ST440/540 Applied Bayesian Analysis

Lab activity for 1/22/2024

(A) QUIZ AND HOMEWORK SOLUTIONS

A1: Download the tyrannosaurid growth curves data attached to this assignment

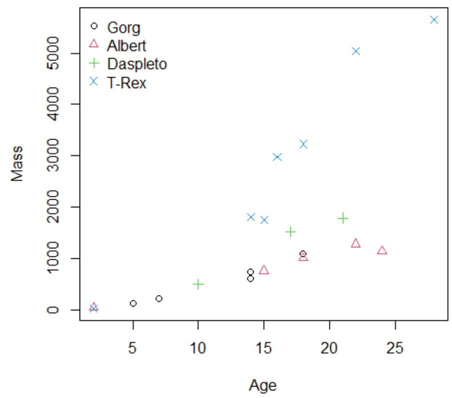
```
> dat <- read.csv("Growth curves data.csv")
> dat[1:2,]
  Taxon.ID       Taxon Age Mass
1        1 Gorgosaurus   5  127
2        1 Gorgosaurus   7  229
id    <- dat[,1]
age   <- dat[,3]
mass  <- dat[,4]
taxon <- c("Gorg","Albert","Daspleto","T-Rex")
```

(a) Write a loop to compute the mean and standard deviation of the mass for each taxon. You must write a loop to get full credit. Put the results in a table and make sure the rows and columns are labelled clearly.

```
> mn <- sd <- rep(0,4)
> for(i in 1:4){
+   mn[i] <- mean(mass[id==i])
+   sd[i] <- sd(mass[id==i])
+ }
> out           <- cbind(mn,sd)
> rownames(out) <- taxon
> colnames(out) <- c("Mean","SD")
> round(out,1)
           Mean     SD
Gorg      563.0  397.2
Albert    849.9  486.2
Daspleto 1268.3  682.6
T-Rex    2929.4 1958.0
```

(b) Make a plot of age versus mass that includes all observations but a different plotting symbol (i.e., the pch option in plot) or color for each taxon. Make sure the axes and legends are clearly labeled.

```
plot(age,mass,pch=id,col=id,xlab="Age",ylab="Mass")
legend("topleft",taxon,pch=1:4,col=1:4,bty="n")
```



(c) Perform a non-Bayesian linear regression of y=log(mass) onto x=log(age) (ignoring taxon), and report and interpret the results.

```
> logage  <- log(age)
> logmass <- log(mass)
> summary(lm(logmass ~ logage))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2946     0.4206   5.456 3.50e-05 ***
logage        1.7538     0.1598  10.977 2.09e-09 ***
```

There is a positive and statistically significant relationship between log age and log mass.

(B) DISCUSSION QUESTIONS

(1) Select an appropriate family of distributions for the following quantities. Justify your choice.

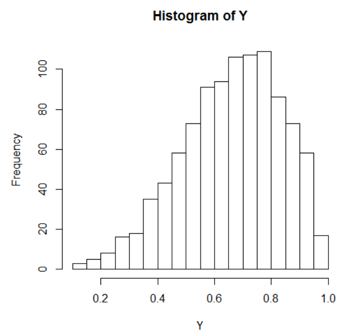
(a) The time between entering Starbucks and having your order: Exponential or gamma because time is continuous and positive.

(b) IQ scores: They are designed to be Normal(100,15^2). If IQ is recorded as an integer, then a normal is still a good approximation.

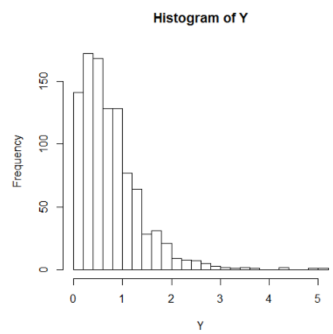
(c) Number of hurricanes in NC in 2022: Since it is a number of events (discrete) in an interval of time so it might be a Poisson random variable.

(d) Number of 100 patients in a vaccine trial that experience an adverse event: Binomial(100, p) since we have a number of success in a fixed number of trials.

(e) The data plotted in this histogram: Beta distribution because it's continuous (although a histogram makes it look discrete) between 0 and 1.



(f) The data plotted in this histogram: Chi-square or gamma distribution because its support is (0,inf) and it is continuous (though again, the histogram makes it look discrete).

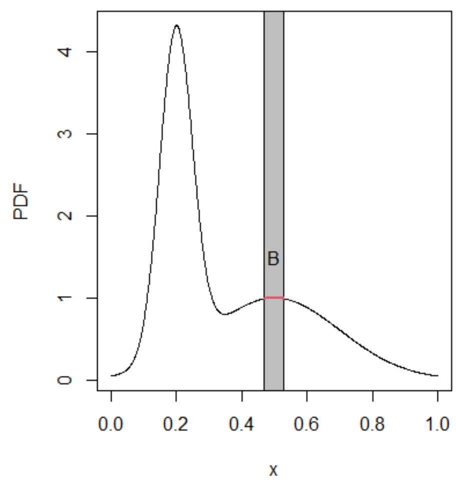


(g) The data in this table: Bernoulli because the support is {0,1}.

```
Y
   0   1
798 202
```

(2) Argue that if X is a continuous random variable, then P(X=x) must be zero for any single value of x.

If we consider a tiny interval like B in the plot below, then it is reasonable to assuming the PDF is constant in the interval, so P(X=x) = c for any x in B. But there are infinitely-many x's in B, and if each has probability c then the total probability in B is infinity * c = infinity. Therefore, c must be zero.



(3) Say that 80% of cabs in the city are blue and 20% are green, and that hit-and-run witnesses can correctly identify the color of a car 90% of the time, regardless of the true color. Given that a witness claims a car is green, what is the probability that the car is truly green? What assumptions are you making in this calculation? (this problem is adapted from Thinking Fast and Slow)

This analysis assumes green and blue cars are equally to be in a crash. Under this assumption, say

$\theta = 1$ if the car is green and $\theta = 0$ otherwise and

$Y = 1$ if the witness sees green and $Y = 0$ otherwise.

The problem gives $P(\theta=1) = 0.2$, $P(\theta=0) = 0.8$, $P(Y=1|\theta=1) = 0.9$ and $P(Y=1|\theta=0) = 0.1$. We want $P(\theta=1|Y=1)$. Bayes Theorem gives

$$\begin{aligned}P(\theta=1|Y=1) &= P(Y=1|\theta=1)P(\theta=1)/P(Y=1)\\ &= P(Y=1|\theta=1)P(\theta=1)/[P(Y=1|\theta=0)P(\theta=0) + P(Y=1|\theta=1)P(\theta=1)]\\ &= 0.9*0.2/(0.1*0.8+0.9*0.2) = 0.69.\end{aligned}$$

This can also be approximated using Monte Carlo sampling,

```
> S <- 100000
> theta <- rbinom(S,1,0.2)
> prob1 <- ifelse(theta==1,0.9,0.1)
> Y     <- rbinom(S,1,prob1)
> table(theta,Y)
     Y
theta     0     1
    0 72225  7933
    1  1962 17880
> mean(theta[Y==1])
[1] 0.6926742
```

(4) Using the fact that f(x,y) = f(x|y)f(y) and f(x,y) = f(y|x)f(x), prove Bayes' Theorem.

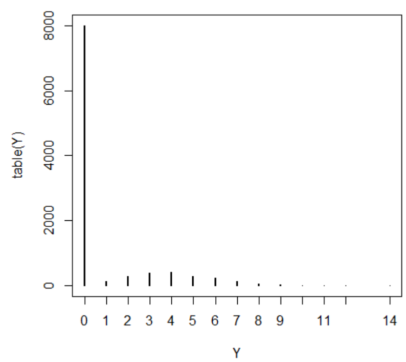
We set them equal giving f(x|y)f(y) = f(y|x)f(x), and divided by f(y) gives

f(x|y) = f(y|x)f(x)/f(y)

which proves Bayes' Theorem.

(5) If Y follows the zero-inflated Poisson (ZIP) distribution, it can be written in terms of two variables Y=U*V, where U ~ Bernoulli(p) independent of V ~ Poisson(lambda). Here is an example with p=0.2 and lambda=4.

```
> U <- rbinom(10000,1,0.2)
> V <- rpois(10000,4)
> Y <- U*V
> plot(table(Y))
```




```
> round(dpois(0:10,4),2) # might be helpful
 [1] 0.02 0.07 0.15 0.20 0.20 0.16 0.10 0.06 0.03 0.01 0.01
```

(a) What is a real-life example of a variable that might follow this distribution? In your example, what are the interpretations of U and V?

Population of a species with a small ecological niche. U indicates that the location is conducive to the species and V is be the number of individuals in a location where the species is present.

(b) What is the probability that Y=0?

P(Y = 0) = P(U=0) + P(U=1)*P(V=0) = 0.8 + 0.2*0.02 = 0.804 (0.02 is from the R table above).

(c) What is the probability that Y=4?

P(Y=4) = P(U=1)*P(V=4) = 0.2*0.2 = 0.04.

(d) What is the probability that U=1 given Y=0?

Bayes rule says P(U=1|Y=0) = P(Y=0|U=1)*P(U=1)/P(Y=0). We know P(Y=0|U=1) = P(V=0) = 0.02 and the prior for U is P(U=1) = 0.2. P(Y=0) is computed above to be 0.804. So

P(U=1|Y=0) = 0.02*0.20/0.804 = 0.005.

(e) What is the probability that U=1 given Y=4?

If Y>0 then we know U=1, so P(U=1|Y=4) = 1.0.