Lecture 9:

Inference About Independence

Inference About Independence

In this lecture, 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 them?

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 will be the subject of the next lecture.

We'll write: $Y \perp Z$ = Y and Z are independent (\perp in LaTeX, compare it to $\amalg$ = \amalg from the book) and $Y \asymp Z$ = Y and Z are dependent (\asymp symbol, different from the coil in the book)

Two Binary Variables

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 it as a 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} = # of observations for which $Y = i$ and $Z = j$

Dotted subscripts denote sums: $X_{i\cdot} = \sum_j X_{ij}, \;\; X_{\cdot j} = \sum_i X_{ij}$

and $n = X_{\cdot\cdot} = \sum_{i,j} X_{ij}$

Two Binary Variables

Denote the corresponding probabilities $p_{ij} = \mathbb{P}(Z = i, Y = j)$

	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

Let $X = (X_{00}, X_{01}, X_{10}, X_{11})$ denote the vector of counts. Then $X \sim$ Multinomial(n, p) where $p = (p_{00}, p_{01}, p_{10}, p_{11})$. Introduce

Definition: The **odds ratio** is $\psi = \dfrac{p_{00}p_{11}}{p_{01}p_{10}}$ and the **log odds ratio** is $\gamma = \log \psi$.

Two Binary Variables

Theorem: The following statements are equivalent:

$$Y \perp Z \quad \text{and} \quad \psi = 1 \quad \text{and} \quad \gamma = 1 \quad \text{and} \quad p_{ij} = p_{i\cdot} p_{\cdot j}$$

Now consider testing $H_0 : Y \perp Z$ versus $H_1 : Y \asymp Z$.

First consider the likelihood ratio test:

Under H_1, we have $X \sim$ Multinomial(n, p) and the MLE is $\widehat{p} = X/n$.

Under H_0, we again have $X \sim$ Multinomial(n, p) but the restricted MLE is computed under $p_{ij} = p_{i\cdot} p_{\cdot j}$ – a constraint.

Two Binary Variables

So:

Theorem: The likelihood ratio test statistic for the above is

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

under H_0, $T \xrightarrow{d} \chi^2_1$. So, the approx. level-$\alpha$ test is obtained by rejecting H_0 when $T > \chi^2_{1,\alpha}$

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

Two Binary Variables

Theorem: Pearson's χ^2 test statistic for independence is

$$U = \sum_{i=0}^{1} \sum_{j=0}^{1} \frac{(X_{ij} - E_{ij})^2}{E_{ij}} \quad \text{where } E_{ij} = \frac{X_{i\cdot} X_{\cdot j}}{n}$$

Under H_0, $U \xrightarrow{d} \chi^2_1$. So, an approx. level-$\alpha$ test is obtained by rejecting H_0 when $U > \chi^2_{1,\alpha}$

Some intuition this: Under H_0, $p_{ij} = p_{i\cdot} p_{\cdot j}$, so the MLE of p_{ij} under H_0 is $\widehat{p}_{ij} = \widehat{p}_{i\cdot} \widehat{p}_{\cdot j} = \frac{X_{i\cdot}}{n} \frac{X_{\cdot j}}{n}$. So the expected # of observations in (i, j)-th cell is $E_{ij} = n \widehat{p}_{ij}$, and U compares observed to expected.

Two Binary Variables

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

Theorem: The MLEs of ψ and γ are $\widehat{\psi} = \dfrac{X_{00}X_{11}}{X_{01}X_{10}}$ and $\hat{\gamma} = \log \widehat{\psi}$.

The asymptotic (delta-method) standard errors are

$$\hat{\text{se}}(\hat{\gamma}) = \sqrt{\frac{1}{X_{00}} + \frac{1}{X_{01}} + \frac{1}{X_{10}} + \frac{1}{X_{11}}} \text{ and } \hat{\text{se}}(\widehat{\psi}) = \widehat{\psi}\,\hat{\text{se}}(\hat{\gamma})$$

Remark: For small sample sizes, $\widehat{\psi}$ and $\hat{\gamma}$ can have very large variance, so a modified estimator $\widehat{\psi} = \dfrac{(X_{00}+1/2)(X_{11}+1/2)}{(X_{01}+1/2)(X_{10}+1/2)}$ can be used.

Two Binary Variables

Another test for independence is the Wald test for $\gamma = 0$ given by

$$W = (\widehat{\gamma} - 0)/\widehat{\text{se}}(\widehat{\gamma}).$$

A $(1-\alpha)$-confidence interval for ψ can be obtained in two ways:

1) Use $\widehat{\psi} \pm z_{\alpha/2}\, \widehat{\text{se}}(\widehat{\psi})$

2) Since $\psi = e^{\gamma}$, use $\exp\left(\widehat{\gamma} \pm z_{\alpha/2}\, \widehat{\text{se}}(\widehat{\gamma})\right)$

Second method is usually more accurate.

Two Discrete Variables

Two Discrete Variables

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

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

Consider testing $H_0 : Y \perp Z$ versus $H_1 : Y \asymp Z$

Theorem: The likelihood ratio test statistic for this is

$T = 2\sum_{i,j} X_{ij} \log\left(\frac{X_{ij}X_{..}}{X_{i.}X_{.j}}\right)$, the limiting distribution of T under the null hyp. (of independence) is χ^2_ν where $\nu = (I-1)(J-1)$.

Two Discrete Variables

Accordingly,

Theorem: Pearson's χ^2 test statistic is

$$U = \sum_{i,j} \frac{(X_{ij} - E_{ij})^2}{E_{ij}}$$

which, asymptotically, under H_0, has χ^2_ν distribution with $\nu = (I-1)(J-1)$

Two Continuous Variables

Two Continuous Variables

Suppose Y and Z are both continuous. If we assume the joint distribution of (Y, Z) is bivariate Normal, then the measure of dependence is the correlation coeff. $\rho = \dfrac{\mathbb{E}((Y-\mu_Y)(Z-\mu_Z))}{\sigma_Y \sigma_Z}$.

For testing independence, one has compute the confidence interval for ρ: for that there's 1) Delta-method, 2) a method due to Fischer:

1. Compute $\hat{\rho}$ – sample correlation coeff. From it, compute

$$\hat{\theta} = \frac{1}{2}\left(\log(1+\hat{\rho}) - \log(1-\hat{\rho})\right)$$

(denote this function $f(x) = \dfrac{1}{2}\log\dfrac{1+x}{1-x}$, then $f^{-1}(x) = \dfrac{e^{2x}-1}{e^{2x}+1}$)

Two Continuous Variables

2. The approx. standard error of $\widehat{\theta}$ is

$$\hat{\text{se}}(\widehat{\theta}) = \frac{1}{\sqrt{n-3}} \text{ (where } n \text{ is the sample size)}$$

3. An approx. $(1-\alpha)$-confidence interval for θ is

$$(a, b) = \left(\widehat{\theta} - \frac{z_{\alpha/2}}{\sqrt{n-3}}, \widehat{\theta} + \frac{z_{\alpha/2}}{\sqrt{n-3}} \right)$$

4. Confidence interval for ρ is $\left(f^{-1}(a), f^{-1}(b)\right)$

Two Continuous Variables

If we do not assume Normality, we can still use this to test correlation.

However, if we conclude that ρ is 0, we can't conclude that Y and Z are independent – only that they are uncorrelated.

Fortunately, the other direction is valid: if Y and Z are correlated, we can conclude they are dependent!

One Continuous, One Discrete

One Continuous, One Discrete

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

Theorem: If so, then $Y \perp Z$ iff (= if and only if) $F_1 = \ldots = F_I$!

So, if we need to test for independence, we test

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

For simplicity, let's consider the case $I = 2$. To test the null hypothesis that $F_1 = F_2$ we'll use **two sample Kolmogorov-Smirnov test**.

One Continuous, One Discrete

Let n_1 = (# of observations when $Y = 1$) and n_2 – same for $Y = 2$.

Let $\widehat{F}_1(z) = \dfrac{1}{n_1} \sum_{i=1}^{n} I(Z_i \leq z)\, I(Y_i = 1)$ and $\widehat{F}_2(z)$ resp.

The test statistic is $D = \sup_x |\widehat{F}_1(x) - \widehat{F}_2(x)|$

Theorem: Let $H(t) = 1 - 2\sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2t^2}$. Under the null hypothesis that $F_1 = F_2$, $\lim_{n\to\infty} \mathbb{P}\left(\sqrt{\dfrac{n_1 n_2}{n_1 + n_2}} D \leq t\right) = H(t)$.

So a level-α test is: reject H_0 if $\sqrt{n_1 n_2/(n_1 + n_2)}\, D > H^{-1}(1 - \alpha)$