CHAPTER 10- Categorical Data

Monday, 22 December 2025 3:53 PM

* In this chapter we are comparing 2 of the **binomial Parameter (Proportion P)**
* we will be using

- [1] Normal-Theory method
  - Critical value method
  - P-value method
- [2] Contingency-Table method [using Chi-square table]
  - 2x2 contingency table
  - RXC contingency table

[1] Normal theory method
we are testing the difference $\hat{P}_1 - \hat{P}_2$ (we almost always work on unpaired samples)

* if the difference is far from 0 (H_1 is true) then H_0 is rejected

* we always use normal distribution (no t-tables)

→ we make a $\hat{P}_1 - \hat{P}_2$ normally distributed distribution defined as

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim N(0,1)$$

since it isn't known we estimate it by P_1 and P_2 using Pooled Proportion

$$\hat{p} = \frac{n_1\hat{P}_1 + n_2\hat{P}_2}{n_1+n_2} = \frac{x_1+x_2}{n_1+n_2}$$

you use one of them according to what is given in the question

* For the hypotheses $H_0: A = P_2$ vs $H_1: P_1 \neq P_2$
  $H_0: P_1 - P_2 = 0$ vs $H_1: P_1 - P_2 \neq 0$
  - note that we will be using 2-Sided tests

① Critical value method :-

- the critical value

$$Z_{1-\alpha/2}$$

- Test statistic = $\frac{|\hat{P}_1 - \hat{P}_2| - \left(\frac{1}{2n_1}+\frac{1}{2n_2}\right)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$

(it doesn't matter which you use $\hat{P}_1 - \hat{P}_2$ or $\hat{P}_2 - \hat{P}_1$ we will be using the absolute value of the difference)

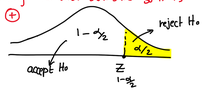
where $\hat{p} = \frac{n_1\hat{P}_1 + n_2\hat{P}_2}{n_1+n_2} = \frac{x_1+x_2}{n_1+n_2}$ (the number of events in the first and second samples)

and $\hat{q} = 1-\hat{p}$

- if T.S > C.V ⇒ reject H_0

T.S < C.V ⇒ Accept H_0

↳ - C.V value won't be needed since we are using the absolute value in calculating the test statistic so it is always +



② P-value method :

- P-value = $2P(Z > t.s)$ (we are always working on t.s > 0)

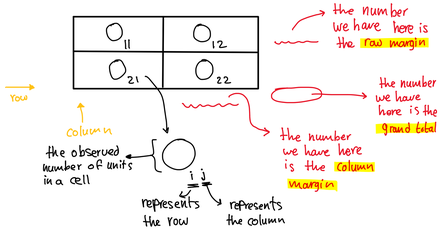
⇒ P-value > α ⇝ accept H_0
P-value < α ⇝ reject H_0

* Note that we use normal theory methods when $n_1\hat{P}\hat{q} \geq 5$ and $n_2\hat{P}\hat{q} \geq 5$

[2] contingency-Table method
↳ it is an appropriate way to display data that can be classified by two different variables. One variable is assigned to the rows and the other is to the column

(to test weather the 2 variables are independent or not)

⊛ 2x2 Contingency-Table
↳ it is made of 2 rows, 2 columns } 4 cells



* using the contingency table we can form an expected table when $H_1: P_1 = P_2 = P$ is true

E_{11}	E_{12}
E_{21}	E_{22}

(the sum of these expected values should be >5 to use Chi-square test)

$$E = \frac{\text{row margin of the cell} \times \text{column margin of the cell}}{\text{grand total}}$$

⊛ RXC contingency Table
it is made of R rows and C columns

O_{11}	O_{12}	O_{13}	O_{14}
O_{21}	O_{22}	O_{23}	O_{24}

↓ the expected table

E_{11}	E_{12}	E_{13}	E_{14}
E_{21}	E_{22}	E_{23}	E_{24}

$$E = \frac{\text{row margin of the cell} \times \text{column margin of the cell}}{\text{grand total}}$$

* Test for homogeneity of binomial Proportion : test weather the proportions of 2 independent samples are the same ⇒ the normal theory method

* Test of independance : test whether there is some association between 2 reported measures of a characteristic ⇒ the contingency table method

Yates-corrected Chi-square test statistic for 2x2 contingency table (because we subtract 0.5)

$$T.S = \chi^2 = \frac{(|O_{11}-E_{11}|-0.5)^2}{E_{11}} + \frac{(|O_{12}-E_{12}|-0.5)^2}{E_{12}} + \frac{(|O_{21}-E_{21}|-0.5)^2}{E_{21}} + \frac{(|O_{22}-E_{22}|-0.5)^2}{E_{22}}$$

* when finding the critical value for 2x2 table

$$\chi^2_{1,1-\alpha}$$ (we don't square the answer (it is called Chi-squared))

degree of freedom $(R-1)(C-1)$

$(2-1)(2-1) = 1$ ⇒ so it is always 1 in 2x2 table

Chi-square test statistic for RXC Contingency table

$$T.S = \chi^2 = \frac{(O_{11}-E_{11})^2}{E_{11}} + \cdots + \frac{(O_{RC}-E_{RC})^2}{E_{RC}}$$

* when finding the critical value for RXC table

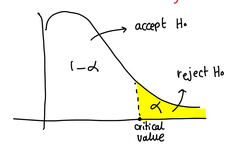
$$\chi^2_{(R-1)(C-1),\,1-\alpha}$$

degree of freedom

(*) using the critical value method :-

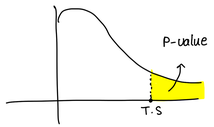
- if the test statistic > critical value ⇒ reject H_0
- if the test statistic ≤ critical value ⇒ accept H_0

(we either use rules of 2x2 or RXC)



(*) using P-value method :-

- if P-value = $P(\chi^2 > t.s) > \alpha$ ⇒ accept H_0
- if P-value = $P(\chi^2 > t.s) < \alpha$ ⇒ reject H_0



* we are always working on the right Side

Question

if given : R = 4, C = 6 T.S = X^2 = 11.04

Find P-value :-

P-value = $P(X^2 > T.S)$
= $P(X^2 > 11.04)$
= $1 - P(X^2 < 11.04)$
↳ from the table (df = (4-1)*(6-1) = 15)

TABLE 6 Percentage points of the chi-square distribution $G_{\nu}^{2}(P)$

ν	.005	.01	.025	.05	.10	.90	.75	.90	.95	.975	.99	.995
1	0.0^4393^a	0.0^3157^a	0.0^3982^a	0.0^2393^a	0.0158	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	0.0100	0.0201	0.0506	0.103	0.211	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60	13.81
3	0.072	0.115	0.216	0.352	0.584	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	0.207	0.297	0.484	0.711	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	0.412	0.554	0.831	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	0.676	0.872	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	0.989	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.95	26.12
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.76	31.26
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	4.60	5.23	6.27	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80	37.70

(*) Chi Square goodness-of-fit test

↳ used to check whether observed data fits a claimed or expected distribution using Chi-square test

we will check weather a set of data is normally distributed or not

→ given a frequency distribution :-
① find the number of groups g

② identify (k) → the number of Parameters that the data could be distributed in
↳ for normal distribution we have
- mean
- variance

⇒ k = 2

③ Find the t.s using the given data

④ compare t.s to the critical value $\chi^2_{g-k-1,\,1-\alpha}$ (degree of freedom)

T.S > C.V reject H_0 ↓ the distribution isn't normal

T.S < C.V accept H_0 ↓ the distribution is normal

* An externally specified model is when the Parameters of the Probability model were specified a priori, without using the present sample data, in this case k = 0 and $(X^2 \sim \chi^2_{g-1})$ follows Chi-square distribution of g-1 degree of freedom

Question

Test for goodness of fit of the normal Probability model :-

Frequency distribution of mean diastolic blood pressure for adults 30–69 years old in a community-wide screening program in East Boston, Massachusetts

Group (mm Hg)	Observed frequency	Expected frequency	Group	Observed frequency	Expected frequency
1 <50	57	69.0	5 ≥80, <90	4604	4538.6
2 ≥50, <60	330	502.5	6 ≥90, <100	2119	2545.9
3 ≥60, <70	2132	2018.4	7 ≥100, <110	659	740.4
4 ≥70, <80	4584	4200.9	8 ≥110	251	120.2
			Total	14,736	14,736

① k = 2
② g = number of groups = 8
③ T.S = $\sum \frac{(O_i - E_i)^2}{E_i} = 326.2$ under H_0 (that the data is normally distributed)
④ Critical value = $\chi^2_{g-k-1,\,1-\alpha} = \chi^2_{5,\,0.95} = 11.07$ (from the table)

⑤ decide ⇒ since T.S > C.V then reject H_0 ⇒ the data can't be normally distributed