Stat 5421 (Geyer) Fall 2020
Homework Assignment 5
Due Wednesday, December 15, 2021

Problem 5.1. The data are the data in Table 10.9 in Agresti. These data can be read into R as follows

```
u <- "http://www.stat.umn.edu/geyer/5421/data/table-10.9.txt"
foo <- read.table(u, header = TRUE)
sapply(foo, class)

##    survival   gestation     smoking         age      counts
## "character" "character" "character" "character"   "integer"
```

If you had trouble doing the above (like in homework 4), then download the file in a separate step and then read it, as discussed in an announcement on the course home page.

(a) Which single hierarchical model has the lowest AIC?

(b) Which group of hierarchical models has the lowest AIC and has Akaike weights adding up to 0.95?

(c) Which models listed in part (b) are graphical?

(d) Which single graphical model has the lowest AIC?

(e) Which group of graphical models has the lowest AIC and has Akaike weights adding up to 0.95?

(f) For the five graphical models with the lowest AIC, draw the graphs and interpret them by giving the implied conditional independence relationships

In all parts it may be helpful to know that the result of `summary.glmbb` is a list, the `results` component of which is the printed data frame. For example

```
gout <- glmbb(...)
sout <- summary(gout)
```

then `sout$results$criterion` is the vector of criteria (AIC, BIC, or AICc, as the case may be), `sout$results$weight` is the vector of weights, and `sout$results$formula` is the vector of formulas expressed as character

strings. The R function `as.formula` converts one character string to a formula. The R function `isGraphical` in the R package `glmbb` tells whether a formula corresponds to a graphical model.

Problem 5.2.

For each of the top five models (according to AIC) in part (a) of problem 1 on this homework, what do the "observed equals expected" and "maximum entropy" principles say about the maximum likelihood estimates?

Problem 5.3.

For the horseshoe crab data

```
library(CatDataAnalysis)
data(table_4.3)
names(table_4.3)

## [1] "color"  "spine"  "width"  "satell" "weight" "y"

sapply(table_4.3, class)

##     color     spine     width    satell    weight         y
## "integer" "integer" "numeric" "integer" "integer" "integer"

table_4.3 <- transform(table_4.3, color = as.factor(color))
```

in problem 3.2 we did a Bayesian analysis of the model having formula

```
satell ~ 0 + color + weight
```

If instead we do a frequentist analysis

```
gout <- glm(satell ~ 0 + color + weight, family = poisson,
    data = table_4.3)
summary(gout)

##
## Call:
## glm(formula = satell ~ 0 + color + weight, family = poisson,
##     data = table_4.3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9833  -1.9272  -0.5553   0.8646   4.8270
##
```

```
## Coefficients:
##          Estimate Std. Error z value Pr(>|z|)
## color2 -4.978e-02  2.331e-01  -0.214   0.8309
## color3 -2.549e-01  1.974e-01  -1.291   0.1967
## color4 -4.996e-01  1.959e-01  -2.551   0.0108 *
## color5 -5.018e-01  2.156e-01  -2.328   0.0199 *
## weight  5.462e-04  6.811e-05   8.019 1.07e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1050.8  on 173  degrees of freedom
## Residual deviance:  551.8  on 168  degrees of freedom
## AIC: 917.1
##
## Number of Fisher Scoring iterations: 6
```

What does the "observed equals expected" principle say about this model? What are the submodel canonical sufficient statistics? Calculate their expected values. Interpret these submodel canonical sufficient statistics. What is their meaning?