Association between categorical variables

Income	NotTooHappy	PrettyHappy	VeryHappy	Total
AboveAverage	21(7%)	159(55%)	110(38%)	290(100%
Average	53(8%)	372(58%)	221(34%)	646(100%
BelowAverage	94(22%)	249(59%)	83(19%)	426(100%

source: Franklin and Agresti, 2007, p. 486.

Q: Are Income and Happiness independent?

Independence between categorical variables

Income	NotTooHappy	PrettyHappy	VeryHappy	Total
Above Average	12%	57%	31%	100%
Average	12%	57%	31%	100%
Below Average	12%	57%	31%	100%

Two categorical variables are independent (have no relationship) if the population conditional distributions for one of them are identical at each category of the other.

Test independence between two categorical variables

Gender	believe	not believe	total
male	60(60%)	40(40%)	100(100%)
female	150(75%)	50(25%)	200(100%)
total	210	90	300

Expected cell count under independence

Under independence:

expected cell count= $\frac{\text{Row Total*Column Total}}{\text{Table Total}}$

expected cell count for

cell(1,1)=row 1 total*column 1 total/table total= $\frac{100*210}{300} = 70.$

cell (1, 2)= $\frac{100*90}{300} = 30$

cell(2,1)= $\frac{200*210}{300} = 140$

cell(2,2)= $\frac{200*90}{300} = 60.$

Expected cell count

Gender	believe	not believe	total
male	70(70%)	30(30%)	100(100%)
female	140(70%)	60(30%)	200(100%)
total	210	90	300

chi-square test

chi-square test statistic

$$\chi^2 = \sum \frac{(\text{Observed Cell Count - Expected Cell Count})^2}{\text{Expected Cell Count}} = \sum \frac{(O-E)^2}{E}$$

d.f.=(r-1)(c-1), where r=number of rows, c=number of columns.

Sample size requirement: each expected cell count ≥ 5.

In this example,

H_0 : Gender and Belief are independent.

H_1 : Gender and Belief are not independent.

$$\chi^2 = \frac{(60-70)^2}{70} + \frac{(40-30)^2}{30} + \frac{(150-140)^2}{140} + \frac{(50-60)^2}{60} = 7.14$$

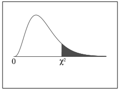
d.f.= $(2-1)*(2-1) = 1$.

The P-value $= P(\chi^2 > 7.14) < 0.01$.

Reject H_0. Data show that gender and belief are dependent (associated).

chi-square distribution table

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Exercise

	Internet	Internet
Community Type	Broadband	No broadband
Urban	300(0.52)	276(0.48)
Suburban	521(0.49)	542(0.51)
Rural	174(0.31)	387(0.69)

Q: Are Internet Connection Type and Community Type independent?

Check your understanding exercise

	heart attack	no heart attack	total
placebo	28	656	684
aspirin	18	658	676
total	46	1314	1360

Test the null hypothesis that having a heart attack is independent of whether one takes placebo or aspirin. Use $\alpha = 0.05$.

Solutions

H_0: Having a heart attack is independent of whether one takes placebo or aspirin.
H_1 : Having a heart attack is NOT independent of whether one takes placebo or aspirin.
Expected counts for the 4 cells are:
cell (1, 1): $\frac{684*46}{1360} = 23.1$
cell(1, 2): $\frac{684*1314}{1360} = 660.9$
cell (2, 1): $\frac{676*46}{1360} = 22.9$
cell (2, 2): $\frac{676*1314}{1360} = 653.1$
$\chi^2 = \frac{(28-23.1)^2}{23.1} + \frac{(656-660.9)^2}{660.9} + \frac{(18-22.9)^2}{22.9} + \frac{(658-653.1)^2}{653.1} = 2.16$
d.f.=$(2-1)(2-1) = 1$.
The p-value =$P(\chi^2 > 2.16) > 0.10$. We fail to reject H_0. There is not sufficient evidence that having a heart attack depends on whether one takes placebo or aspirin.

Testing $H_0 : p_1 = p_2$

Placebo: 28 had heart attacks out of 684 people.
Aspirin: 18 had heart attacks out of 676 people.
p_1 : probability of getting heart attacks for a person who takes placebo,
p_2 : probability of getting heart attacks for a person who takes aspirin.
$H_0 : p_1 = p_2$ (independence)
$H_1 : p_1 \neq p_2$ (dependence)
$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}}$, $\hat{p}_1, \hat{p}_2$ are sample proportions
Pooled proportion: $\hat{p} = \frac{x_1+x_2}{n_1+n_2}$, x_i: number of successes in sample i.

Example continued

$\hat{p}_1 = 28/684 = 0.0409, \hat{p}_2 = 18/676 = 0.0266, \hat{p} = \frac{28+18}{684+676} = 0.0338,$
$z = \frac{0.0409 - 0.0266}{\sqrt{0.0338 * 0.9612(\frac{1}{684} + \frac{1}{676})}} = 1.46$
P-value = $2P(z > 1.46) = 0.144$.

exercise

Among a random sample of 160 men, 55 had nightmare often,
among a random sample of 192 women, 60 had nightmare often.
Test $H_0 : p_M = p_W$
vs $H_1 : p_M \neq p_W$
using both the chi square test and the z test.

solutions

$\hat{p}_M = 55/160 = 0.344, \hat{p}_W = 60/192 = 0.313, \hat{p} = \frac{55+60}{160+192} = 0.327.$
$z = \frac{0.344-0.313}{\sqrt{0.327*0.673(1/160+1/192)}} = 0.62.$
The P-value = $2P(z > 0.62) = 0.535.$

solutions

	nightmare often	not often	total
man	55(52.3)	105(107.7)	160
woman	60(62.7)	132(129.3)	192
total	115	237	352

The expected cell counts are in parenthesis.
$\chi^2 = \frac{(55-52.3)^2}{52.3} + \frac{(105-107.7)^2}{107.7} + \frac{(60-62.7)^2}{62.7} + \frac{(132-129.3)^2}{129.3} = 0.38.$
d.f.=1, p-value is between 0.1 and 0.9.

Z test for the believe in heaven example

Gender	believe	not believe	total
male	60(60%)	40(40%)	100(100%)
female	150(75%)	50(25%)	200(100%)
total	210	90	300

$H_0 : p_M = p_W$
$H_1 : p_M \neq p_w$.
$\hat{p}_M = 60/100 = 0.60, \hat{p}_W = 150/200 = 0.75, \hat{p} = 210/300 = 0.7.$
$z = \frac{0.60-0.75}{\sqrt{0.7*0.3*(1/100+1/200)}} = -2.67,$
p-value = $2p(z < -2.67) = 0.0076.$

Short cut formula for 2 by 2 table:

	Level 1	Level 2	Total
Level 1	a	b	a+b
Level 2	c	d	c+d
Total	a+c	b+d	

$\chi^2 = \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$ where $N = a + b + c + d$