# Homework 1

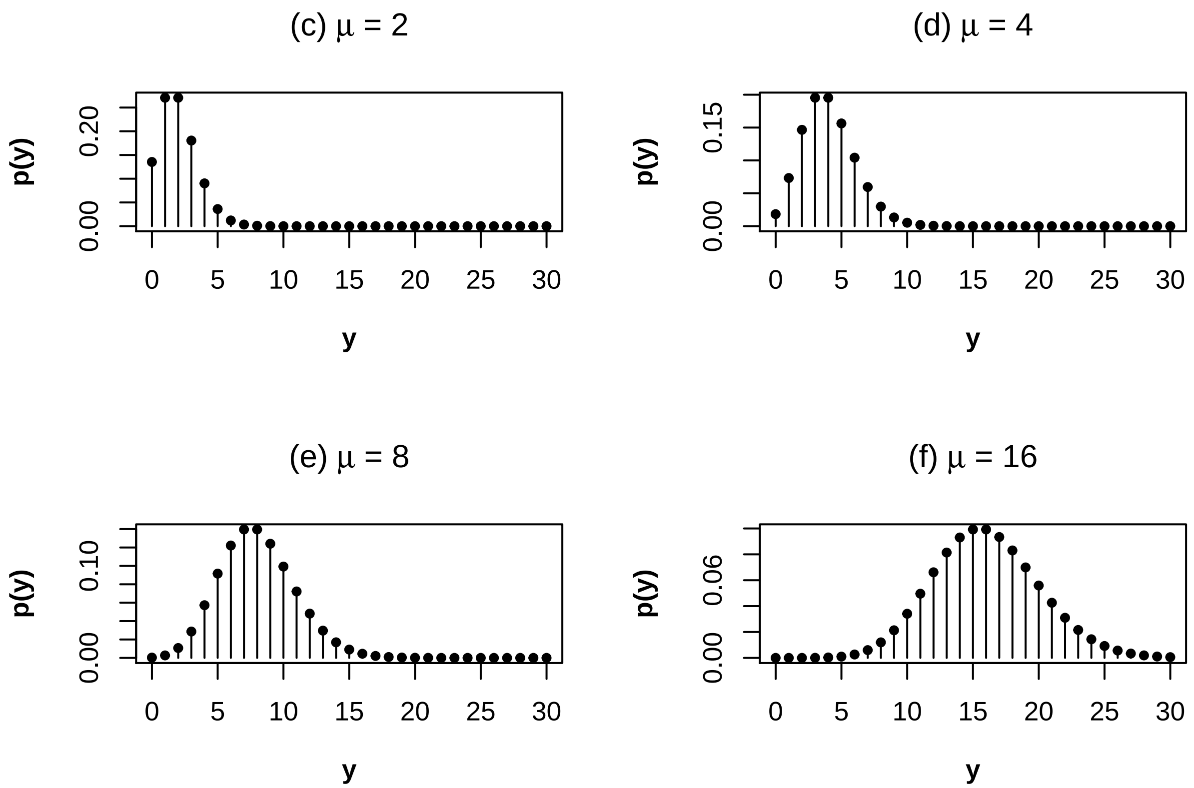
Joshua Ingram

8/30/2020

## Problem 1

### 1.

```r
seq_vals <- seq(0, 30, 1)
plamb_2 <- dpois(seq_vals, 2)
plamb_4 <- dpois(seq_vals, 4)
plamb_8 <- dpois(seq_vals, 8)
plamb_16 <- dpois(seq_vals, 16)
layout(matrix(c(1,2,3, 4),ncol=2), )
plot(seq_vals, plamb_2, pch=20, main=expression(paste("(c) ", mu, " = 2")),
     xlab =expression(bold("y")), ylab=expression(bold("p(y)")))
segments(seq_vals, 0, seq_vals, plamb_2)
plot(seq_vals, plamb_8, pch=20, main=expression(paste("(e) ", mu, " = 8")),
     xlab =expression(bold("y")), ylab=expression(bold("p(y)")))
segments(seq_vals, 0, seq_vals, plamb_8)
plot(seq_vals, plamb_4, pch=20, main=expression(paste("(d) ", mu, " = 4")),
     xlab =expression(bold("y")), ylab=expression(bold("p(y)")))
segments(seq_vals, 0, seq_vals, plamb_4)
plot(seq_vals, plamb_16, pch=20, main=expression(paste("(f) ", mu, " = 16")),
     xlab =expression(bold("y")), ylab=expression(bold("p(y)")))
segments(seq_vals, 0, seq_vals, plamb_16)
```




**Note: I was able to make the graphs have 1x1 aspect ratios like in the slides, but the graphs were too small in the pdf output for some reason. If you would like to see the code so they are exactly the same (besides the size), I will be happy to provide it.**

## 2.

```r
# mu = 2
ppois(10, 2, lower.tail = TRUE) - ppois(5,2, lower.tail = TRUE)
```

```
## [1] 0.0165553
```

```r
# mu = 4
ppois(10, 4, lower.tail = TRUE) - ppois(5,4, lower.tail = TRUE)
```

```
## [1] 0.2120298
```

```r
# mu = 8
ppois(10, 8, lower.tail = TRUE) - ppois(5,8, lower.tail = TRUE)
```

```
## [1] 0.6246497
```

```r
# mu = 16
ppois(10, 16, lower.tail = TRUE) - ppois(5,16, lower.tail = TRUE)
```

```
## [1] 0.07601223
```

$P(X \in [5, 10])$, where $X \sim Pois(2)$, is 0.0166

$P(X \in [5, 10])$, where $X \sim Pois(4)$, is 0.212

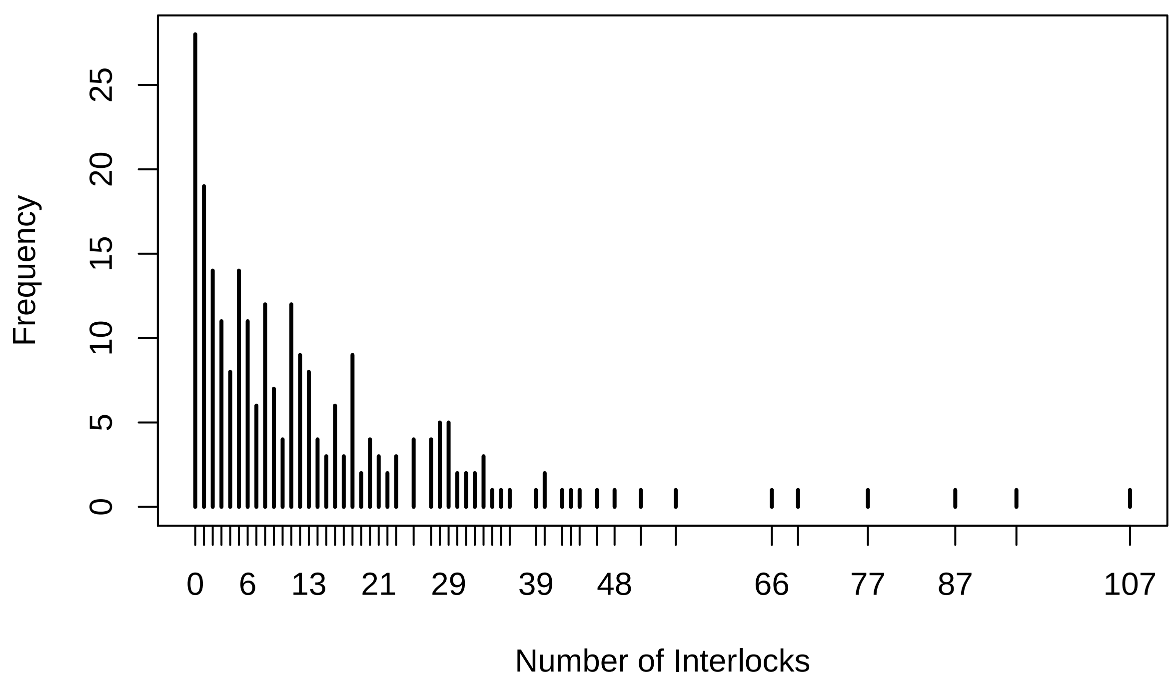
$P(X \in [5, 10])$, where $X \sim Pois(8)$, is 0.625

$P(X \in [5, 10])$, where $X \sim Pois(16)$, is 0.076

### 3.

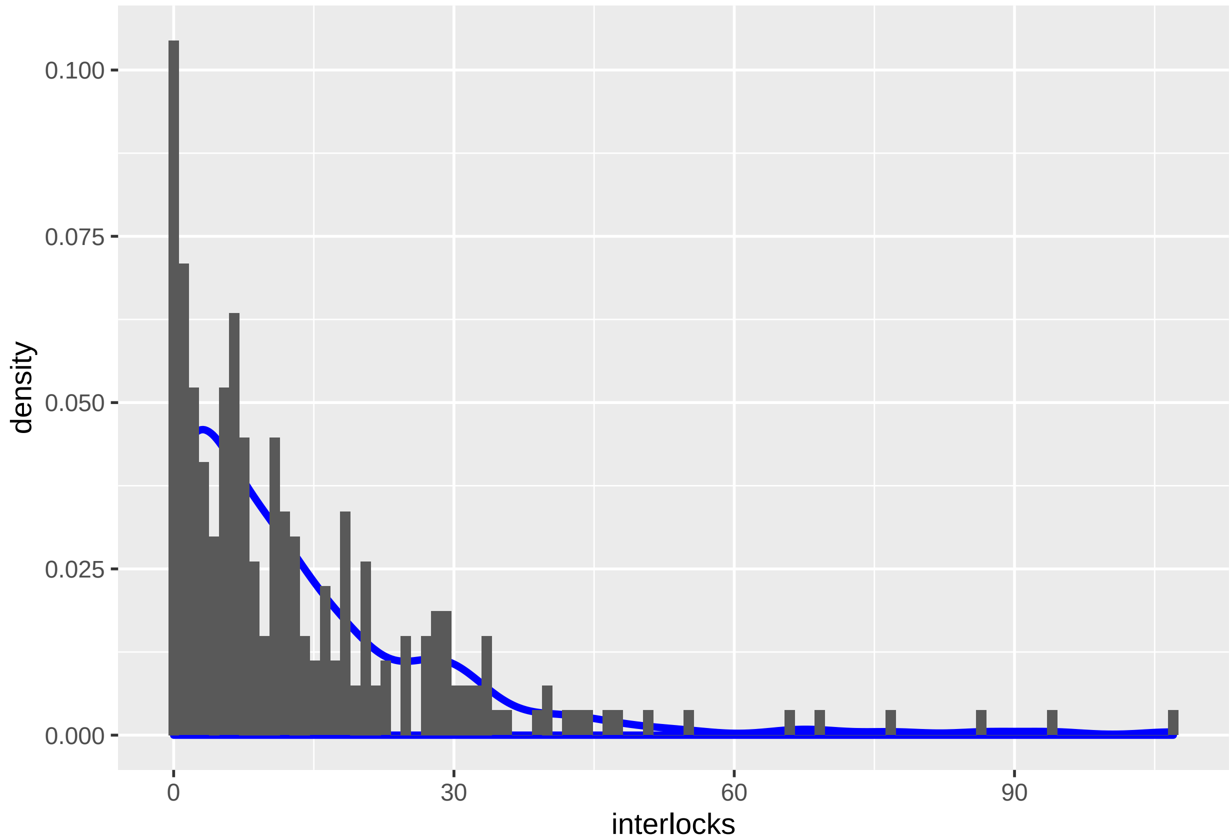
We are working with count data for the number of interlocks. Clearly right-skewed.

```r
plot(table(Ornstein$interlocks), xlab="Number of Interlocks", ylab="Frequency")
```




Since we want to find $\hat{\mu}$ using MLE, we need to find the value of $\mu$ that maximizes the likelihood of observing our distribution. We can get an idea by looking at the the density plot.

```r
ggplot(data = ornstein, aes(x=interlocks, y = ..density..))+
  geom_density( col = "blue", size = 1.3) +
  geom_histogram(bins = 100)
```




Now, we could estimate $\mu$ by simply finding the mean of the distribution as it is the value where our data would be most likely to occur. That value would be: 13.58065

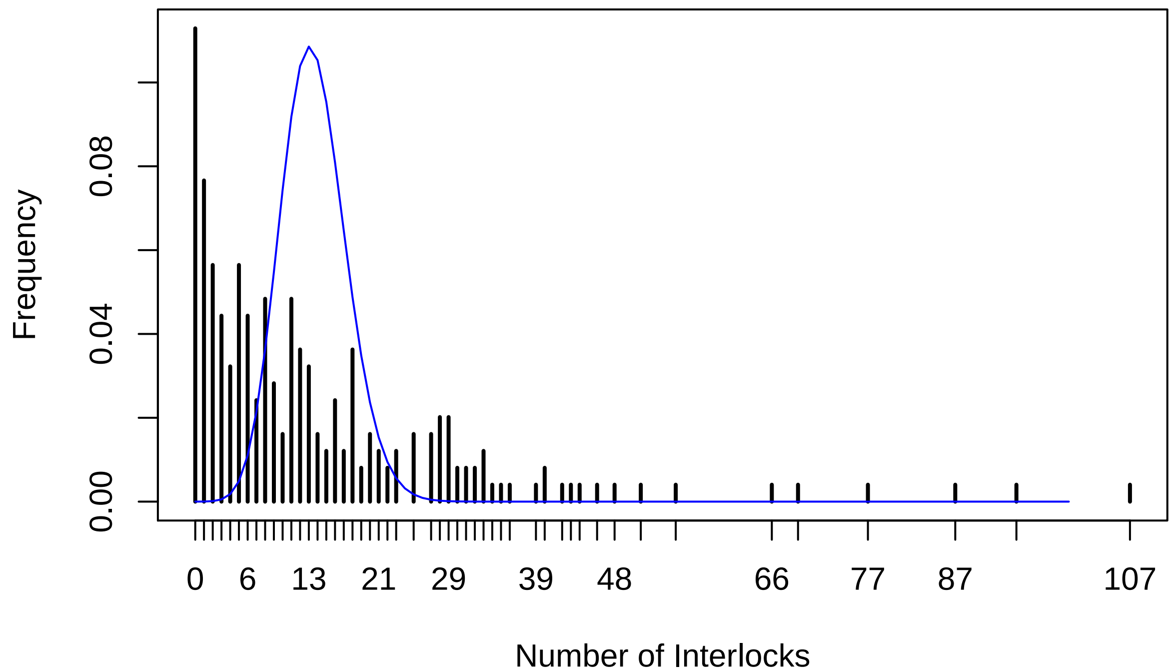
```
mean(ornstein$interlocks)
```

```
## [1] 13.58065
```

We can then plot the poisson distribution given $\hat{\mu}$ (notice the scale for y-axis).

```
seq_vals <- seq(0, 100)
estimates <- dpois(seq_vals, 13.58065)
df <- data.frame(seq_vals, estimates)
colnames(df) <- c("values", "estimates")

plot(table(Ornstein$interlocks)/length(Ornstein$interlocks), xlab="Number of Interlocks", ylab="Frequenc
lines(df$values, df$estimates, col ="blue")
```

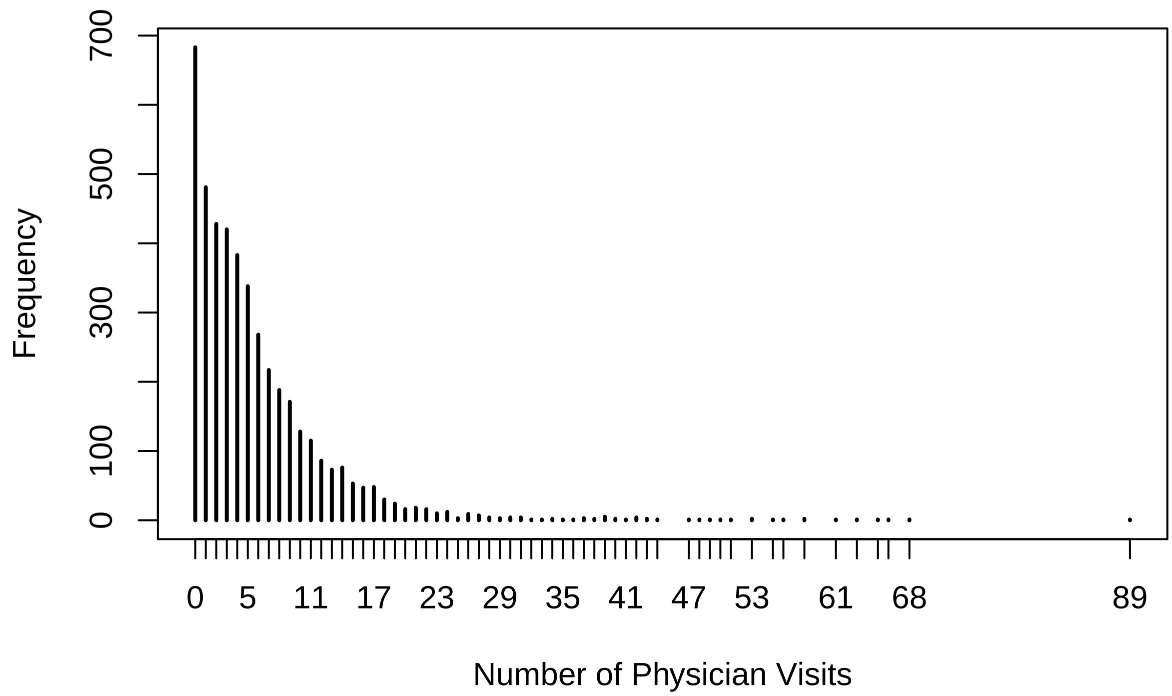



This doesn't seem like a great way to estimate the marginal distribution, as there is a lot more data closer to 0 and the distribution is right-skewed, meaning the mean rate of our data is highly affected by the skew. Zero-inflated model or a power-law?!

## Problem 2

### 1.

```
plot(table(nmes$visits), xlab="Number of Physician Visits", ylab="Frequency")
```




The most common number of visits is 0 with the counts reaching numbers as high as 68 and 89. Right-skewed. Zeros seem to be overrepresented.

**2.**

$$(Y|x = chronic_i, age_i, gender_i, income_i, insurance_i) \sim_{ind.} Pois(\mu_i), i = 1, 2..., 4406$$

$$log(\mu_i) = \beta_0 + \beta_1 chronic_i + \beta_2 age_i + \beta_3 I_{gender,i} + \beta_4 income_i + \beta_5 I_{insurance,i}$$

$$I_{chronic} \in \{0 = female, 1 = male\},\ I_{insurance} \in \{0 = no, 1 = yes\}$$

**3.**

```
nmes_fit <- glm(visits ~ chronic + age + gender + income + insurance, family=poisson, data=nmes)
summary(nmes_fit)
```

```
##
## Call:
## glm(formula = visits ~ chronic + age + gender + income + insurance,
##     family = poisson, data = nmes)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.0349  -2.0695  -0.7102   0.7390  17.6511
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.491e+00  7.728e-02  19.296  < 2e-16 ***
## chronic       2.038e-01  4.113e-03  49.562  < 2e-16 ***
## age          -3.278e-02  1.009e-02  -3.249  0.00116 **
## gendermale   -1.154e-01  1.304e-02  -8.849  < 2e-16 ***
## income       -5.927e-05  2.163e-03  -0.027  0.97814
## insuranceyes  2.464e-01  1.620e-02  15.210  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 26943  on 4405  degrees of freedom
## Residual deviance: 24438  on 4400  degrees of freedom
## AIC: 37225
##
## Number of Fisher Scoring iterations: 5
```

We use the Likelihood Ratio Test to test the full model significance.

$H_0: \ \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

$H_A: \ \{\exists \ \beta_j \neq 0 \mid j = 1, ..., 5\}$

$\alpha = 0.05$

Null Model: $Y_i \sim_{ind.} Pois(\mu_i), \ log(\mu_i) = \beta_0$

LRT statistic $= 2log(\frac{L_1}{L_0}) = G_0^2 \sim \chi^2_{5-1}$

p-value: $P(\chi^2_{5-1} \geq G_0^2)$

```r
nmes_null <- glm(visits ~ 1,
                 family=poisson, data=nmes)
anova(nmes_null, nmes_fit, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: visits ~ 1
## Model 2: visits ~ chronic + age + gender + income + insurance
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      4405      26943
## 2      4400      24438  5   2505.2 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the LRT, we receive a p-value of basically 0. This gives us significant evidence to reject the null hypothesis and our overall model is statistically significant.

### 4.

```r
Anova(nmes_fit)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: visits
##           LR Chisq Df Pr(>Chisq)
## chronic    2255.50  1  < 2.2e-16 ***
## age          10.62  1   0.001119 **
## gender       78.97  1  < 2.2e-16 ***
## income        0.00  1   0.978132
## insurance   242.47  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Income is not statistically significant.

### 5.

```r
nmes_fit2 <- glm(visits ~ chronic + age + gender + insurance, family=poisson, data=nmes)
summary(nmes_fit2)
```

```
##
## Call:
## glm(formula = visits ~ chronic + age + gender + insurance, family = poisson,
##     data = nmes)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.0350  -2.0693  -0.7102   0.7392  17.6512
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.490913   0.076971  19.370  < 2e-16 ***
## chronic       0.203841   0.004109  49.614  < 2e-16 ***
## age          -0.032771   0.010080  -3.251  0.00115 **
## gendermale   -0.115410   0.012955  -8.908  < 2e-16 ***
## insuranceyes  0.246359   0.016067  15.333  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 26943  on 4405  degrees of freedom
## Residual deviance: 24438  on 4401  degrees of freedom
## AIC: 37223
##
## Number of Fisher Scoring iterations: 5
```

$$log(\mu_i) = 1.5 + 0.2chronic_i - 0.03age_i - .12gendermale_i + 0.25insuranceyes_i$$

## 6.

```r
summary_fit2 <- summary(nmes_fit2)
1- summary_fit2$deviance/summary_fit2$null.deviance
```

```
## [1] 0.09298111
```

9.3% of the variation in our data is explained by the model.

## 7.

### a.

chronic:

For every 1 additional chronic condition, the number of physician office visits will increase by a factor of $e^{0.2}$, on average, ceteris paribus. (or "will multiply by $e^{0.2}$")

insurance:

For people with insurance, the number of physician office visits are $e^0.25$ times greater than those without insurance, on average, ceteris paribus.

### b.

```r
round(confint(nmes_fit2),3)
```

```
## Waiting for profiling to be done...
```

```
##                2.5 % 97.5 %
## (Intercept)    1.340  1.642
## chronic        0.196  0.212
## age           -0.053 -0.013
## gendermale    -0.141 -0.090
## insuranceyes   0.215  0.278
```

chronic:

For every 1 additional chronic condition, we are 95% confident that the number of physician office visits will increase by between a factor of $e^{0.196}$ and $e^{0.212}$, on average, ceteris paribus. (or "will multiply by between $e^{0.196}$ and $e^{0.212}$")

insurance:

For people with insurance, we are 95% confident that the number of physician office visits are between $e^{0.215}$ and $e^{0.278}$ times greater than those without insurance, on average, ceteris paribus.