Lecture #17

Recall: • A discrete r.v. pair (X, Y) are independent $\Leftrightarrow$ $\mathbb{P}((X, Y) = (x, y))$ (★)

$= \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y) \quad \forall x, y$

We can also write this as $\mathbb{P}(X = x \cap Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y)$.

Via the defining formula, we can make an equivalent definition in terms of conditional distributions: $\mathbb{P}(X = x \mid Y = y) = \dfrac{\mathbb{P}(X = x \cap Y = y)}{\mathbb{P}(Y = y)}$

(★) ↘

$= \dfrac{\mathbb{P}(X = x)\,\mathbb{P}(Y = y)}{\mathbb{P}(Y = y)}$

$= \mathbb{P}(X = x), \quad \forall x, y.$

• When working with continuous r.v. (X, Y), we formulate (★) in terms of the joint density: $\rho_{X,Y}(x, y) = \rho_X(x)\,\rho_Y(y)$

Q: Given observations $\{(x_i, y_i)\}_{i=1}^{n}$ from the joint distribution (X, Y), how can we assess the independence of X and Y?

• We shall build up a testing framework by starting with simple discrete r.v. and working our way up.

• **Notation:** • $X \perp\!\!\!\perp Y$ means X, Y are independent.

• $X \not\perp\!\!\!\perp Y$ means not $X \perp\!\!\!\perp Y$, i.e. X and Y are dependent (at least somewhat).

(X, Y) binary: Suppose $X = \begin{cases} 0, \text{ prob} = p_0 \\ 1, \text{ prob} = p_1 \end{cases}$, $Y = \begin{cases} 0, \text{ prob} = q_0 \\ 1, \text{ prob} = q_1 \end{cases}$

We don't know any of the parameters (p_0, p_1, q_0, q_1) nor whether X, Y are independent. We observe $\{(x_i, y_i)\}_{i=1}^n$ from the joint distribution (X, Y).

Let us write N_{kj} = # observations of the form (k, j). We can build a table recording the results as follows:

	$X=0$	$X=1$	
$Y=0$	N_{00}	N_{10}	$N_{\cdot 0}$
$Y=1$	N_{01}	N_{11}	$N_{\cdot 1}$
	$N_{0\cdot}$	$N_{1\cdot}$	n

where $N_{\cdot j}$ = # observations where $Y = j$

and similarly for $N_{k\cdot}$.

This table of frequencies allows us to estimate p_0, p_1, q_0, q_1 as

$\hat{p}_j \approx \frac{N_{j\cdot}}{n}$ and similarly for $\hat{q}_k$.

Importantly, we can also estimate $r_{jk} = \mathbb{P}(X = j, Y = k)$ as

$\hat{r}_{jk} = \frac{N_{jk}}{n}$.

This yields a table of estimates

	$X=0$	$X=1$	
$Y=0$	$\hat{r}_{00}$	$\hat{r}_{10}$	$\hat{q}_0$
$Y=1$	$\hat{r}_{01}$	$\hat{r}_{11}$	$\hat{q}_1$
	$\hat{p}_0$	$\hat{p}_1$	

- Now, given perfect knowledge of the underlying parameters, we can develop a quick test through the so-called odds ratio.

Defn: For [redacted] $\{r_{jk}\}_{j,k=0}^{1}$ as above, (i) Let $\psi = \frac{r_{00}\, r_{11}}{r_{01}\, r_{10}}$ be the odds ratio

(ii) Let $\gamma = \log(\psi) = \log\left(\frac{r_{00}\, r_{11}}{r_{01}\, r_{10}}\right)$ be the log odds ratio.

In our case, ψ and γ are cute tools, but they generalize in a useful way to more complicated settings.

Theorem (Independence of Binary r.v.): Let (X,Y) be as above. Then the following are equivalent: (i) $X \perp\!\!\!\perp Y$

(ii) $\psi = 1$

(iii) [redacted] $\gamma = 0$

(iv) [redacted] $r_{jk} = p_j q_k \quad \forall (j,k) \in \{0,1\}^2$.

Proof: (HW), just follow the definitions.

- Of course, the challenge is to address the fact that $\{r_{jk}\}_{j,k=0}^{1}$ must be estimated by $\{\hat{r}_{jk}\}_{j,k=0}^{1}$.

- Let $H_0 : (X, Y)$ are independent $(X \perp\!\!\!\perp Y)$

  $H_1 : (X, Y)$ are not independent $(X \not\perp\!\!\!\perp Y)$.

Of course, if we compute something like an "empirical" odds ratio, we will get that (X, Y) appear (weakly) dependent, even if H_0 holds. So, we need to know how much deviation from $\psi = 0$ is reasonable under the null.

Theorem (Pearson χ^2 Test for Independence): Let $U = \sum_{k,j=0}^{1} \frac{[N_{jk} - E_{jk}]^2}{E_{jk}}$

where $E_{jk} = \frac{(N_{j\cdot})\cdot(N_{\cdot k})}{n}$. Then under $H_0 : X \perp\!\!\!\perp Y$, the statistic U converges in distribution to a χ^2 r.v. with 1 d.o.f.

Proof: Exercise ... model using the proof of connection between multinomial and χ^2 distributions.

ex: Suppose we observe the following values:

	Smoker	Non Smoker
Lung C.	11	3
No Lung C.	105	141

$N_{00} = 11$ $N_{01} = 3$

$N_{10} = 105$ $N_{11} = 141$

Then $n = 260$

$N_{0\cdot} = 14$

$N_{1\cdot} = 246$

$N_{\cdot 0} = 116$

$N_{\cdot 1} = 144$

H_0: Smoking $\perp\!\!\!\perp$ lung cancer

H_1: Smoking $\not\perp\!\!\!\perp$ lung cancer

So, our test statistic is

$$U = \frac{\left[11 - \frac{14 \cdot 116}{260}\right]^2}{\frac{14 \cdot 116}{260}} + \frac{\left[105 - \frac{246 \cdot 116}{260}\right]^2}{\frac{246 \cdot 116}{260}} + \frac{\left[3 - \frac{14 \cdot 144}{260}\right]^2}{\frac{14 \cdot 144}{260}} + \frac{\left[141 - \frac{246 \cdot 144}{260}\right]^2}{\frac{246 \cdot 144}{260}}$$

$= 6.9044$

Compared to $\chi^2_{df,\alpha}$ for 1 deg, $\alpha = .05$, we reject H_0 if

$6.9044 > \chi^2_{DC}$

$= 3.8415$ ✓