Lecture 11

Tuesday, February 18, 2020 1:59 PM

Agenda:

- Chapter 13
  - contingency tables (G-O-F tests)
  - 1-sample binomial (13.3)
  - r-sample
  - Test of common p
    - specified
    - unspecified

Goodness of Fit Tests:

- given that we have a certain model for the data, how well does the data support the model assumption?

Recall Example from prior lecture...

	F	NF	Total
D	16	28	44
C	24	32	56
T	40	60	100

- let's look at one column at a time and treat it like a Binomial.
  - starting with column F.
    - let p_1 be the true proportion of female cat owners (FCO for short)
    - then $(1-p_1)$ is the proportion of female dog owners (FDO for short)

      $H_0: p_1 = p_{10}$, $H_A: p_1 \neq p_{10}$

    - let X be the count of FCO from n.
      - Then, $E(X) = n \cdot p_{10}$ under H_0 and

        $E(40-X) = n \cdot (1-p_{10})$ under H_0

        ... because under the null, $X \sim Bin(n=40, p=p_{10})$

  Options:

  - we could do a conservative test based on Binomial (Chapter 12)
  - we could also approximate a test based on the Normal.

Recall:

$$Z = \frac{x - np_{10}}{\quad} \sim N(0,1)$$

$$Z = \frac{x - np_{10}}{\sqrt{np_{10}(1-p_{10})}} \sim N(0,1)$$

(under some conditions)

$$Z^2 = \chi^2(1)$$

- consider Z^2 . . .

$$Z^2 = \frac{(x - np_{10})^2}{np_{10}(1-p_{10})}$$

$$= (x - np_{10})^2 \left[\frac{1}{np_{10}} + \frac{1}{n(1-p_{10})} \right]$$

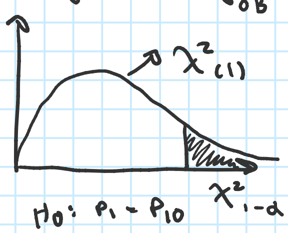
$$= \frac{(x - np_{10})^2}{np_{10}} + \frac{(x - np_{10})^2}{n(1-p_{10})}$$

$$= \text{''} + \frac{(x - n + n - np_{10})^2}{n(1-p_{10})}$$

$$= \text{''} + \frac{(n - x - n(1-p_{10}))^2}{n(1-p_{10})}$$

$$\chi^2_{OB} = \sum_{j=1}^{2} \frac{(O_j - e_j)^2}{e_j} \quad \text{where} \begin{cases} O_j = \text{observed count } j \\ e_j = \text{expected count } j \end{cases}$$

- we reject if $\chi^2_{OB} > \chi^2_{1-\alpha}(1)$



Since we are trying to approximate a categorical variable, a discontinuity error may be introduced and is corrected for by:

$$\chi^2_{OB} = \sum_{j=1}^{2} \frac{(|O_j - e_j| - 0.5)^2}{e_j}$$ "Yate's correction for continuity" (1934)

Let's do a quick example pertaining to the female-dog-cat data from earlier...

Ex FCO vs FDO

$H_0: p_1 = 0.5$

$H_A: p_1 \neq 0.5$

$H_A: p_i \neq 0.5$

$p_{10} = 40(0.5) = n(1-p_{10})$

- $z^2 = \dfrac{(16-20)^2}{20} + \dfrac{(24-20)^2}{20}$

$= 16$

- then, $\chi^2_{1-0.05}(1) = 3.84$

$\therefore$ we fail to reject H_0 ✓

r-Sample Binomial:

- Suppose $X_i \sim Bin(n_i, p_i)$; $i=1,\ldots,r$
- $H_0: p_i = p_{10}$, $i=1,\ldots,r$ => implies there are r restrictions

  $H_A: p_i \neq p_{10}$ for at least some i
- Recall that if some RV $M \sim \chi^2(m)$ and some RV $N \sim \chi^2(n)$ and $M \perp N$, then $M+N \sim \chi^2(m+n)$

Thus:

$$\chi^2_{OB} = \sum_{j=1}^{2}\sum_{i=1}^{r} \frac{(o_{ij}-e_{ij})^2}{e_{ij}} \sim \chi^2(r)$$

And we reject if:

$$\chi^2_{OB} > \chi^2_{1-\alpha}(r)$$

Now let us reconsider the earlier example...

Pet \ Owner	F	NF	Total
D	16	28	44
C	24	32	56
B	33	17	50
Total	73	77	150

Question: Is the proportion of female pet owners (FPO) the same across pet types and is this proportion equal to 0.5 for all pet types??

In other words, is pet type independent of owner type??

Solution: Let x_i be the count of female pet owners in the ith pet category.

Then, $x_i \sim Bin(n_i, p_i = 0.5)$ under the null hypotheses.

$H_0: p_i = p_{i0} = 0.5$ for $i = 1, 2, 3$ (r)

- $E(x_1) = 44(0.5) = 22$

- $E(n_1 - x_1) = 22$
- $E(x_2) = 56(0.5) = 28 = E(n_2 - x_2)$
- $E(x_3) = 50(0.5) = 25 = E(n_3 - x_3)$

Then, $$\chi^2_{OB} = \left[\frac{(16-22)^2}{22} + \frac{(28-22)^2}{22}\right] + \left[\frac{(24-28)^2}{28} + \frac{(32-28)^2}{28}\right] + \left[\frac{(33-25)^2}{25} + \frac{(17-25)^2}{25}\right] \approx 9$$

and $\chi^2_{1-0.05}(3) = 12.59$

$\therefore$ Then because $9 > 12.59$, we fail to reject the null. ✓

Test of Common P: (P₀ unknown under the null)

$H_0: P_1 = P_2 = P_3 = P_0$

$$\hat{P}_0 = \frac{\sum_{i=1}^{3} X_i}{n_1 + n_2 + n_3} = \frac{\sum_{j=1}^{3} O_{1j}}{\sum_{j=1}^{n} n_j} = \frac{16 + 24 + 33}{150} \approx 0.49$$

Then
$$\begin{cases} \hat{e}_{11} = 44(0.49) = 21.56 \\ \hat{e}_{21} = 56(0.49) = 27.44 \\ \hat{e}_{31} = 50(0.49) = 24.5 \end{cases}$$

Then $$\chi^2_{OB} = \left[\frac{(16-21.56)^2}{21.56} + \frac{(28-22.44)^2}{22.44}\right] + \left[\frac{(24-27.44)^2}{27.44} + \frac{(32-18.56)^2}{18.56}\right] + \left[\frac{(33-24.5)^2}{24.5} + \frac{(17-25.5)^2}{25.5}\right]$$

$$= [1.43 + 1.38] + [0.43 + 9.73] + [2.95 + 2.83] = 18.87$$

$\chi^2_{0.95}(2) = 5.99$

$\therefore$ $18.87 > 5.99$, so we reject the null hypothesis.