# STA 250/MTH 342 – Intro to Mathematical Statistics

Lecture 21

# Sampling models on contingency tables

- The test of independence on contingency tables as we have studied so far rely on the multinomial sampling model: the $n$ individual cases are distributed over the $r \times c$ cells of the table independently according to $r \times c$ cell probabililities.
- So the sampling distribution of the $r \times c$ cell counts are multinomial.
- This is called *full multinomial sampling* (FMS).

- Sometimes the counts in a table do not arise in this fashion.
- For example, one or both sets of the marginal totals may be fixed by design.
- If one set of marginal totals (either the row totals or the column totals) are fixed, we have a *product multinomial sampling* (PMS) model.
- If both row totals and column totals are fixed, we have *hypergeometric sampling* model.
- Let us learn these through examples.

# Example A: Social influence on right-handedness?

Counts of 1180 art works showing activity that can be categorized as left- or right-handed, by geographic location.

| | Right | Left | Total | % Right |
|---|---|---|---|---|
| Central Europe | 312 | 23 | 335 | 93% |
| Medit. Europe | 300 | 17 | 317 | 95% |
| Middle East | 85 | 4 | 89 | 96% |
| Africa | 105 | 12 | 117 | 90% |
| Central Asia | 93 | 8 | 101 | 92% |
| Far East | 126 | 13 | 139 | 91% |
| Americas | 72 | 10 | 82 | 88% |
| Total | 1093 | 87 | 1180 | 92.6% |

From Coren and Porac (1977). Science.

- Is the degree of right-handedness socially determined?

- The row totals are essentially *fixed*, not actually random.
- What is the null hypothesis we want to test?
- $H_0$: The proportion of right-handed is the same across all the geographic locations.
- Under $H_0$, each row is a multinomial vector (in this particular example a Binomial($n_i; \theta_i$) vector).
- The rows are independent. So the joint probability of the cell counts are the product of the probability of each row.
- Let us look at another example.

# Example B: Like father, like son?

- A study was carried out to investigate the starting words in writing samples from British economists James Mill and John Stuart Mill.

| First Word: | But | Where | This/ItThus/And | A/By | All others | Totals |
|---|---|---|---|---|---|---|
| James Mill | 39 | 26 | 339 | 33 | 638 | 1075 |
| John Stuart Mill | 38 | 16 | 112 | 11 | 274 | 451 |
| Totals | 77 | 42 | 451 | 44 | 912 | 1526 |

From O'Brien and Darnell (1982).

- Do the two have similar style in choosing the start of a sentence?

- Similar to the previous example, the row totals are essentially *fixed*.
- The null hypothesis: the proportion of each word is the same across the rows.
- Again, the joint probability of the cell counts are the product of the two multinomials.
- This is called *product multinomial sampling*.

# Product multinomial sampling

When the row totals are fixed,

- For each row $i$,

$$(X_{i1}, X_{i2}, \ldots, X_{ic}) \sim \text{multinomial}(X_{i+} = n_i; \theta_{i1}, \theta_{i2}, \ldots, \theta_{ic}).$$

- The rows are independent multinomial vectors.

We want to test

$$H_0 : (\theta_{i1}, \theta_{i2}, \ldots, \theta_{ic}) = (\theta_1, \theta_2, \ldots, \theta_c) \text{ for all } i$$

vs

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

That is, the vectors $(\theta_{i1}, \theta_{i2}, \ldots, \theta_{ic})$ are not the same across all $i$.

- This is called the *test of homogeneity*.

- Let us try to find out what tests can we use for this purpose.
- Note that this still lies inside the general problem of testing two composite hypotheses. So again we can try to construct the (generalized) LR test and perhaps even find a corresponding $\chi^2$ test as an approximation to the LR test.
- The likelihood under $H_0$ is

$$L(\theta_1, \theta_2, \ldots, \theta_c) = \prod_{i=1}^{r} \left( \frac{X_{i+}!}{\prod_{j=1}^{c} X_{ij}!} \theta_1^{X_{i1}} \theta_2^{X_{i2}} \cdots \theta_c^{X_{ic}} \right).$$

- So we can solve for the restricted MLE (exercise!):

$$\hat{\theta}_j = \frac{X_{+j}}{X_{++}} = \frac{X_{+j}}{n}$$

- and by the invariance property

$$\hat{m}_{ij} = \frac{X_{i+} X_{+j}}{n}. \quad \text{(Does this look familiar?)}$$

- With out any restrictions, the likelihood is

$$L(\theta_{11}, \theta_{12}, \ldots, \theta_{rc}) = \prod_{i=1}^{r} \left( \frac{X_{i+}!}{\prod_{j=1}^{c} X_{ij}!} \theta_{i1}^{X_{i1}} \theta_{i2}^{X_{i2}} \cdots \theta_{ic}^{X_{ic}} \right).$$

- The global MLE for the $\theta$'s are now

$$\hat{\theta}_{ij} = \frac{X_{ij}}{X_{i+}}.$$

- Accordingly, the generalized LR is

$$\Lambda = \frac{L(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_c)}{L(\hat{\theta}_{11}, \hat{\theta}_{12}, \ldots, \hat{\theta}_{rc})} = \prod_{i=1}^{r}\prod_{j=1}^{c}\left(\frac{\hat{m}_{ij}}{X_{ij}}\right)^{X_{ij}}.$$

- Thus

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

- Look back at the test of *independence* under *full multinomial sampling*.
- This is exactly the same LR test statistic!

- Consequently, after applying Taylor's expansion, we get exactly the same $\chi^2$ test as well!

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

- Now the test statistics are the same as those for the test of independence. What are about their sampling distribution under the null hypothesis?
- We know the null sampling distribution is $\chi^2$. What's the degrees of freedom?

- What is the number of *free* parameters under $H_0$, that is in $\Theta_0$?
- We have $(c-1)$ free column marginal probabilities. So the total is
$$(c-1).$$
- What is the number of *free* parameters overall?
- We have $(c-1)$ free column probabilities *each row*. So the total is
$$r(c-1).$$
- Therefore the degrees of freedom for the approximate sampling distribution is
$$r(c-1)-(c-1)=(r-1)(c-1).$$
- We have exactly the same sampling distribution as in the case of testing independence!

# Example A: The 1970 draft lottery

- In 1970 the US conducted a draft lottery to determine the order of induction of mails aged 19-26. The 366 possible birthdates were randomly drawn one by one without replacement.
- The order in which they were drawn was their "drawing number".

| Drawing numbers | Month Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec | Totals |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1-122 | 9 | 7 | 5 | 8 | 9 | 11 | 12 | 13 | 10 | 9 | 12 | 17 | 122 |
| 123-244 | 12 | 12 | 10 | 8 | 7 | 7 | 7 | 7 | 15 | 15 | 12 | 10 | 122 |
| 245-366 | 10 | 10 | 16 | 14 | 15 | 12 | 12 | 11 | 5 | 7 | 6 | 4 | 122 |
| Totals | 31 | 29 | 31 | 30 | 31 | 30 | 31 | 31 | 30 | 31 | 30 | 31 | 366 |

From Feinberg (1971). Science.

- This presents an extreme case of deviation from the multinomial sampling model.
- Both the column totals and row totals are *fixed* by design.
- This is called *hypergeometric sampling* model, because the joint distribution of the cell counts under the null hypothesis is called the *hypergeometric distribution*.
- What is the null hypothesis we want to test?
- The null hypothesis: all possible assignment of the numbers 1 to 366 to the birthdates are equally likely.
- Under this null hypothesis, what is the probability of observing the previous table?

$$
\begin{aligned}
P(\text{Table}) &= P(Jan)P(Feb|Jan)P(Mar|Jan,Feb)\cdots P(Dec|Jan,Feb,\ldots,Nov) \\
&= \frac{\binom{122}{9}\binom{122}{12}\binom{122}{10}}{\binom{366}{31}}\frac{\binom{113}{7}\binom{110}{12}\binom{112}{10}}{\binom{335}{29}}\cdots\frac{\binom{17}{17}\binom{10}{10}\binom{4}{4}}{\binom{31}{31}} \\
&= \frac{\frac{122!122!122!}{9!12!10!7!12!10!\cdots 17!10!4!}}{\frac{366!}{31!29!\cdots 31!}}.
\end{aligned}
$$