Homework #1

Håkon Sandaker Vincent Wilmet

```
library(fitdistrplus)
```

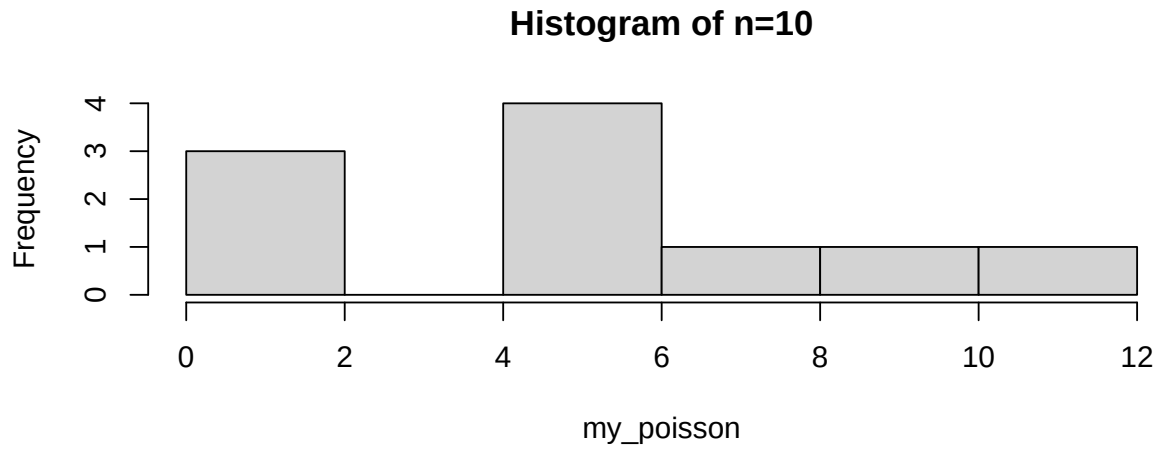
```
## Loading required package: MASS
```

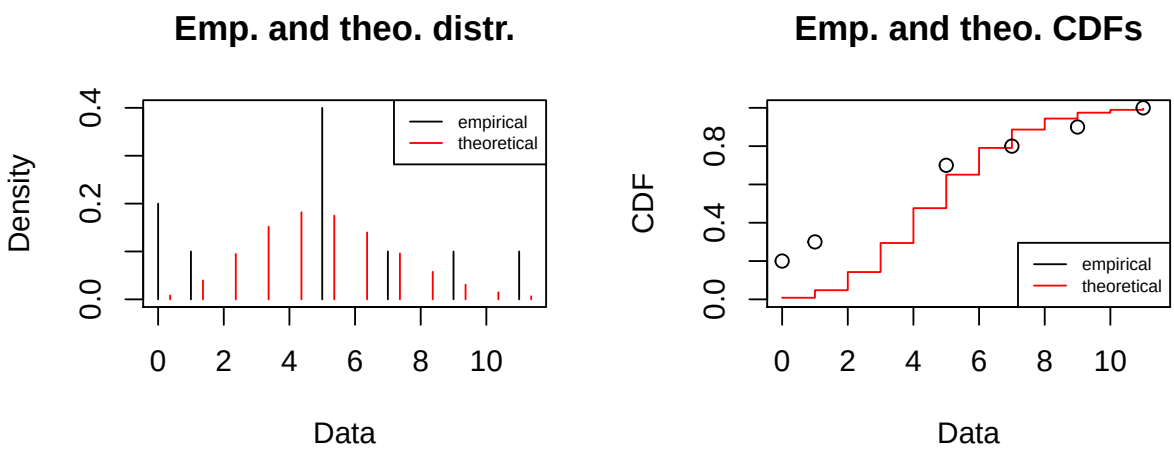
```
## Loading required package: survival
```

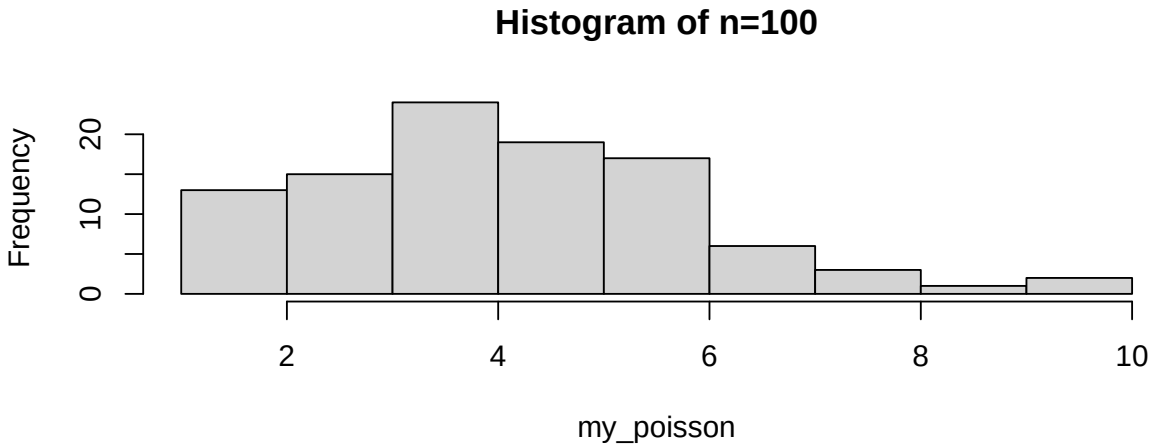
Q1 Simulation, Asymptotic behavior

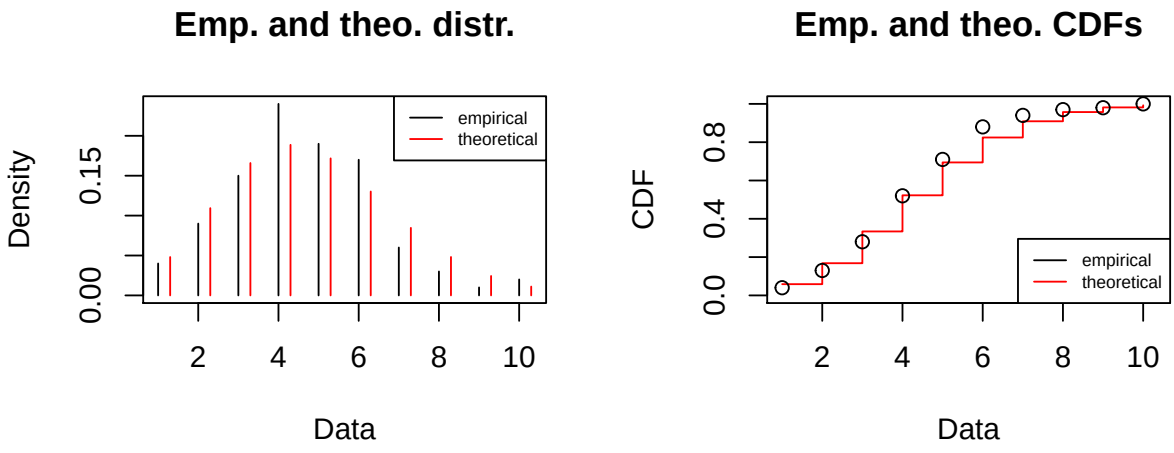
a)

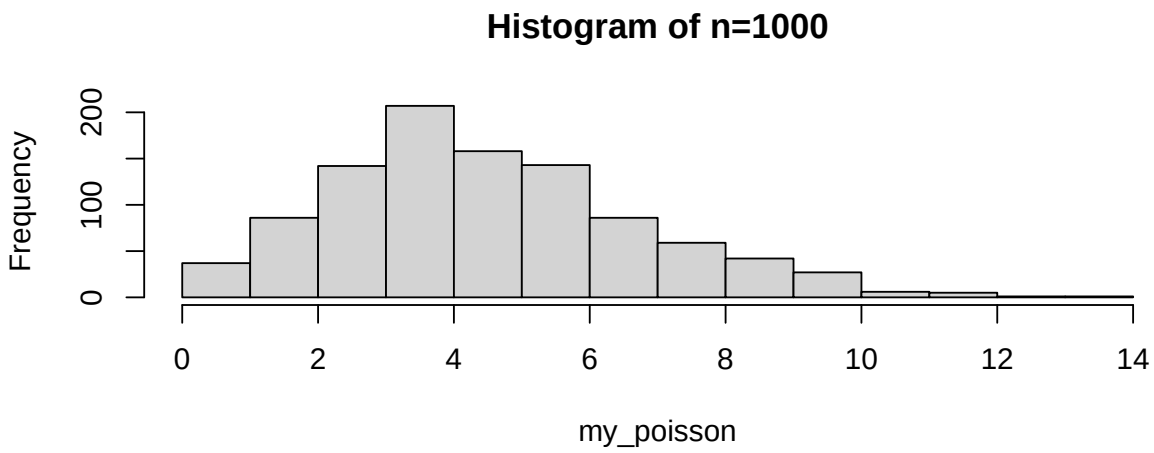
```
library(MASS)
library(survival)
library(fitdistrplus)

for (n in c(10, 100, 1000))
{
  my_poisson <- rpois(n, 5)
  hist(my_poisson, main=paste("Histogram of n=", n, sep=""))
  plot(fitdist(my_poisson, "pois"))
}
```

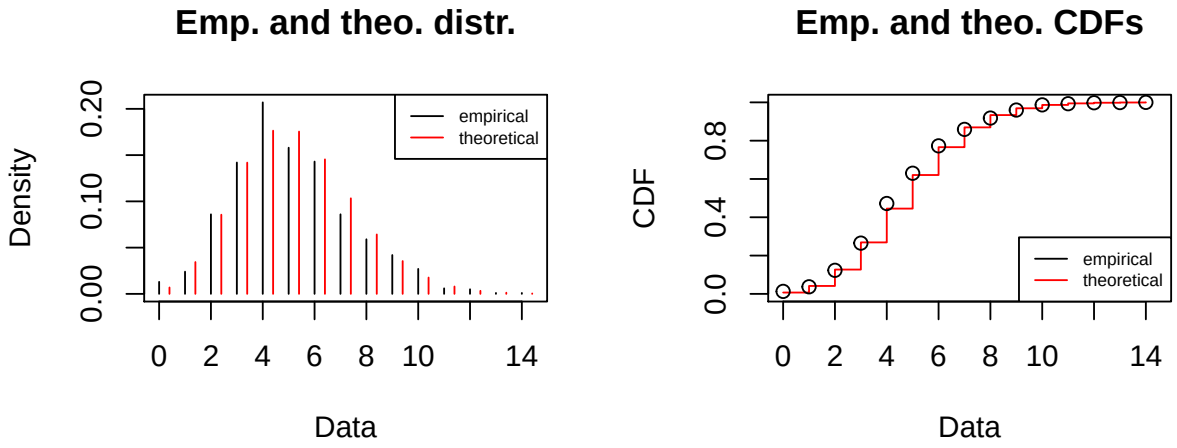



Emp. and theo. distr.
Density
0.0
0.2
0.4
0
2
4
6
8
10
Data
empirical
theoretical
Emp. and theo. CDFs
CDF
0.0
0.4
0.8
0
2
4
6
8
10
Data
empirical
theoretical


Histogram of n=100
Frequency
0
10
20
2
4
6
8
10
my_poisson

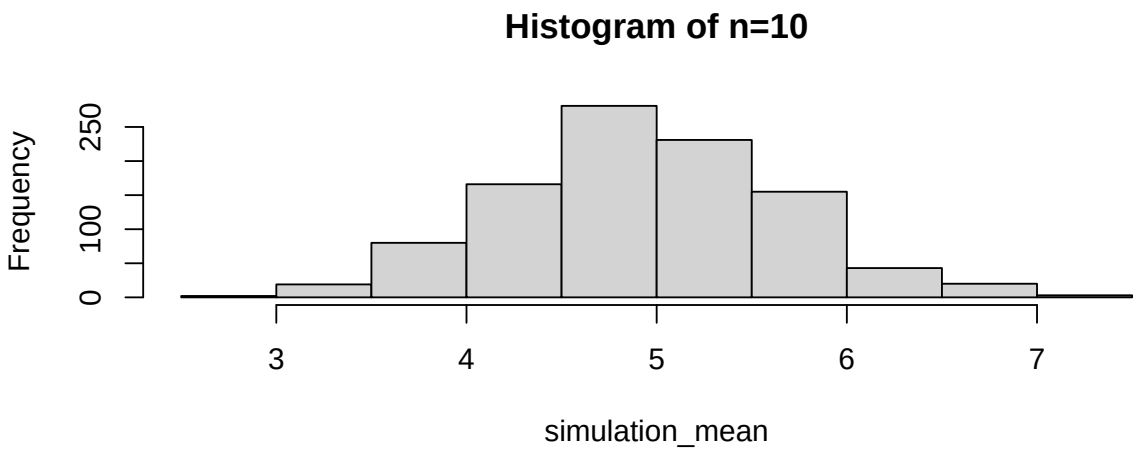

Emp. and theo. distr.
Density
0.00
0.15
2
4
6
8
10
Data
empirical
theoretical
Emp. and theo. CDFs
CDF
0.0
0.4
0.8
2
4
6
8
10
Data
empirical
theoretical

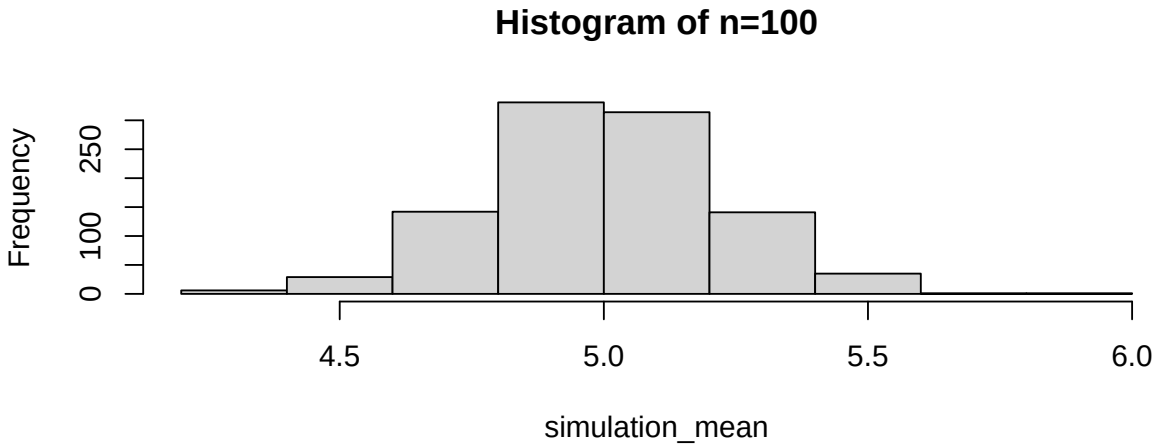


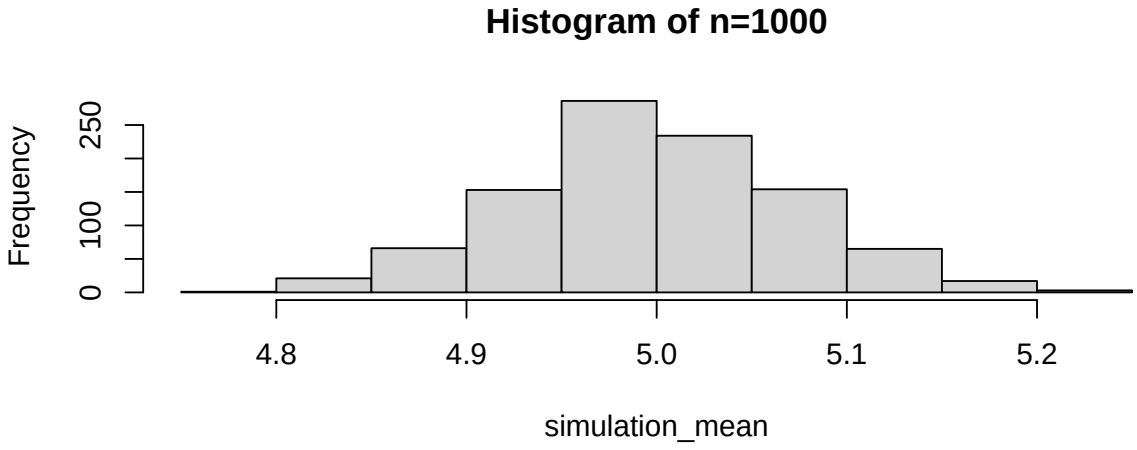


b)

```r
for (n in c(10, 100, 1000))
{
  simulation <- lapply(1:1000, function(x) rpois(n, 5))
  simulation_mean <- sapply(simulation, mean)
  hist(simulation_mean, main=paste("Histogram of n=", n, sep=""))
}
```


Histogram of n=10
Frequency
0
100
250
3
4
5
6
7
simulation_mean


Histogram of n=100
Frequency
0
100
250
4.5
5.0
5.5
6.0
simulation_mean


Histogram of n=1000
Frequency
0
100
250
4.8
4.9
5.0
5.1
5.2
simulation_mean

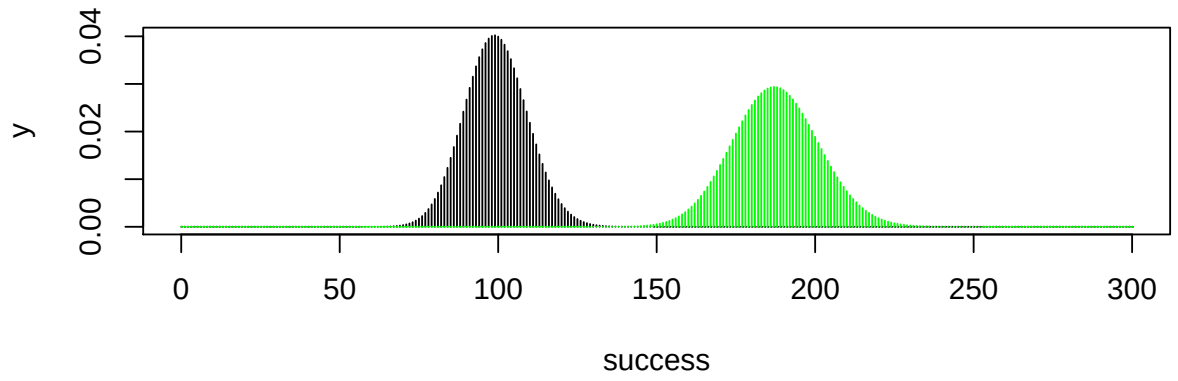
c)

As N increases, we can see it converges towards the Normal Distribution. Confirms the Central Limit Theorem.

Statistical approach and model

First group: $X_1(x), ..., X_n(x) \sim Binomial(n, p)$, where n=11,034 and $p = \frac{189}{11034} = 0.017$.
Second group: $Y_1(x), ..., Y_n(x) \sim Binomial(n, p)$, where n=11,037 and $p = \frac{104}{11037} = 0.009$.

```
success <- 0:300
x <- dbinom(success, size=11034, prob=.017)
y <- dbinom(success, size=11037, prob=.009)
plot(success, y, type='h')
lines(success, x, type='h', col="green")
```



We can see from the graph that the distributions look like two distinct distributions. With this, qualitatively we can say that taking aspirin has a significant and independent effect, but to quantify this we used several methods: zscore and confidence interval. We could have also done a chi squared test with one degree of freedom, using the summation of both distributions.

Z-score method

We say that our control group is our 'true' group, that converged close to the true values of μ and σ.

$$H_0 = \text{we dont observe a difference}$$

$$H_a = \text{there is a difference}$$

$$
\begin{aligned}
P(X \leq 104) \sim Bin(11034, 0.017) &= P(Z \leq \frac{104-\mu}{\sigma}) \\
&= P(Z \leq \frac{104-189}{\sqrt{n*p*q}}) \\
&= P(Z \leq \frac{104-189}{\sqrt{11034*0.017*\frac{(11034-189)}{11034}}}) \\
&= P(Z \leq \frac{104-189}{13.615}) \\
&= P(Z \leq -6.243) \\
&\approx 0 <= \alpha(0.05) \text{ we can reject the null hypothesis.}
\end{aligned}
$$

Taking an aspirin is statistically significant to not taking aspirin, we can reject the null hypothesis.

Q2 Confidence Interval method

$n_1 = 104, n = 11037, \hat{p} = \frac{104}{11037} = 0.0094, z = 1.96$ when $\alpha = 0.05$.

Wald Theorem:

$$
\begin{aligned}
\hat{p} \pm Z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= 0.0094 \pm 1.96\sqrt{(\frac{0.0094(0.9906)}{11037})} \\
&= (0.007599, 0.112) \text{ with } 95\%\alpha = 0.05 \\
&= (85.855, 123.58) \text{ multiplied by } 11037
\end{aligned}
$$

Thus, we can say with 95% certainty that the true value of H_a is between (85.855, 123.58), far away from the value 189.

Algorithm for simulation

a)

General Quantile Formula.

$$Q(P) = inf\{x \in R : P \leq F(x)\}$$

Quantile Function of the Uniform Distribution.

$$F^{-1}(U) = Q(U) = inf\{x \in R : U \leq F(x)\}$$

b)

Question is asking how to utilize a sample from a uniform distribution in an exponential one.

The Exponential Distribution,

$$F(X) = 1 - e^{-\lambda x}$$

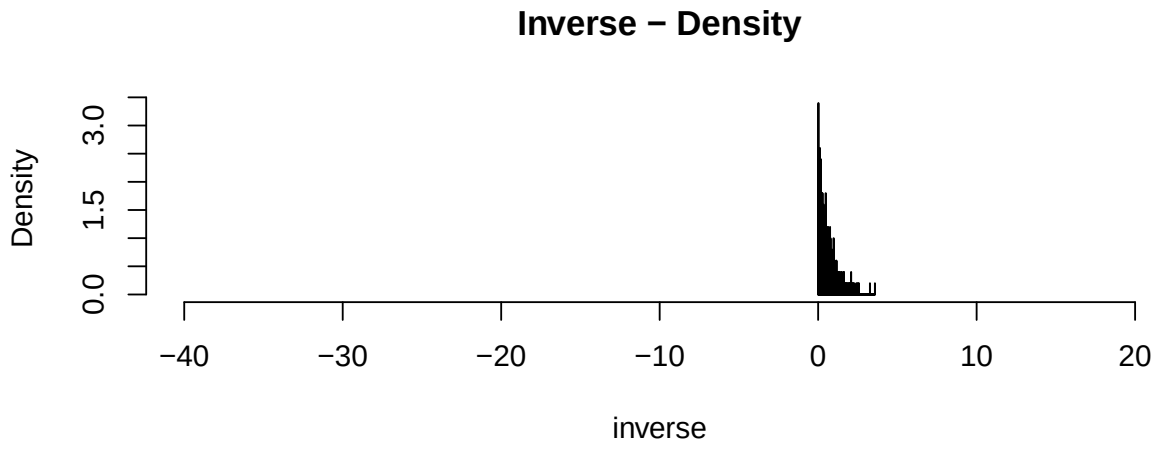
We have that,

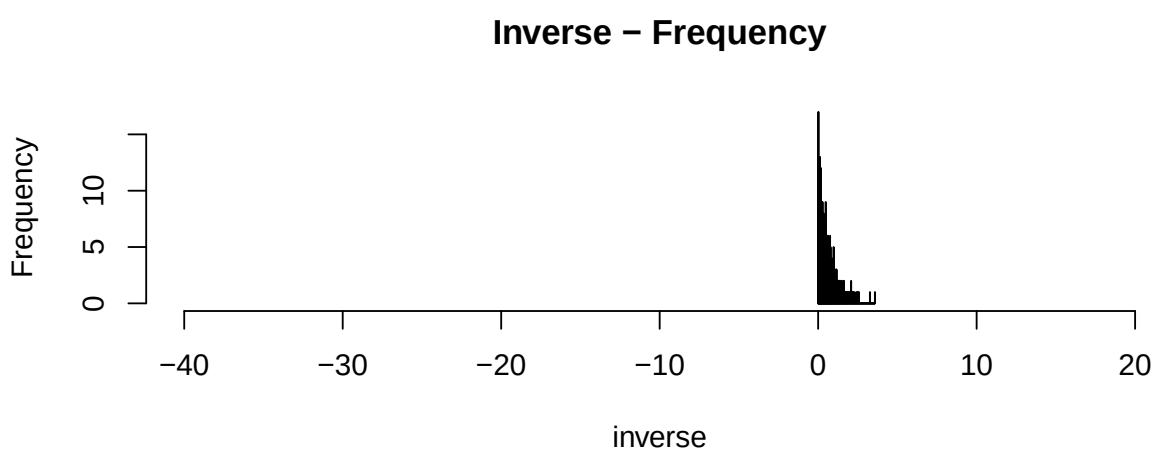
$$
\begin{aligned}
F^{-1}(U) &= inf\{x \in R : U \leq F(x)\} \\
&= inf\{x \in R : U \leq 1 - e^{-\lambda x}\} \\
&= inf\{x \in R : 1 - U \geq e^{-\lambda x}\} \\
&= inf\{x \in R : log(1 - U) \geq log(e^{-\lambda x})\} \\
&= inf\{x \in R : log(1 - U) \geq -\lambda x\} \\
&= inf\{x \in R : \frac{log(1 - U)}{\lambda} \geq -x\} \\
&= inf\{x \in R : -\frac{log(1 - U)}{\lambda} \leq x\} \\
&= -\frac{log(1 - U)}{\lambda}
\end{aligned}
$$

Our Algorithm 1. Create a Sample from a Uniform Distribution.
2. Use the sample data as input for the $F^{-1}(U)$.

c)

```
exp <- function(lambda=2, size=1000)
{
  # Create sample size
  sample <- runif(size)
  # Create sample using the inverse
  inverse <- -(log(1-sample)/lambda)
  exp <- rexp(size, rate=5)
  #hist(inverse, prob=TRUE, main="Inverse")

  # Density Plot
  hist(inverse, prob=TRUE, main="Inverse - Density", breaks=1000, xlim=c(-40, 20))

  # Frequency
  hist(inverse, freq=TRUE, main="Inverse - Frequency", breaks=1000, xlim=c(-40, 20))
}
exp(2, 1000)
```





d)

Our Algorithm 1. Create a Sample from a Uniform Distribution.
2. Use the sample data as input for the $F^{-1}(U)$.

We have the PMF of the Cauchy distribution

$$f(x) = \frac{1}{\pi\gamma}\frac{\gamma^2}{(x-x_0)^2+\gamma^2}$$

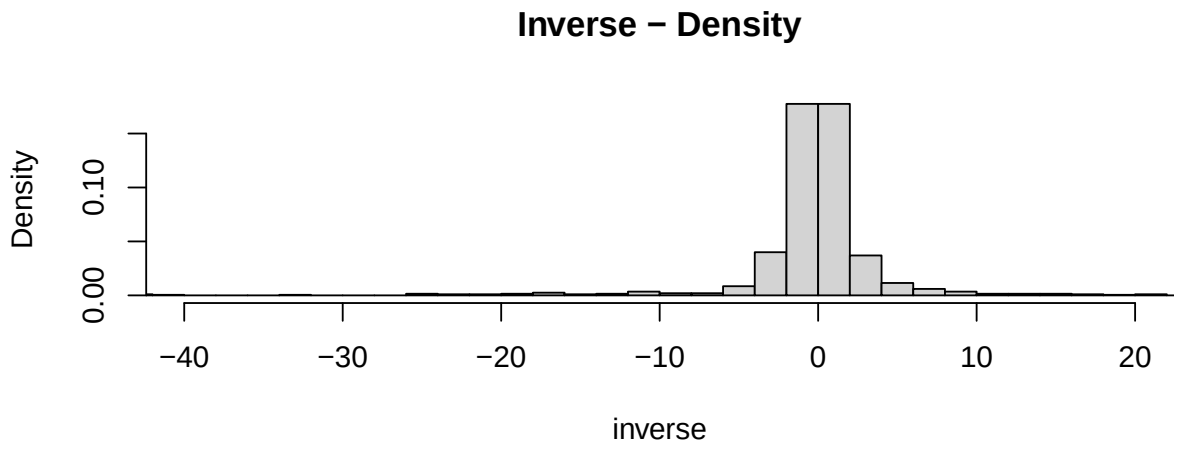
CDF of the Cauchy distribution

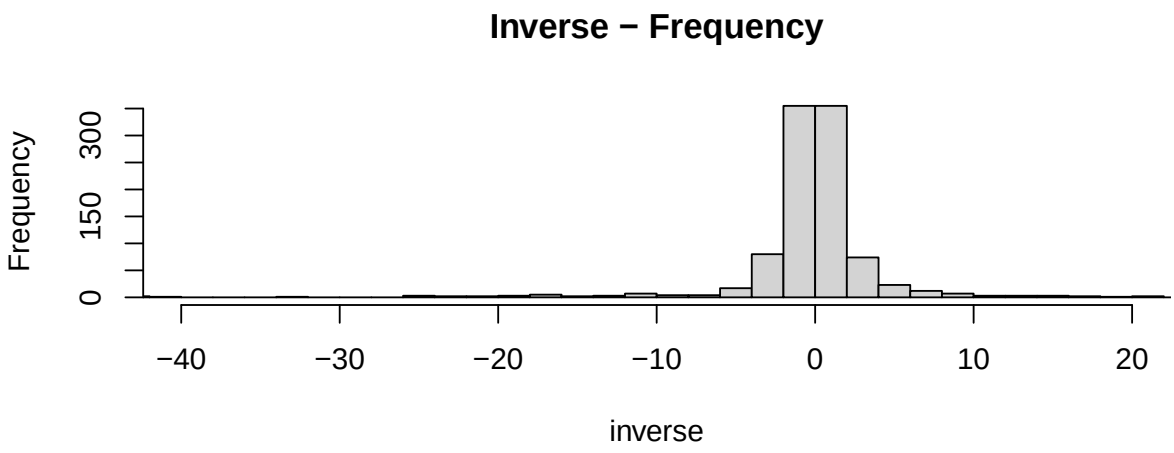
$$F(x) = \frac{1}{\pi}arctan(\frac{x-x_0}{\gamma}) + \frac{1}{2}$$

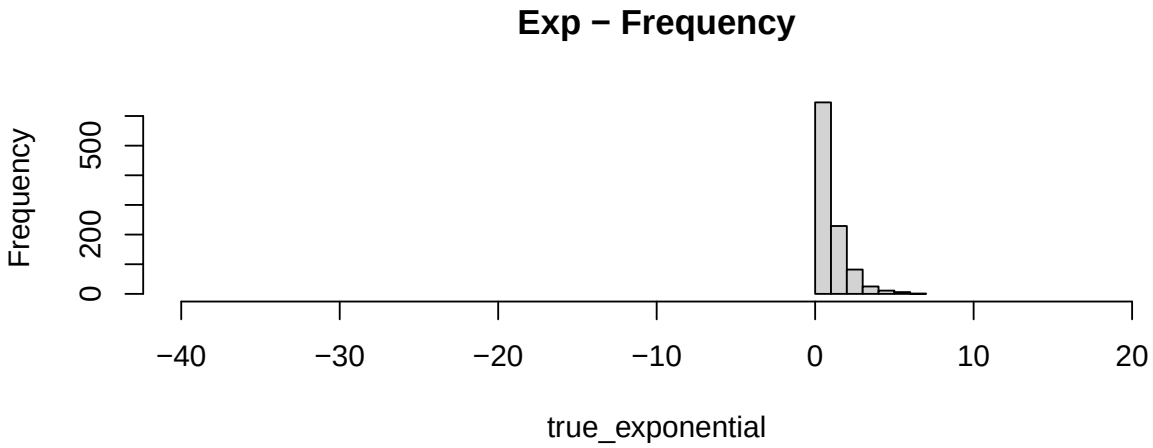
We have that,

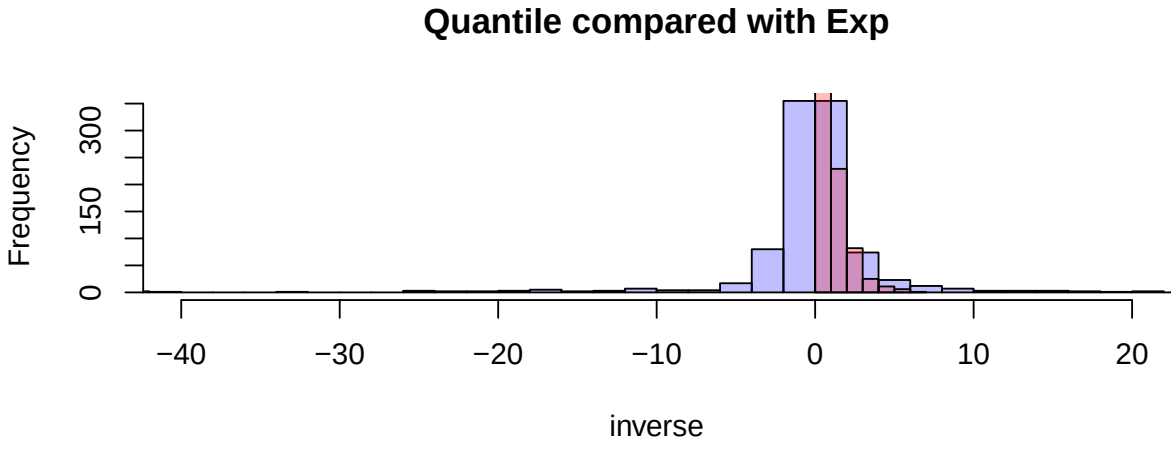
$$\begin{aligned}
F^{-1}(U) &= inf\{x \in R : U \leq F(x)\} \\
&= inf\{x \in R : U \leq \frac{1}{\pi}arctan(\frac{x-x_0}{\gamma}) + \frac{1}{2}\} \\
&= inf\{x \in R : U - \frac{1}{2} \leq \frac{1}{\pi}arctan(\frac{x-x_0}{\gamma})\} \\
&= inf\{x \in R : \pi(U - \frac{1}{2}) \leq arctan(\frac{x-x_0}{\gamma})\} \\
&= inf\{x \in R : tan(\pi(U - \frac{1}{2})) \leq tax(arctan(\frac{x-x_0}{\gamma}))\} \\
&= inf\{x \in R : tan(\pi(U - \frac{1}{2})) \leq \frac{x-x_0}{\gamma}\} \\
&= inf\{x \in R : \gamma tan(\pi(U - \frac{1}{2})) \leq x - x_0\} \\
&= inf\{x \in R : \gamma tan(\pi(U - \frac{1}{2})) + x_0 \leq x\} \\
&= \gamma tan(\pi(U - \frac{1}{2})) + x_0
\end{aligned}$$

```r
cauchy <- function(x_0=0, gamma=1.0, size=1000)
{
  # Create uniform sample size
  sample <- runif(size)

  # Create sample using the inverse
  inverse <- gamma * tan(pi*(sample - 1/2)) + x_0

  # True Exponential distribution
  true_exponential <- rexp(size, rate=gamma) + x_0

  # Density Plot
  hist(inverse, prob=TRUE, main="Inverse - Density", breaks=1000, xlim=c(-40, 20))

  # Frequency
  p1 <- hist(inverse, freq=TRUE, main="Inverse - Frequency", breaks=1000, xlim=c(-40, 20))
  p2 <- hist(true_exponential, freq=TRUE, main="Exp - Frequency", breaks=5, xlim=c(-40, 20))

  # Comparisson
  plot(p1, col=rgb(0,0,1,1/4), xlim=c(-40, 20), main="Quantile compared with Exp")
  plot(p2, col=rgb(1,0,0,1/4), xlim=c(-40, 20), add=T)
}

cauchy(0, 1.0, 1000)
```




Inverse − Frequency
Frequency
0
150
300
−40
−30
−20
−10
0
10
20
inverse


Exp − Frequency
Frequency
0
200
500
−40
−30
−20
−10
0
10
20
true_exponential


Quantile compared with Exp
Frequency
0
150
300
−40
−30
−20
−10
0
10
20
inverse

Q3 Estimation of a agricultural area

a)

$$\{N(\mu, \sigma^2), (\mu, \sigma^2) \in R^2\}$$

From the fact that σ^2 needs to be greater than zero, we can conclude that the model is identifiable. It implies that it is injective.

b) Quantity to be estimated (area of the field):

$$X_n = distance + \epsilon$$

The mean is our distance, thus we have a distribution $N(distance, \sigma^2)$.

$$Area = Distance^2$$

Thus, our area is

$$h(x) = x^2$$

Our mean distance is,

$$E[Distance] = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2}{2}$$

Thus, our area estimator is of two observations is:

$$h(E[Distance]) = (\frac{x_1 + x_2}{2})^2$$

We choose this as our estimator because, although s4 is unbiased compared to s5, s5 has a much faster rate of convergence. We can see, through an application of the continuous map theorem, where the MSE of s4 is

$$\frac{\sigma^2}{n}$$

and s5 =

$$\frac{\sigma^4}{n^2}$$

$$4\sigma^2\mu^2$$

, assuming there is no degenerate solution

$$\mu = 0$$

(in which case we just use the standard gaussion distribution under CLT with

$$N(0, \sigma^2)$$

), c) Generalized estimator:

$$(\frac{1}{n}\sum_{i=1}^{n} x_i)^2 = \frac{1}{n}(\sum_{i=1}^{n} x_i)^2$$

Since our sample is finite, and the fact that the distance $E[Distance] = distance$ and area $h(x) = x^2$ we know that $T(\hat{F}_n) \to T(F_n)$ asymptotically. Thus, our estimator is strongly consistent.
Since both the distance and area is differentiable and we have an strongly consistent estimator we can use the Delta Method to tell that our estimator is asymptotically normal.
We have the Delta Method theorem:

$$\sqrt{(}n)\{g(\bar{X_n}) - g(\mu)\} \to N\{0, \sigma^2 g'(\mu)^2\}$$

Thus, its asymptotic variance is

$$Var(X) = 4\sigma^2\mu^2$$

.

Q4 Descriptive statistics

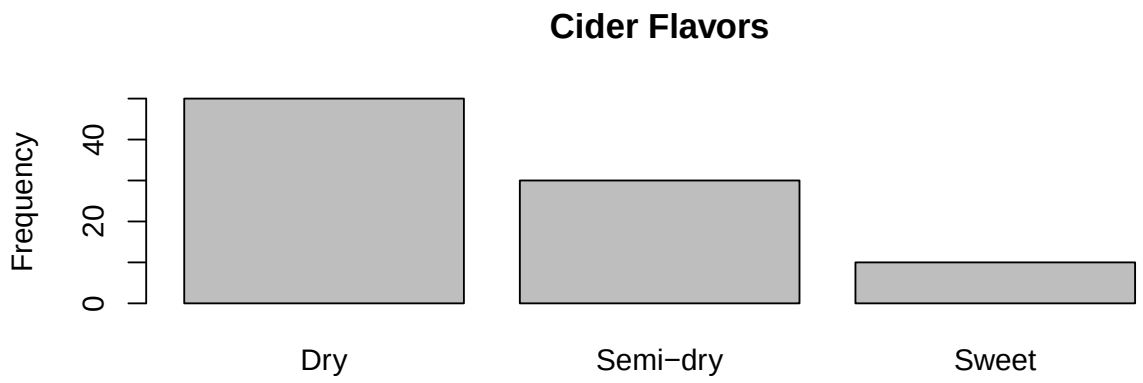
a)

```
df <- read.csv("cider.csv")
head(df)
```

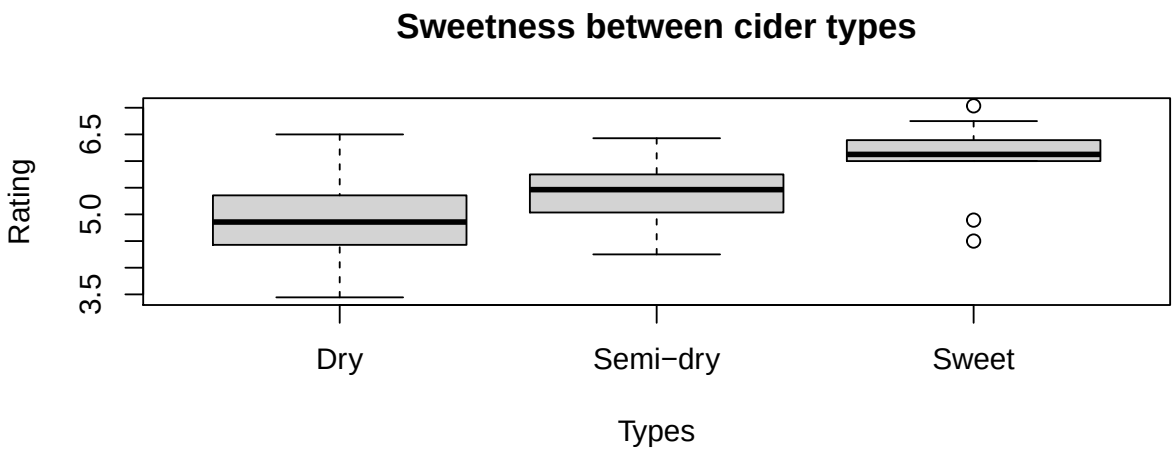
```
##   Type Sweetness     Acid Bitterness Astringent
## 1  Dry  4.678571 3.392857   5.857143  2.3214286
## 2  Dry  5.285714 4.285714   4.857143  2.7857143
## 3  Dry  6.500000 4.642857   2.357143  0.7142857
## 4  Dry  5.035714 5.714286   4.642857  3.8214286
## 5  Dry  4.571429 4.607143   4.321429  2.1785714
## 6  Dry  6.071429 3.250000   3.428571  1.1428571
```

b)

```
barplot(table(df$Type), main = "Cider Flavors", ylab = "Frequency")
```

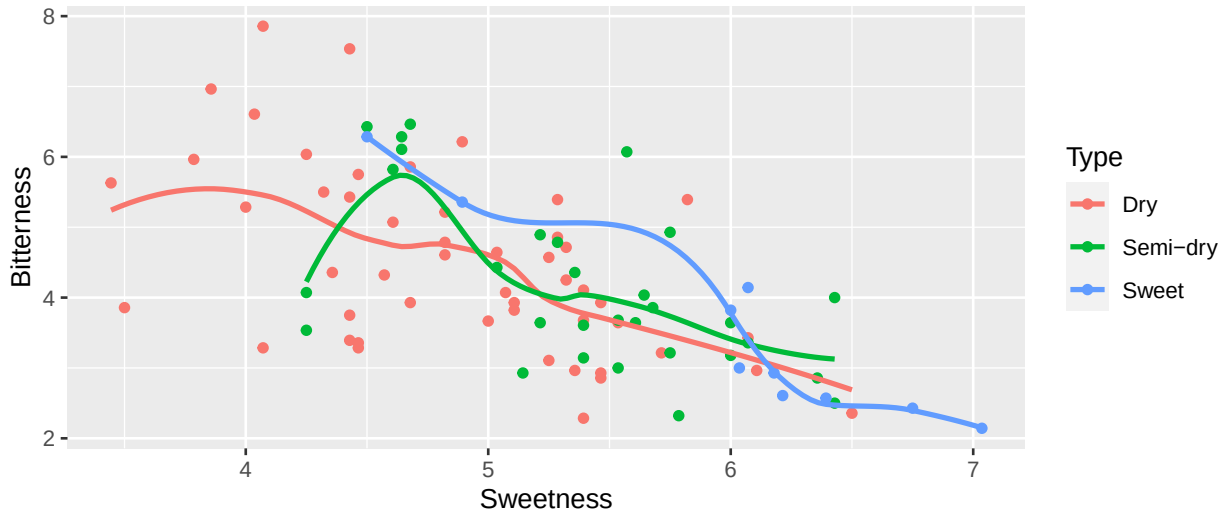



```
boxplot(df$Sweetness~df$Type, main = "Sweetness between cider types", xlab = "Types", ylab = "Rating")
```



d)

```
library(ggplot2)
sweet <- ggplot(df, aes(x = Sweetness, y = Bitterness, color = Type)) +
  geom_point() +
  geom_smooth(method="loess", se=FALSE, fullrange=FALSE, level=0.95, formula="y ~ x")
sweet
```



e) We can see that the correlation between bitterness and sweetness is negativly correlated. The sweeter, the less bitter.