MA 576 — Qualifying Exam

Spring 2019

1. Evolutionary genetics studies allele—a type of gene—frequencies in populations. The sampling distribution of allelic frequencies under natural selection was proposed by Ewens: given a sample of n genes, the frequency partition $(m_1, \ldots, m_n)$ of the population is such that m_j is the number of alleles appearing exactly j times, for $j = 1, \ldots, n$, and so $\sum_{j=1}^{n} j \cdot m_j = n$. The probability of an allelic partition given a parameter $\alpha > 0$ is then given by *Ewens sampling formula*,

$$\mathbb{P}(m_1, \ldots, m_n) = \frac{n!}{\alpha(\alpha+1)\cdots(\alpha+n-1)} \prod_{j=1}^{n} \frac{\alpha^{m_j}}{(j!)^{m_j} m_j!}.$$

   (a) If $Y = \sum_{j=1}^{n} m_j$ is the number of different alleles in the population, show that the distribution of Y belongs to the exponential family. What is its dispersion?

   (b) Identify the canonical parameter θ and, using the recurrence relation

$$\Gamma(x+1) = x\Gamma(x), \tag{*}$$

   where Γ is the gamma function, show that the cumulant function is

$$b(\theta) = \log\Gamma(e^\theta + n) - \log\Gamma(e^\theta).$$

   (c) Using the cumulant function, find the mean $\mu = \mathbb{E}[Y]$ as a function of the digamma function $\Psi(x) = d(\log\Gamma(x))/dx$. Deduce from the recurrence (*) that $\Psi(x+1) = \Psi(x) + 1/x$ and so $\mu > 1$ for all $\alpha > 0$, as expected.

   (d) Write down the variance function $V(\mu)$ of Y as a function of the trigamma function $\Psi'(x) = d\Psi(x)/dx$. Using once more the recurrence (*), deduce that $\Psi'(x+1) = \Psi'(x) - 1/x^2$ and so $V(\mu)$ is pointwise *smaller* than the variance function of a Poisson distribution.

   (e) Show that $Z = Y/n$ also belongs to the exponential family but with a *weighted* dispersion. With $\Psi(x) \approx \log x$, show that the *inverse* canonical link for Z is

$$\mathbb{E}[Z] \approx e^\eta \log(1 + e^{-\eta})$$

   with $\eta = \theta - \log n$, that is, with an offset.

2. The *Voynich* manuscript is a mysterious codex hand-written in an unknown writing system. While attempts to decipher the text have not been successful, statistical analyses suggest that the language used is "compatible with natural languages and incompatible with random texts". A simple empirical law describing the frequency of words in a random text corpus as a function of their frequency ranks is the *Zipf-Mandelbrot* (ZM) law. If there are n different words in the text with the i-th word having rank r_i, the ZM law prescribes that its relative frequency f_i is

$$f_i = \frac{(r_i + q)^{-s}}{\sum_{j=1}^{n} (r_j + q)^{-s}},$$

   for (rank) shift parameter q and power parameter s.

(a) If $Y = (Y_1, \ldots, Y_n)$, where Y_i is the absolute frequency of the i-th word in the Voynich manuscript, argue that Y follows a multinomial distribution and so, assuming that the shift parameter q is *known*, explain how a Poisson log-linear GLM can be used to estimate the power parameter s. In particular, show that the deviance from this Poisson GLM coincides with the deviance from a multinomial GLM.

Taking the word frequencies Y in the Voynich manuscript and assuming $q = 10.67$ (covariate `r` contains the word ranks), we fit a Poisson log-linear model to obtain the following output:

```
Call:
glm(formula = Y ~ log(q + r), family = poisson)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  8.935897   0.017337   515.4   <2e-16
log(q + r)  -1.078714   0.003081  -350.2   <2e-16

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 103204.70  on 6636  degrees of freedom
Residual deviance:    340.58  on 6635  degrees of freedom
```

(b) Many natural language corpora have $s = 1.07$. Conduct a Wald test to assess if the language in the Voynich manuscript can be classified as a natural language. State the test statistic and its distribution under the null.

(c) Is the model dispersed? Formally assess if that is the case by assuming a dispersed model with $\text{Var}[Y_i] = \sigma^2 V(\mu_i)$ and then conducting a two-sided test. As usual, state the test statistic and its distribution under the null.

(d) An expert linguist claims that the most frequent word in the manuscript (e.g, top ranked) is 10% more frequent than the second most frequent word. Conduct another Wald test, this time under the dispersed model, to verify this claim. As before, state test statistic and its distribution under the null. How does the test statistic depend on an estimate for σ^2?

(e) If q is unknown, does Y belong to the exponential family? How would you estimate the shift parameter q? Be as specific as possible, describing a numerical fitting procedure and how you would jointly estimate s.