

Chapter 13 Categorical Data Analysis

13.1 Categorical Data and the Multinomial Experiment

Recall **Variable:** _____(numerical) variable (i.e. # of students, temperature, height,).

_____ (non-numerical, categorical) variable (i.e. color of hair; brand of car; type of blood; etc)

Binomial experiment: Categorical variable only has _____possible outcomes (success or failure).

Multinomial experiment: Categorical variable has _____possible outcomes.

• **Properties of the Multinomial experiment:**

1. The experiment consist of _____ trials,
2. There are _____outcomes to each trial. These outcomes are sometimes called _____, categories, or _____;
3. The probabilities of the k possible outcomes, denoted by _____, remain the same from trial to trial, where $p_1 + p_2 + \dots + p_k = 1$.
4. The trials are _____;
5. The random variables of interest are the _____, n_1, n_2, \dots, n_k , the number of observations that fall in each of the k categories.

Example1. To study the percentage of specific educational level (high school, BS, MS, PHD) of employee in a big company, randomly choose 1000 employee and the result is listed below.
Is this a multinomial experiment?

Ed. level	High school(1)	BS(2)	MS(3)	PHD(4)
counts	55	678	197	70

1. $n =$ _____ identical trials;
2. Each trial has $k =$ _____ possible outcomes;
3. Suppose p_i is the true probability that employee is in level i , $P_1 = 5\%$, $P_2 = 69\%$, $P_3 = 20\%$, $P_4 = 6\%$, then it keeps _____for the 1000 trials;
4. The trials are _____, the education level for one employee does not have effect on the education level for any other employee;
5. The random variables of interest are the _____-of employee who fall into each of the four education levels. We can denote the four cell counts as n_1, n_2, n_3, n_4 .

13.2 Testing categorical Probabilities: one categorical variable (one-way table)

- **One-way table Analysis: Test of a hypothesis about**_____

(_____)

$$H_0 : p_1 = p_{1,0} \quad p_2 = p_{2,0} \quad \dots \quad p_k = p_{k,0} \quad (P_{i,0} \text{ is the hypothesized probability})$$

$$H_a : \text{_____}$$

Test statistic: _____, where $E_i = np_{i,0}$ (**expected cell count**)

Rejection region: _____, where χ^2_α has (_____) df. (Table VII, p798)

Conclusion.

- **The properties of χ^2 distribution:**

1. the area under the curve is _____;
2. _____skewed;
3. df is getting larger, χ^2 curve is close to _____curve.

- **Conditions required for a valid χ^2 test: one-way table**

1. _____ sample---Multinomial experiment
2. large sample size n, for each cell, the expect cell count_____.

TABLE VII Critical Values of χ^2

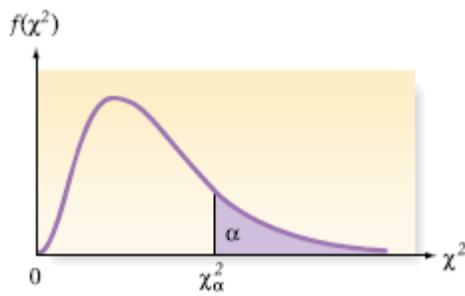


TABLE VII Continued

Degrees of Freedom	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	2.70554	3.84146	5.02389	6.63490	7.87944
2	4.60517	5.99147	7.37776	9.21034	10.5966
3	6.25139	7.81473	9.34840	11.3449	12.8381
4	7.77944	9.48773	11.1433	13.2767	14.8602
5	9.23635	11.0705	12.8325	15.0863	16.7496
6	10.6446	12.5916	14.4494	16.8119	18.5476
7	12.0170	14.0671	16.0128	18.4753	20.2777
8	13.3616	15.5073	17.5346	20.0902	21.9550
9	14.6837	16.9190	19.0228	21.6660	23.5893
10	15.9871	18.3070	20.4831	23.2093	25.1882
11	17.2750	19.6751	21.9200	24.7250	26.7569
12	18.5494	21.0261	23.3367	26.2170	28.2995
13	19.8119	22.3621	24.7356	27.6883	29.8194
14	21.0642	23.6848	26.1190	29.1413	31.3193
15	22.3072	24.9958	27.4884	30.5779	32.8013
16	23.5418	26.2962	28.8454	31.9999	34.2672
17	24.7690	27.5871	30.1910	33.4087	35.7185
18	25.9894	28.8693	31.5264	34.8053	37.1564
19	27.2036	30.1435	32.8523	36.1908	38.5822
20	28.4120	31.4104	34.1696	37.5662	39.9968
21	29.6151	32.6705	35.4789	38.9321	41.4010
22	30.8133	33.9244	36.7807	40.2894	42.7956
23	32.0069	35.1725	38.0757	41.6384	44.1813
24	33.1963	36.4151	39.3641	42.9798	45.5585
25	34.3816	37.6525	40.6465	44.3141	46.9278
26	35.5631	38.8852	41.9232	45.6417	48.2899
27	36.7412	40.1133	43.1944	46.9630	49.6449
28	37.9159	41.3372	44.4607	48.2782	50.9933
29	39.0875	42.5569	45.7222	49.5879	52.3356
30	40.2560	43.7729	46.9792	50.8922	53.6720
40	51.8050	55.7585	59.3417	63.6907	66.7659
50	63.1671	67.5048	71.4202	76.1539	79.4900
60	74.3970	79.0819	83.2976	88.3794	91.9517
70	85.5271	90.5312	95.0231	100.425	104.215
80	96.5782	101.879	106.629	112.329	116.321
90	107.565	113.145	118.136	124.116	128.299
100	118.498	124.342	129.561	135.807	140.169

Example 1: there are three candidates are running for the same elective position. We do a survey to determine the voting preferences of a random sample of 150 voters.

Results of voter-preference survey

candidate	1	2	3
count	61	53	36

a. Describe the qualitative variable of interest in the study. Give the levels associated with the variable.

b. At $\alpha = 0.05$, do the sample data provide sufficient evidence that the voters have a preference for any of the candidates?

c. Construct a 95% confidence interval for the true proportion of preference of candidate 1?

SPSS OUTPUT for Example1: **Voter-preference: Chi-Square Test** **Frequencies**

count			
	Observed N	Expected N	Residual
36	36	50.0	-14.0
53	53	50.0	3.0
61	61	50.0	11.0
Total	150		

Test Statistics	
	count
Chi-Square(a)	6.520
df	2
Asymp. Sig.	.038
a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 50.0.	

Example2: Violent crimes The U.S. FBI published the information in Crime in the United States. The distribution of violent crimes in 1995 is given below. A random sample of 500 violent-crime reports from last year yielded the frequency distribution shown also in the table.

Type of crime	Relative freq. of 1995	Freq. of last year
Murder	0.012	9
Forcible rape	0.054	26
robbery	0.323	144
Aggressive assault	0.611	321

Do the data provide sufficient evidence to conclude that last year's distribution of violent crimes has changed from the 1995 distribution? Use $\alpha = 0.01$.

$$H_0 : p_1 = \text{---} , p_2 = \text{---} , p_3 = \text{---} , p_4 = \text{---}$$

$$H_a : \text{-----}$$

(Last year's distribution is different from the 1995's distribution)

Example 3, An Edge in Roulette? An American roulette wheel contains 18 red numbers, 18 black numbers, and 2 green numbers. The following table shows the frequency with which the ball landed on each color in 200 trials.

Color	Red	Black	Green
Frequency	88	102	10

a. At the 5% significance level, do the data suggest that the wheel is out of balance?

b. If the wheel is on balance, how many times will be the ball expected to land on black numbers out of the 200 trials?

13.3 Testing categorical probabilities: Two categorical variables

Classification with respect to _____ **categorical variables (row and column variables)**

Interest of test: **the two categorical variables are** _____.

Example: Education and Gender:

	High school	BS	MS	PHD
male	26	373	126	49
female	29	305	71	21

- Observed counts of two-way (contingency) table:

	Col. 1	Col. 2	...	total
Row. 1				
Row. 2				
....				
total				

- Probabilities of contingency table:

	Col. 1	Col. 2	...	total
Row. 1				
Row. 2				
...				
total				

Observed cell count: _____,

Marginal count: _____.

Total count: $n =$ _____

Cell probability: _____

Marginal probability: _____

- **Basic theory:** If the two categorical variables are _____, the _____ probability is the product of the corresponding _____ probabilities.

Now let's find the **expected cell count**.

For example, $E_{11} = \underline{\hspace{2cm}}$, P_{11} is the true cell probability,

If the two categorical variables are independent, we have $P_{11} = \underline{\hspace{2cm}}$,

Then, $E_{11} = np_{11} = np_{r1}p_{c1}$

The $\underline{\hspace{2cm}}$ expected cell count: $\hat{E}_{11} = n\hat{p}_{11} = np_{r1}p_{c1} = n\left(\frac{r_1}{n}\right)\left(\frac{c_1}{n}\right) = \underline{\hspace{2cm}}$

Similarly, $\hat{E}_{12} = \underline{\hspace{1cm}}$, $\hat{E}_{21} = \underline{\hspace{1cm}}$, $\hat{E}_{22} = \underline{\hspace{1cm}}$

So the general formula for the **estimated expected cell count** is: $\hat{E}_{ij} = \underline{\hspace{2cm}}$

- Two-way table analysis: **Chi-Square test for $\underline{\hspace{4cm}}$ of two categorical variables**

H_0 : the two categorical variables are $\underline{\hspace{2cm}}$ ($P_{ij} = P_{ri}P_{cj}$)

H_a : the two categorical variables are $\underline{\hspace{2cm}}$

Test statistic: $\chi^2 = \underline{\hspace{2cm}}$, where $\hat{E}_{ij} = \underline{\hspace{1cm}}$

Rejection region: $\chi^2 > \chi^2_{\alpha}$, where χ^2_{α} has $\underline{\hspace{2cm}}$ df. (Table VII, p798)

Conclusion.

- **Conditions required for a valid χ^2 test: contingency tables**

1. The n observed counts are $\underline{\hspace{2cm}}$ sample from the population of interest,
2. n will be large enough so that the expected count for each cell $E(n_{ij}) \geq 5$.

Example1: Hiring status and Gender, Take a random sample of 80 job applicants at Mega-mart. The result is listed below. Consider hiring status and gender. At $\alpha = 0.05$, conduct a test of hypothesis to determine if gender and hiring status are dependent?

	Hired	Not Hired	
Male	14	32	
Female	14	20	

SPSS output for Example 2: Hiring status and Gender,

Case Processing Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
gender * hire	80	100.0%	0	.0%	80	100.0%

gender * hire Crosstabulation					
			hire		Total
			no	yes	
gender	male	Count	32	14	46
		Expected Count	29.9	16.1	46.0
	female	Count	20	14	34
		Expected Count	22.1	11.9	34.0
Total		Count	52	28	80
		Expected Count	52.0	28.0	80.0

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.992(b)	1	.319		
Continuity Correction(a)	.576	1	.448		
Likelihood Ratio	.988	1	.320		
Fisher's Exact Test				.351	.224
N of Valid Cases	80				
a Computed only for a 2x2 table					
b 0 cells (.0%) have expected count less than 5. The minimum expected count is 11.90.					

Example2: Education level and Gender:

A survey result about education level and gender based on 1000 employee in a company is listed below. At $\alpha = 0.10$, is there evidence to indicate that gender and education level are dependent?

	High school	BS	MS	PHD
male	26	373	126	49
female	29	305	71	21