Math 170S
Lecture Notes Section 9.2 \*†

Contingency tables

Instructor: Swee Hong Chan

---

**NOTE:** Materials that appear in the textbook but do not appear in the lecture notes might still be tested. Please send me an email if you find typos.

\*Version date: Tuesday 8\textsuperscript{th} December, 2020, 22:11.

†This notes is based on Hanbaek Lyu’s and Liza Rebrova’s notes from the previous quarter, and I would like to thank them for their generosity. “*Nanos gigantum humeris insidentes* (I am but a dwarf standing on the shoulders of giants)”.

---
1 Motivating example

Two instructors are teaching Math 170S for $n = 50$ students, with final grade:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friendly instructor (FI)</td>
<td>8</td>
<td>13</td>
<td>16</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>Evil instructor (EI)</td>
<td>4</td>
<td>9</td>
<td>14</td>
<td>16</td>
<td>7</td>
</tr>
</tbody>
</table>

- Null hypothesis $H_0$: The two instructors follow the same grading scheme.

- Alternative hypothesis $H_1$: The two instructors do not follow the grading scheme.

Can we reject $H_0$ with significance level $\alpha = 0.05$?
2 Notation: Motivating examples

\[ p_1 := \text{probability to get A in FI's class}; \]
\[ p_2 := \text{probability to get B in FI's class}; \]
\[ \vdots \]
\[ p_5 := \text{probability to get F in FI's class}. \]

\[ p'_1 := \text{probability to get A in EI's class}; \]
\[ p'_2 := \text{probability to get B in EI's class}; \]
\[ \vdots \]
\[ p'_5 := \text{probability to get F in EI's class}. \]

These are \textbf{unknown parameters.}
\[ Y_1 := \text{number of people getting A in FI’s class}; \]
\[ Y_2 := \text{number of people getting B in FI’s class}; \]
\[ \vdots \]
\[ Y_5 := \text{number of people getting F in FI’s class}. \]

\[ Y_1' := \text{number of people getting A in EI’s class}; \]
\[ Y_2' := \text{number of people getting B in EI’s class}; \]
\[ \vdots \]
\[ Y_5' := \text{number of people getting F in EI’s class}. \]

These are known parameters.
The hypothesis can then be rewritten as

- \( H_0: \ p_i = p'_i \) for all \( i \in \{1, \ldots, 5\} \);

- \( H_1: \ p_i \neq p'_i \) for some \( i \in \{1, \ldots, 5\} \).
3 How to test the hypothesis

1. We already knew the sample mean is a good approximation for the unknown probabilities:

\[ Y_1 \approx np_1; \quad Y_1' \approx np_1', \quad \text{and} \quad \frac{Y_1 + Y_1'}{2n} \approx \frac{p_1 + p_1'}{2}. \]

Let us write

\[ \hat{p}_1 := \frac{Y_1 + Y_1'}{2n}. \]

2. Now note that, if \( p_1 \approx p_1' \), then

\[ p_1 \approx \frac{p_1 + p_1'}{2}. \]

3. Combining all these observations, if \( p_1 \approx p_1' \):

\[ Y_1 \approx np_1 \approx n\frac{p_1 + p_1'}{2} \approx n\frac{Y_1 + Y_1'}{2n} = n\hat{p}_1. \]
4. Hence we have, if \( p_1 \approx p'_1 \), then

\[ Y_1 - n \hat{p}_1 \approx 0. \]

By CLT, we in fact have, if \( p_1 \approx p'_1 \):

\[ \frac{(Y_1 - n \hat{p}_1)^2}{n \hat{p}_1} \]

is small.

5. On the other hand, if \( p_1 \) is very far away from \( p'_1 \), (e.g., \( p_1 = 0 \) and \( p'_1 = 1 \)), then

\[ Y_1 = 0; \quad \hat{p}_1 = \frac{1}{2}, \]

so

\[ \frac{(Y_1 - n \hat{p}_1)^2}{n \hat{p}_1} = \frac{(0 - n/2)^2}{n/2} = n/2 = \text{very big}. \]
6. So we conclude that

\[
\frac{(Y_1 - n \hat{p}_1)^2}{n \hat{p}_1} \text{ is small if and only if } p_1 \approx p_1'.
\]

By the same reasoning, we have

\[
\frac{(Y_1' - n \hat{p}_1)^2}{n \hat{p}_1} \text{ is small if and only if } p_1 \approx p_1'.
\]

7. To provide balance, we add these two tests together:

\[
\frac{(Y_1 - n \hat{p}_1)^2}{n \hat{p}_1} + \frac{(Y_1' - n \hat{p}_1)^2}{n \hat{p}_1} \text{ is small if and only if } p_1 \approx p_1'.
\]

8. By the same reasoning, for all \(i = \{1, 2, 3, 4, 5\}\),

\[
\frac{(Y_i - n \hat{p}_i)^2}{n \hat{p}_i} + \frac{(Y_i' - n \hat{p}_i)^2}{n \hat{p}_i} \text{ is small if and only if } p_i \approx p_i'.
\]
9. We want to test all five parameters \textit{simultaneously}. So we add all the tests up together:

\[
Q := \left[ \frac{(Y_1 - n \hat{p}_1)^2}{n \hat{p}_1} + \frac{(Y'_1 - n \hat{p}_1)^2}{n \hat{p}_1} \right] + \left[ \frac{(Y_2 - n \hat{p}_2)^2}{n \hat{p}_2} + \frac{(Y'_2 - n \hat{p}_2)^2}{n \hat{p}_2} \right] + \ldots + \left[ \frac{(Y_5 - n \hat{p}_5)^2}{n \hat{p}_5} + \frac{(Y'_5 - n \hat{p}_5)^2}{n \hat{p}_5} \right].
\]

We have

\[Q \text{ is small if and only if } H_0 \text{ is true.}\]

10. It can be shown that \(Q\) is approximately a \(\chi^2\) random variable with 4 degrees of freedom.

\textbf{Conclusion}: we reject \(H_0\) if and only if \(Q \geq \chi^2_{\alpha}(4)\).
4 Answer: motivating examples

We have from the sample data that

\[ Y_1 = 8; \quad Y_2 = 13; \quad Y_3 = 16; \quad Y_4 = 10; \quad Y_5 = 3; \]
\[ Y_1' = 4; \quad Y_2' = 9; \quad Y_3' = 14; \quad Y_4' = 16; \quad Y_5' = 7, \]

and

\[ \hat{p}_1 = 0.12; \quad \hat{p}_2 = 0.22; \quad \hat{p}_3 = 0.30; \]
\[ \hat{p}_4 = 0.26; \quad \hat{p}_5 = 0.10. \]

Then \( Q \) is equal to

\[
Q = \left[ \frac{(8 - (50)(0.12))^2}{(50)(0.12)} \right] + \left[ \frac{(4 - (50)(0.12))^2}{(50)(0.12)} \right] + \left[ \frac{(13 - (50)(0.22))^2}{(50)(0.22)} \right] + \left[ \frac{(9 - (50)(0.22))^2}{(50)(0.22)} \right] + \ldots + \left[ \frac{(3 - (50)(0.10))^2}{(50)(0.10)} \right] + \left[ \frac{(7 - (50)(0.10))^2}{(50)(0.10)} \right] = 5.18.
\]
On the other hand, $\chi^2_\alpha(4)$ is equal to

$$\chi^2_\alpha(4) = \chi^2_{0.05}(4) = 9.488.$$ 

Since $Q$ is smaller than $\chi^2_\alpha(4)$, we conclude that the test is inconclusive.
5 Settings: equality in distribution

Object:

- $X^{(1)}, X^{(2)}, \ldots, X^{(h)}$ are **independent** random variables with **unknown distribution**.

- $k$ mutually exclusive, exhaustive events $A_1, \ldots, A_k$, and we write

$$p_i^{(j)} := \text{probability of the event } A_i \text{ to occur for } X^{(j)},$$

for $i \in \{1, 2, \ldots, k\}$ and $j \in \{1, \ldots, h\}$. 
Hypotheses:

• **Null Hypothesis** $H_0$: $X^{(1)}, X^{(2)}, \ldots X^{(h)}$ have the same distribution, i.e.,

\[ p_1^{(1)} = p_1^{(2)} = \ldots = p_1^{(h)}; \text{ and} \]
\[ p_2^{(1)} = p_2^{(2)} = \ldots = p_2^{(h)}; \text{ and} \]
\[ \vdots \quad \vdots \quad \vdots \]
\[ p_k^{(1)} = p_k^{(2)} = \ldots = p_k^{(h)} . \]

• **Alternative Hypothesis** $H_1$: The null hypothesis is false.
**Input:** \( n^{(1)} \) many random samples for \( X^{(1)} \), \( n^{(2)} \) many random samples for \( X^{(2)} \), \ldots, \( n^{(h)} \) many random samples for \( X^{(h)} \), and significance level \( \alpha \).

**Methodology:**

- Compute \( Y_i^{(j)} \) for \( i \in \{1, 2, \ldots, k\} \) and \( j \in \{1, \ldots, h\} \),

  \[ Y_i^{(j)} := \text{number of times } A_i \text{ occurs in samples for } X^{(j)}. \]

- Compute \( \hat{p}_1, \ldots, \hat{p}_k \) given by

  \[ \hat{p}_i := \frac{Y_i^{(1)} + Y_i^{(2)} + \ldots + Y_i^{(h)}}{n^{(1)} + n^{(2)} + \ldots + n^{(h)}}. \]

- Compute \( Q \) given by

  \[ Q := \sum_{j=1}^{h} \sum_{i=1}^{k} \frac{(Y_i^{(j)} - n^{(j)} \hat{p}_i)^2}{n^{(j)} \hat{p}_i}. \]

- Reject \( H_0 \) if \( Q \geq \chi^2(2)((h-1)(k-1)) \), and the test is inconclusive otherwise.
6 Example: equality in distribution

A survey was conducted, asking for the education level and the media preference for news sources:

<table>
<thead>
<tr>
<th>Education Level</th>
<th>Newspaper</th>
<th>Television</th>
<th>Radio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade school</td>
<td>45</td>
<td>22</td>
<td>6</td>
</tr>
<tr>
<td>High School</td>
<td>94</td>
<td>115</td>
<td>30</td>
</tr>
<tr>
<td>College</td>
<td>49</td>
<td>52</td>
<td>13</td>
</tr>
</tbody>
</table>

Let $X^{(1)}$ be the (random) media preference for grade schoolers, $X^{(2)}$ the (random) media preference for high schoolers, and $X^{(3)}$ be the (random) media preference for college schoolers.

Can we reject the hypothesis $X^{(1)} = X^{(2)} = X^{(3)}$ with significance level $\alpha = 0.05$?
Answer: equality in distribution

From the sample data, we have

\[ n^{(1)} = 45 + 22 + 6 = 73; \]
\[ n^{(2)} = 94 + 115 + 30 = 239; \]
\[ n^{(3)} = 49 + 52 + 13 = 114, \]

and

\[ Y_1^{(1)} = 45; \quad Y_2^{(1)} = 22; \quad Y_3^{(1)} = 6; \]
\[ Y_1^{(2)} = 94; \quad Y_2^{(2)} = 115; \quad Y_3^{(2)} = 30; \]
\[ Y_1^{(3)} = 49; \quad Y_2^{(3)} = 52; \quad Y_3^{(3)} = 13. \]

Note that here \( h = k = 3 \).
So \( \hat{p}_i \)’s are given by

\[
\begin{align*}
\hat{p}_1 & := \frac{\text{first column}}{\text{total samples}} = \frac{45 + 94 + 49}{73 + 239 + 114} = \frac{188}{426}; \\
\hat{p}_2 & := \frac{\text{second column}}{\text{total samples}} = \frac{22 + 115 + 52}{73 + 239 + 114} = \frac{189}{426}; \\
\hat{p}_3 & := \frac{\text{third column}}{\text{total samples}} = \frac{6 + 30 + 13}{73 + 239 + 114} = \frac{49}{426}.
\end{align*}
\]

Then \( Q \) is equal to

\[
Q = \sum_{j=1}^{h} \sum_{i=1}^{k} \frac{(Y_{(j)} - n_{(j)} \hat{p}_i)^2}{n_{(j)} \hat{p}_i}
\]

\[
= \frac{((45) - (73)\left(\frac{188}{426}\right))^2}{(73)\left(\frac{188}{426}\right)} + \frac{((22) - (73)\left(\frac{189}{426}\right))^2}{(73)\left(\frac{189}{426}\right)} + \frac{((6) - (73)\left(\frac{49}{426}\right))^2}{(73)\left(\frac{49}{426}\right)}
\]

\[
+ \frac{((94) - (239)\left(\frac{188}{426}\right))^2}{(239)\left(\frac{188}{426}\right)} + \frac{((115) - (239)\left(\frac{189}{426}\right))^2}{(239)\left(\frac{189}{426}\right)} + \frac{((30) - (239)\left(\frac{49}{426}\right))^2}{(239)\left(\frac{49}{426}\right)}
\]

\[
+ \frac{((49) - (114)\left(\frac{188}{426}\right))^2}{(114)\left(\frac{188}{426}\right)} + \frac{((52) - (114)\left(\frac{189}{426}\right))^2}{(114)\left(\frac{189}{426}\right)} + \frac{((13) - (114)\left(\frac{49}{426}\right))^2}{(114)\left(\frac{49}{426}\right)}
\]

\[
= \frac{27503239}{3026142} + \frac{235591}{105399} + \frac{354817}{4725756} = \frac{939839042381}{82450264932} \approx 11.40.
\]
On the other hand, $\chi^2_\alpha((h - 1)(k - 1))$ is equal to

$$\chi^2_\alpha((h - 1)(k - 1)) = \chi^2_{0.05}(4) = 9.488.$$ 

Since $Q$ is greater than $\chi^2_\alpha(4)$, we reject the null hypothesis.

**Remark 1.** The textbook uses different notations. The $Y_{ij}^{(j)}$ here is written $Y_{ij}$ in the textbook, $n^{(j)}$ here is written $n_j$ in the textbook, and $p_{ij}^{(j)}$ is written $p_{ij}$ in the textbook.
8 Example: Contingency tables

Four hundred UCLA undergraduate students are classified according to their college and their gender:

<table>
<thead>
<tr>
<th></th>
<th>Bsns</th>
<th>Engnrg</th>
<th>Lib. Arts</th>
<th>Nursing</th>
<th>Phrmcy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>21</td>
<td>16</td>
<td>145</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Female</td>
<td>14</td>
<td>4</td>
<td>175</td>
<td>13</td>
<td>4</td>
</tr>
</tbody>
</table>

Test at $\alpha = 0.01$ whether gender and choice of college are independent.
9 Setting: Contingency tables

Object:

- $X$ is an unknown random variables.

- Two different attributes:
  - $k$ mutually exclusive, exhaustive events $A_1, \ldots, A_k$;
  - $h$ mutually exclusive, exhaustive events $B_1, \ldots, B_h$;

Hypotheses:

- **Null Hypothesis** $H_0$: The two attributes are independent, i.e., for $i \in \{1, \ldots, k\}$, $j \in \{1, \ldots, h\}$:

  \[ P[A_i \cap B_j] = P[A_i]P[B_j]. \]

- **Alternative Hypothesis** $H_1$: The two attributes are not independent.
**Input:** $n$ many random samples for $X$, and significance level $\alpha$.

**Methodology:**

- Compute $Y(A_i, B_j)$ for $i \in \{1, 2, \ldots, k\}$ and $j \in \{1, \ldots, h\}$ by

  $$Y(A_i, B_j) := \text{number of times } A_i \text{ and } B_j \text{ occurs in samples}.$$ 

- Compute $a_1, \ldots, a_k$ given by

  $$a_i := \frac{1}{n}(\text{number of times } A_i \text{ occurs in the samples}).$$

- Compute $b_1, \ldots, b_h$ given by

  $$b_j := \frac{1}{n}(\text{number of times } B_j \text{ occurs in the samples}).$$
• Compute $Q$ given by

$$Q := \sum_{j=1}^{h} \sum_{i=1}^{k} \frac{(Y(A_i, B_j) - na_i b_j)^2}{na_i b_j}.$$  

• Reject $H_0$ if $Q \geq \chi^2_\alpha((h - 1)(k - 1))$, and the test is inconclusive otherwise.
10 Answer: Contingency tables

Let $A_1, \ldots, A_5$ be the event that a student is in the college of business, engineering, liberal arts, nursing, pharmacy, respectively.

Let $B_1$ be the event that the student is male, and let $B_2$ is the event that the student is female.

We have from the sample data that

$$a_1 = \frac{35}{400}; \quad a_2 = \frac{20}{400}; \quad a_3 = \frac{320}{400}; \quad a_4 = \frac{15}{400}; \quad a_5 = \frac{10}{400};$$

and

$$b_1 = \frac{190}{400}; \quad b_2 = \frac{210}{400}. $$
So $Q$ is equal to

$$
Q = \frac{((21) - (400)(\frac{35}{400})(\frac{190}{400}))^2}{(400)(\frac{35}{400})(\frac{190}{400})} + \frac{((145) - (400)(\frac{320}{400})(\frac{190}{400}))^2}{(400)(\frac{320}{400})(\frac{190}{400})} + \frac{((6) - (400)(\frac{10}{400})(\frac{190}{400}))^2}{(400)(\frac{10}{400})(\frac{190}{400})} + \frac{((4) - (400)(\frac{20}{400})(\frac{210}{400}))^2}{(400)(\frac{20}{400})(\frac{210}{400})} + \frac{((13) - (400)(\frac{15}{400})(\frac{210}{400}))^2}{(400)(\frac{15}{400})(\frac{210}{400})} + \frac{((175) - (400)(\frac{320}{400})(\frac{210}{400}))^2}{(400)(\frac{320}{400})(\frac{210}{400})} + \frac{((4) - (400)(\frac{10}{400})(\frac{210}{400}))^2}{(400)(\frac{10}{400})(\frac{210}{400})} + \frac{((14) - (400)(\frac{35}{400})(\frac{210}{400}))^2}{(400)(\frac{35}{400})(\frac{210}{400})} + \frac{((2) - (400)(\frac{15}{400})(\frac{190}{400}))^2}{(400)(\frac{15}{400})(\frac{190}{400})}
$$

$$
= 18.93.
$$

On the other hand, $\chi^2_\alpha(((h - 1)(k - 1))$ is equal to

$$
\chi^2_\alpha(((h - 1)(k - 1)) = \chi^2_{0.01}(4) = 13.28.
$$

Since $Q$ is greater than $\chi^2_\alpha(4)$, we reject the null hypothesis.
Remark 2. This test can be extended to test more than two attributes. Check the textbook for exercises.

Remark 3. The textbook uses different notations. The $Y(A_i, B_j)$ here is written $Y_{ij}$ in the textbook, $a_i$ here is written $Y_{i.}/n$ in the textbook, and $b_j$ is written $Y_{.j}/n$ in the textbook.