Math 170S

Lecture Notes Section 9.1 [*†]

Chi-square tests

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: Monday 7th December, 2020, 08:29.

[†]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

Your friendly instructor was asked by a casino to determine if a die is fair. So he rolled the die n times, and

$$Y_i \; := \; \text{number of times the outcome of the die roll is } i\,.$$

Note that, for $i \in \{1, 2, 3, 4, 5, 6\}$,

- Y_i is a binomial random variable with a **known** number of trials n **unknown** success probability p_i.
- $Y_1, \ldots, Y_6$ are **not independent** random variables since $Y_1 + Y_2 + \ldots + Y_6 = n$.
- Also note that $p_1 + \ldots + p_6 = 1$.

The friendly instructor makes these two hypothesis:

- **Null Hyp.** H_0: $p_1 = p_2 = \ldots = p_6 = \frac{1}{6}$;

- **Alternative Hyp.** H_1: there is i with $p_i \neq \frac{1}{6}$.

2 The difference: Level 1

How should the friendly instructor test this hypothesis?

- **What we usually dealt with**: Usually we have **one** unknown parameter, e.g., mean only, median only, or variance only.

- **What we are dealing with now**: We now care about **six** unknown parameters together: $p_1, \ldots, p_6$.

3 Rationale: Chi-square tests

- If $p_1 = \frac{1}{6}$, then Y_1 has mean $np_1 = \frac{n}{6}$.

  So the (random) square difference $(Y_1 - \frac{n}{6})^2$ is very close to 0 if $p_1 = \frac{1}{6}$.

  Then, by the central limit theorem, the rescaled

$$\frac{(Y_1 - \frac{n}{6})^2}{n/6} \qquad \text{is still very small.}$$

- If $p_1 \neq \frac{1}{6}$ (e.g., $p_1 = 0$), then the (random) ratio

$$\frac{(Y_1 - \frac{n}{6})^2}{n/6} \approx \frac{(0 - \frac{n}{6})^2}{n/6} \approx \frac{n}{6} \qquad \text{is very big.}$$

- So we conclude that

$$\frac{(Y_1 - \frac{n}{6})^2}{n/6} \begin{cases} \text{is very small} & \text{if} \quad p_1 \approx \frac{1}{6}, \\ \text{is very big} & \text{if} \quad p_1 \text{ is far from } \frac{1}{6}. \end{cases}$$

- By the same reasoning, for every $i \in \{1, \ldots, 6\}$,

$$\frac{(Y_i - \frac{n}{6})^2}{n/6} \begin{cases} \text{is very small} & \text{if} \quad p_i \approx \frac{1}{6}, \\ \text{is very big} & \text{if} \quad p_i \text{ is far from } \frac{1}{6}. \end{cases}$$

- We add them up so we can **track all these parameters simultaneously**:

$$Q_5 := \frac{(Y_1 - \frac{n}{6})^2}{n/6} + \frac{(Y_2 - \frac{n}{6})^2}{n/6} + \ldots + \frac{(Y_6 - \frac{n}{6})^2}{n/6}.$$

  Since each term in the sum is nonnegative, we have

$$Q_5 \begin{cases} \text{is very small} & \text{if} \quad p_1, \ldots, p_6 \approx \frac{1}{6}; \\ \text{is very big} & \text{if} \quad \text{some } p_i \text{ are far from } \frac{1}{6}. \end{cases}$$

- Here Q_5 has (six - one) degrees of freedom.

  The six degrees is because we have six parameters $p_1, p_2, \ldots, p_6$.

  The minus one degree is because these parameters satisfy one equation $p_1 + \ldots + p_6 = 1$.

- It can be shown that Q_5 is approximately a χ^2 random variable with 5 degrees of freedom.

4 Setting: Chi-square tests

Object:

- X is an RV with **unknown distribution**.
- There are k mutually exclusive events $A_1, \ldots, A_k$,

  $$p_i := P[A_i] \quad i \in \{1, \ldots, k\}, \qquad p_1 + \ldots + p_k = 1,$$

  where $p_1, \ldots, p_k$ are **unknown constants**.

Hypotheses: Given $c_1, \ldots, c_k$,

- **Null Hypothesis** H_0:

  $$p_1 = c_1, \quad p_2 = c_2, \quad \ldots, \quad p_k = c_k.$$

- **Alternative Hypothesis** H_1:

  $$p_i \neq c_i \quad \text{for some } i.$$

Input: Random samples $X_1, \ldots, X_n$ for X and significance level α.

Methodology:

- Compute $Y_1, \ldots, Y_k$ by

  $$Y_i \ := \ \text{number of times } A_i \text{ occurs in } \ X_1, \ldots, X_n.$$

- Compute Q_{k-1} by

  $$Q_{k-1} \ := \ \sum_{i=1}^{k} \frac{(Y_i - nc_i)^2}{nc_i}.$$

- Reject H_0 if $Q_{k-1} \ \geq \ \chi^2_\alpha(k-1)$, and the test is inconclusive otherwise.

  The value $\chi^2_\alpha(k-1)$ can be found from the Table IV Appendix B in the textbook.

5 Example: Chi-square tests, Level 1

The friendly instructor rolls the dice 60 times, and get

$$
\begin{aligned}
Y_1 &= 12; & Y_2 &= 11; & Y_3 &= 9; \\
Y_4 &= 7; & Y_5 &= 10; & Y_6 &= 11.
\end{aligned}
$$

Can the instructor reject the null hypothesis with significance level $\alpha = 0.05$?

6 Answer: Chi-square tests

We have

$$
\begin{aligned}
Q_{k-1} &= \sum_{i=1}^{6} \frac{(Y_i - nc_i)^2}{nc_i} \\
&= \frac{(12-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(9-10)^2}{10} + \frac{(7-10)^2}{10} \\
&\quad + \frac{(10-10)^2}{10} + \frac{(11-10)^2}{10} \\
&= \frac{8}{5} = 1.6.
\end{aligned}
$$

On the other hand,

$$\chi^2_\alpha(k-1) = \chi^2_{0.05}(5) = 11.07.$$

The test is therefore inconclusive.

7 The difference: Level 2

- **What we usually dealt with**: X is a known random variable (e.g., normal, Bernoulli, chi-square).

- **What we are dealing with now**: X is a random variable with **unknown distribution**, which we want to guess.

8 Setting: Chi-square tests, Level 2

Object:

- X is an RV with **unknown distribution**.
- There are k mutually exclusive events $A_1, \ldots, A_k$

$$P[A_1 \cup A_2 \cup \ldots \cup A_k] = 1.$$

Hypotheses: Given some density function f_θ,

- **Null Hypothesis** H_0: density of X is f_θ **for some** θ.
- **Alternative Hypothesis** H_1: density of X is not f_θ **for any** θ.

Input: Random samples $X_1, \ldots, X_n$ for X and significance level α.

Methodology:

- Compute the best guess $\widehat{\theta}$ for θ. (This is usually the MLE $\widehat{\theta}$ for the density f_θ.)

- Compute the probability of the events $A_1, \ldots, A_k$:

$$c_i \;:=\; P[A_i] \;=\; \begin{cases} \sum_{x \in A_i} f_{\widehat{\theta}}(x) & \text{if discrete;} \\ \int_{x \in A_i} f_{\widehat{\theta}}(x)\, dx & \text{if continuous.} \end{cases}$$

- Compute $Y_1, \ldots, Y_k$ given by

$$Y_i := \text{number of times } A_i \text{ occurs in } X_1, \ldots, X_n.$$

- Compute Q_{k-1} given by

$$Q_{k-1} \;:=\; \sum_{i=1}^{k} \frac{(Y_i - nc_i)^2}{nc_i}.$$

- Reject H_0 if $Q_{k-1} \geq \chi^2_\alpha(k-\mathbf{2})$, and the test is inconclusive otherwise. Note that the degree of freedom drops by another one degree because we spent it on estimating $\widehat{\theta}$.

9 Example: Level 2

Let X be an RV with the following $n = 50$ samples:

$$
\begin{array}{cccccccccc}
7 & 4 & 3 & 6 & 4 & 4 & 5 & 3 & 5 & 3 \\
5 & 5 & 3 & 2 & 5 & 4 & 3 & 3 & 7 & 6 \\
6 & 4 & 3 & 11 & 9 & 6 & 7 & 4 & 5 & 4 \\
7 & 3 & 2 & 8 & 6 & 7 & 4 & 1 & 9 & 8 \\
4 & 8 & 9 & 3 & 9 & 7 & 7 & 9 & 3 & 10
\end{array}
$$

The hypothesis are:

- Null hypotheses H_0: X is a Poisson RV;
- Alternative hypotheses H_1: X is not a Poisson RV.

The $k = 6$ events we want to test are

- $A_1 = \{0, 1, 2, 3\}$;
- $A_2 = \{4\}$;
- $A_3 = \{5\}$;
- $A_4 = \{6\}$;
- $A_5 = \{7\}$;
- $A_6 = \{8, 9, 10, \ldots\}$.

Can we reject the null hypothesis at $\alpha = 0.05$ significance level?

10 Answer: Level 2

Recall that the Poisson RV with mean λ has density

$$f_\lambda(x) \ = \ \frac{\lambda^x e^{-\lambda}}{x!}.$$

Also recall that the MLE $\widehat{\lambda}$ for the Poisson RV is the sample mean, so

$$\widehat{\lambda} = \overline{\mathrm{x}} = 5.4.$$

We now compute the probability of the events $A_1, \ldots, A_6$:

$$
\begin{aligned}
c_1 &= P[A_1] = \sum_{x=0}^{3} \frac{(5.4)^x e^{-(5.4)}}{x!} = 0.213 \\
c_2 &= P[A_2] = \frac{(5.4)^4 e^{-(5.4)}}{4!} = 0.160 \\
c_3 &= P[A_3] = \frac{(5.4)^5 e^{-(5.4)}}{5!} = 0.173 \\
c_4 &= P[A_4] = \frac{(5.4)^6 e^{-(5.4)}}{6!} = 0.156 \\
c_5 &= P[A_5] = \frac{(5.4)^7 e^{-(5.4)}}{7!} = 0.120 \\
c_6 &= P[A_6] = \sum_{x=8}^{\infty} \frac{(5.4)^x e^{-(5.4)}}{x!} = 0.178.
\end{aligned}
$$

We now compute $Y_1, \ldots, Y_6$ from the sample mean:

$$
Y_1 = 13; \qquad Y_2 = 9; \qquad Y_3 = 6;
$$

$$
Y_4 = 5; \qquad Y_5 = 7; \qquad Y_6 = 10.
$$

We now compute Q_{k-1} given by

$$
\begin{aligned}
Q_{k-1} &= \sum_{i=1}^{k} \frac{(Y_i - nc_i)^2}{nc_i} \\
&= \frac{(13-(50)(0.213))^2}{(50)(0.213)} + \frac{(9-(50)(0.160))^2}{(50)(0.160)} \\
&\quad + \frac{(6-(50)(0.173))^2}{(50)(0.173)} + \frac{(5-(50)(0.156))^2}{(50)(0.156)} \\
&\quad + \frac{(7-(50)(0.120))^2}{(50)(0.120)} + \frac{(10-(50)(0.178))^2}{(50)(0.178)} \\
&= 2.763.
\end{aligned}
$$

On the other hand,

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

The test is therefore inconclusive.

Remark 1. If X has density determined by d parameters (e.g., normal random variable $N(\mu, \sigma^2)$ with $d = 2$ parameter), then the degree of freedom of the chi-square random variable will drop to $k - 1 - d$.