Chi-squared Test (1)

- H_0: population follows distribution
- Divide the observations into k bins
- O_i = observed frequency in i-th bin
- E_i = expected frequency in i-th bin
- Test statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Chi-squared Test (2)

- χ^2 follows (approx) a chi-square distribution with $k\text{-}p\text{-}1$ degrees of freedom where p is the number of estimated parameters
- Reject H_0 if $\chi^2 > \chi^2_{\alpha,k-p-1}$.

How many categories?

- Expected frequency in each bin should be at least 3. Can be as small as 1 or 2 if other bins are at least 5. If number is too small, lump with adjacent bin.

Example: Uniform Birthdays

- H_0: birthdays follow a uniform distribution
- Construct a chi-squared test for this hypothesis

Contingency Table Test

- Goal: to find out if two methods of classification are statistically indept

$$\hat{p}_i = \frac{1}{n}\sum_{i=1}^{r} O_{ij}; \hat{q}_j = \frac{1}{n}\sum_{j=1}^{c} O_{ij}$$

$$E_{ij} = n\hat{p}_i\hat{q}_j$$

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

• χ^2 follows a chi-squared distribution with *(r-1)(c-1)* degrees of freedom.

• Reject the null hypothesis if χ^2 exceeds $\chi^2_{\alpha,(r-1)(c-1)}$

Example: Smoking & Gender

- Test the hypothesis (at the 1% level) that whether a person smokes is independent of the person's gender.