Categorical Data

Data such that each observation can be classified as belonging to one of a finite number of categories.

Nonparametric Problem / Methods

N.P. problem: problem in which possible distr of observation are not restricted to a specific parametric family

N.P. method: statistical methods applicable to N.P. problems.

χ^2 Test for Goodness-of-fit

Suppose observation has k categories. Let p_i be the probability that a sample belongs to category i.

H_0 : $p_i = p_i^0$ for $i = 1 \dots k$

H_1 : $p_i \neq p_i^0$ for some i

Suppose n samples are taken, N_i being number of category i.

χ^2 test statistic: $Q = \sum_{i=1}^{k} \frac{(N_i - n p_i^0)^2}{n p_i^0}$

- If H_0 is true and $n \to \infty$, Q converge in distr to χ^2_{k-1}.

Test procedure: reject H_0 if $Q \geq c$

Choose c given α_0: $c = 1 - \alpha_0$ quantile of χ^2_{k-1}

χ^2 test for continuous distr can be done by discretization into discrete distr.

χ^2 for Composite Hypo

Suppose for a vector of params $\vec{\theta} = (\theta_1 \cdots \theta_s)$, $\exists$ function $\pi_i, \forall i$

H_0: $\exists\, \vec{\theta} \in \Omega$, s.t. $p_i = \pi_i(\theta)$ $\qquad$ H_1: o/w

Interpretation: $\vec{\theta}$ is an encoding of prob distribution over categories.

π maps encoding to prob distribution. H_0: Ω contains true prob distr.

Test statistics:

First find $\hat{\theta}$, the MLE of θ given observations.

$$Q = \sum_{i=1}^{k} \frac{[N_i - n\pi_i(\hat{\theta})]}{n\,\pi_i(\hat{\theta})}$$

- If H_0 is true, $n \to \infty$, $Q \to \chi^2_{k-1-s}$

Test procedure: reject H_0 if $Q \geq c$.

Testing Whether a Distr is Normal

Consider using χ^2 test on discretized distr w/ interval (a_i, b_i)

$\theta = (\mu, \sigma^2)$. $\quad \pi_i = \Phi\left(\frac{b_i - \mu}{\sigma}\right) - \Phi\left(\frac{a_i - \mu}{\sigma}\right)$

Use the χ^2 test above.

χ^2 Test of Independence

Contingency table: table in which observations are classified in ≥ 2 ways.

E.g.,

	ill	not ill
smoke	a	b
not smoke	c	d

Parameters: p_{ij} : true possibility of a grid. N_{ij}: # observation

$p_{i+} = \sum_j p_{ij}$ $\quad p_{+j} = \sum_i p_{ij}$

H_0: $p_{ij} = p_{i+} p_{+j}$, $\forall i,j$ (independence) $\quad H_1$: o/w

Test statistic:

Find $\hat{E}_{ij}$ which is MLE of N_{ij} if H_0 is true.

$$\hat{E}_{ij} = n \cdot \frac{N_{i+}}{n} \cdot \frac{N_{+j}}{n} = \frac{N_{i+} N_{+j}}{n}$$

$$Q = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{N_{ij} - \hat{E}_{ij}}{\hat{E}_{ij}}$$

- If H_0 is true, $n \to \infty$, $Q \to \chi^2$ w/ $(R-1)(C-1)$ dof.

Test of Homogeneity

Test if distr of observed RV is same among multiple populations

Suppose there are R populations, C categories.

H_0: $p_{1j} = p_{2j} = \cdots = p_{Rj}$ for $j = 1 \cdots c$.

Test statistic:

Suppose we know p_{ij}, then for population i,
we can use χ test for goodness of fit to test $p_{i1} = \cdots = p_{ic}$.
The χ^2 statistic used:

$$\sum_{j=1}^{c} \frac{(N_{ij} - N_{i+}p_{ij})^2}{N_{i+}\,p_{ij}}$$

Sum this statistic over all population gives a statistic that tests H_0:

$$Q = \sum_{i=1}^{R} \sum_{j=1}^{c} \frac{(N_{ij} - N_{i+}p_{ij})^2}{N_{i+}\,p_{ij}}$$

- For large samples, $Q \to \chi^2_{R(c-1)}$

However, p_{ij} is unknown. But we can use MLE as substitute:

$$\hat{p}_{ij} = \frac{N_{+j}}{n}$$ (MLE assuming H_0)

Now, let $\hat{E}_{ij} = \frac{N_{i+}N_{+j}}{n}$,

$$Q = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(N_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$$

- H_0, $n \to \infty \Rightarrow Q \to \chi^2_{(R-1)(C-1)}$

Note that the form of Q is very similar to χ^2 of independence.