Count Data

September 17, 2017

Multinomial Data

- $X_1, X_2, \ldots, X_n$ are n independent random variables each of which can take one of k mutually exclusive outcomes.
- probability of outcome type i is $P(\textit{type } i) = p_i$ for $i = 1, 2, \ldots, k$
- n_i is the number of the X's that are of type i for $i = 1, 2, \ldots, k$
- $\sum_{i=1}^{k} p_i = 1, \sum_{i=1}^{k} n_i = n$

Multinomial Data

- $n_1, n_2, \ldots, n_k$ are realizations of a multinomial distribution
- The joint pdf of $N_1, N_2, \ldots, N_k$ is

$$P(N_1 = n_1, N_2 = n_2 \ldots, N_k = n_k) = \frac{n!}{n_1! n_2! \cdots n_k!} p_1^{n_1} p_2^{p_2} \cdots p$$

Multinomial Data

- To test $H_0 : p_1 = p_{10}, p_2 = p_{20}, \ldots, p_k = p_{k0}$ versus an alternative that at least one of the p_i's is not equal to p_{i0}.
- compute the expected number of each type if H_0 is true, $e_i = np_{i0}$ for $i = 1, \ldots, k$
- compute test statistic:

$$T = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}$$

- T is the squared discrepancy between observed and expected, scaled by the magnitude of the expected
- For large n, $T \to \chi^2_{k-1}$
- Reject H_0 if T is too big
- If too many n_i's are small (less than 5) the approximation may not be good. (combine categories)

Multinomial Data - Example

We wish to test whether our random number generator is uniformly generating digits from 0 to 9. We observe $n = 100$ digits. Our null hypothesis is $H_0 : p_i = 0.1$ for $i = 0, \ldots, 9$, that is that each digit is equally likely to occur. Under this hypothesis $e_i = 100 * 0.01 = 10$

Multinomial Data - Example

Data

Digit	0	1	2	3	4	5	6	7	8	9
n_i	12	7	12	7	13	13	7	13	6	10
e_i	10	10	10	10	10	10	10	10	10	10
T	.4	.9	.4	.9	.9	.9	.9	.9	1.6	0

- test statistic = sum of last row = T=7.40
- Compare to χ^2_9 distribution
- p-value $P(\chi^2_9 > 7.40) = 0.596$
- p-value is not small so we do not reject the null hypothesis
- Random number generator is behaving as it should

Goodness of Fit Tests

- Goodness of fit tests - how well a fitted or assumed model fits the observed data
- Null hypothesis - model fits the observed; Alternative - model does not fit
- O_i for $i = 1, \ldots, k$ are the observed values for each of k outcomes
- E_i is the expected number of outcomes of each type if the model is correct
- Test statistic

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

Goodness of Fit Tests

- $\sum_{i=1}^{k} O_i = n$
- for large n and no cell count under 5, $T \to \chi^2_{k-1}$

Goodness of Fit Tests

Example

We want to test whether the count data below come from a Poisson distribution

i	0	1	2	3	4	5
O_i	678	227	56	28	8	14

- We find the mean of the data and use this to estimate the parameter λ which is the mean of a Poisson distribution.
- for this data we find $\bar{X} = \hat{\lambda} = 0.5193$
- $P(X = x) = \dfrac{\hat{\lambda}^x e^{-\hat{\lambda}}}{x!} \quad x = 0, 1, \ldots$
- $E_i = nP(X = i)$ where $n = 1011$ is the sum of the observed

Goodness of Fit - Example

i	0	1	2	3	4	5
O_i	678	227	56	28	8	14
$\hat{p}_i$	0.5949	0.3089	0.0802	0.0139	0.0018	0.0002
E_i	601.49	312.35	81.1	13.24	1.82	0.19

- All of the O_i are greater than 5 but two of the E_i's are less than 5
- Using the expected counts as they are can make the chi-square approximation poor
- Combine the categories 3,4,5 into one.

Goodness of Fit - Example

i	O_i	$\hat{p}_i$	E_i	T
0	678	0.5949	601.49	9.73
1	227	0.3089	312.35	23.32
2	56	0.0802	81.1	7.77
3+	50	0.0159	16.05	71.82
Total	1011			112.64

- We compare $T = 112.64$ with a Chi-square distribution
- to get the degrees of freedom
  - the number of categories is 4
  - subtract 1 degree of freedom because of the fixed totals
  - subtract 1 degree of freedom because we estimated $\hat{\lambda}$
  - degrees of freedom = 4 - 1 - 1 = 2
- p-value = $P(\chi^2_2 > 112.64) \approx 0$
- reject H_0, conclude the data is not from a Poisson distribution

Contingency Tables

Count data that is organized into two-way classification can be expressed in a contingency table as shown in the $r \times c$ table below

n_{11}	n_{12}	$\ldots$	n_{1c}
n_{21}	n_{22}	$\ldots$	n_{2c}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_{r1}	n_{r2}	$\ldots$	n_{rc}

- Define row sums $n_1, n_2, \ldots, n_r$
- Define column sums $m_1, m_2, \ldots, m_c$
- $N = \sum_{i=1}^{r} n_i = \sum_{i=1}^{c} m_i$

Contingency Tables

Independence Test

- Assume N is fixed
- $p_{ij} = P(row = i, column = j)$ for $i = 1, \ldots, r$, $j = 1, \ldots, c$
- Marginal probabilities

$$p_{i\cdot} = \sum_{j=1}^{c} p_{ij} = P(row = i)$$

$$p_{\cdot j} = \sum_{i=1}^{r} p_{ij} = P(column = j)$$

Contingency Tables

Independence Test

- If the row and column assignments are independent

$$P(row = i, column = j) = P(row = i)P(column = j)$$

$$p_{ij} = p_{i\cdot}p_{\cdot j}$$

- We can estimate $p_{i\cdot}$ by $\frac{n_i}{N}$ and $p_{\cdot j}$ by $\frac{m_j}{N}$
- Under the assumption of independence, we expect cell counts of

$$E_{ij} = Np_{ij} = N\frac{n_i}{N}\frac{m_j}{N} = \frac{n_i m_j}{N}$$

Contingency Tables

Independence Test

If we let O_{ij} be the observed cell count in cell ij then a test statistic is given by

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Reject the null hypothesis if T is too big. For large N and not too many cells with count below 5

$$T \to \chi^2_{(r-1)(c-1)}$$

Contingency Tables

Test of Homogeneity

If we consider row totals, $n_1, n_2, \ldots, n_r$ as fixed, then the cell probabilities are defined to be

$$p_{ij} = P(column = j | row = i)$$

- An assumption of homogeneity is equivalent to assuming that the probability of being in column j is the same for each row.
- Under this assumption we can estimate the probability of being in column j as $\frac{m_j}{N}$
- The expected cell counts are

$$E_{ij} = n_i \times \frac{m_j}{N}$$

Contingency Tables

Test of Homogeneity

- This results in the same test statistic used in the test for independence
- For large N and not too many cell counts under 5, $T \rightarrow \chi^2_{(r-1)(c-1)}$