15
Inference About Independence

In this chapter we address the following questions:

(1) How do we test if two random variables are independent?

(2) How do we estimate the strength of dependence between two random variables?

When Y and Z are not independent, we say that they are **dependent** or **associated** or **related.** If Y and Z are associated, it does **not** imply that Y causes Z or that Z causes Y. Causation is discussed in Chapter 16.

Recall that we write $Y \amalg Z$ to mean that Y and Z are independent and we write $Y \not\amalg Z$ to mean that Y and Z are dependent.

15.1 Two Binary Variables

Suppose that Y and Z are both binary and consider data $(Y_1, Z_1), \ldots, (Y_n, Z_n)$. We can represent the data as a two-by-two table:

	$Y = 0$	$Y = 1$	
$Z = 0$	X_{00}	X_{01}	$X_{0\cdot}$
$Z = 1$	X_{10}	X_{11}	$X_{1\cdot}$
	$X_{\cdot 0}$	$X_{\cdot 1}$	$n = X_{\cdot\cdot}$

where

$$X_{ij} = \text{number of observations for which } Y = i \text{ and } Z = j.$$

The dotted subscripts denote sums. Thus,

$$X_{i\cdot} = \sum_j X_{ij}, \quad X_{\cdot j} = \sum_i X_{ij}, \quad n = X_{\cdot\cdot} = \sum_{i,j} X_{ij}.$$

This is a convention we use throughout the remainder of the book. Denote the corresponding probabilities by:

	$Y=0$	$Y=1$	
$Z=0$	p_{00}	p_{01}	$p_{0\cdot}$
$Z=1$	p_{10}	p_{11}	$p_{1\cdot}$
	$p_{\cdot 0}$	$p_{\cdot 1}$	1

where $p_{ij} = \mathbb{P}(Z = i, Y = j)$. Let $X = (X_{00}, X_{01}, X_{10}, X_{11})$ denote the vector of counts. Then $X \sim \text{Multinomial}(n, p)$ where $p = (p_{00}, p_{01}, p_{10}, p_{11})$. It is now convenient to introduce two new parameters.

15.1 Definition. *The* **odds ratio** *is defined to be*

$$\psi = \frac{p_{00}p_{11}}{p_{01}p_{10}}. \tag{15.1}$$

The **log odds ratio** *is defined to be*

$$\gamma = \log(\psi). \tag{15.2}$$

15.2 Theorem. *The following statements are equivalent:*

1. $Y \amalg Z$.
2. $\psi = 1$.
3. $\gamma = 0$.
4. *For* $i, j \in \{0, 1\}$, $p_{ij} = p_{i\cdot}p_{\cdot j}$.

Now consider testing

$$H_0 : Y \amalg Z \text{ versus } H_1 : Y \not\amalg Z. \tag{15.3}$$

First we consider the likelihood ratio test. Under H_1, $X \sim \text{Multinomial}(n, p)$ and the MLE is the vector $\widehat{p} = X/n$. Under H_0, we again have that $X \sim \text{Multinomial}(n, p)$ but the restricted MLE is computed under the constraint $p_{ij} = p_{i\cdot}p_{\cdot j}$ This leads to the following test:

15.3 Theorem. *The likelihood ratio test statistic for (15.3) is*

$$T = 2\sum_{i=0}^{1}\sum_{j=0}^{1} X_{ij}\log\left(\frac{X_{ij}X_{..}}{X_{i\cdot}X_{\cdot j}}\right). \tag{15.4}$$

Under H_0, $T \rightsquigarrow \chi^2_1$. Thus, an approximate level α test is obtained by rejecting H_0 when $T > \chi^2_{1,\alpha}$.

Another popular test for independence is Pearson's χ^2 test.

15.4 Theorem. *Pearson's χ^2 test statistic for independence is*

$$U = \sum_{i=0}^{1}\sum_{j=0}^{1}\frac{(X_{ij}-E_{ij})^2}{E_{ij}} \tag{15.5}$$

where

$$E_{ij} = \frac{X_{i\cdot}X_{\cdot j}}{n}.$$

Under H_0, $U \rightsquigarrow \chi^2_1$. Thus, an approximate level α test is obtained by rejecting H_0 when $U > \chi^2_{1,\alpha}$.

Here is the intuition for the Pearson test. Under H_0, $p_{ij} = p_{i\cdot}p_{\cdot j}$, so the maximum likelihood estimator of p_{ij} under H_0 is

$$\widehat{p}_{ij} = \widehat{p}_{i\cdot}\widehat{p}_{\cdot j} = \frac{X_{i\cdot}}{n}\frac{X_{\cdot j}}{n}.$$

Thus, the expected number of observations in the (i,j) cell is

$$E_{ij} = n\widehat{p}_{ij} = \frac{X_{i\cdot}X_{\cdot j}}{n}.$$

The statistic U compares the observed and expected counts.

15.5 Example. The following data from Johnson and Johnson (1972) relate tonsillectomy and Hodgkins disease. [1]

	Hodgkins Disease	No Disease	
Tonsillectomy	90	165	255
No Tonsillectomy	84	307	391
Total	174	472	646

[1] The data are actually from a case-control study; see the appendix for an explanation of case-control studies.

We would like to know if tonsillectomy is related to Hodgkins disease. The likelihood ratio statistic is $T = 14.75$ and the p-value is $\mathbb{P}(\chi^2_1 > 14.75) = .0001$. The χ^2 statistic is $U = 14.96$ and the p-value is $\mathbb{P}(\chi^2_1 > 14.96) = .0001$. We reject the null hypothesis of independence and conclude that tonsillectomy is associated with Hodgkins disease. This does not mean that tonsillectomies cause Hodgkins disease. Suppose, for example, that doctors gave tonsillectomies to the most seriously ill patients. Then the association between tonsillectomies and Hodgkins disease may be due to the fact that those with tonsillectomies were the most ill patients and hence more likely to have a serious disease. ■

We can also estimate the strength of dependence by estimating the odds ratio ψ and the log-odds ratio γ.

15.6 Theorem. *The* MLE*'s of* ψ *and* γ *are*

$$\widehat{\psi} = \frac{X_{00}X_{11}}{X_{01}X_{10}}, \quad \widehat{\gamma} = \log \widehat{\psi}. \tag{15.6}$$

The asymptotic standard errors (computed using the delta method) are

$$\widehat{\mathsf{se}}(\widehat{\gamma}) = \sqrt{\frac{1}{X_{00}} + \frac{1}{X_{01}} + \frac{1}{X_{10}} + \frac{1}{X_{11}}} \tag{15.7}$$

$$\widehat{\mathsf{se}}(\widehat{\psi}) = \widehat{\psi}\, \widehat{\mathsf{se}}(\widehat{\gamma}). \tag{15.8}$$

15.7 Remark. For small sample sizes, $\widehat{\psi}$ and $\widehat{\gamma}$ can have a very large variance. In this case, we often use the modified estimator

$$\widehat{\psi} = \frac{\left(X_{00} + \frac{1}{2}\right)\left(X_{11} + \frac{1}{2}\right)}{\left(X_{01} + \frac{1}{2}\right)\left(X_{10} + \frac{1}{2}\right)}. \tag{15.9}$$

Another test for independence is the Wald test for $\gamma = 0$ given by $W = (\widehat{\gamma} - 0)/\widehat{\mathsf{se}}(\widehat{\gamma})$. A $1 - \alpha$ confidence interval for γ is $\widehat{\gamma} \pm z_{\alpha/2}\widehat{\mathsf{se}}(\widehat{\gamma})$.

A $1 - \alpha$ confidence interval for ψ can be obtained in two ways. First, we could use $\widehat{\psi} \pm z_{\alpha/2}\widehat{\mathsf{se}}(\widehat{\psi})$. Second, since $\psi = e^{\gamma}$ we could use

$$\exp\left\{\widehat{\gamma} \pm z_{\alpha/2}\widehat{\mathsf{se}}(\widehat{\gamma})\right\}. \tag{15.10}$$

This second method is usually more accurate.

15.8 Example. In the previous example,

$$\widehat{\psi} = \frac{90 \times 307}{165 \times 84} = 1.99$$

and

$$\widehat{\gamma} = \log(1.99) = .69.$$

So tonsillectomy patients were twice as likely to have Hodgkins disease. The standard error of $\widehat{\gamma}$ is

$$\sqrt{\frac{1}{90}+\frac{1}{84}+\frac{1}{165}+\frac{1}{307}} = .18.$$

The Wald statistic is $W = .69/.18 = 3.84$ whose p-value is $\mathbb{P}(|Z| > 3.84) = .0001$, the same as the other tests. A 95 per cent confidence interval for γ is $\widehat{\gamma} \pm 2(.18) = (.33, 1.05)$. A 95 per cent confidence interval for ψ is $(e^{.33}, e^{1.05}) = (1.39, 2.86)$. ■

15.2 Two Discrete Variables

Now suppose that $Y \in \{1, \ldots, I\}$ and $Z \in \{1, \ldots, J\}$ are two discrete variables. The data can be represented as an $I \times J$ table of counts:

	$Y=1$	$Y=2$	$\cdots$	$Y=j$	$\cdots$	$Y=J$	
$Z=1$	X_{11}	X_{12}	$\cdots$	X_{1j}	$\cdots$	X_{1J}	$X_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$Z=i$	X_{i1}	X_{i2}	$\cdots$	X_{ij}	$\cdots$	X_{iJ}	$X_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$Z=I$	X_{I1}	X_{I2}	$\cdots$	X_{Ij}	$\cdots$	X_{IJ}	$X_{I\cdot}$
	$X_{\cdot 1}$	$X_{\cdot 2}$	$\cdots$	$X_{\cdot j}$	$\cdots$	$X_{\cdot J}$	n

where

$$X_{ij} = \text{number of observations for which } Z = i \text{ and } Y = j.$$

Consider testing

$$H_0 : Y \amalg Z \quad \text{versus} \quad H_1 : Y \not\amalg Z. \tag{15.11}$$

15.9 Theorem. *The likelihood ratio test statistic for (15.11) is*

$$T = 2\sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij} \log\left(\frac{X_{ij}\, X_{\cdot\cdot}}{X_{i\cdot}\, X_{\cdot j}}\right). \tag{15.12}$$

The limiting distribution of T under the null hypothesis of independence is χ^2_ν where $\nu = (I-1)(J-1)$. Pearson's χ^2 test statistic is

$$U = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(X_{ij} - E_{ij})^2}{E_{ij}}. \tag{15.13}$$

Asymptotically, under H_0, U has a χ^2_ν distribution where $\nu = (I-1)(J-1)$.

15.10 Example. These data are from Dunsmore et al. (1987). Patients with Hodgkins disease are classified by their response to treatment and by histological type.

Type	Positive Response	Partial Response	No Response	
LP	74	18	12	104
NS	68	16	12	96
MC	154	54	58	266
LD	18	10	44	72

The χ^2 test statistic is 75.89 with $2 \times 3 = 6$ degrees of freedom. The p-value is $\mathbb{P}(\chi^2_6 > 75.89) \approx 0$. The likelihood ratio test statistic is 68.30 with $2 \times 3 = 6$ degrees of freedom. The p-value is $\mathbb{P}(\chi^2_6 > 68.30) \approx 0$. Thus there is strong evidence that response to treatment and histological type are associated. ■

15.3 Two Continuous Variables

Now suppose that Y and Z are both continuous. If we assume that the joint distribution of Y and Z is bivariate Normal, then we measure the dependence between Y and Z by means of the correlation coefficient ρ. Tests, estimates, and confidence intervals for ρ in the Normal case are given in the previous chapter in Section 14.2. If we do not assume Normality then we can still use the methods in Section 14.2 to draw inferences about the correlation ρ. However, if we conclude that ρ is 0, we cannot conclude that Y and Z are independent, only that they are uncorrelated. Fortunately, the reverse direction is valid: if we conclude that Y and Z are correlated than we can conclude they are dependent.

15.4 One Continuous Variable and One Discrete

Suppose that $Y \in \{1, \ldots, I\}$ is discrete and Z is continuous. Let $F_i(z) = \mathbb{P}(Z \leq z | Y = i)$ denote the CDF of Z conditional on $Y = i$.

15.11 Theorem. *When $Y \in \{1, \ldots, I\}$ is discrete and Z is continuous, then $Y \amalg Z$ if and only if $F_1 = \cdots = F_I$.*

It follows from the previous theorem that to test for independence, we need to test

$$H_0 : F_1 = \cdots = F_I \quad \text{versus} \quad H_1 : \text{not } H_0.$$

For simplicity, we consider the case where $I = 2$. To test the null hypothesis that $F_1 = F_2$ we will use the **two sample Kolmogorov-Smirnov test**. Let n_1 denote the number of observations for which $Y_i = 1$ and let n_2 denote the number of observations for which $Y_i = 2$. Let

$$\widehat{F}_1(z) = \frac{1}{n_1} \sum_{i=1}^{n} I(Z_i \leq z) I(Y_i = 1)$$

and

$$\widehat{F}_2(z) = \frac{1}{n_2} \sum_{i=1}^{n} I(Z_i \leq z) I(Y_i = 2)$$

denote the empirical distribution function of Z given $Y = 1$ and $Y = 2$ respectively. Define the test statistic

$$D = \sup_x |\widehat{F}_1(x) - \widehat{F}_2(x)|.$$

15.12 Theorem. *Let*

$$H(t) = 1 - 2 \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2 t^2}. \tag{15.14}$$

Under the null hypothesis that $F_1 = F_2$,

$$\lim_{n \to \infty} \mathbb{P}\left(\sqrt{\frac{n_1 n_2}{n_1 + n_2}} D \leq t \right) = H(t).$$

It follows from the theorem that an approximate level α test is obtained by rejecting H_0 when

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}} D > H^{-1}(1 - \alpha).$$

15.5 Appendix

Interpreting The Odds Ratios. Suppose event A as probability $\mathbb{P}(A)$. The odds of A are defined as $\text{odds}(A) = \mathbb{P}(A)/(1 - \mathbb{P}(A))$. It follows that

$\mathbb{P}(A) = \text{odds}(A)/(1 + \text{odds}(A))$. Let E be the event that someone is exposed to something (smoking, radiation, etc) and let D be the event that they get a disease. The odds of getting the disease given that you are exposed are:

$$\text{odds}(D|E) = \frac{\mathbb{P}(D|E)}{1 - \mathbb{P}(D|E)}$$

and the odds of getting the disease given that you are not exposed are:

$$\text{odds}(D|E^c) = \frac{\mathbb{P}(D|E^c)}{1 - \mathbb{P}(D|E^c)}.$$

The *odds ratio* is defined to be

$$\psi = \frac{\text{odds}(D|E)}{\text{odds}(D|E^c)}.$$

If $\psi = 1$ then disease probability is the same for exposed and unexposed. This implies that these events are independent. Recall that the log-odds ratio is defined as $\gamma = \log(\psi)$. Independence corresponds to $\gamma = 0$.

Consider this table of probabilities and corresponding table of data:

	D^c	D	
E^c	p_{00}	p_{01}	$p_{0\cdot}$
E	p_{10}	p_{11}	$p_{1\cdot}$
	$p_{\cdot 0}$	$p_{\cdot 1}$	1

	D^c	D	
E^c	X_{00}	X_{01}	$X_{0\cdot}$
E	X_{10}	X_{11}	$X_{1\cdot}$
	$X_{\cdot 0}$	$X_{\cdot 1}$	$X_{\cdot\cdot}$

Now

$$\mathbb{P}(D|E) = \frac{p_{11}}{p_{10} + p_{11}} \quad \text{and} \quad \mathbb{P}(D|E^c) = \frac{p_{01}}{p_{00} + p_{01}},$$

and so

$$\text{odds}(D|E) = \frac{p_{11}}{p_{10}} \quad \text{and} \quad \text{odds}(D|E^c) = \frac{p_{01}}{p_{00}},$$

and therefore,

$$\psi = \frac{p_{11}p_{00}}{p_{01}p_{10}}.$$

To estimate the parameters, we have to first consider how the data were collected. There are three methods.

Multinomial Sampling. We draw a sample from the population and, for each person, record their exposure and disease status. In this case, $X = (X_{00}, X_{01}, X_{10}, X_{11}) \sim \text{Multinomial}(n, p)$. We then estimate the probabilities in the table by $\widehat{p}_{ij} = X_{ij}/n$ and

$$\widehat{\psi} = \frac{\widehat{p}_{11}\widehat{p}_{00}}{\widehat{p}_{01}\widehat{p}_{10}} = \frac{X_{11}X_{00}}{X_{01}X_{10}}.$$

Prospective Sampling. (Cohort Sampling). We get some exposed and unexposed people and count the number with disease in each group. Thus,

$$\begin{aligned} X_{01} &\sim \text{Binomial}(X_{0\cdot}, \mathbb{P}(D|E^c)) \\ X_{11} &\sim \text{Binomial}(X_{1\cdot}, \mathbb{P}(D|E)). \end{aligned}$$

We should really write $x_{0\cdot}$ and $x_{1\cdot}$ instead of $X_{0\cdot}$ and $X_{1\cdot}$ since in this case, these are fixed not random, but for notational simplicity I'll keep using capital letters. We can estimate $\mathbb{P}(D|E)$ and $\mathbb{P}(D|E^c)$ but we cannot estimate all the probabilities in the table. Still, we can estimate ψ since ψ is a function of $\mathbb{P}(D|E)$ and $\mathbb{P}(D|E^c)$. Now

$$\widehat{\mathbb{P}}(D|E) = \frac{X_{11}}{X_{1\cdot}} \quad \text{and} \quad \widehat{\mathbb{P}}(D|E^c) = \frac{X_{01}}{X_{0\cdot}}.$$

Thus,

$$\widehat{\psi} = \frac{X_{11}X_{00}}{X_{01}X_{10}}$$

just as before.

Case-Control (Retrospective) Sampling. Here we get some diseased and non-diseased people and we observe how many are exposed. This is much more efficient if the disease is rare. Hence,

$$\begin{aligned} X_{10} &\sim \text{Binomial}(X_{\cdot 0}, \mathbb{P}(E|D^c)) \\ X_{11} &\sim \text{Binomial}(X_{\cdot 1}, \mathbb{P}(E|D)). \end{aligned}$$

From these data we can estimate $\mathbb{P}(E|D)$ and $\mathbb{P}(E|D^c)$. Surprisingly, we can also still estimate ψ. To understand why, note that

$$\mathbb{P}(E|D) = \frac{p_{11}}{p_{01}+p_{11}}, \quad 1-\mathbb{P}(E|D) = \frac{p_{01}}{p_{01}+p_{11}}, \quad \text{odds}(E|D) = \frac{p_{11}}{p_{01}}.$$

By a similar argument,

$$\text{odds}(E|D^c) = \frac{p_{10}}{p_{00}}.$$

Hence,

$$\frac{\text{odds}(E|D)}{\text{odds}(E|D^c)} = \frac{p_{11}p_{00}}{p_{01}p_{10}} = \psi.$$

From the data, we form the following estimates:

$$\widehat{P}(E|D) = \frac{X_{11}}{X_{\cdot 1}}, \quad 1-\widehat{P}(E|D) = \frac{X_{01}}{X_{\cdot 1}}, \quad \widehat{\text{odds}}(E|D) = \frac{X_{11}}{X_{01}}, \quad \widehat{\text{odds}}(E|D^c) = \frac{X_{10}}{X_{00}}.$$

Therefore,

$$\widehat{\psi} = \frac{X_{00}X_{11}}{X_{01}X_{10}}.$$

So in all three data collection methods, the estimate of ψ turns out to be the same.

It is tempting to try to estimate $\mathbb{P}(D|E) - \mathbb{P}(D|E^c)$. In a case-control design, this quantity is not estimable. To see this, we apply Bayes' theorem to get

$$\mathbb{P}(D|E) - \mathbb{P}(D|E^c) = \frac{\mathbb{P}(E|D)\mathbb{P}(D)}{\mathbb{P}(E)} - \frac{\mathbb{P}(E^c|D)\mathbb{P}(D)}{\mathbb{P}(E^c)}.$$

Because of the way we obtained the data, $\mathbb{P}(D)$ is not estimable from the data. However, we can estimate $\xi = \mathbb{P}(D|E)/\mathbb{P}(D|E^c)$, which is called the **relative risk**, under the **rare disease assumption.**

15.13 Theorem. *Let* $\xi = \mathbb{P}(D|E)/\mathbb{P}(D|E^c)$. *Then*

$$\frac{\psi}{\xi} \to 1$$

as $\mathbb{P}(D) \to 0$.

Thus, under the rare disease assumption, the relative risk is approximately the same as the odds ratio and, as we have seen, we can estimate the odds ratio.

15.6 Exercises

1. Prove Theorem 15.2.
2. Prove Theorem 15.3.
3. Prove Theorem 15.6.
4. The *New York Times* (January 8, 2003, page A12) reported the following data on death sentencing and race, from a study in Maryland: [2]

	Death Sentence	No Death Sentence
Black Victim	14	641
White Victim	62	594

Analyze the data using the tools from this chapter. Interpret the results. Explain why, based only on this information, you can't make causal conclusions. (The authors of the study did use much more information in their full report.)

[2] The data here are an approximate re-creation using the information in the article.

5. Analyze the data on the variables Age and Financial Status from:

   http://lib.stat.cmu.edu/DASL/Datafiles/montanadat.html

6. Estimate the correlation between temperature and latitude using the data from

   http://lib.stat.cmu.edu/DASL/Datafiles/USTemperatures.html

   Use the correlation coefficient. Provide estimates, tests, and confidence intervals.

7. Test whether calcium intake and drop in blood pressure are associated. Use the data in

   http://lib.stat.cmu.edu/DASL/Datafiles/Calcium.html