L17

Wednesday, December 5, 2018
10:43 AM

<<Lecture17_Chi_SS.pdf>>

Stat 324 – Introduction to Statistics for Engineers

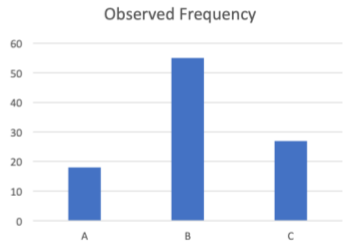
LECTURE 17: CHI SQUARED TESTS OF CATEGORICAL RELATIONSHIPS

OTT AND LONGNECKER: 10.4-10.5 PG 501-5

Categorical Classifications

Ex 1: The offspring produced by a cross between two types of plants can be any of the three genotypes denoted by A, B, and C. A theoretical model of gene inheritance suggests that the offspring of types A, B, and C should be in the ratio 1:2:1. For experimental verification, 100 plants are bred by crossing the two given types. Their genetic classifications are recorded in the table. Do these data contradict the genetic model?

Classification of Crossbred Plants				
Genotype	A	B	C	Total
Observed Frequency	18	55	27	100



What tests could we use to "sort of" answer the question of interest?

Ideally we would like to perform 1 test that combines these statements into one hypothesis.

In general, when we wish to do tests that help us decide whether categorical data is consistent with a particular population model, we call these **Goodness of Fit tests**.

Classification Into Categories

...offspring of types A, B, and C should be in the ratio 1:2:1. Do these data contradict the genetic model?

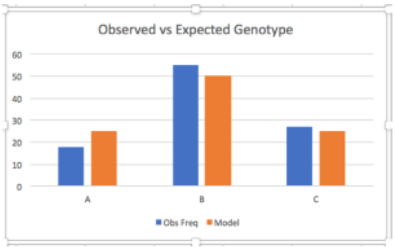
1+2+1 = 4

Our hypotheses: $H_o: \pi_a = \frac{1}{4}, \pi_b = \frac{2}{4}, \pi_c = \frac{1}{4}$; H_a: the data do not match the prescribed model.

We start by computing counts we would *expect under the null*: $\mathrm{EV}_{count\ for\ outcome\ i} =$ n * π_i

Where n is the total sample size and π_i is the probability of outcome i from the null.

Classification of Crossbred Plants				
Genotype	A	B	C	Total
Observed Frequency	18	55	27	100
Expected Frequency	$\frac{1}{4} \times 100 = 25$	$\frac{2}{4} \times 100 = 50$	$\frac{1}{4} \times 100 = 25$	100



(*Notice, Expected Counts DO NOT NEED TO BE WHOLE NUMBERS- do not round them)

Classification Into Categories

$H_o: \pi_a = \frac{1}{4}, \pi_b = \frac{2}{4}, \pi_c = \frac{1}{4}$; H_a: the data do not match the prescribed model.

To get an overall summary of fit for the 3 (k) categories, we need a new test statistic.

"Chi Squared" $\chi^2 = \sum_{i=1}^{k} \frac{(Observed\ Count - Expected\ Count)^2}{Expected\ Count}$

Classification of Crossbred Plants				
Genotype	A	B	C	Total
Observed Frequency	18	55	27	100
Expected Frequency	1/4 (100)=25	2/4(100)=50	1/4 (100)=25	100
$\frac{(Obs - Exp)^2}{Exp}$	$\frac{(18-25)^2}{25}$ 1.96	$\frac{(55-50)^2}{50}$ 0.5	$\frac{(27-25)^2}{25}$ 0.16	2.62

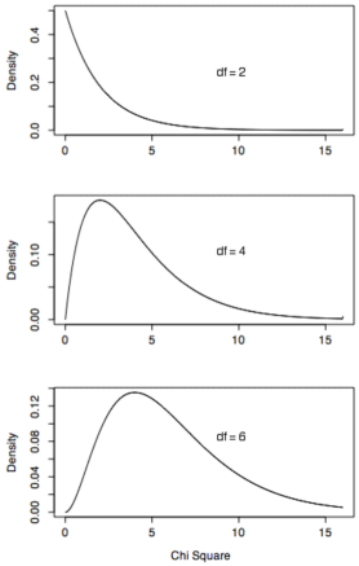
$\chi^2_{obs} =$ 1.96 + 0.5 + 0.16 = 2.62

The chi-squared goodness-of-fit test based on k levels for 1 variable, will have k-1 degrees of freedom. 3-1

Obtaining a p value from a chi-squared test statistic

Large chi-squared statistic constitute strong evidence against the null because the chi-squared statistic is large when the observed and expected counts do not agree closely.

To find a p value, we compare our observed test statistic to an approximate chi-squared distribution . The approximation improves as n increases and will be adequate if no E_i is <1 and no more than 20% of the E_i are less than 5.



Pvalue: $P(\chi^2_2 \geq 2.62) =$ $0.25 < \text{pvalue} < 0.75$ $= 0.269$ from R

low evidence against null

low evidence to suggest the data come from a population with a different genetic model

*Chi-squares are what you get when you square standard normals and add them up. The square of one standard normal is a chi-square on 1df, the sum of two squared standard normals is a chi-square on 2df, etc. They look a little like F distributions - they can't be smaller than zero, and they have a long right tail for small degrees of freedom.

In R:

```
> chisq.test(x=c(18,55,27), p=c(1/4, 2/4, 1/4), correct=FALSE)

        Chi-squared test for given probabilities

data:  c(18, 55, 27)
X-squared = 2.62, df = 2, p-value = 0.2698
```

```
> pchisq(2.62, df=2)
[1] 0.7301799
```

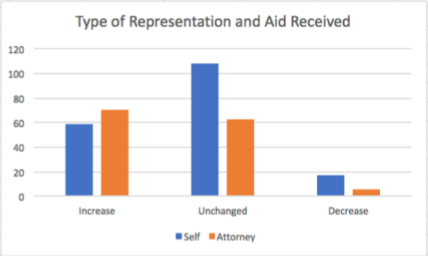
```
> pchisq(2.62, df=2, lower.tail=FALSE)
[1] 0.2698201
```

```
> 1-pchisq(2.62, df=2)
[1] 0.2698201
```

Relating Two Characteristics from 1 sample

Ex 2: Applicants for public assistance are allowed an appeals process when they feel unfairly treated. At such a hearing, the applicant may choose self-representation, or representation by an attorney. The appeal may result in an increase, decrease, or no change of the aid recommendation. Court records of 320 appeals cases are cross-classified below:

Type of	Amount of Aid		
Representation	Increased	Unchanged	Decreased
Self	59	108	17
Attorney	70	63	6



Does there appear to be a relationship between the type of representation and amount of aid?

When we did our independence calculations earlier, we had assumed that the data we had was the whole population of interest. With sample data, there usually appears to be some degree of dependence – we want a hypothesis test – is the perceived dependence in the same data more likely a result of variability rather than real dependence?

Relating Two Characteristics from 1 sample

We'll calculate the counts we would expect to see under the null of independence, and build the same χ^2 statistic.

H_o: The Amount of Aid and Type of Representation are independent

H_a: The Amount of Aid and Type of Representation are dependent(associated)

Observed Counts:

Type of Representation	Amount of Aid			
	Increased	Unchanged	Decreased	Total
Self	59	108	17	184
Attorney	70	63	6	139
Total	129	171	23	323

Expected Counts (under null):

P(Increased and Self)= $P(\text{Increased}) * P(\text{Self}) \approx \frac{129}{323} * \frac{184}{323}$

Expected Count of Increased and Self: $\frac{129}{323} * \frac{184}{323} * 323 = \frac{129 * 184}{323} = 73.49$

$\frac{(\text{Increased column Total}) * (\text{Self Row total})}{\text{Table total}}$

Relating Two Characteristics from 1 sample

We'll calculate the counts we would expect to see under the null, and build the same χ^2 statistic.

H_o: The Amount of Aid and Type of Representation are independent

Observed (Expected) Counts:

Type of Representation	Amount of Aid			
	Increased	Unchanged	Decreased	Total
Self	59 (73.49)	108 (97.41)	17 (13.10)	184
Attorney	70 (55.51)	63 (73.59)	6 (9.90)	139
Total	129	171	23	323

Expected Counts Calculations:

$55.51 = \frac{129 * 139}{323}$

$55.51 = 129 - 73.49$

self & decrease $23 - 9.9 = 13.10$

$\frac{23 * 184}{323} = 13.10$

* can use column or row totals in addition to $\frac{(\text{Row Tot})(\text{Col tot})}{\text{Table tot}}$ to get EV counts

Relating Two Characteristics from 1 sample

χ^2 components by each cell: $\frac{(Obs\ Count - Exp\ Count)^2}{Exp\ Count}$

$\chi^2_{obs} = \sum_{i=1}^{k} \frac{(Obs\ Count - Exp\ Count)^2}{Expected\ Count} = 12.02$

Type of Representation	Amount of Aid			
	Increased	Unchanged	Decreased	Total
Self	2.86	1.153593	1.161069	5.175206
Attorney	3.788	1.523958	1.536	6.84861
Total	6.648832	2.677551	2.697433	12.02382

Inc self: $\frac{(59 - 73.49)}{73.49}$

} 2

DF=(number of columns-1)*(number of rows-1)

*Excluding total row/columns

For our example: DF= $(3-1)(2-1) = 2*1 = 2$

Pvalue: $P(\chi^2_2 \geq 12.02) = 0.0025$

Conclusion: strong evidence the null, evidence suggest association between aid and representation

$0.0025 < 0.05$

Df corresponds to the number of cells that are free to vary,

ie the cell could take any value (within the limits implied by the totals), but then all remaining cells would be determined by the totals

```
> (Ex2.test<-chisq.test(ex2data, correct=FALSE))

        Pearson's Chi-squared test

data:  ex2data
X-squared = 12.005, df = 2, p-value
= 0.002473
```

More on χ^2

- Sample Size: no expected count (E) is <1 and no more than 20% of Es are less than 5
  - when some E's too small, combine categories in way that makes sense or do an exact test (Ex: Fisher's Exact)

- For **χ^2 test of independence:**
  - Data result from a single random sample from the whole population
  - Rejection of null hypothesis indicates only that the apparent association is not reasonably attributed to chance. It doesn't indicate about the strength or type of association.
  - H_o: $Independence\ of\ Row\ and\ Column\ Variables$

We can apply χ^2 to a slightly different sampling procedure, the **test of homogeneity**

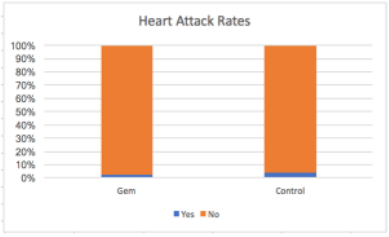
- Several separate random samples are taken from the subpopulations defined by row or columns.
- We want to know if the set of proportions are the same (up to random chance) within each subpopulation
- H_o: $Population\ distributions\ are\ the\ same\ (populations\ are\ homogeneous)$

Comparing two Independent Population Proportions with GOF (revisiting lecture 12 data)

Ex 3: Independent Samples Classified in Several Categories- Test of Homogeneity

Hypothesis tests and confidence intervals for $\pi_g - \pi_c$, the difference in population proportions of heart attacks in gemfibrozil and control patients are desired.... H_o: $\pi_g - \pi_c = 0$; H_a: $\pi_g - \pi_c \neq 0$ $\widehat{p_g} = \frac{56}{2051}$ $\widehat{p_C} = \frac{84}{2030}$

π_g: proportion of gemfibrozil patients that have heart attack; π_c: proportion of control patients that have HA



Heart Attack	Yes	No	Total
gemfibrozil	56	1995	2051
control	84	1946	2030

We can express our hypotheses as:

H_o: The distribution of heart attacks in gemfibrozil and control are the same or

$$\pi_{gY} = \pi_{CY} = \pi_Y; \qquad \pi_{gN} = \pi_{CN} = \pi_N$$

(yes) (no)

H_a: The distribution of gemfibrozil and control heart attacks are somehow different

Comparing two Independent Population Proportions with GOF

Observed Counts:

Heart Attack	Yes	No	Total
gemfibrozil	56	1995	2051
control	84	1946	2030
Total	140	3941	4081

Expected Counts:

Heart Attack	Yes	No	Total
gemfibrozil	$2051*\pi_Y$	$2051*\pi_N$	2051
control	$2030*\pi_Y$	$2030*\pi_N$	2030
Total	$140=\pi_Y * 4081$	$3941=\pi_N*4081$	4081

So we estimate $\widehat{\pi_Y} = \frac{140}{4081}$ and $\widehat{\pi_N} = \frac{3941}{4081}$

So Expected Count of Gemfibrozil and Yes $= \frac{140}{4081} * 2051 = \frac{(\text{Row Tot})(\text{Col Tot})}{\text{Table Tot}}$

SAME CALCULATION as independence test!

Comparing two Independent Population Proportions with GOF

Ex3: Independent Samples Classified in Several Categories- Test of Homogeneity

Observed (Expected) Counts:

Heart Attack	Yes	No	Total
gemfibrozil	56 (70.36)	1995 (1980.64)	2051
control	84 (69.64)	1946 (1960.36)	2030
Total	140	3941	4081

Chi Squared Components:

Heart Attack	Yes	No	Total
gemfibrozil	2.193	0.1041126	3.034892
control	2.96108	0.105	3.06627
Total	5.891859	0.209	6.101162

$\chi^2_{obs} = \sum_{cells} \frac{(O-E)^2}{E} =$ 6.10

d.f=(Number of rows-1)(Number of Col-1) *Not counting Total Row/Cols

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

$P(\chi^2_{df=1} > 6.10) =$ 0.01358

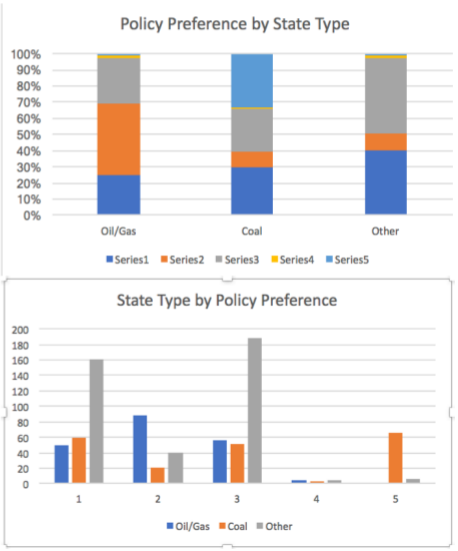
From Lecture 12 when we did a 2 independent sample z test for difference in proportions:
*z=−2.47, p=0.006946851 (from 1 sided test), How related?

$Z^2 = \chi^2_{obs}$ (chi squared)

1 sided p value from z test is half of chi squared (χ^2) pvalue

Practice: Random samples of 200 individuals from major oil-producing and natural gas-producing states, 200 from coal states, and 400 from other states participate in a poll of attitudes toward five possible energy policies. Each respondent indicates the most preferred alternative shown in the table below. Conduct a test of whether the distribution of policy choice preference is the same across the three populations

Policy Choice	Oil/Gas State	Coal State	Other State	Total
1	50	59	161	270
2	88	20	40	148
3	56	52	188	296
4	4	3	5	12
5	2	66	6	74
Totals	200	200	400	800



Practice

Policy Choice	Oil/Gas State	Coal State	Other State	Total
1	50	59	161 [135]	270
2	88 [37]	20	40	148
3	56	52 [74]	188 [148]	296
4	4 [3]	3 [3]	5 [6]	12
5	2 [18.5]	66	6 [37]	74
Totals	200	200	400	800

Test of homogeneity:

T.S.: $\chi^2 = \frac{(50-67.5)^2}{67.5} + \cdots + \frac{(6-37)^2}{37} = 289.22$. pvalue:

Conclusion:

The χ^2 test has limited (but important purpose). This test can only assess whether the data indicate a statistically detectable relation among categories. A weak relation in a large data set may be detectable; a strong relation in a small data set may be non significant.

Checking our work in R:

```
> ex3data<-matrix(c(56, 1995, 84, 1946), nrow=2, byrow=TRUE)
> chisq.test(ex3data, correct=FALSE)

        Pearson's Chi-squared test

data:  ex3data
X-squared = 6.1013, df = 1, p-value = 0.01351
```

```
ex4data<-matrix(c(50,59,161,88,20,40,56,52,188,4,3,5,2,66,6),
nrow=5, byrow=TRUE)
```

```
> chisq.test(ex4data, correct=FALSE)

        Pearson's Chi-squared test

data:  ex4data
X-squared = 289.22, df = 8, p-value < 2.2e-16

Warning message:
In chisq.test(ex4data, correct = FALSE) :
  Chi-squared approximation may be incorrect

  > ex4.test$expected
       [,1] [,2] [,3]
  [1,] 67.5 67.5  135
  [2,] 37.0 37.0   74
  [3,] 74.0 74.0  148
  [4,]  3.0  3.0    6
  [5,] 18.5 18.5   37
  > .2*10
  [1] 2
```