This is hypothesis testing for categorical data

Ex. A wildlife conservationist is interested in assessing the distribution of geese nesting sites throughout the KW region.

The conservationist locates 80 different nesting sites and wants to determine whether they're evenly distributed throughout the region.

Location	Observed Frequency	Expected Frequency
North	21	20
South	18	20
East	24	20
West	17	20

▶ **Question**: how might we model these data?

Let Y_j = number of nesting sites in location j

Let θ_j = probability that some nesting site is in location j

▶ We have

$$P(Y_1 = y_1, \ldots, Y_4 = y_4) = \frac{n!}{y_1! \ldots y_4!} \theta_1^{y_1} \ldots \theta_4^{y_4}$$

with

$$\sum_{j=1}^{4} \theta_j = 1,\ 0 < \theta_j < 1$$

and

$$\sum_{j=1}^{4} y_j = n,\ y_j = 0, 1, \ldots$$

Since we have 4 nesting sites, a good null hypothesis would be $\theta_j = 0.25$ for all j.

What if we want to test this hypothesis?

Using the *likelihood ratio test statistic*:

$$L(\theta_1, \theta_2, \theta_3, \theta_4) = \theta_1^{y_1} \cdot \theta_2^{y_2} \cdot \theta_3^{y_3} \cdot \theta_4^{y_4}$$

We want to maximize this over the constraint $\theta_j = 1$ for all j, which, using Lagrange multipliers (no need to show this in STAT 231) gives us $\hat{\theta}_j = \frac{y_j}{n}$

- Note that, rather than estimating k parameters θ_1 through θ_k, we only need estimate $k-1$ parameters since $\sum_{j=1}^{k} \theta_j = 1$
- Suppose that the $theta_j$'s are related in some way, i.e. can each be expressed in terms of unknown parameter α:

$$H_0 : \theta_j = \theta_j(\alpha),\ j = 1, ..., k$$

where $\alpha = (\alpha_1, ..., \alpha_p)$, $\underline{p < k-1}$

: $k = 4$

In the above example: $H_0 = \theta_1 = \theta_2 = \theta_3 = \theta_4 = 0.25$, so no need to work with the alphas

---

Aside: Some other form of null hypothesis would be

$$\theta_1 = \alpha_1, \quad \theta_2 = \alpha_1 + \alpha_2, \quad \theta_3 = \alpha_1 + 2\alpha_2, \quad \theta_4 = 2 - \alpha_2, \quad \alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$$

$$\Rightarrow \Lambda(\theta_0) = -2\log(R(\theta_0)) = -2\log\left[\prod_{j=1}^{k} \left(\frac{E_j}{Y_j}\right)^{Y_j}\right]$$

$$\Rightarrow E_j = n \cdot \underbrace{\theta_j(\tilde{\alpha})}_{\text{function}}$$: expected value - function of α

- Using some rules of logarithms, can further simplify this to

$$\Lambda(\theta_0) = 2\sum_{j=1}^{k} Y_j \log\left(\frac{Y_j}{E_j}\right)$$

which, for a given dataset, will have an observed value of

$$\lambda(\theta_0) = 2\sum_{j=1}^{k} y_j \log\left(\frac{y_j}{e_j}\right),\ e_j = n\theta_j(\hat{\alpha})$$

(goose example)

: $2\sum_{j=1}^{k} y_j \log\left(\frac{y_j}{20}\right) = 1.482$

$p = P(W \geq 1.482)$

- In the context of a multinomial problem, if n is large and H_0 is true, then

$$\Lambda(\theta_0) = 2\sum_{j=1}^{k} Y_j \log\left(\frac{Y_j}{E_j}\right) \sim \chi^2(k-1-p)$$

Steps:

1. Get $\theta_1, \ldots, \theta_k$ (possibly using α's)
2. Calculate $e_j = n \cdot \theta_j(\alpha)\ \forall j$
3. Calculate sum $2\sum_{j=1}^{k} y_j \log\left(\frac{y_j}{e_j}\right) = w$, $p = P(W \geq w)$
   $W \sim \chi^2(k-1-p)$

Ex.

- Suppose individuals in a population can have their blood type classified as MM, MN, or NN
- Let Y_1 = number of MM types observed, Y_2 = number of MN types observed, and Y_3 = number of NN types observed
  - Respective proportions: $\theta_1, \theta_2, \theta_3$
  - $\sum_{j=1}^{k=3} \theta_j = 1$
- The joint probability function of Y_1, Y_2, Y_3 is Multinomial$(n; \theta_1, \theta_2, \theta_3)$ with $k = 3$

- In genetic theory, the θ_j's can be expressed in terms of a single parameter α:

$$H_0 : \theta_1 = \alpha^2,\ \theta_2 = 2\alpha(1-\alpha),\ \theta_3 = (1-\alpha)^2$$

- Suppose data on 100 persons gave $y_1 = 20$, $y_2 = 43$, and $y_3 = 37$
- Our likelihood function, in terms of α, is then

$$L_1(\alpha) = L(\theta_1(\alpha), \theta_2(\alpha), \theta_3(\alpha)) \propto (\alpha^2)^{20}(2\alpha(1-\alpha))^{43}((1-\alpha)^2)^{37}$$

$\Rightarrow L_1(\alpha) \propto \alpha^{83}(1-\alpha)^{117}$ → $43 + 2(37)$

↳ $83 = 2(20) + 43$

↳ from $(2\alpha)^{43}$

$\Rightarrow \hat{\alpha} = 0.415$

Note: do not truncate values

$e_j = n\theta_j \rightarrow e_1 = n\hat{\alpha}^2$

$e_2 = n \cdot 2\hat{\alpha}(1-\hat{\alpha})$

$e_3 = n \cdot (1-\hat{\alpha})^2$

sub in $\hat{\alpha} = 0.415$, $n = 100$

Goodness of Fit Test

Checks whether a certain distribution fits a dataset

Ex. Suppose we want to check whether the number of goals scored by the Toronto Maple Leafs follows a Poisson distribution.

- Consider the following (hypothetical) data from a season of 82 games:

Goals	0	1	2	3	4	5	6	≥ 7
Games	2	17	21	18	15	7	1	1

- Our data are multinomial in nature
  - 82 events, each event fits into one of 8 categories
- Let θ_0 = the probability of a game having 0 goals, θ_1 = the probability of a game having 1 goal, etc.
- If we let Y_j = the number of games from a sample of n, then

$$Y_j \sim \text{Multinomial}(\theta_0, \theta_1, \ldots, \theta_7)$$

Then: $H_0 : \theta_j = \dfrac{\theta^j e^{-\theta}}{j!}$ and $\hat{\theta} = \hat{\alpha} = 2.695$: Poisson → MLE is $\bar{y}$

$\Rightarrow$ Expected goal counts: $e_j = (n)\left[\dfrac{\hat{\theta}^j e^{-\hat{\theta}}}{j!}\right]$

If we calculate the expected counts for each j, the values of e6 and e7 are below 5.

As such, bin those columns into column e5, and redo the computation from there

Pearson Goodness of Fit Statistic

- An alternative test statistic for multinomial data is the Pearson goodness of fit test statistic:

$$D = \sum_{j=1}^{k} \frac{(Y_j - E_j)^2}{E_j} \sim \chi^2(k-1-p)$$

with observed value

$$d = \sum_{j=1}^{k} \frac{(y_j - e_j)^2}{e_j} \sim \chi^2(k-1-p)$$

and p-value

$$P(D \geq d),\ D \sim \chi^2(k-1-p)$$