MA 576 — Qualifying Exam

Spring 2017

1. A discrete random variable Y follows a *generalized binomial* distribution with index $m \in \mathbb{N}^+$ and parameters π and ϕ, $Y \sim \mathsf{GenBin}(m, \pi, \phi)$, with $0 \le \pi \le 1$ and $0 \le \phi < 1/\pi$, if we have
$$\mathbb{P}(Y = y) = \frac{m}{m + \phi y} \binom{m + \phi y}{y} \pi^y (1 - \pi)^{m + \phi y - y}, \quad y = 0, 1, 2, \ldots$$
   (a) What is the distribution of Y when $\phi = 0$ and $\phi = 1$? Show that, in general, the generalized binomial distribution belongs to the exponential family only if ϕ is fixed. Argue that, similar to the distributions when $\phi = 0$ and $\phi = 1$, the dispersion is unit but *weighted*, that is, $a(\phi) = 1/w$; what is the weight w?
   (b) Identify the canonical parameter θ as a function of π and ϕ and the cumulant $b(\pi(\theta))$.
   (c) Using the cumulant, find the mean $\mu = \mathbb{E}[Y]/m$ of $\mathsf{GenBin}(m, \pi, \phi)/m$. Since you specified the cumulant as a function of π above, first deduce that $\partial\pi/\partial\theta = \pi(1 - \pi)/(1 - \phi\pi)$.
   (d) Write down the variance function $V(\mu)$ of a $\mathsf{GenBin}(m, \pi, \phi)$. Compare your result to the variance functions when $\phi = 0$ and $\phi = 1$.
   (e) Find the canonical link for this distribution.

2. A botanist researches the incidence of leaf blotch (a type of fungus) on barley. She designs an experiment with ten *varieties* of barley and at nine experimental *sites* and collects data on *proportion* of leaf area affected by the blotch. For simplicity, we focus on three sites (5, 6, and 7) and varieties (5, 6, and 7):

proportion (%)	site 5	site 6	site 7	variety 5	variety 6	variety 7
2.5	1	0	0	1	0	0
8.0	1	0	0	0	1	0
16.5	1	0	0	0	0	1
5.0	0	1	0	1	0	0
5.0	0	1	0	0	1	0
10.0	0	1	0	0	0	1
50.0	0	0	1	1	0	0
10.0	0	0	1	0	1	0
50.0	0	0	1	0	0	1

Since the response is a proportion, the researcher attempts to fit a binomial model. However, given the experimental design, she wants to achieve an ANOVA-like variance breakdown and so she adopts a *constant-information* link instead of the canonical logit link. A constant-information link g is defined as a function of the variance function $V(\mu)$:
$$g(\mu) = \int_c^\mu \frac{ds}{\sqrt{V(s)}},$$
where the constant c is arbitrary.

(a) Show that a constant-information link for the binomial distribution is

$$g(\mu) = 2 \arcsin \sqrt{\mu} - \frac{\pi}{2}$$

and so that $g(\text{logit}^{-1}(0)) = 0$.

The researcher fits the binomial regression using the link above, obtaining the following output:

```
Call:
glm(formula = proportion ~ site + variety - 1, data = leafblotch,
    family = binomial(link = asinsqrt))

Coefficients:
         Estimate Std. Error z value Pr(>|z|)
site5     -0.9637     0.7454  -1.293    0.196
site6     -1.0452     0.7454  -1.402    0.161
site7     -0.2871     0.7454  -0.385    0.700
variety6  -0.2079     0.8165  -0.255    0.799
variety7   0.2133     0.8165   0.261    0.794

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5.92258  on 9  degrees of freedom
Residual deviance: 0.39218  on 4  degrees of freedom
```

(b) Provide point estimates for the proportions of affected leaf area in all sites (5, 6, 7) for variety **5**.

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

(d) Now conduct a test for no difference in proportions across sites and varieties under the dispersed model. As before, state test statistic and its distribution under the null. How does the test statistic depend on an estimate for σ^2?

(e) Interestingly, the standard errors in the summary above have a "block" structure: all the standard errors for the site coefficient estimates are the same, as are the standard errors for the variety coefficient estimates. To investigate this, first compute the *information matrix* for the coefficient estimates $\widehat{\beta}$, and confirm that it is indeed constant, that is, that it does not depend on $\widehat{\beta}$. Now argue for the shared standard errors in the site and variety blocks.