Lecture #15

- Last time, we considered MLE for multinomial distributions.
- That is, we consider $\vec{X} \sim \text{Multi}(p_1, \dots, p_k; n)$ for unknown parameters $(p_1, \dots, p_k)$.

This means $\vec{X}$ outputs a random vector $(x_1, \dots, x_k)$, where each x_i is a non-negative integer and $\sum_{i=1}^{k} x_i = n$, for a fixed n. The distribution function is given by $\mathbb{P}(\vec{X} = (x_1, \dots, x_k)) = \frac{n!}{x_1! \cdots x_k!} \cdot \prod_{i=1}^{k} p_i^{x_i}$.

Remarks:
- This generalizes the binomial distribution, which is exactly when $k=2$.
- We think of n as fixed and known to us, while the probability vector $(p_1, \dots, p_k)$ is the unknown parameter(s) to estimate.

- So, suppose we observe $X = (x_1, \dots, x_k)$. We saw that solving the constrained likelihood maximization yields the (very reasonable) MLE estimates

$$(\hat{p}_1, \dots, \hat{p}_k) = \left(\frac{x_1}{n}, \dots, \frac{x_k}{n}\right).$$

Q: How can we do a hypothesis test on multinomial distributions? Well, we need to be able to say something about how $\hat{p}_i - p_i$ behaves.

- Well, let's just fix i and see what happens in a single coordinate:

$$\hat{p}_i - p_i$$

$$= \frac{X_i}{n} - p_i$$

. We would expect the variance for this to scale something like $\left(\sqrt{np_i(1-p_i)}\right)$, based on the binomial case. Then the standard error should be like $\sqrt{\frac{p_i(1-p_i)}{n}}$, so that by CLT, $\frac{\frac{X_i}{n} - p_i}{\sqrt{\frac{p_i(1-p_i)}{n}}} \rightsquigarrow N(0,1)$

$$\Leftrightarrow \frac{X_i - np_i}{\sqrt{p_i(1-p_i)n}} \rightsquigarrow N(0,1)$$

$$\Leftrightarrow \frac{X_i - np_i}{\sqrt{p_i n}} \rightsquigarrow N(0,[1-p_i]),$$

i.e. $\frac{X_i - np_i}{\sqrt{p_i n}}$ is asymptotically limiting to a mean 0, variance $(1-p_i)$ Gaussian.

. Now, there is a constraint! Indeed, $X_1 + X_2 + \dots + X_k = n$

$$np_1 + np_2 + \dots + np_k = n \quad (\star)$$

So, we should consider the random vector

$(Z_1, \dots, Z_k)$, where $\frac{X_i - np_i}{\sqrt{p_i n}} =: Z_i$.

Q: What is the distribution of this random vector? The tricky point is that due to (A), Z_i and Z_j are not (in general) independent.

What we can show (exercise) is that

$$(Z_1, \dots, Z_K) \sim \mathcal{N}(0, \Sigma), \quad \Sigma \in \mathbb{R}^{K \times K}$$

where the covariance matrix Σ is given by

$$\Sigma = \begin{pmatrix} 1-p_1 & -\sqrt{p_1 p_2} & \cdots & -\sqrt{p_1 p_K} \\ -\sqrt{p_1 p_2} & 1-p_2 & & \\ \vdots & & \ddots & \\ -\sqrt{p_1 p_K} & & & 1-p_K \end{pmatrix}$$

- We want to analyze this covariance matrix. Exercise: show Σ is a symmetric matrix with eigenvalues 1 (multiplicity $K-1$) and 0 (multiplicity 1).

This means we can diagonalize $\Sigma = U \Lambda U^T$ for an orthogonal matrix U (i.e. U has U^T as its inverse, $UU^T = U^T U = I$) and diagonal $\Lambda = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 0 \end{pmatrix}$ eigenvalue matrix.

- So, if $Z = (Z_1, \dots, Z_K) \sim \mathcal{N}(0, \Sigma)$, then $Y := U^T Z \sim \mathcal{N}(0, U^T \Sigma U)$.

But $U^T \Sigma U = [U^T][U \Lambda U^T] U$

$= [U^T U] \Lambda [U^T U]$

$= \Lambda$

So, Y is a vector $(Y_1, \dots, Y_k)$, where $Y_i \sim N(0,1)$, $i = 1, \dots, k-1$

$Y_k \equiv 0$

Finally, let $f((z_1, \dots, z_k)) = \sum_{i=1}^{k} z_i^2 = \|Z\|_2^2$

$$\Rightarrow f((y_1, \dots, y_k)) = f(U^T Z) = \|U^T Z\|_2^2$$

$$\sum_{i=1}^{k} y_i^2 \qquad = \|Z\|_2^2, \text{ by } U \text{ orthogonal}$$

Hence, $f(z_1, \dots, z_k)$ is distributed as $\sum_{i=1}^{r} y_i^2$, where $y_1 \equiv 0$, $y_i \sim N(0,1)$

So, where does this leave us? If we consider the random variable

$$f(z_1, \dots, z_k) = \sum_{i=1}^{k} z_i^2 = \sum_{i=1}^{k} \left(\frac{x_i - np_i}{\sqrt{np_i}} \right)^2 = \sum_{i=1}^{k} \frac{[x_i - np_i]^2}{np_i}, \text{ then}$$

this has a distribution given by the sum of $(k-1)$ squared Gaussians (at least in the asymptotic limit):

$$\sum_{i=1}^{k} \frac{[x_i - np_i]^2}{np_i} \sim \boxed{[k-1] \text{ squared Gaussian sum.}}$$

We call this a χ^2 distribution with $(k-1)$ degrees of freedom

Defn: A χ^2 distribution with k degrees of freedom is the r.v. $\chi^2 = \sum_{i=1}^{k} Z_i^2$, where each Z_i is an i.i.d. standard normal. The density of $\chi^2(k)$ is

$$f_{\chi^2(k)}(x) = \begin{cases} \dfrac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \cdot \Gamma\left(\frac{k}{2}\right)}, & x > 0 \\ 0, & \text{else} \end{cases}$$

So, our above calculation establishes that if $(X_1, \dots, X_k) \sim \text{Multi}(p_1, \dots, p_k, n)$, then in the asymptotic limit $n \to \infty$, $Z = (z_1, \dots, z_k)$

$$= \left(\frac{[X_1 - np_1]}{\sqrt{np_1}}, \dots, \frac{[X_k - np_k]}{\sqrt{np_k}} \right)$$

converges to some crazy k-dimensional Gaussian. More practically,

$$\sum_{i=1}^{k} \frac{[x_i - np_i]^2}{np_i}$$

converges in the asymptotic limit to $\chi^2(k-1)$.

- We can use this to do inference!