$

Performing the division, we get the simplification

$$
\hat{\theta} = \frac{n_2 + 2n_3}{2(n_3 + n_1 + n_2)}.
$$

ANSWER

Observe that $n_1 + n_2 + n_3$ is n, thus

$$
\hat{\theta} = \frac{n_2 + 2n_3}{2n}.
$$

10

Part (b)

Suppose we observe data $n_1 = 40$, $n_2 = 50$, $n_3 = 10$ from a sample of size $n = 100$. Compute the mle $\hat{\theta}$ and the estimated category probabilities $\pi_0(\hat{\theta})$.

Using the MLE equation, we compute the result using R:

```r
n = c(40,50,10)
mle <- (n[2] + 2*n[3])/sum(n)
```

ANSWER

We see that
$$\hat{\theta} = 0.7.$$

Plugging in this result,
$$\vec{\pi}_o(\hat{\theta}) = ((1-\hat{\theta})^2, 2\hat{\theta}(1-\hat{\theta}), \hat{\theta}^2)'$$

which we compute with R:

```r
pi0.hats <- c((1-mle)^2,
              2*mle*(1-mle),
              mle^2)
```

We see that $\vec{\pi}_o(\hat{\theta}) = (0.09, 0.42, 0.49)'$.

Part (c)

Compute the statistic G^2 and the p-value for testing the Hardy-Weinberg model. Provide an interpretation of your result, stated in the context of the problem.

ANSWER

The parameter space of the Hardy-Weinberg model is a function of a single parameter θ and thus has a dimension of 1. The parameter space of the full model is 2. Thus, the reference distribution is χ^2 with $df = 2 - 1 = 1$ degrees of freedom.

The following R code computes the realization of G^2, denoted by G_0^2, and its p-value:

```r
hardy_test <- multinomial_goodness_of_fit(obs=n,pi.0=pi0.hats,t=1)$G2
hardy_test
```

```
## $stat
## [1] 104.983
##
## $p
## [1] 1.231867e-24
##
## $df
## [1] 1
```

We see that $G_0^2 = 104.983$ with a p-value less than 0.001, which we consider to be very strong evidence against the Hardy-Weinberg.

Library of test functions

```r
# function for performing a multinomial goodness of fit hypothesis test
# with a specified null model pi.0 in which pi.0 is the result of t parameters.
```

```r
#
# input:
#     obs: vector of counts
#     pi.0: vector of hypothesized probabilites,
#     t: dimension of the parameter space under the null hypothesis.
#
# outputs: two lists, one for X2 and one for G2. each list has:
#     X2: list relevant results for X^2 test
#     G2: list of relevant results for G^2 test, e.g.:
#        list(stat,p,df,reject) where
#          p: p-value of the obs under the null model
#          df: degrees of freedom of the reference distribution (chi-square)
multinomial_goodness_of_fit <- function(obs, pi.0, t=0)
{
  n <- sum(obs)                   # number of observations
  t.full <- length(obs)-1         # dim of parameter space for the full model.
  df <- t.full-t                  # degrees of freedom for the ref distribution
  m.0 <- n*pi.0                   # expected observations (for given n)

  G2 <- 2*sum(obs*log(obs/m.0)) # log-likelihood ratio test statistic
  G2.p <- pchisq(q=G2,df=df,lower.tail=FALSE)

  X2 <- sum((obs-m.0)^2/m.0)    # test statistic
  X2.p <- pchisq(q=X2,df=df,lower.tail=FALSE)

  # return relevant test statistics
  list(m.0=m.0,pi.0=pi.0,
       X2=list(stat=X2,p=X2.p,df=df),
       G2=list(stat=G2,p=G2.p,df=df))
}


# obs is an IxJ matrix representing cross-sectional data (observed counts).
# tests for independence P(X,Y) = P(X)P(Y).
multinomial_goodness_of_fit_independence_test <- function(obs)
{
  I <- dim(obs)[1]
  J <- dim(obs)[2]
  t <- (I-1)+(J-1)

  n <- sum(obs)
  row.sum <- apply(obs,1,sum)
  col.sum <- apply(obs,2,sum)
  m.0 <- matrix(row.sum) %*% t(matrix(col.sum))/n

  multinomial_goodness_of_fit(obs=obs,pi.0=m.0/n,t=t)
}


independence_test_monotonic_association <- function(obs)
{
  # to compute the estimate of gamma and the standard error, we need MESS
  library("MESS")

  # gkgamma takes an IxJ matrix as an input
  gk.result <- gkgamma(obs)
```

```r
  # compute gamma.hat and its standard error
  gamma.hat <- gk.result$estimate
  se <- gk.result$se1

  # compute the z statistic and p-value for testing gamma=0 (independence)
  z <- gamma.hat / se
  p.value <- 2*pnorm(abs(z),lower.tail = FALSE)

  list(gamma=gamma.hat,ase=se,stat=z,p=p.value)
}
```