Regression and Linear Models (37252)

Lecture 9 - Analysis of Categorical RVs

Lecturer: Joanna Wang
Notes adopted from Scott Alexander

School of Mathematical and Physical Sciences, UTS

Autumn 2024

Acknowledgement

These notes are based, in part, on earlier versions prepared by Dr Ed Lidums and Prof. James Brown.

9. Lecture outline

Topics:

- introduction
- multinomial distribution
- Chi-square goodness-of-fit test
- two-way tables
  - joint and marginal distributions
  - conditional distributions
- Chi-square independence test
- relative risks
- odds and odds ratios

For the sections on Chi-square tests and two-way table analysis we have relied on Chapter 14 of Wackerly et al. (2008).

9. Introduction

In the remainder of the course we will look at logistic regression, a tool for modelling categorical dependent variables.

The dependent variable of interest will be the log of **odds**.

In this lecture we lay the groundwork and describe how odds, and the related **odds ratios**, arise in the context of **two-way tables**.

We also described the necessary statistical tools used in the analysis of two-way tables.

9. Multinomial distribution

Consider a categorical RV Z taking K possible states $z_1, \ldots, z_K$.

Define $p_k = \text{Prob}(Z = z_k)$ with the usual conditions $0 \leq p_k \leq 1$ and

$$\sum_{k=1}^{K} p_k = 1.$$

Let $Z^{(1)}, \ldots, Z^{(N)}$ be N independent RVs drawn from this distribution and define

$$N_k = \#\{n \in \{1, \ldots, N\} | Z^{(n)} = z_k\}$$

as the number of these RVs in state z_k.

It should be clear that

$$\sum_{k=1}^{K} N_k = N.$$

9. Multinomial distribution

The RV $(N_1, \ldots, N_K)$ is said to possess a **multinomial** distribution and we write

$$(N_1, \ldots, N_K) \sim \text{Multinom}(N, p_1, \ldots, p_K).$$

Each N_k possess a binomial distribution; i.e.

$$N_k \sim \text{Bin}(N, p_k) \tag{1}$$

for $k \in \{1, \ldots, K\}$.

We can use such RVs to describe a **multinomial experiment** characterised by:

- N independent trials
- the result of each trial falling into one of K distinct categories or **cells**
- the quantity of interest being the number of trials N_k falling in the k-th cell.

9. Chi-square goodness-of-fit test

$X \sim B(N, p)$ $E(X) = Np$

Using $\mathbb{E}[N_k] = p_k N$, which follows from (1), it can be shown that for large enough N the RV

$$X^2 = \sum_{k=1}^{K} \frac{\left(N_k - \mathbb{E}[N_k]\right)^2}{\mathbb{E}[N_k]} \quad \sum\left(\frac{\text{observed} - \text{expected}}{\text{expected}}\right)^2$$

$$= \sum_{k=1}^{K} \frac{\left(N_k - p_k N\right)^2}{p_k N} \tag{2}$$

is approximately Chi-squared–distributed (we also need $\mathbb{E}[N_k] \geq 5$ or thereabouts – see pg. 715 Wackerly et al. (2008).

The degrees of freedom necessary to parameterise the distribution depends on the form of the hypothesis test in which it is used.

9. Chi-square goodness-of-fit test

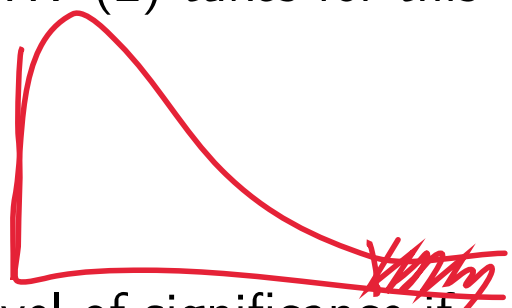
A test that can be constructed using (2) as a test statistic involves hypothesised values of p_k.

Consider the hypotheses

$$H_0: \; p_1 = p_1^*, \ldots, p_K = p_K^*$$
$$H_A: \; \text{at least one } p_k \neq p_k^*.$$

Let x_*^2 be the test statistic, which is the value the RV (2) takes for this particular test.

Also let $\chi^2 \sim \text{ChiSquare}(k-1)$.



The null hypothesis H_0 can be rejected at the α level of significance if

$$p = \text{Prob}(\chi^2 > x_*^2) < \alpha.$$

We will use the results of this section to develop the Chi-square test of independence.

9. Two-way tables

The following table collects data from the British Household Panel Survey, a survey collecting health-related data for respondents in Britain.

The data is broken down into two categories:

- self-reported health ($Health = 1$ for very good, ..., $Health = 5$ for very poor)
- visits to a GP ($GP = 0$ for infrequent, $GP = 1$ for frequent).

Sample data

		GP		
		0	**1**	**Total**
Health	**1**	1351	192	1543
	2	2051	936	2987
	3	507	766	1273
	4	51	291	342
	5	6	72	78
Total		3966	2257	6223

9. Two-way tables – joint and marginal distributions

Define $N_{i,j}$ as the **observed cell count** corresponding to $Health = i$, $GP = j$; e.g. $N_{1,1} = 1351$.

Also define

$$N_{i,\cdot} = \sum_j N_{i,j} \quad \text{and} \quad N_{\cdot,j} = \sum_i N_{i,j}$$

so that, for instance, $N_{1,\cdot} = 1543$ and $N_{\cdot,1} = 3966$.

Set the total observations $N = 6223$ and convert the frequency data into percentages of total observations N.

Sample joint and marginal probabilities

Health	GP = 0	GP = 1	Total
1	0.2171	0.0309	0.2480
2	0.3296	0.1504	0.4800
3	0.0815	0.1231	0.2046
4	0.0082	0.0468	0.0550
5	0.0010	0.0116	0.0125
Total	0.6373	0.3627	1.0000

9. Two-way tables – joint and marginal distributions

The previous table defines the **joint distribution** of $(Health, GP)$ as the probabilities

$$p_{i,j} := \text{Prob}(Health = i, GP = j)$$

so that, for instance, $p_{1,1} = 0.2171$.

The table also defines the **marginal distribution** of $Health$ via

$$p_{i,\cdot} := \text{Prob}(Health = i) = \sum_j p_{i,j}$$

and the **marginal distribution** of GP via the probabilities

$$p_{\cdot,j} := \text{Prob}(GP = j) = \sum_i p_{i,j}$$

so that, for example, $p_{1,\cdot} = 0.2480$ and $p_{\cdot,1} = 0.6373$.

Obviously these marginal probabilities sum to one.

9. Two-way tables – conditional distributions

If we normalise the previous table so that the row probabilities sum to one we get the following.

Sample conditional row probabilities

Health	GP 0	GP 1	Total
1	0.8756	0.1244	1.0000
2	0.6866	0.3134	1.0000
3	0.3983	0.6017	1.0000
4	0.1491	0.8509	1.0000
5	0.0769	0.9231	1.0000

This table defines the **conditional distribution** of $GP|Health$; i.e.

$$\text{Prob}(GP = j|Health = i) = \frac{\text{Prob}(Health = i, GP = j)}{\text{Prob}(Health = i)} = \frac{p_{i,j}}{p_{i,\cdot}}$$

so that, for example,

$$\text{Prob}(GP = 0|Health = 1) = \frac{p_{1,0}}{p_{1,\cdot}} = \frac{0.2171}{0.2480} = 0.8756.$$

9. Two-way tables – conditional distributions

Of course, we can do the same thing with the columns.

Sample conditional column probabilities

		GP	
		0	1
Health	1	0.3406	0.0851
	2	0.5171	0.4147
	3	0.1278	0.3394
	4	0.0129	0.1289
	5	0.0015	0.0319
Total		1.0000	1.0000

This table defines the **conditional distribution** of $Health|GP$; i.e.

$$\text{Prob}(Health = i|GP = j) = \frac{\text{Prob}(Health = i, GP = j)}{\text{Prob}(GP = j)} = \frac{p_{i,j}}{p_{\cdot,j}}$$

so that, for example,

$$\text{Prob}(Health = 1|GP = 0) = \frac{p_{1,0}}{p_{\cdot,0}} = \frac{0.2171}{0.6373} = 0.3406.$$

9. Chi-square independence test

The **Chi-square test of independence** relies on a fundamental property of independent RVs.

Let X, Y be independent categorical (or discrete numerical) RVs. Then the joint probability of X, Y equals the multiple of their marginal probabilities; i.e.

$$\text{Prob}(X = x, Y = y) = \text{Prob}(X = x)\,\text{Prob}(Y = y).$$

(Note: for continuous numerical RVs we need a slight modification.)

We can construct the table of joint probabilities Prob(*Health*, *GP*) under the assumption of independence from the table *Sample joint and marginal probabilities* table.

Define

$$p^*_{i,j} := \text{Prob}(Health = i) \times \text{Prob}(GP = j) = p_{i,\cdot} \times p_{\cdot,j}$$

as the joint probability under the assumption of independence.

9. Chi-square independence test

For example, under the assumption of independence

$$p^*_{1,1} = p_{1,\cdot} \times p_{\cdot,1} = 0.2480 \times 0.6373 = 0.1580.$$

Proceeding in this manner results in the following table.

Independent joint and marginal probabilities

		GP		
		0	1	Total
Health	1	0.1580	0.0899	0.2480
	2	0.3059	0.1741	0.4800
	3	0.1304	0.0742	0.2046
	4	0.0350	0.0199	0.0550
	5	0.0080	0.0045	0.0125
Total		0.6373	0.3627	1.0000

9. Chi-square independence test

These probabilities can then be used to calculate the expected value of the cell frequencies under independence, which we define as

$$N^*_{i,j} := \mathbb{E}[N_{i,j}|Health, GP \text{ independent}] \equiv \mathbb{E}[N_{i,j}] = N \times p^*_{i,j}.$$

For example, the expected number of observations corresponding to $(Health = 1, GP = 0)$ is

$$N^*_{1,1} = N \times p^*_{1,1} = 6223 \times 0.1580 \approx 983.$$

This procedure produces the following table.

Expected cell counts under independence

		GP		
		0	1	Total
Health	1	983	560	1543
	2	1904	1083	2987
	3	811	462	1273
	4	218	124	342
	5	50	28	78
Total		3966	2257	6223

9. Chi-square independence test

Under the assumption that *Health* and *GP* are independent we have

$$N^*_{i,j} \sim \text{Multinom}(N, p^*_{i,j}).$$

The test statistic, derived from (2), takes the value

$$\begin{aligned} x^2_* &= \sum_{i,j} \frac{\left(N_{i,j} - \mathbb{E}[N_{i,j}]\right)^2}{\mathbb{E}[N_{i,j}]} = \sum_{k=1}^{K} \frac{\left(N_{i,j} - N^*_{i,j}\right)^2}{N^*_{i,j}} \\ &\approx 1184 \end{aligned}$$

and is, under the assumption of independence, from a Chi-square distribution with $(5-1)(2-1) = 4$ degrees of freedom.

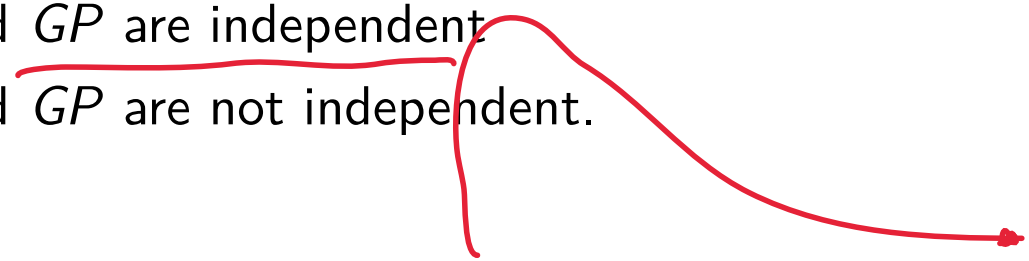
9. Chi-square independence test

We now formalise the **Chi-square independent test**.

Define the hypotheses

H_0: *Health* and *GP* are independent

H_A: *Health* and *GP* are not independent.



Let $\chi^2 \sim \text{ChiSquare}(4)$.

As

$$0.0 \approx \text{Prob}(\chi^2 > 1184) < 0.05$$

we can reject H_0 and conclude that *Health* and *GP* are not independent.

9. Relative risks

We can use the conditional probabilities, calculated above, to calculate **Relative Risk (RR)**.

Reconsider the conditional probabilities Prob($GP|Health$).

Sample conditional row probabilities

		GP		
		0	**1**	**Total**
Health	**1**	0.8756	0.1244	1.0000
	2	0.6866	0.3134	1.0000
	3	0.3983	0.6017	1.0000
	4	0.1491	0.8509	1.0000
	5	0.0769	0.9231	1.0000

The RR for those with very poor health ($Health = 5$) visiting the GP frequently (GP=1) compared to those with very good health ($Health = 1$) is

$$\frac{\text{Prob}(GP = 1|Health = 5)}{\text{Prob}(GP = 1|Health = 1)} = \frac{0.9231}{0.1244} \approx 7.42.$$

9. Relative risks

Repeating this calculation for all $Health = i$ gives the following table.

Relative risks

		Prob(GP=1 | Health=i)	Relative risk
Health	1	0.1244	1.00
	2	0.3134	2.52
	3	0.6017	4.84
	4	0.8509	6.84
	5	0.9231	7.42

This gives the multiple over the baseline $Health = 1$ of the likelihood of visiting the GP frequently.

9. Odds and odds ratios

We can also use the conditional probabilities to define **odds** and **odds ratios**.

Define

$$p_i := \text{Prob}(GP = 1 | Health = i).$$

Then given $Health = i$, the odds of visiting the GP frequently over infrequently is given by

$$odds_i := \frac{p_i}{1 - p_i}. \tag{3}$$

The odds ratio for $Health = i$ over the reference group $Health = 1$ is given by

$$oddsRatio_{i,1} := \frac{odds_i}{odds_1}. \tag{4}$$

9. Odds and odds ratios

Applying this rule gives the following table.

Odds and odds ratios

		GP			
		0	1	Odds	Odds ratio
Health	1	0.8756	0.1244	0.1421	1.0000
	2	0.6866	0.3134	0.4564	3.2112
	3	0.3983	0.6017	1.5108	10.6310
	4	0.1491	0.8509	5.7059	40.1492
	5	0.0769	0.9231	12.0000	84.4375

We see, for instance, that the odds of those with very poor health visiting the GP frequently are more than 84 times those with very good health.

9. Odds and odds ratios

Using $Health = 1$ as the reference group set

$$odds_1 = \frac{p_1}{1 - p_1} = e^{\beta_0}.$$

Also set the odds ratio for the group $Health = 5$

$$oddsRatio_{5,1} = e^{\beta_1}.$$

Then by (4)

$$odds_5 = odds_1 \times oddsRatio_{5,1} = e^{\beta_0} \times e^{\beta_1}.$$

This gives rise to the **multiplicative model**

$$\frac{p_i}{1 - p_i} = e^{\beta_0} \times e^{\beta_1 x_i}, \quad x_i \in \{0, 1\},$$

the log of which

$$\ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_i, \quad x_i \in \{0, 1\},$$

is the basis for the logistic regression with a two-state categorical predictor.

9. Odds and odds ratios

Recall that with the Chi-square test of independence we were able to reject the null hypothesis that *Health* and *GP* were independent.

If we were not able to reject this hypothesis, then the conditional probabilities

$$\text{Prob}(GP = j | Health = i) = \frac{\text{Prob}(Health = i, GP = j)}{\text{Prob}(Health = i)}$$

$$= \frac{\text{Prob}(Health = i)\,\text{Prob}(GP = j)}{\text{Prob}(Health = i)}$$

$$= \text{Prob}(GP = j) = \begin{cases} 0.6373, & j = 0 \\ 0.3627, & j = 1 \end{cases}$$

contradicting those in the table *Sample conditional row probabilities*.

Using these probabilities, $odds_i$ would be the the same for all i and $oddsRatio_{i,1} = 1$ for all i, giving us nothing to model with a logistic regression.

References I

Wackerly, D., Mendenhall, W., and Scheaffer, R.L. (2008). *Mathematical Statistics with Applications.* Thomson Brooks/Cole, Belmont, CA, 7 edition.

References I

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*. 2nd edition.