Hypothesis Testing 6

Day 16 (3/10/20)

4.7 Chi-square Tests of Homogeneity and Independence

In this section, we compare frequency distributions of two or more populations.

Example A sample of $n_1 = 873$ subjects from Canada and $n_2 = 624$ subjects from the U.S. were classified according to their BMI category. The data is summarized below. We want to test whether the frequency distributions are the same for Canada and US.

	Canada	US
Underweight	297	156
Normal	498	349
Overweight	61	75
Obese	17	44
Total	873	624

For each country, the observed frequencies are realizations of a multinomial distribution. Let $(p_{11}, p_{21}, p_{31}, p_{41})$ denote the multinomial probabilities for Canada, and $(p_{12}, p_{22}, p_{32}, p_{42})$ denote the probabilities for the US. We may summarize as follows:

	Canada	US
Underweight	p_{11}	p_{12}
Normal	p_{21}	p_{22}
Overweight	p_{31}	p_{32}
Obese	p_{41}	p_{42}
Total	1.0	1.0

If n_1 and n_2 are large and the two samples are independent, then

$$Q = \sum_i^4 \sum_j^2 \frac{(X_{ij} - n_j p_{ij})^2}{n_j p_{ij}} = \sum_i^4 \frac{(X_{i1} - n_1 p_{i1})^2}{n_1 p_{i1}} + \sum_i^4 \frac{(X_{i2} - n_2 p_{i2})^2}{n_2 p_{i2}} \tag{1}$$

is the sum of two independent χ^2_{k-1} random variables, and therefore has a χ^2_{2k-2} distribution. How do we compute the null expected counts $E_{ij} = n_j p_{ij}$? Under the null hypothesis,

$$\begin{pmatrix} p_{11} \\ p_{21} \\ p_{31} \\ p_{41} \end{pmatrix} = \begin{pmatrix} p_{12} \\ p_{22} \\ p_{32} \\ p_{42} \end{pmatrix} \equiv \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix}$$

which can be estimated along the right margin of the table

	Canada	US	Total	
Underweight	297	156	451	$\rightarrow \hat{p}_1 = .302$
Normal	498	349	847	$\hat{p}_2 = .566$
Overweight	61	75	136	$\hat{p}_3 = .091$
Obese	17	44	61	$\hat{p}_4 = .041$
Total	n_1=873	n_2=624	n=1497	

(2)

The expected counts under homogeneity are

	Canada	US			
Underweight	873(.302)	624(.302)		264.2	188.8
Normal	873(.566)	624(.566)	or	493.9	353.1
Overweight	873(.091)	624(.091)		79.3	56.7
Obese	873(.041)	624(.041)		35.6	25.4
Total	873	624			

Finally,

$$\begin{aligned} Q &= \frac{(297-264.2)^2}{264.2} + \frac{(156-188.8)^2}{188.8} + \cdots + \frac{(17-35.6)^2}{35.6} + \frac{(44-25.4)^2}{25.4} \\ &= 4.08 + 5.71 + 0.03 + 0.05 + 4.23 + 5.91 + 9.70 + 13.57 \\ &= 43.271 \end{aligned}$$

What is the null distribution of Q? From equation (1), we see that Q is the sum of two χ^2_3 random variables, for a total of 6 degrees of freedom. However, from (2) we need to subtract 3 degrees of freedom because we estimated 3 nuisance parameters (the fourth is determined by the first three). Since $Q = 43.71$ is larger than $\chi^2_{.05,3\ df} = 7.814$, we reject the null hypothesis of homogeneity of frequency distributions. The p-value is $P[\chi^2_{.05,3\ df} > 43.71] < .0001$.

In general:

- $H_0 : (p_{11}, \ldots, p_{k1}) = (p_{12}, \ldots, p_{k2})$ vs H_1 : Not equal
- $E_{ij} = n_j \hat{p}_i = n_j \frac{X_{i1}+X_{i2}}{n_1+n_2}$
- $Q = \sum_{i=1}^k \sum_{j=1}^2 \frac{(X_{ij}-E_{ij})^2}{E_{ij}}$
- Reject H_0 if $Q > \chi^2_{\alpha,\ k-1\ df}$

since the degrees of freedom is $(2k-2)-(k-1) = k-1$.

5 Chi-square test of independence

Consider a random sample of observations where each observation is classified by two attributes. We want to test for independence of the two attributes.

Example A sample of $n = 22,361$ Scottish children were cross-classified according to color of hair and color of eyes. Is eye color independent of hair color, or are they associated? The data $\{X_{ij}\}$ is presented in an $a \times b$ contingency table below.

	Fair	Red	Medium	Dark	Black
Blue	1368	170	1041	398	1
Light	2577	474	2703	932	11
Medium	1390	420	3826	1842	33
Dark	454	255	1848	2506	112

Let p_{ij} denote the probability that an individual falls in the (ij)th cell. Write the probabilities in a cross classification table.

	Fair	Red	Medium	Dark	Black	Total
Blue	p_{11}	p_{12}	p_{13}	p_{14}	p_{15}	$p_{1.}$
Light	p_{21}	p_{22}	p_{23}	p_{24}	p_{25}	$p_{2.}$
Medium	p_{31}	p_{32}	p_{33}	p_{34}	p_{35}	$p_{3.}$
Dark	p_{41}	p_{42}	p_{43}	p_{44}	p_{45}	$p_{4.}$
Total	$p_{.1}$	$p_{.2}$	$p_{.3}$	$p_{.4}$	$p_{.5}$	1.0

There are $ab = 20$ random variables X_{ij}, then

$$Q = \sum_{i=1}^{4}\sum_{j=1}^{5} \frac{(X_{ij} - np_{ij})^2}{np_{ij}}$$

has an approximate chi-square distribution with (20-1) degrees of freedom. Since the $\{p_{ij}\}$ are unspecified, they need to be estimated. First, estimate the marginal probabilities $\{p_{i.}\}$ of falling in the ith row or and the marginal probabilities $\{p_{.j}\}$ of falling in the jth column.

	Fair	Red	Medium	Dark	Black	Total	
Blue	1368	170	1041	398	1	2978	$\to \hat{p}_{1.} = .13$
Light	2577	474	2703	932	11	6697	$\hat{p}_{2.} = .30$
Medium	1390	420	3826	1842	33	7511	$\hat{p}_{3.} = .34$
Dark	454	255	1848	2506	112	5175	$\hat{p}_{2.} = .23$
Total	5789	1319	9418	5678	157	22361	
	$\downarrow$						
	$\hat{p}_{.1} = .26$	$\hat{p}_{.2} = .06$	$\hat{p}_{.3} = .42$	$\hat{p}_{.4} = .25$	$\hat{p}_{.5} = .01$		

The null hypothesis of independence may be written as $H_0 : p_{ij} = p_{i.}p_{.j}$, so the probabilities $\{p_{ij}\}$ are estimated as $\hat{p}_{ij} = (\hat{p}_{i.})(\hat{p}_{.j})$

	Fair	Red	Medium	Dark	Black
Blue	(.13)(.26)	(.13)(.06)	(.13)(.42)	(.13)(.25)	(.13)(.01)
Light	(.30)(.26)	(.30)(.06)	(.30)(.42)	(.30)(.25)	(.30)(.01)
Medium	(.34)(.26)	(.34)(.06)	(.34)(.42)	(.34)(.25)	(.34)(.01)
Dark	(.23)(.26)	(.23)(.06)	(.23)(.42)	(.23)(.25)	(.23)(.01)

Multiplying by $n = 22,361$ the expected cell counts are

	Fair	Red	Medium	Dark	Black
Blue	771.0	175.7	1254.3	756.2	20.9
Light	1733.8	395.0	2820.6	1700.5	47.0
Medium	1944.5	443.0	3163.5	1907.2	52.7
Dark	1339.7	305.3	2179.6	1314.1	36.3

and finally

$$Q = \frac{(1368 - 771.0)^2}{771.0} + \frac{(170 - 175.7)^2}{175.7} + \cdots + \frac{(2506 - 1314.1)^2}{1314.1} + \frac{(112 - 36.30)^2}{36.3}$$
$$= 3683.9$$

What is the null distribution of Q? Before estimating nuisance parameters, we started with (20-1) degrees of freedom. We estimated (4-1) row margin probabilities and (5-1) column margin probabilities and end up with (20-1)-(4-1)-(5-1)=12 degrees of freedom. For a general $a \times b$ contingency table, the degrees of freedom are

$$df = (ab-1) - (a-1) - (b-1) = (a-1)(b-1)$$

```
> #### Test of homogeneity in R
> can<-c(297,498,61,17)      #### Observed counts Canada
> us<-c(156,349,75,44)       #### Observed counts US
> n1<-sum(can)
> n2<-sum(us)
> n<-n1+n2
> n
[1] 1497
> phom<-(can+us)/n            #### Probs under homogeneity
> phom
[1] 0.30260521 0.56579826 0.09084836 0.04074816
> ecan<-n1*phom              #### Expected counts Canada
> ecan
[1] 264.17435 493.94188  79.31062  35.57315
> eus<-n2*phom               #### Expected counts US
> eus
[1] 188.82565 353.05812  56.68938  25.42685
> qhom<-sum((can-ecan)^2/ecan) + sum((us-eus)^2/eus)  # Test statistic
> qhom
[1] 43.27108
> qchisq(.95, df=3)          #### Critical value
[1] 7.814728
> 1-pchisq(qhom,df=3)        #### P-value
[1] 2.15553e-09
> (can-ecan)^2/ecan          #### Post-hoc analysis
[1] 4.07883426 0.03334058 4.22741425 9.69725198
> (us-eus)^2/eus
[1]  5.70644600  0.04664475  5.91431513 13.56682849
>
> canmat<-cbind(can,us)      #### Using built-in chisq.test()
> canmat
     can  us
[1,] 297 156
[2,] 498 349
[3,]  61  75
[4,]  17  44
> chisq.test(canmat)
```

```
Pearson's Chi-squared test

data:  canmat
X-squared = 43.271, df = 3, p-value = 2.156e-09
```

Test of independence in R

```
> #### Test of independence in R
> c1<-c(1368,2577,1390,454)
> c2<-c(170,474,420,255)
> c3<-c(1041,2703,3826,1848)
> c4<-c(398,932,1842,2506)
> c5<-c(1,11,33,112)
> ind_mat<-cbind(c1,c2,c3,c4,c5)
> ind_mat
       c1  c2   c3   c4  c5
[1,] 1368 170 1041  398   1
[2,] 2577 474 2703  932  11
[3,] 1390 420 3826 1842  33
[4,]  454 255 1848 2506 112
> rowtotal<-apply(ind_mat,1,sum)
> coltotal<-apply(ind_mat,2,sum)
> rowtotal
[1] 2978 6697 7511 5175
> coltotal
  c1   c2   c3   c4   c5
5789 1319 9418 5678  157
> n<-sum(rowtotal)
> n
[1] 22361
> prow<-rowtotal/n           #### Row marginal probabilities
> pcol<-coltotal/n           #### Col marginal probabilities
> pij<-prow%*%t(pcol)        #### Cell probabilities assuming independence
> pij
             c1          c2         c3         c4           c5
[1,] 0.03447830 0.007855739 0.05609200 0.03381720 0.0009350652
[2,] 0.07753565 0.017666180 0.12614108 0.07604895 0.0021027978
[3,] 0.08695987 0.019813451 0.14147314 0.08529247 0.0023583865
[4,] 0.05991443 0.013651259 0.09747351 0.05876562 0.0016249035
> mat_exp<-n*pij              #### Matrix of expected counts
> mat_exp
            c1       c2       c3        c4       c5
[1,]  770.9692 175.6622 1254.273  756.1864 20.90899
[2,] 1733.7746 395.0335 2820.641 1700.5307 47.02066
[3,] 1944.5096 443.0486 3163.481 1907.2250 52.73588
```

```
[4,] 1339.7467 305.2558 2179.605 1314.0580 36.33447
> matq<-(ind_mat-ind_exp)^2/ind_exp                  #### (Obs-Exp)^2/Exp
> matq
           c1        c2         c3          c4         c5
[1,] 462.3347  0.182511  36.264408  169.663849  18.956820
[2,] 410.1047 15.785286   4.906448  347.326505  27.593998
[3,] 158.1277  1.199048 138.749519    2.230623   7.385957
[4,] 585.5937  8.273866  50.450401 1081.174405 157.571408
> q<-sum(matq)                                 #### Test statistic
> q
[1] 3683.876
> qchisq(.95,df=12)                            #### Critical value
[1] 21.02607
> 1-pchisq(q,df=12)                            #### P-value
[1] 0
> chisq.test(ind_mat)                          #### Using built-in chisq.test()

Pearson's Chi-squared test

data:  ind_mat
X-squared = 3683.9, df = 12, p-value < 2.2e-16

>
```