STA 250/MTH 342 – Intro to Mathematical Statistics

Lecture 20

Goodness-of-fit test: Weldon's Data

No. of Dice showing 5 or 6	Observed (X_i)	Theory (m_i)	$X_i - m_i$
0	185	203	-18
1	1149	1217	-68
2	3265	3345	-80
3	5475	5576	-101
4	6114	6273	-159
5	5194	5018	176
6	3067	2927	140
7	1331	1254	77
8	403	392	11
9	105	87	18
10-12	*18*	*14*	*4*
Total	26,306	26,306	0

After the grouping, we have $k = 11$ categories, and *we carry out the tests with this new k*.

- The χ^2 statistic is

$$Q = \frac{(-18)^2}{203} + \frac{(-68)^2}{1217} + \frac{(-80)^2}{3345} + \cdots + \frac{18^2}{87} + \frac{4^2}{14} \approx 35.9.$$

- What is the sampling distribution of Q under H_0?
- The rejection region corresponding to level α is

$$\mathscr{R}(\alpha): \quad Q > F^{-1}_{\chi^2_{10}}(1-\alpha).$$

- The p-value is

$$1 - F_{\chi^2_{10}}(35.9) \approx .0000876 < .01.$$

- So we will reject H_0 at level $\alpha = .01$ or $.05$ etc.

Back to Weldon's data

No. of Dice showing 5 or 6	Observed (X_i)	Theory (m_i)	$X_i - m_i$
0	185	203	-18
1	1149	1217	-68
2	3265	3345	-80
3	5475	5576	-101
4	6114	6273	-159
5	5194	5018	176
6	3067	2927	140
7	1331	1254	77
8	403	392	11
9	105	87	18
10	14	13	1
11	4	*1*	3
12	0	*0*	0
Total	26,306	26,306	0

- The data provides strong evidence against Weldon's original null hypothesis, that the dice show "5" or "6" with 1/3 chance.
- The pattern in the table seems to suggest that the chance for getting "5" or "6" may be a a little larger than 1/3.
- Weldon accepted Pearson's response, and asked whether the dice are at least independently tossed with the same probability θ for "5" or "6".
- This is the null hypothesis that the number of "5" or "6" that one gets from each roll (of 12 dice) follow a Binomial(12,p) distribtuion, as opposed to Binomial(12,1/3) as he originally suggested.

Formulate the testing problem

- Weldon's new problem falls into the following general form.
- For multinomial$(n;\theta_1, \theta_2, \ldots, \theta_k)$ data $(X_1, X_2, \ldots, X_k)$, we are interested in testing the null hypothesis

$$H_0 : \theta_1 = a_1(p), \theta_2 = a_2(p), \ldots, \theta_k = a_k(p)$$

  where $a_1(p), a_2(p), \ldots, a_k(p)$ are functions of an unknown (possibly vector-valued) parameter p, versus the alternative

$$H_1 : \text{otherwise.}$$

  That is, at least one θ_i is not equal to $a_i(p)$.
- For Weldon's data, θ is the probability for a die to show "5" or "6".

- The slight complication here is that now the null hypothesis is also composite.
- But we have learned how to test a composite null versus a composite alternative!

- Recall our earlier result: under some "smoothness" conditions, when the sample size is large

$$-2\log\Lambda \sim_{approx} \chi^2_h$$

  where the degrees of freedom $h = \mathrm{Dim}(\Theta_0 \cup \Theta_1) - \mathrm{Dim}(\Theta_0)$.
- $\mathrm{Dim}(\Theta_0 \cup \Theta_1)$ is the number of *free* parameters estimated in computing the global MLE, and $\mathrm{Dim}(\Theta_0)$ is the number of *free* parameters estimated in computing the restricted MLE.
- This is just a special case of our earlier general result about $-2\log\Lambda$.
- Similarly,

$$Q \sim_{approx} \chi^2_h.$$

- For Weldon's problem, the degree of freedom is $h = k-1-1 = k-2$. Recall that when the null hypothesis is simple, the degree of freedom is $k-1-0 = k-1$.
- There is a reduction in the degree of freedom.

Specific form of the tests for Weldon's data

- In this case, the new null hypothesis is

$$\Theta_0 = \{(\theta_1, \theta_2, \ldots, \theta_k) : \theta_i = a_i(p) \text{ for } i = 1, 2, \ldots, k\}$$

and

$$\Theta_1 = \Theta \backslash \Theta_0$$

where $a_i(p) = \binom{12}{i} p^i (1-p)^{12-i}$ and

$$\Theta = \left\{ (\theta_1, \theta_2, \ldots, \theta_k) : \theta_i \geq 0 \text{ for } i = 1, 2, \ldots, k, \sum_{i=1}^{k} \theta_i = 1 \right\}$$

- Now we reject the null when

$$\Lambda = \frac{\max_{(\theta_1, \theta_2, \ldots, \theta_k) \in \Theta_0} L(\theta_1, \theta_2, \ldots, \theta_k)}{\max_{(\theta_1, \theta_2, \ldots, \theta_k) \in \Theta} L(\theta_1, \theta_2, \ldots, \theta_k)} < K.$$

- Now the restricted MLE (by the invariance property) is

$$\hat{\theta}_0 = (\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_k) = (a_1(\hat{p}), a_2(\hat{p}), \ldots, a_k(\hat{p})).$$

- The unrestricted MLE is still

$$\hat{\theta} = \left(\frac{X_1}{n}, \frac{X_2}{n}, \ldots, \frac{X_k}{n} \right)$$

- The LR becomes

$$\begin{aligned}
\Lambda &= \frac{L(\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_k)}{L(\frac{X_1}{n}, \frac{X_2}{n}, \ldots, \frac{X_k}{n})} \\
&= \frac{\frac{n!}{X_1!X_2!\cdots X_k!} \hat{a}_1^{X_1} \hat{a}_2^{X_2} \cdots \hat{a}_k^{X_k}}{\frac{n!}{X_1!X_2!\cdots X_k!} (\frac{X_1}{n})^{X_1} (\frac{X_2}{n})^{X_2} \cdots (\frac{X_k}{n})^{X_k}} \\
&= \left(\frac{n\hat{a}_1}{X_1} \right)^{X_1} \left(\frac{n\hat{a}_2}{X_2} \right)^{X_2} \cdots \left(\frac{n\hat{a}_k}{X_k} \right)^{X_k} \\
&= \left(\frac{\hat{m}_1}{X_1} \right)^{X_1} \left(\frac{\hat{m}_2}{X_2} \right)^{X_2} \cdots \left(\frac{\hat{m}_k}{X_k} \right)^{X_k}
\end{aligned}$$

where $\hat{m}_i = n\hat{a}_i$ is the MLE of the expected value of X_i under the null hypothesis H_0.

The LR test

- So the LR test statistic becomes rejecting when

$$-2\log\Lambda = -2\sum_{i=1}^{k} X_i \log\left(\frac{\hat{m}_i}{X_i}\right) = \sum_{i=1}^{k} 2X_i \log\left(\frac{X_i}{\hat{m}_i}\right) > C.$$

- Similarly, the χ^2 test rejects when

$$Q = \sum_{i=1}^{k} \frac{(X_i - \hat{m}_i)^2}{\hat{m}_i} > C.$$

- What is the approximate sampling distribution of $-2\log\Lambda$ and Q under H_0?

- Weldon computed the MLE for p, the probability of getting a "5" or "6" from the data he has, under H_0.
- How to do this?
- By brute-force, write down the joint likelihood, take log, and derivative …
- A shortcut: There are a total of

$$12 \times 26,306 = 315,672$$

  dice tosses.
- Out of these,

$$0 \times 185 + 1 \times 1149 + \cdots + 11 \times 4 + 12 \times 0 = 106,602$$

  tosses show "5" or "6".
- Treating each toss as a Bernoulli(p) trial, we have the MLE for p

$$\hat{p} = \frac{106,602}{315,672} = .337699.$$

- Under this new “theory”, we can compute the MLE of the expected counts (by what principle?)

$$\begin{aligned}\hat{m}_i = n \cdot a_i(\hat{p}) &= 26,306 \times \binom{12}{i} \hat{p}^i (1-\hat{p})^{12-i} \\ &= 26,306 \times \binom{12}{i} (.337699)^i (1-.337699)^{12-i}\end{aligned}$$

for $i = 0, 1, 2, \ldots, 12$.

Under the new "theory"

No. of Dice showing 5 or 6	Observed (X_i)	Theory ($\hat{m}_i$)	$X_i - \hat{m}_i$
0	185	187.4	-2.4
1	1149	1146.5	2.5
2	3265	3215.2	49.8
3	5475	5464.7	10.3
4	6114	6269.3	-155.3
5	5194	5114.7	79.3
6	3067	3042.5	24.5
7	1331	1329.7	1.3
8	403	423.8	-20.8
9	105	96.0	9.0
10	14	14.7	-0.7
11	4	1.4	2.6
12	0	0.1	-0.1
Total	26,306	26,306	0

Again let's group the last three categories—a total of 11 categories.

- Let us compute the χ^2 statistic as before (after combining the last three categories)

$$Q = \frac{(-2.4)^2}{187.4} + \frac{2.5^2}{1146.5} + \cdots + \frac{1.8^2}{16.2} = 8.2.$$

- Similarly, we can compute the LR statistic $-2\log\Lambda$.
- The corresponding p-value is .51, so we still do not reject the null hypothesis for all levels no more than .51.

Example: Galton's height data

		Wife: Tall	Medium	Short	Total
	Tall	18	28	14	60
Husband:	Medium	20	51	28	99
	Short	12	25	9	46
	Total	50	104	51	205

- This is an example of a *contingency table*, a cross-classification of a number of "individual cases" according to two (or more) different categorical measurements.
- These data are a list of counts of observations, organized in a *rectangular list* whose margins correspond to the classifying factors.
- What is the sampling unit (i.e. an observation) here?
- What are the individual cases and the "classifying factors" here?

		Factor B			
		B_1	B_2	B_3	Total
	A_1	X_{11}	X_{12}	X_{13}	X_{1+}
Factor A	A_2	X_{21}	X_{22}	X_{23}	X_{2+}
	A_3	X_{31}	X_{32}	X_{33}	X_{3+}
	Total	X_{+1}	X_{+2}	X_{+3}	$X_{++} = n$

- We use X_{ij} to denote the count in the (i,j)th cell—that is the count in row i and column j.
- We use X_{i+} to denote the marginal total count for the ith row, and X_{+j} the marginal total count for the jth column, and X_{++} the overall total count, which is equal to n, the sample size.

		Factor B				
		B_1	B_2	$\cdots$	B_c	Total
Factor A	A_1	X_{11}	X_{12}	$\cdots$	X_{1c}	X_{1+}
	A_2	X_{21}	X_{22}	$\cdots$	X_{2c}	X_{2+}
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	A_r	X_{r1}	X_{r2}	$\cdots$	X_{rc}	X_{r+}
	Total	X_{+1}	X_{+2}	$\cdots$	X_{+c}	$X_{++} = n$

- More generally, we may have a table with r rows and c columns.
- We have

$$X_{i+} = \sum_{j=1}^{c} X_{ij}, \quad X_{+j} = \sum_{i=1}^{r} X_{ij}, \quad \text{and} \quad X_{++} = \sum_{i=1}^{r} \sum_{j=1}^{c} X_{ij} = n.$$

- In some problems, the rc counts X_{ij} can be modeled by a multinomial distribution with n being the total number of trials and $\theta_{ij} = P(A_i \cap B_j)$ the probability of the (i,j)th catogory.

Multinomial model on a contingency table

		Factor B				
		B_1	B_2	$\cdots$	B_c	Total
Factor A	A_1	θ_{11}	θ_{12}	$\cdots$	θ_{1c}	θ_{1+}
	A_2	θ_{21}	θ_{22}	$\cdots$	θ_{2c}	θ_{2+}
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	A_r	θ_{r1}	θ_{r2}	$\cdots$	θ_{rc}	θ_{r+}
	Total	θ_{+1}	θ_{+2}	$\cdots$	θ_{+c}	$\theta_{++} = 1$

- Correspondingly, we use θ_{i+}, θ_{+j} to denote the total row and column marginal probabilities:

$$\theta_{i+} = \sum_{j=1}^{c} \theta_{ij} = P(A_i), \quad \theta_{+j} = \sum_{i=1}^{r} \theta_{ij} = P(B_j),$$

and

$$\theta_{++} = \sum_{i=1}^{r} \sum_{j=1}^{c} \theta_{ij} = 1.$$

Testing the independence null hypothesis

- Now consider testing the null hypothesis that the two categories are independent of each other: $P(A_i \cap B_j) = P(A_i)P(B_j)$ for all i and j.
- Under this null hypothesis, the cell probabilities are completely specified by the marginal probabilities θ_{i+} and θ_{+j} for all i and j.
- Formally, we want to test

$$H_0 : \theta_{ij} = a_{ij}(\theta_{1+}, \theta_{2+}, \ldots, \theta_{r+}, \theta_{+1}, \theta_{+2}, \ldots, \theta_{+c}) = \theta_{i+}\theta_{+j}$$

for all i and j, versus the alternative

$$H_1 : \text{otherwise},$$

That is, at least one of the θ_{ij} is not equal to $\theta_{i+}\theta_{+j}$. (This is called the *test of independence*.)

- Now we are back in the situation of testing goodness-of-fit with the null hypothesis determined by some unknown parameters.
- Recall that we can apply either the LR test or the χ^2 test.
- To apply these tests, we can first find the MLE of the unknown parameters $(\theta_{1+}, \theta_{2+}, \ldots, \theta_{r+}, \theta_{+1}, \theta_{+2}, \ldots, \theta_{+c})$ under the null hypothesis.
- The MLE's are the corresponding observed proportions (show!):

$$\hat{\theta}_{i+} = \frac{X_{i+}}{n} \quad \text{and} \quad \hat{\theta}_{+j} = \frac{X_{+j}}{n}$$

  for all i and j.
- Thus the MLE for θ_{ij} under the null is

$$\hat{\theta}_{ij} = \hat{a}_{ij} = \hat{\theta}_{i+}\hat{\theta}_{+j} = \frac{X_{i+}X_{+j}}{n^2}$$

  by the *invariance property* of MLEs.

- Accordingly, the MLE of the expected counts under the null are

$$\hat{m}_{ij} = m_{ij}(\hat{\theta}) = n\hat{\theta}_{ij} = n\hat{a}_{ij} = \frac{X_{i+}X_{+j}}{n}.$$

- The LR is

$$\Lambda = \frac{\frac{n!}{\prod_{i=1}^r \prod_{j=1}^c X_{ij}} \prod_{i=1}^r \prod_{j=1}^c \hat{a}_{ij}^{X_{ij}}}{\frac{n!}{\prod_{i=1}^r \prod_{j=1}^c X_{ij}} \prod_{i=1}^r \prod_{j=1}^c \left(\frac{X_{ij}}{n}\right)^{X_{ij}}}$$

$$= \prod_{i=1}^{r} \prod_{j=1}^{c} \left(\frac{n\hat{a}_{ij}}{X_{ij}}\right)^{X_{ij}} = \prod_{i=1}^{r} \prod_{j=1}^{c} \left(\frac{\hat{m}_{ij}}{X_{ij}}\right)^{X_{ij}}$$

- So the LR test statistic is

$$-2\log\Lambda = \sum_{i=1}^{r} \sum_{j=1}^{c} 2X_{ij} \log\left(\frac{X_{ij}}{\hat{m}_{ij}}\right).$$

- The χ^2 test statistic is

$$Q = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(X_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(X_{ij} - \frac{X_{i+}X_{+j}}{n}\right)^2}{\frac{X_{i+}X_{+j}}{n}}.$$

- What is the degrees of freedom for their approximate sampling distribution under the null hypothesis?
- What is the number of *free* parameters under H_0, that is in Θ_0?
- We have $(r-1)$ free row marginal probabilities and $(c-1)$ free column marginal probabilities. So the total is

$$(r-1)+(c-1).$$

- What is the number of *free* parameters overall?
- We have a total of $(rc-1)$ free cell probabilities.
- Therefore the degrees of freedom for the approximate sampling distribution is

$$(rc-1)-((r-1)+(c-1)) = rc-r-c+1 = (r-1)(c-1).$$

Back to Galton's data

		Wife: Tall	Medium	Short	Total
Husband:	Tall	18 (14.6)	28 (30.4)	14 (14.9)	60
	Medium	20 (24.1)	51 (50.2)	28 (24.6)	99
	Short	12 (11.2)	25 (23.3)	9 (11.4)	46
	Total	50	104	51	205

- The MLE of the expected count $\hat{m}_{ij}$ are given in the parentheses.

$$Q = \frac{(18-14.6)^2}{14.6} + \frac{(28-30.4)^2}{30.4} + \cdots + \frac{(9-11.4)^2}{11.4} = 2.91.$$

- The degrees of freedom is $(r-1)(c-1) = 2 \times 2 = 4$.
- The p-value is $1 - F_{\chi^2_4}(2.91) \approx .573$.
- So we do not reject the null hypothesis of independence under any level less than .573. There is no evidence for rejecting the null that the spouses were selected without regard to height.