STAT 5703 Homework 1

Shijie He(sh3975), Yunjun Xia(yx2569), Shuyu Huang(sh3967)

2/6/2020

Exercise 5

Part 1

The goal of the paper is to investigate that whether new discoveries or inventions are inevitable. The author aims to demonstrate that techno-scientific contributions are probabilistic rather than deterministic by using the Poisson model. He addresses three topics:
1. The theoretical basis for the Poisson distribution to predict invetions.
2. The accuracy of the Poisson model to describe the empirical data.
3. The implications for techno-scientific creativity by the Poisson model.
The author employs the exponential approximation of binomial expansion to define the probability function and identify the parameters, which is the Poisson distribution.

The Poisson model seems to be a reasonable choice to us because techno-scientific contributions are rare events in the history. That is, the probability for one inventions to occur is small. Also, the number of researchers who have a probability to be successful is large. The properties of the parameters p and n fit the properties of a poisson distribution intuitively. Moreover, although we lack the comprehensive empirical data of "complete failures", the Poisson model can avoid the problem of ignorance of the number of total failures and single successes because we only have to find one parameter μ. For the poisson distribution, μ represents both the mean and the variance of the distribution.

Part 2

As mentioned in part 1, one problem is our ignorance of the number of complete failures and singletons in the history. Therefore, by using the truncated Poisson model, when we estimate the parameter μ, we do not have to consider cases of complete failures or singletons. Therefore, we can avoid the problem of lacking empirical data.

Part 3

We have: $P(Y = k) = \frac{e^{-\mu}\mu^k}{k!}\frac{1}{1-e^{-\mu}-\mu e^{-\mu}}$, for $k \geq 2$.

$$
\begin{aligned}
\mathbb{E}[Y] =& \sum_{k=2}^{\infty} k \frac{e^{-\mu}\mu^k}{k!} \frac{1}{1-e^{-\mu}-\mu e^{-\mu}} \\
=& \frac{1}{1-e^{-\mu}-\mu e^{-\mu}} \sum_{k=2}^{\infty} k \frac{e^{-\mu}\mu^k}{k!} \\
=& \frac{1}{1-e^{-\mu}-\mu e^{-\mu}} [\sum_{k=0}^{\infty} k \frac{e^{-\mu}\mu^k}{k!} - \sum_{k=0}^{1} k \frac{e^{-\mu}\mu^k}{k!}] \\
=& \frac{1}{1-e^{-\mu}-\mu e^{-\mu}} [\sum_{k=0}^{\infty} k \frac{e^{-\mu}\mu^k}{k!} - \mu e^{-\mu}] \\
=& \frac{1}{1-e^{-\mu}-\mu e^{-\mu}} [\mu e^{-\mu} \sum_{k=1}^{\infty} \frac{\mu^{k-1}}{(k-1)!} - \mu e^{-\mu}] \\
=& \frac{1}{1-e^{-\mu}-\mu e^{-\mu}} [\mu e^{-\mu} \sum_{j=0}^{\infty} \frac{\mu^{j}}{j!} - \mu e^{-\mu}] \\
=& \frac{1}{1-e^{-\mu}-\mu e^{-\mu}} (\mu e^{-\mu} e^{\mu} - \mu e^{-\mu}) \\
=& \frac{1}{1-e^{-\mu}-\mu e^{-\mu}} (\mu - \mu e^{-\mu})
\end{aligned}
$$

We have: $\text{Var}[Y] = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2$

$$
\begin{aligned}
\mathbb{E}[Y^2] =& \sum_{k=2}^{\infty} k^2 \frac{e^{-\mu}\mu^k}{k!} \frac{1}{1-e^{-\mu}-\mu e^{-\mu}} \\
=& \frac{1}{1-e^{-\mu}-\mu e^{-\mu}} [\sum_{k=0}^{\infty} k^2 \frac{e^{-\mu}\mu^k}{k!} - \mu e^{-\mu}] \\
=& \frac{1}{1-e^{-\mu}-\mu e^{-\mu}} [\mu e^{-\mu} \sum_{k=1}^{\infty} k \frac{\mu^{k-1}}{(k-1)!} - \mu e^{-\mu}] \\
=& \frac{1}{1-e^{-\mu}-\mu e^{-\mu}} [\mu e^{-\mu} (\mu e^{\mu} + e^{\mu}) - \mu e^{-\mu}] \\
=& \frac{1}{1-e^{-\mu}-\mu e^{-\mu}} (\mu^2 + \mu - \mu e^{-\mu}) \\
=& \frac{\mu}{1-e^{-\mu}-\mu e^{-\mu}} (\mu + 1 - e^{-\mu})
\end{aligned}
$$

Therefore, $\text{Var}[Y] = \frac{\mu}{1-e^{-\mu}-\mu e^{-\mu}} (\mu + 1 - e^{-\mu}) - [\frac{\mu}{1-e^{-\mu}-\mu e^{-\mu}} (1 - e^{-\mu})]^2$

Part 4

Let the sample size be n. Let the observed sample be $y_1, y_2, ..., y_n$.

$$\mathcal{L}(\mu; y_1, y_2, ..., y_n) = \prod_{i=1}^{n} [\frac{e^{-\mu}\mu^{y_i}}{y_i!} \frac{1}{1 - e^{-\mu} - \mu e^{-\mu}}]$$

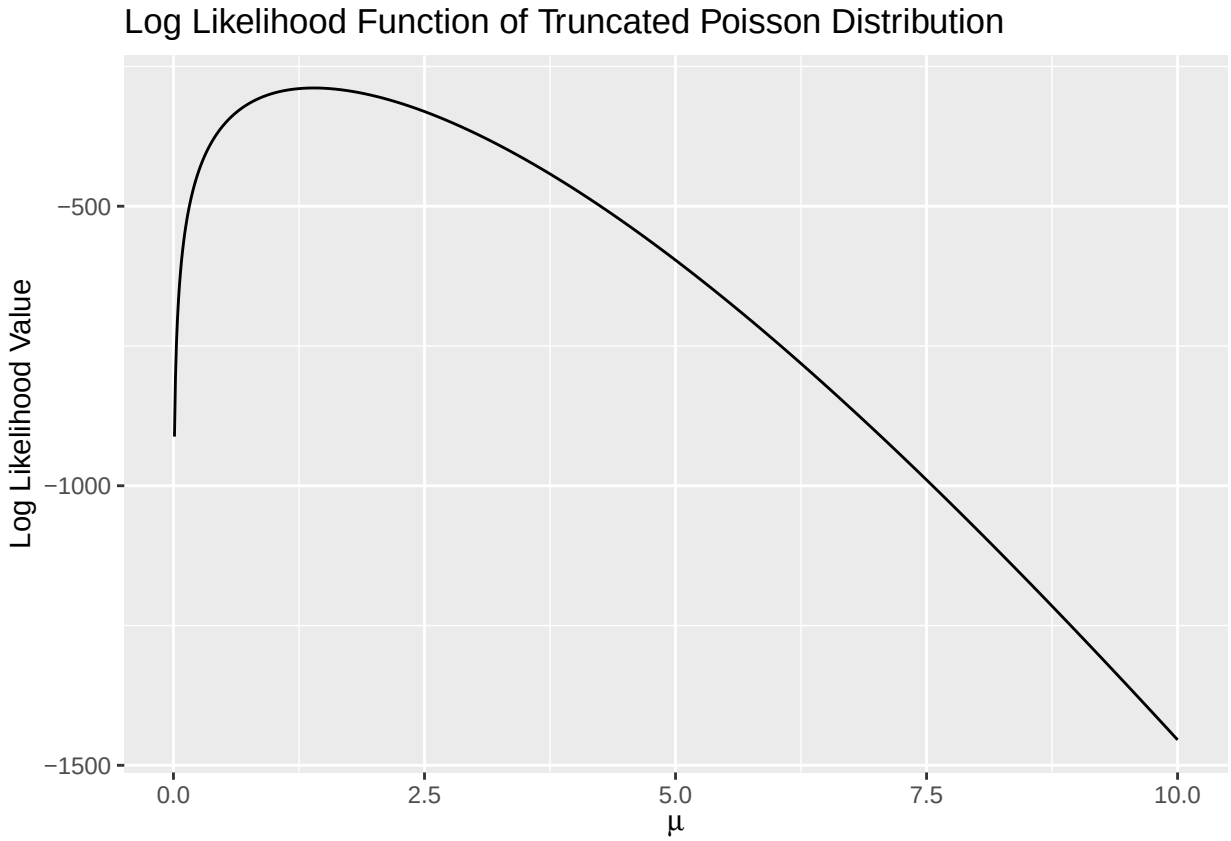
$$= (1 - e^{-\mu} - \mu e^{-\mu})^{-n} e^{-\mu n} \mu^{\sum_{i=1}^{n} y_i} \prod_{i=1}^{n} \frac{1}{y_i!}$$

$$\ln \mathcal{L}(\mu; y_1, y_2, ..., y_n) = -n \ln(1 - e^{-\mu} - \mu e^{-\mu}) - \mu n + \sum_{i=1}^{n} y_i \ln(\mu) - \sum_{i=1}^{n} \ln(y_i!)$$

```r
obs_freq <- c(179,51,17,6,8,1,0,2)
k <- seq(2,9,1)
table1_data <- data.frame('k'=k, 'obs_freq'=obs_freq)
obs_data <- numeric()
for (i in 1:8){
  obs_data <- c(obs_data, rep(table1_data[i,1],table1_data[i,2]))
}
```

```r
log_likelihood <- function(data, mu){
  n <- length(data)
  log_llh <- -n*log(1-exp(-mu)-mu*exp(-mu))-mu*n+sum(data*log(mu))-sum(log(factorial(data)))
}
```

```r
list_mu <- seq(0.01,10,0.01)
list_log_llh <- numeric()
for (i in 1:length(list_mu)){
  temp <- log_likelihood(obs_data, list_mu[i])
  list_log_llh <- c(list_log_llh,temp)
}
df <- data.frame('mu'=list_mu, 'llh'=list_log_llh)
ggplot(df, aes(x=mu,y=llh))+
  geom_line()+
  ggtitle('Log Likelihood Function of Truncated Poisson Distribution')+
  labs(x=expression(mu), y='Log Likelihood Value')
```



Part 5

```
neg_log_likelihood <- function(mu){
  result <- -log_likelihood(data=obs_data, mu)
  return(result)
}
mle(neg_log_likelihood, start = list(mu=1), method = "BFGS")
```

```
##
## Call:
## mle(minuslogl = neg_log_likelihood, start = list(mu = 1), method = "BFGS")
##
## Coefficients:
##       mu
## 1.398391
```

Therefore, the maximum likelihood estimate of μ for the data in table 1 is approximately 1.3984.

Part 6

Since the truncated poisson distribution is under regularity condistions, by using the theorem, we have:

$$\hat{\mu}_{MLE} \xrightarrow[n\to\infty]{\mathcal{P}} \mu$$
$$\sqrt{n}(\hat{\mu}_{MLE} - \mu) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, I(\mu)^{-1})$$

where $I(\mu)$ is Fisher Information.

Therefore, the asymptotic distribution for $\hat{\mu}_{MLE}$ is:

$$\hat{\mu}_{MLE} \approx \mathcal{N}(\mu, \frac{1}{nI(\mu)})$$

To computer Fisher Information for this distribution:
$\ln \mathcal{L}(\mu; Y_1) = -\ln(1 - e^{-\mu} - \mu e^{-\mu}) - \mu + y_1 \ln(\mu) - \ln(y_1!)$

$$\begin{aligned} I(\mu) =& -\mathbb{E}[\frac{\partial^2 \ln \mathcal{L}(\mu; Y_1)}{(\partial \mu)^2}] \\ =& \text{Var}(\frac{\partial}{\partial \mu} \ln \mathcal{L}(\mu; Y_1)) \\ =& \frac{1}{\mu^2} \text{Var}(Y) \end{aligned}$$

$\text{Var}[Y] = \frac{\mu}{1-e^{-\mu}-\mu e^{-\mu}}(\mu + 1 - e^{-\mu}) - [\frac{\mu}{1-e^{-\mu}-\mu e^{-\mu}}(1 - e^{-\mu})]^2$

Therefore, we have:

$$I(\mu) = \frac{(1 - e^{-\mu})^2 - \mu^2 e^{-\mu}}{\mu(1 - e^{-\mu} - \mu e^{-\mu})^2}$$

By plugging in $\hat{\mu}_{MLE} = 1.398391$:

```
mu_mle <- 1.398391
fisher_info <- ((1-exp(-mu_mle))^2-mu_mle^2 * exp(-mu_mle))/
  (mu_mle * (1-exp(-mu_mle)-mu_mle*exp(-mu_mle))^2)
fisher_info
```

```
## [1] 0.3616344
```

Part 7

By using the asymptotic distribution of $\hat{\mu}_{MLE}$ from part 6, we can derive the 95% asymptotic confidence interval for μ:
$[\hat{\mu}_{MLE} - z_{0.025}\frac{1}{\sqrt{nI(\hat{\mu}_{MLE})}}, \hat{\mu}_{MLE} + z_{0.025}\frac{1}{\sqrt{nI(\hat{\mu}_{MLE})}}]$

```
z <- 1.96
n <- 264
lower <- mu_mle-z/sqrt(n*fisher_info)
upper <- mu_mle+z/sqrt(n*fisher_info)
c(lower, upper)
```

```
## [1] 1.197796 1.598986
```

Plug in the value of Fisher information and MLE of μ, the confidence interval is: (1.197796 1.598986).

Part 8

This seems like a reasonable approach to fit the desired model, because the true parameter μ can vary a great deal across different groups. Therefore, the ad-hoc technique can increase the accuary of the estimate. However, it's hard to derive the properties of this estimator mathematically.

Part 9

The estimate obtained by Simonton's technique is $\mu = 1.4$ whereas the maximum likelihood estimate computed above is $\hat{\mu}_{MLE} = 1.398391$. These two results are quite similar which shows that Simonton's technique is as accurate as our MLE. Thus, both of the two estimates can show that the techno-scientific contributions are not deterministic or inevitable, but they are probabilistic such that inventions quite depend on luck.