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 8th 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:

	A	B	C	D	F
Friendly instructor (FI)	8	13	16	10	3
Evil instructor (EI)	4	9	14	16	7

- 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

$$
\begin{aligned}
p_1 &:= \text{probability to get A in FI's class;}\\
p_2 &:= \text{probability to get B in FI's class;}\\
&\vdots\\
p_5 &:= \text{probability to get F in FI's class.}
\end{aligned}
$$

$$
\begin{aligned}
p'_1 &:= \text{probability to get A in EI's class;}\\
p'_2 &:= \text{probability to get B in EI's class;}\\
&\vdots\\
p'_5 &:= \text{probability to get F in EI's class.}
\end{aligned}
$$

These are **unknown parameters**.

$$
\begin{aligned}
Y_1 &:= \text{number of people getting A in FI's class;}\\
Y_2 &:= \text{number of people getting B in FI's class;}\\
&\ \vdots\\
Y_5 &:= \text{number of people getting F in FI's class.}
\end{aligned}
$$

$$
\begin{aligned}
Y_1' &:= \text{number of people getting A in EI's class;}\\
Y_2' &:= \text{number of people getting B in EI's class;}\\
&\ \vdots\\
Y_5' &:= \text{number of people getting F in EI's class.}
\end{aligned}
$$

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:

$$Y_1 \approx n p_1; \qquad Y_1' \approx n p_1', \qquad \text{and} \qquad \frac{Y_1 + Y_1'}{2n} \approx \frac{p_1 + p_1'}{2}.$$

   Let us write

$$\widehat{\mathrm{p}}_1 := \frac{Y_1 + Y_1'}{2n}.$$

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 $p_1 \approx p_1'$:

$$Y_1 \approx n p_1 \approx n \frac{p_1 + p_1'}{2} \approx n \frac{Y_1 + Y_1'}{2n} = n\, \widehat{\mathrm{p}}_1 \,.$$

4. Hence we have, if $p_1 \approx p_1'$, then

$$Y_1 - n\,\widehat{\mathrm{p}}_1 \;\approx\; 0.$$

By CLT, we in fact have, if $p_1 \approx p_1'$:

$$\frac{(Y_1 - n\,\widehat{\mathrm{p}}_1)^2}{n\,\widehat{\mathrm{p}}_1} \quad \text{is small.}$$

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

$$Y_1 = 0; \qquad \widehat{\mathrm{p}}_1 = \frac{1}{2},$$

so

$$\frac{(Y_1 - n\,\widehat{\mathrm{p}}_1)^2}{n\,\widehat{\mathrm{p}}_1} \;=\; \frac{(0 - n/2)^2}{n/2} \;=\; n/2 \;=\; \text{very big.}$$

6. So we conclude that

$$\frac{(Y_1 - n\,\widehat{\mathrm{p}}_1)^2}{n\,\widehat{\mathrm{p}}_1} \text{ is small} \qquad \text{if and only if} \qquad p_1 \approx p_1'.$$

By the same reasoning, we have

$$\frac{(Y_1' - n\,\widehat{\mathrm{p}}_1)^2}{n\,\widehat{\mathrm{p}}_1} \text{ is small} \qquad \text{if and only if} \qquad p_1 \approx p_1'.$$

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

$$\frac{(Y_1 - n\,\widehat{\mathrm{p}}_1)^2}{n\,\widehat{\mathrm{p}}_1} + \frac{(Y_1' - n\,\widehat{\mathrm{p}}_1)^2}{n\,\widehat{\mathrm{p}}_1} \text{ is small}$$
$$\text{if and only if} \qquad p_1 \approx p_1'.$$

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

$$\frac{(Y_i - n\,\widehat{\mathrm{p}}_i)^2}{n\,\widehat{\mathrm{p}}_i} + \frac{(Y_i' - n\,\widehat{\mathrm{p}}_i)^2}{n\,\widehat{\mathrm{p}}_i} \text{ is small}$$
$$\text{if and only if} \qquad p_i \approx p_i'.$$

9. We want to test all five parameters **simultaneously**. So we add all the tests up together:

$$
\begin{aligned}
Q := & \left[\frac{(Y_1 - n\,\widehat{\mathrm{p}}_1)^2}{n\,\widehat{\mathrm{p}}_1} + \frac{(Y_1' - n\,\widehat{\mathrm{p}}_1)^2}{n\,\widehat{\mathrm{p}}_1}\right] \\
& + \left[\frac{(Y_2 - n\,\widehat{\mathrm{p}}_2)^2}{n\,\widehat{\mathrm{p}}_2} + \frac{(Y_2' - n\,\widehat{\mathrm{p}}_2)^2}{n\,\widehat{\mathrm{p}}_2}\right] \\
& + \ldots + \left[\frac{(Y_5 - n\,\widehat{\mathrm{p}}_5)^2}{n\,\widehat{\mathrm{p}}_5} + \frac{(Y_5' - n\,\widehat{\mathrm{p}}_5)^2}{n\,\widehat{\mathrm{p}}_5}\right].
\end{aligned}
$$

We have

Q is small if and only if H_0 is true.

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

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

$$\widehat{\mathrm{p}}_1 = 0.12; \quad \widehat{\mathrm{p}}_2 = 0.22; \quad \widehat{\mathrm{p}}_3 = 0.30;$$

$$\widehat{\mathrm{p}}_4 = 0.26; \quad \widehat{\mathrm{p}}_5 = 0.10.$$

Then Q is equal to

$$\begin{aligned} Q = & \left[\frac{((8) - (50)(0.12))^2}{(50)(0.12)} + \frac{((4) - (50)(0.12))^2}{(50)(0.12)}\right] \\ & + \left[\frac{((13) - (50)(0.22))^2}{(50)(0.22)} + \frac{((9) - (50)(0.22))^2}{(50)(0.22)}\right] \\ & + \ldots + \left[\frac{((3) - (50)(0.10))^2}{(50)(0.10)} + \frac{((7) - (50)(0.10))^2}{(50)(0.10)}\right] = 5.18. \end{aligned}$$

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

  $p_i^{(j)} :=$ probability of the event A_i 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.,

$$\begin{array}{lllllll} p_1^{(1)} & = & p_1^{(2)} & = & \ldots & = & p_1^{(h)}; \quad \text{and} \\ p_2^{(1)} & = & p_2^{(2)} & = & \ldots & = & p_2^{(h)}; \quad \text{and} \\ \vdots & & \vdots & & \vdots & & \\ p_k^{(1)} & = & p_k^{(2)} & = & \ldots & = & p_k^{(h)}. \end{array}$$

- **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 α.

Methodology:

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

  $Y_i^{(j)} :=$ number of times A_i occurs in samples for $X^{(j)}$.

- Compute $\widehat{\mathrm{p}}_1, \ldots \widehat{\mathrm{p}}_k$ given by

$$\widehat{\mathrm{p}}_i := \frac{Y_i^{(1)} + Y_i^{(2)} + \ldots + Y_i^{(h)}}{n^{(1)} + n^{(2)} + \ldots + n^{(h)}}.$$

- Compute Q given by

$$Q := \sum_{j=1}^{h} \sum_{i=1}^{k} \frac{(Y_i^{(j)} - n^{(j)}\,\widehat{\mathrm{p}}_i)^2}{n^{(j)}\,\widehat{\mathrm{p}}_i}.$$

- Reject H_0 if $Q \geq \chi^2_\alpha((h-1)(k-1))$, 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:

	Newspaper	Television	Radio
Grade school	45	22	6
High School	94	115	30
College	49	52	13

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 $\alpha = 0.05$?

7 Answer: equality in distribution

From the sample data, we have

$$\begin{aligned} n^{(1)} &= 45 + 22 + 6 = 73; \\ n^{(2)} &= 94 + 115 + 30 = 239; \\ n^{(3)} &= 49 + 52 + 13 = 114, \end{aligned}$$

and

$$\begin{aligned} Y_1^{(1)} &= 45; & Y_2^{(1)} &= 22; & Y_3^{(1)} &= 6; \\ Y_1^{(2)} &= 94; & Y_2^{(2)} &= 115; & Y_3^{(2)} &= 30; \\ Y_1^{(3)} &= 49; & Y_2^{(3)} &= 52; & Y_3^{(3)} &= 13. \end{aligned}$$

Note that here $h = k = 3$.

So $\widehat{\mathrm{p}}_i$'s are given by

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

Then Q is equal to

$$Q = \sum_{j=1}^{h}\sum_{i=1}^{k} \frac{(Y_i^{(j)} - n^{(j)}\,\widehat{\mathrm{p}}_i)^2}{n^{(j)}\,\widehat{\mathrm{p}}_i}$$

$$\begin{aligned}
&= \frac{((45)-(73)(\frac{188}{426}))^2}{(73)(\frac{188}{426})} + \frac{((22)-(73)(\frac{189}{426}))^2}{(73)(\frac{189}{426})} + \frac{((6)-(73)(\frac{49}{426}))^2}{(73)(\frac{49}{426})}\\
&+ \frac{((94)-(239)(\frac{188}{426}))^2}{(239)(\frac{188}{426})} + \frac{((115)-(239)(\frac{189}{426}))^2}{(239)(\frac{189}{426})} + \frac{((30)-(239)(\frac{49}{426}))^2}{(239)(\frac{49}{426})}\\
&+ \frac{((49)-(114)(\frac{188}{426}))^2}{(114)(\frac{188}{426})} + \frac{((52)-(114)(\frac{189}{426}))^2}{(114)(\frac{189}{426})} + \frac{((13)-(114)(\frac{49}{426}))^2}{(114)(\frac{49}{426})}\\
&= \frac{27503239}{3026142} + \frac{235591}{105399} + \frac{354817}{4725756} = \frac{939839042381}{82450264932} \approx 11.40.
\end{aligned}$$

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 $Y_i^{(j)}$ here is written Y_{ij} in the textbook, $n^{(j)}$ here is written n_j in the textbook, and $p_i^{(j)}$ 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:

	Bsns	Engnrg	Lib. Arts	Nursing	Phrmcy
Male	21	16	145	2	6
Female	14	4	175	13	4

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 α.

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

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

- 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

$$b_1 = \frac{190}{400}; \qquad b_2 = \frac{210}{400}.$$

So Q is equal to

$$
\begin{aligned}
Q \;=\; & \frac{((21)-(400)(\frac{35}{400})(\frac{190}{400}))^2}{(400)(\frac{35}{400})(\frac{190}{400})} + \frac{((16)-(400)(\frac{20}{400})(\frac{190}{400}))^2}{(400)(\frac{20}{400})(\frac{190}{400})} \\
& + \frac{((145)-(400)(\frac{320}{400})(\frac{190}{400}))^2}{(400)(\frac{320}{400})(\frac{190}{400})} + \frac{((2)-(400)(\frac{15}{400})(\frac{190}{400}))^2}{(400)(\frac{15}{400})(\frac{190}{400})} \\
& + \frac{((6)-(400)(\frac{10}{400})(\frac{190}{400}))^2}{(400)(\frac{10}{400})(\frac{190}{400})} + \frac{((14)-(400)(\frac{35}{400})(\frac{210}{400}))^2}{(400)(\frac{35}{400})(\frac{210}{400})} \\
& + \frac{((4)-(400)(\frac{20}{400})(\frac{210}{400}))^2}{(400)(\frac{20}{400})(\frac{210}{400})} + \frac{((175)-(400)(\frac{320}{400})(\frac{210}{400}))^2}{(400)(\frac{320}{400})(\frac{210}{400})} \\
& + \frac{((13)-(400)(\frac{15}{400})(\frac{210}{400}))^2}{(400)(\frac{15}{400})(\frac{210}{400})} + \frac{((4)-(400)(\frac{10}{400})(\frac{210}{400}))^2}{(400)(\frac{10}{400})(\frac{210}{400})} \\
= \; & 18.93.
\end{aligned}
$$

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.