Categorical Data Analysis – Binomial/Multinomial Distributions

Used when variables are of a qualitative or categorical type.

I.e. population classified by sex (M/F) – Binomial distribution (only 2 options)
population classified by marital status (S,M,D,W) - Multinomial distribution (k possible outcomes)

Binomial distributions

1) 2 possible outcomes (i.e. success/falure)
2) Fixed # of trials (n)
3) Trials are independent of each other (therefore results are independent)
4) Probability of success, P(x) = p and it remains constant from trial to trial
   ∴ P(failure) = 1-p = q is also constant ∴ p+q =1
5) If X is a random variable that counts # of successes out of N trials, we say that **X~ Bin(N, p)** where N=# of trials and p=Probability of success
6) **E(X) = np**
   V(X) = npq = np(1-p)

Categorical Data Analysis – Binomial/Multinomial Distributions

Multinomial distributions

1) Fixed # of trials (n)
2) Trials are independent of each other (therefore results are independent)
3) k possible outcomes (when k=2 then it's a Binomial distribution)
4) Probability, P_i, that outcome falls in a given category remains constant from trial to trial where $\sum_{i=1}^{n} \boldsymbol{p_i} = \mathbf{1}$

STEPS:

1) **STATE ASSUMPTIONS:**
   1. random sample from..
   2. Multinomial distribution where…
   3. n is large enough for X^2 distribution to apply (i.e. > 30)

   (Must explicitly STATE how each assumption is satisfied)
2) Random variables, O1, O2 , …. Ok count the # of observations in category i where $i = 1 \ldots k$

$$\sum_{i=1}^{k} O_i = n$$

3) Put values in a table: (i.e. current and previous monthly payments)

	Full	1 Mth	2 Mth	>2mths	Total
Current	287	49	30	34	400
Past (Expected)	320	40	24	16	400

4) $\boldsymbol{E_i = E(O_i) = np_i} = expected\ \#\ of\ observations\ in\ category\ \boldsymbol{i}\ after\ \boldsymbol{n}\ trials$

H_0: $\boldsymbol{p_1 = p_{O_1}}$; $\boldsymbol{p_2 = p_{O_2}}$ … $\boldsymbol{p_k = p_{O_k}}$

H_a: At least one probability is not equal to its prescribed value

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

$$= \frac{(287-320)^2}{320} + \frac{(49-40)^2}{40} + \frac{(30-24)^2}{24} + \frac{(34-16)^2}{16} = \mathbf{27.18}$$

R.R. – REJECT H_0 if $X^2 > X^2_{\alpha(k-1)}$

$X^2_{0.01(3)} = 11.34$

5) **Write conclusion:** We reject H_o given that $X^2 = 27.18 > X^2_{0.01(3)} = 11.34$ and conclude at a 1% level of significance there is evidence to say that current pattern of payment differs from historical (Past) norms. (NO FOLLOWUP REQUIRED)

(other ways data could be presented: N trials, ONLY observed values for k categories given. Must COMPUTE expected as N/k for each category)

Categorical Data Analysis – CONTINGENCY tables

RxC contingency table – population classified by 2 characteristics (I.E POLITICAL PARTY and Voting Ward, or
(ROW X COLUMN)

1. **TEST OF INDEPENDENCE** – ***find out if 2 characteristics are related***

   H_o: 2 Characteristics are independent
   H_a: 2 Characteristics are related ; α = x

   A single random sample of size n is take from the population and then classified by 2 characteristics of interest

   Test Statistic is Chi-squared (X^2)

   Test-Statistic $X^2_{(r-1)(c-1)} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(o_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}}$

   where $\hat{E}_{ij}$ is estimated expected count $\hat{E}_{ij} = \frac{r_1 c_1}{n}$

   R.R. : reject H_o if $X^2_{(r-1)(c-1)} > X^2_{\alpha;(r-1)(c-1)}$

   Assumptions:

   1) a single r sample
   2) from **population classified by 2 characteristics**
   3) n large enough for X^2 to be valid (n > 30)

Categorical Data Analysis – CONTINGENCY tables

EX…. A certain cola company sells 4 types of cola throughout north America. To help determine if the same marketing approach used in US can be used in Canada and Mexico, on of the firms marketing analysts wants to ascertain if there is an association between the type of cola preferred and the nationality of the consumer. A Random sample of 250 cola drinkers from the 3 countries was interviewed and then classified according to the type of cola preferred and nationality. The observed frequency of drinkers falling into each of the possible cells is shown below. Is there evidence of an association between cola preference and nationality, using α=0.01?

	Cola preference								
Nationality	**A**	**E**	**B**	**E**	**C**	**E**	**D**		**Row Totals**
N1	72	48.3	8	12.88	12	19.32	23	34.5	**115**
N2	26	35.7	10	9.52	16	14.28	33	25.5	**85**
N3	7	21	10	5.6	14	8.4	19	15	**50**
Column Totals	**105**		**28**		**42**		**75**		**n= 250**

H_o: Cola preference and nationality are independent

H_a: Cola preference and nationality are related ; α = 0.01

Assumptions:

1) a single **r** sample
2) … from population **classified by 2 characteristics**: cola preference and nationality
3) **n=** 250 total samples ∴ **large enough for X^2 to be valid (i.e. > 30)**

$$X^2_{(r-1)(c-1)} = \sum_{i=1}^{3}\sum_{j=1}^{4}\frac{(o_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}}$$

$$= \frac{(17-48.3)^2}{48.3} + \frac{(8-12.88)^2}{12.88} + \cdots . \frac{(19-15)^2}{15} = 42.75$$

$$\hat{E}_{11} = \frac{r_1 c_1}{n} = 115 * \frac{105}{250} = 48.3$$

(From Tables) $X^2_{0.01\ ;(3)(2)} = 16.8119$

CONCLUSION: Given that $X^2_{(r-1)(c-1)} = 48.3 > X^2_{0.01\ ;(6)} = 16.8119$ we REJECT H_o and conclude that at 1% level of significance there is an evidence that cola preference and nationality are related.

2. TEST OF HOMOGENEITY – *Tests if category proportions are the same for all populations*

Independent random samples are taken from several different populations of multinomial type

H_o: category proportions are the same for all populations

H_a: category proportions differ between populations ; α = x

Assumptions:

1) m independent random samples
2) from m independent populations distributed with MULTINOMIAL DISTRIBUTION
3) sample sizes are large enough for X^2 to apply (i.e. over 30)

Test-Statistic: $X^2_{(r-1)(c-1)} = \sum_{i=1}^{3}\sum_{j=1}^{4}\frac{(o_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}}$

R.R. : reject H_o if $X^2_{(r-1)(c-1)} > X^2_{\alpha;(r-1)(c-1)}$

Ex. A survey of voter sentiment was conducted in 3 city political wards to compare the proportions of voters favoring the 3 candidates A, B, C. Independent random samples of 200 voters were polled in each of the 3 wards with the results shown below. Do the data provide sufficient evidence to indicate that the proportions of voters favoring candidates A, B, and C differ in the 3 wards? Use α=0.05%

	Wards							
Candidate	1	E_R	2	E_R	3	E_R	Row Totals	
A	108	102.33	87	102.33	112	102.33	307	E_A=307/3=102.33
B	52	47.33	51	47.33	39	47.33	142	E_B=142/3=47.33
C	40	50.33	62	50.33	49	50.33	151	E_C=151/3=50.33
Col Totals:	200		200		200		N= 600	

H_o: proportions of voters favoring candidates A, B, C are the same for all 3 wards

H_a: proportions of voters favoring candidates A, B, C DIFFER between the 3 wards ; α = 0.05

Assumptions:

1) **3 independent random samples**
2) **From 3 independent populations** distributed with a **multinomial distribution**
3) **sample sizes are large enough for X^2 to apply** : n1=n2=n3 = 200 (n_x > 30)

$$\hat{E}_{11} = \frac{r_1 c_1}{n} = \frac{307*200}{600} = 102.33 = \hat{E}_{12} = \hat{E}_{13}$$

$$\hat{E}_{21} = \frac{r_1 c_1}{n} = \frac{142*200}{600} = 47.33 = \hat{E}_{22} = \hat{E}_{23}$$

$$\hat{E}_{31} = \frac{r_1 c_1}{n} = \frac{151*200}{600} = 50.33 = \hat{E}_{32} = \hat{E}_{33}$$

Test-Stat.... $X^2_{(r-1)(c-1)} = \sum_{i=1}^{3}\sum_{j=1}^{3}\frac{(o_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}}$

$$\frac{(108-102.33)^2}{102.33} + \frac{(87-102.33)^2}{102.33} + \cdots . \frac{(49-50.33)^2}{50.33} = \mathbf{10.597}$$

(From tables) $X^2_{0.05\ ;(4)}$ = **9.4877**

Conclusion: Because $X^2_{(r-1)(c-1)} = 10.597 > X^2_{0.05\ ;(4)} = 9.4877$ we REJECT Ho and conclude that at a 5% level of significance there is an evidence that the proportions of voters favoring candidates A, B, C, differ from ward to ward.