MIE237

Neil Montgomery
2016-01-26

A taste of categorical data 9.10, 10.13

Testing independence - more motivation

100 people (50 male and 50 female) are asked a question about whether they agree or disagree with some political statement. 80 people agreed. 20 people disagreed. Consider the following possible outcomes:

Table 1	Male	Female	Total
Agree	40	40	80
Disagree	10	10	20
Total	50	50	100

Table 2	Male	Female	Total
Agree	50	30	80
Disagree	0	20	20
Total	50	50	100

Table 3	Male	Female	Total
Agree	39	41	80
Disagree	11	9	20
Total	50	50	100

Testing independence

- Treat marginal totals as fixed
- Compute E_i assuming independence
- With O_i as use the following test statistic:

$$\sum_i \frac{(O_i - E_i)^2}{E_i} \sim^{\text{approx}} \chi^2_{(r-1)(c-1)}$$

- Approximation is good if $E_i \geq 5$ for all i.

"more motivation" table calculations

The table of expected cell counts is:

Expected Cell Counts	Male	Female	Total
Agree	40	40	80
Disagree	10	10	20
Total	50	50	100

p-values

Table 1: $\sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = 0$ and the p-value is $P\left(\chi_1^2 \geq 0\right) = 1$.

Table 2: $\sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = 10^2/40 + 10^2/40 + 10^2/10 + 10^2/10 = 25$ and the p-value is $P\left(\chi_1^2 \geq 25\right) = 0$

Table 3: $\sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = 1^2/40 + 1^2/40 + 1^2/10 + 1^2/10 = 0.25$ and the p-value is $P\left(\chi_1^2 \geq 0.25\right) = 0.617$

Table 4: $\sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = 4^2/40 + 4^2/40 + 4^2/10 + 4^2/10 = 4$ and the p-value is $P\left(\chi_1^2 \geq 4\right) = 0.046$

Gas pipe data

Is leak status independent of pipe size? Here is the table summary of the data (leak status in the rows; pipe size in the columns:

	1	1.5	1.75	Sum
No	329	249	224	802
Yes	93	52	53	198
Sum	422	301	277	1000

The expected cell counts:

	1	1.5	1.75	Sum
No	338.4	241.4	222.2	802
Yes	83.6	59.6	54.8	198
Sum	422.0	301.0	277.0	1000

All the E_i easily exceed 5, so the approximation will be good.

Gas pipe data

The results:

```
chisq.test(pipeline$Leak, pipeline$Size)
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  pipeline$Leak and pipeline$Size
## X-squared = 2.6162, df = 2, p-value = 0.2703
```

Not surprising since the columns in the simulated data really were simulated independently.

Regression

Linear models

- Model: Output = Input + Noise
- Linear model: Output = Linear function of inputs + Noise
- Examples (all with $\varepsilon_i \sim N(0, \sigma^2)$):
- $Y_i = \mu + \varepsilon_i$
- $Y_{ij} = \mu_i + \varepsilon_{ij}, \quad i \in \{1, 2\}$
- $Y_{ij} = \mu_i + \varepsilon_{ij}, \quad i \in \{1, \ldots, k\}$
- $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad x_i \in \{0, 1\}$
- $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad x_i \in \mathbb{R}$
- $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \quad x_{ji} \in \mathbb{R}$

Linear models in matrix form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Linear models in matrix form

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}$$

"Simple" linear regression

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$\varepsilon \sim N(0, \sigma^2)$$

β_1 is the "slope" parameter