In-Class 1. Categorical Data Analysis

Suzer-Gurtekin

January 2025

Overview

1. Review of Distributions

2. Testing Goodness of Fit

Distributions

- Most of 615/685 built upon normal distribution
- This course focuses on what happens when this assumption is violated
- Non-normal distributions for categorical and count data
    - Bernoulli
    - Binomial
    - Multinomial
    - Poisson

Bernoulli

Let the random variable X take on two values, 0 and 1, with the following distribution:

$P(X = 1) = p$
$P(X = 0) = 1 - p$

This is called a *Bernoulli* random variable. It is denoted $X \sim Ber(p)$.

Bernoulli

Bernoulli distribution expected value and variance:

$E[X] = p$
$V(X) = p(1 - p)$

Binomial

Properties of $\hat{p}$ from a Binomial Distribution

1. $\hat{p}$ is a random variable
2. $\hat{p}$ is unbiased
3. $V(\hat{p}) = \frac{p(1-p)}{n}$
4. For large n, $\sqrt{n}(\hat{p} - p) \to N[0, p(1-p)]$
5. The MLE for any one-to-one function of p, $\theta = g(p)$ is $\hat{\theta} = g(\hat{p})$

Multinomial

Let **X** be a $k \times 1$ dimensional random vector which has all its components equal to zero, except for one which is equal to 1.

Further, we assume that:

$\mathbf{X}' = (X_1, \ldots, X_k)$

$P(X_j = 1) = p_j$ for $j = 1, \ldots, k$

$\sum_{i=1}^{k} p_i = 1$

Multinomial

Example of a variable with a multinomial distribution ($k = 5$):

Overall Health Status
1=Excellent
2=Very Good
3=Good
4=Fair
5=Poor

Multinomial

Properties of $\hat{\boldsymbol{p}}$ from a Multinomial Distribution

1. $\hat{\boldsymbol{p}}$ is a random vector
2. $\hat{\boldsymbol{p}}$ is unbiased
3. $V(\hat{\boldsymbol{p}}) = n^{-1}[diag(\boldsymbol{p}) - \boldsymbol{pp'}]$, $V(\hat{p}_i) = \frac{p_i(1-p_i)}{n}$, $C(\hat{p}_i, \hat{p}_j) = \frac{-p_i p_j}{n}$
4. For large n, $\sqrt{n}(\hat{\boldsymbol{p}} - \boldsymbol{p}) \rightarrow N_k(\boldsymbol{0}, diag(\boldsymbol{p}) - \boldsymbol{pp'})$
5. The MLE for any one-to-one function $\theta = g(\boldsymbol{p})$ is $\hat{\theta} = g(\hat{\boldsymbol{p}})$

Multinomial

We can characterize any element of a multinomial distribution as a binomial of the form:

$$Y_i \sim Bin(n, p_i)$$
$$E[Y_i] = np_i$$
$$V(Y_i) = np_i(1 - p_i)$$

Example 1

We have 1301 plants from a plant-breeding experiment. The question is, how often does each variety result from this particular combination of plants. The following are the observed counts of each variety:

Variety	**Count**
Green	773
Golden	231
Green-Striped	238
Golden-Green-Striped	59
Sum	1301

Genetics predicts 9/16, 3/16, 3/16 and 1/16 as the distribution of varieties. We will use these predictions as our H_0.

In-Class Problem

Variety	Count
Green	773
Golden	231
Green-Striped	238
Golden-Green-Striped	59
Sum	1301

What is the probability that a “Green” variety will result?
What is the variance of this estimate?
What is the probability that a “Golden” or “Golden-Green-Striped” will result?
What is the variance of this estimate?

In-Class Problem

What is the probability that a “Green” variety will result?

$Pr(Green) = \hat{p}_1 = \frac{773}{1301} = 0.59$

What is the variance of this estimate?

$V(\hat{p}_1) = \frac{\hat{p}_1(1-\hat{p}_1)}{n} = \frac{.59(1-.59)}{1301} = 0.000185345$

In-Class Problem

What is the probability that a “Golden” or “Golden-Green-Striped” will result?

Pr(Golden or Golden-Green-Striped)$= \hat{p}_2 + \hat{p}_4 = \frac{231+59}{1301} = 0.223$

What is the variance of this estimate?

$$
\begin{aligned}
\hat{V}(\hat{p}_2 + \hat{p}_4) &= \hat{V}(\hat{p}_2) + \hat{V}(\hat{p}_4) + 2C(\hat{p}_2, \hat{p}_4) \\
&= \frac{\hat{p}_2(1-\hat{p}_2)}{n} + \frac{\hat{p}_4(1-\hat{p}_4)}{n} + 2[\frac{-(\hat{p}_2\hat{p}_4)}{n}] \\
&= \frac{.1776(1-.1776)}{1301} + \frac{.0453(1-.0453)}{1301} + 2[\frac{-(.1776[.0453])}{1301} \\
&= 0.000133143
\end{aligned}
$$

In-Class Problem

The following R code can also be used to create a solution:

```
## Example 1
varieties<-c(773,231,238,59)

p1<-varieties[1]/sum(varieties)
var_p1<-(1-p1)*p1/sum(varieties)

p2_4<-sum(varieties[c(2,4)])/sum(varieties)
var_p2_4<-(1-p2_4)*p2_4/sum(varieties)
```

Poisson

Let Y_i be a Poisson random variable with parameter m_i for $i = 1, \ldots, k$. We write $Y_i \sim \mathit{Poisson}(m_i)$. Define $\mathbf{Y}' = (Y_1, \ldots, Y_k)$.

We can write the pmf of Y_i as:

$$Pr(Y_i = y_i) = \frac{e^{-m_i} m_i^{y_i}}{y_i!}$$

for $y_i = 0, 1, \ldots, \infty$.

Poisson

The mean and variance of the Poisson distribution are:

$E[Y_i] = m_i$
$V(Y_i) = m_i$

Poisson

Properties of $\hat{\boldsymbol{m}}$ from Poisson Sampling

1. $\hat{\boldsymbol{m}}$ is a random vector
2. $\hat{\boldsymbol{m}}$ is unbiased
3. $V(\hat{\boldsymbol{m}}) = diag(\boldsymbol{m})$
4. The MLE for any one-to-one function $\theta = g(\boldsymbol{m})$ is $\hat{\theta} = g(\hat{\boldsymbol{m}})$

Tests of Goodness of Fit

Let $(n_1, \ldots, n_k)$ denote the observed responses on a multinomial random vector with parameters n and $\mathbf{p}' = (p_1, \ldots, p_k)$, where $\sum_{i=1}^{k} n_i = n$.

We want to test the hypothesis: H_0:$\mathbf{p} = \mathbf{p_0}$ versus H_A:$\mathbf{p} \neq \mathbf{p_0}$

where we specify the vector $\mathbf{p_0}$.

Tests of Goodness of Fit

We looked two tests of this hypothesis:

1. Pearson Chi-Square Statistic (X^2)
2. Likelihood Ratio Statistic (G^2)

χ^2 Distribution

These statistics (and many others) follow a χ^2 distribution.

They follow this form:

$$\frac{(\widehat{\theta} - \theta_0)^2}{\mathrm{var}(\hat{\theta})} \tag{1}$$

This class of tests are sometimes called *Wald tests*

Pearson Chi-Square Statistic

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where $O_i = n_i$ is the observed count in the i^{th} category, and $E_i = np_{0i}$ is the expected count in the i^{th} category (from H_0).

Likelihood Ratio Statistic

Using the notation from the likelihood:

$$G^2 = 2\sum_{i=1}^{k} n_i \ln\left(\frac{n_i}{np_{0i}}\right)$$

Or, we can write using the “observed” versus “expected” notation from above:

$$G^2 = 2\sum_{i=1}^{k} O_i \ln\left(\frac{O_i}{E_i}\right)$$

Likelihood Ratio Statistic

In practice, X^2 and G^2 are quite close, which makes sense as they both have the same asymptotic distribution.

We can show this by rewriting G^2 in the form of X^2:

$$G^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} - 2 \sum_{i=1}^{k} O_i R\left(\frac{E_i}{O_i}\right) \approx X^2$$

If we assume that $O_i \approx E_i$, then the last part drops out and the two are the same.

Example 2

- We have the 1301 plants from the plant-breeding experiment.
- Genetics gives us expected counts H_0 of each variety. (9/16, 3/16, 3/16, 1/16)
  - $\frac{9}{16} \times 1301 = 731.815$
  - $\frac{3}{16} \times 1301 = 243.938$
  - $\frac{1}{16} \times 1301 = 81.313$

Example 2

Complete the table and interpret the result:

Variety (i)	**Observed**	**Expected**	$\frac{(O_i - E_i)^2}{E_i}$	$2O_i ln\left(\frac{O_i}{E_i}\right)$
Green	773	731.815	?	?
Golden	231	243.9375	?	?
Green-Striped	238	243.9375	?	?
Golden-Green-Striped	59	81.3125	?	?
Sum			?	?

Example 2

The following are the observed and expected quantities:

Variety (i)	**Observed**	**Expected**	$\frac{(O_i - E_i)^2}{E_i}$	$2O_i ln\left(\frac{O_i}{E_i}\right)$
Green	773	731.815	2.317805	84.64551
Golden	231	243.9375	0.686155	-25.17638
Green-Striped	238	243.9375	0.14452	-11.72929
Golden-Green-Striped	59	81.3125	6.122646	-37.84995
Sum			9.271126	9.88988

Compare these values to a χ^2_3 with specified α. These values exceed the cutoff value for $\alpha = 0.05$ (7.815) but not for $\alpha = 0.01$ (11.34).

Example 2

R Demonstration:

```
## Probabilities, Variance Est
varieties<-c(773,231,238,59)

p1<-varieties[1]/sum(varieties)
var_p1<-(1-p1)*p1/sum(varieties)

p2_4<-sum(varieties[c(2,4)])/sum(varieties)
var_p2_4<-(1-p2_4)*p2_4/sum(varieties)
```

Example 2

R Demonstration (cont.):

```
## Hypothesis Test
exp_p<-c(0.5625,0.1875,0.1875,0.0625)

#Pearson chi-sq
test1<-chisq.test(varieties,p=exp_p)

#Likelihood ratio test
sum<-0
for (i in 1:4) sum<-sum+2*(varieties[i]*
log(varieties[i]/(exp_p[i]*1301)))

1-pchisq(sum,3)
1-pchisq(test1$statistic,3)
```