STAT 111 Section 01: Review and Introduction

Al Xin

1/25/2021

Problems

Problem 1 (Cauchy distribution)

(a) Representation of the Cauchy distribution *How is the Cauchy distribution represented?*

Let $N_1, N_2 \overset{i.i.d}{\sim} \mathcal{N}(0,1)$. The r.v. $X \sim \frac{N_1}{N_2}$ has the Cauchy distribution.

(b) Practice with R *Write a function that generates an arbitrary number of observations from a Cauchy random variable using the representation found in (a). Then, plot a histogram of 50, 500, and 1000 observations from your function.*

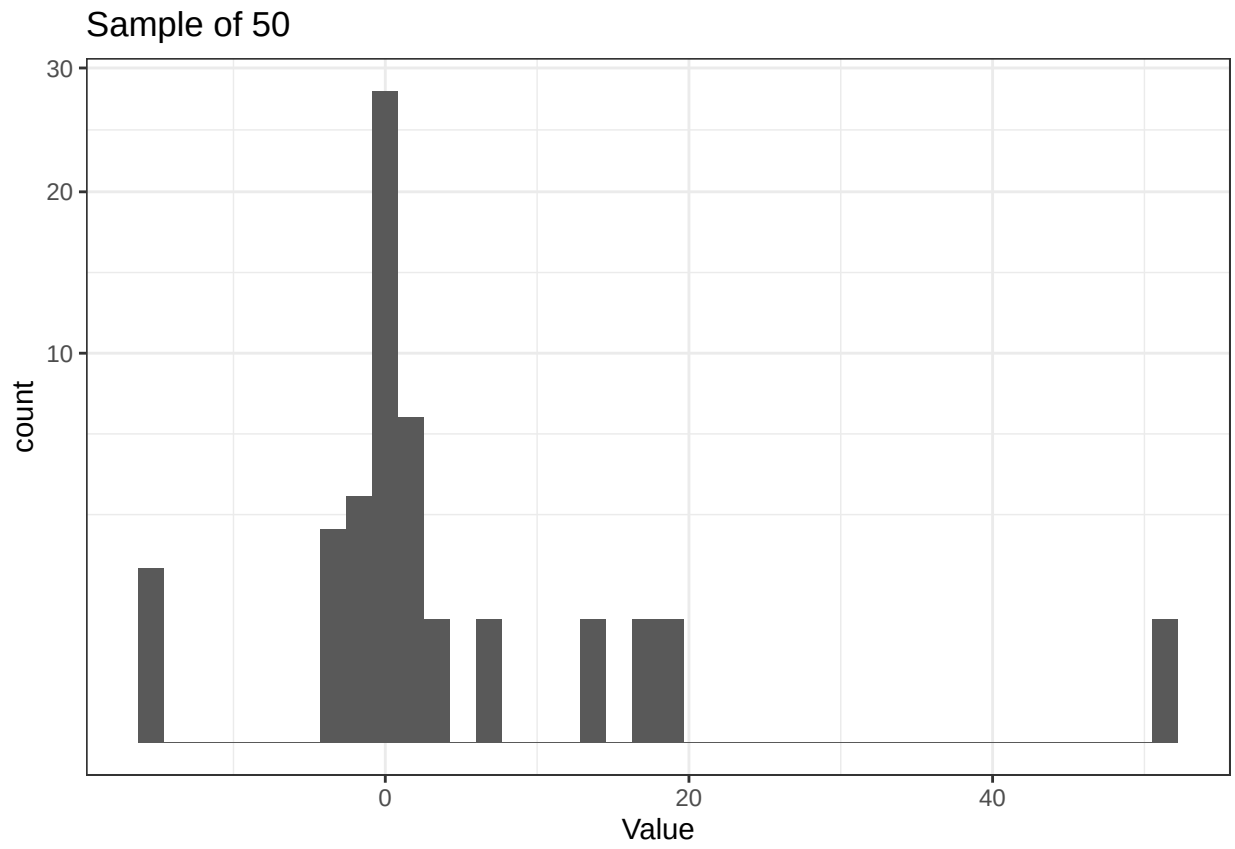
For future reference, the existing function is `rcauchy`*. Try to name a function that doesn't conflict with this one.*

```
# Vector solution (simplest)
rcchy <- function(x) {
  rnorm(x)/rnorm(x)
}

# Vector solution (clarity in r.v.s)
rcchy <- function(x) {
  n1 <- rnorm(x)
  n2 <- rnorm(x)
  n1/n2
}

# Replicate solution
rcchy <- function(x) {
  replicate(x, rnorm(1)/rnorm(1))
}
```

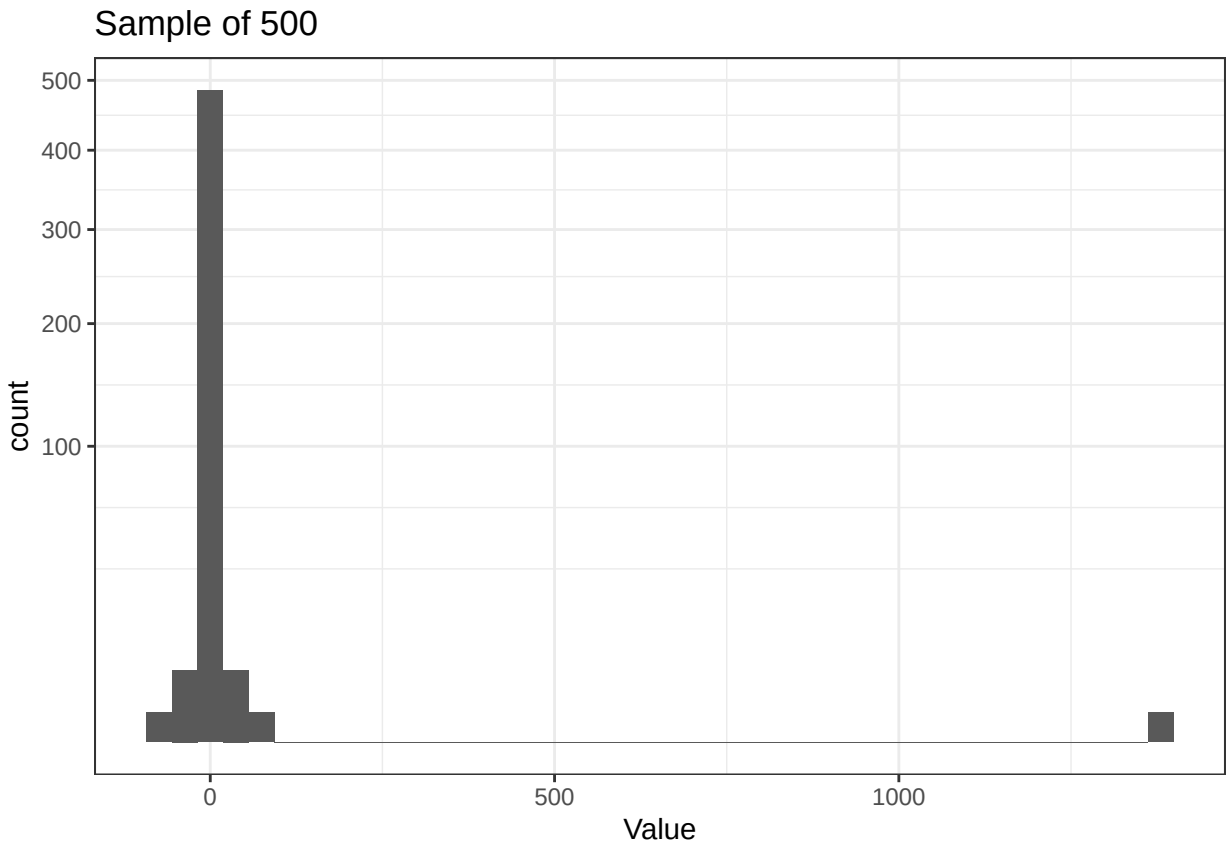
```
n_obs <- c(50, 500, 10000)

lapply(
  n_obs,
  function(x) {
    obs <- data.frame(obs = rcchy(x))
    ggplot(obs, aes(x = obs)) +
      geom_histogram(bins = 40) +
```

```
      labs(title = paste("Sample of", x), x = "Value") +
      scale_y_sqrt() +
      theme_bw()
  }
)
```

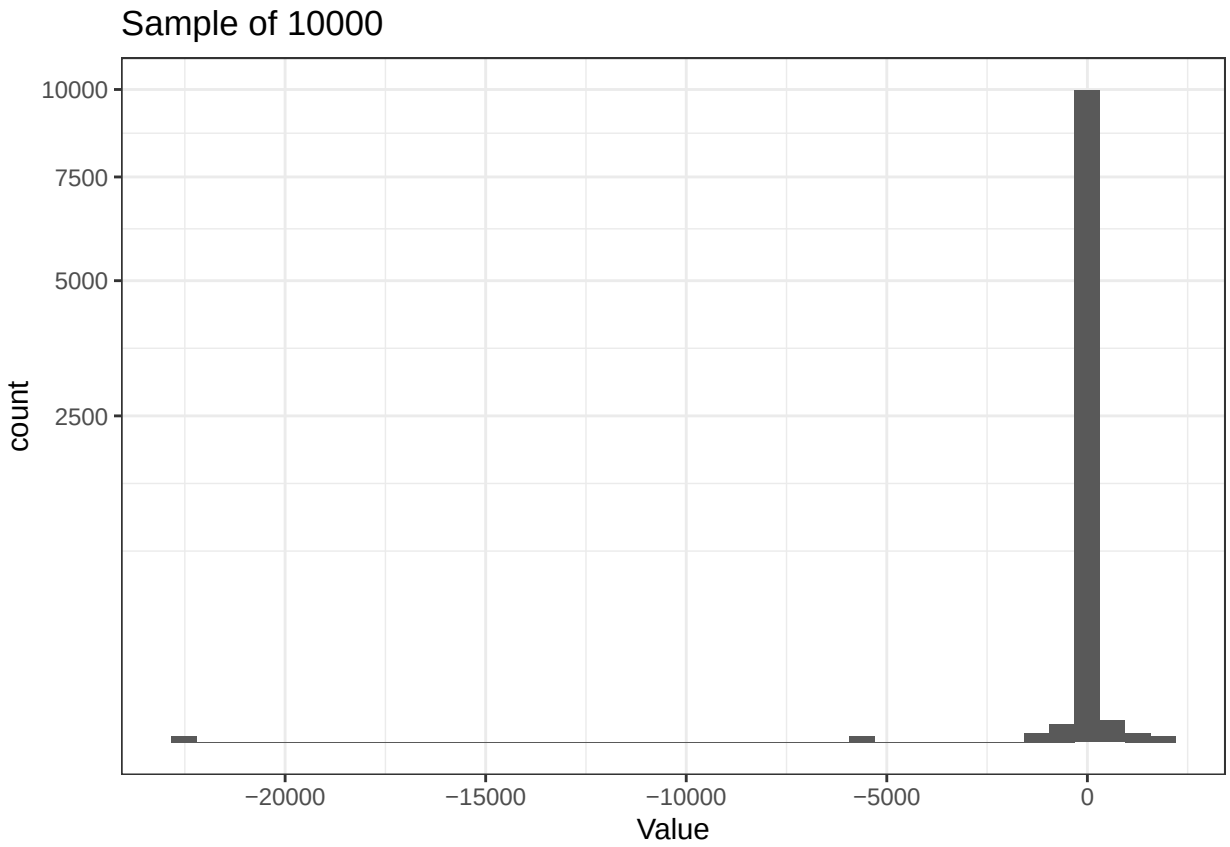
```
## [[1]]
```




```
##
## [[2]]
```



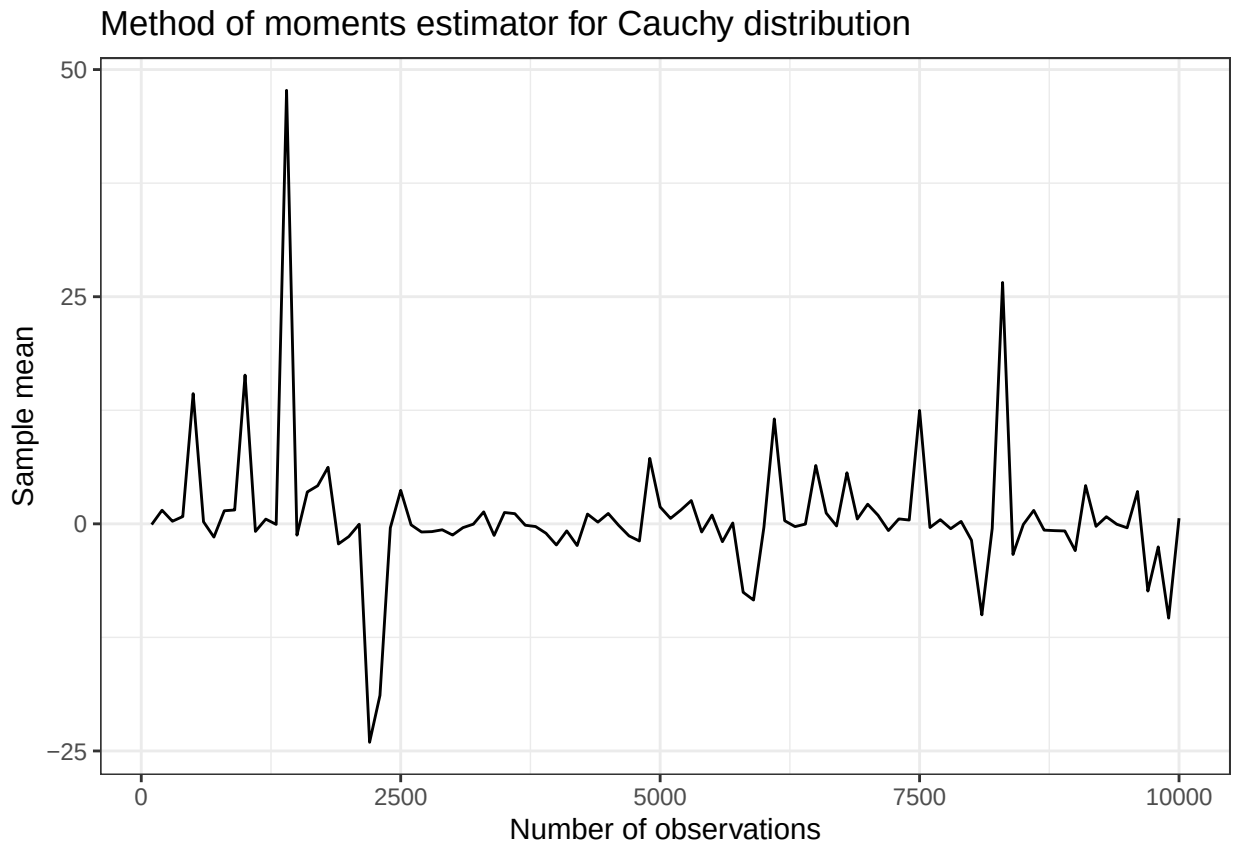

```
##
## [[3]]
```



(c) Visualize expectation of the Cauchy *In class, we discussed unbiased and consistent estimators. For example, by linearity of expectation, the method of moments estimator is unbiased and consistent. Test the behavior of the MoM estimator for the Cauchy distribution.*

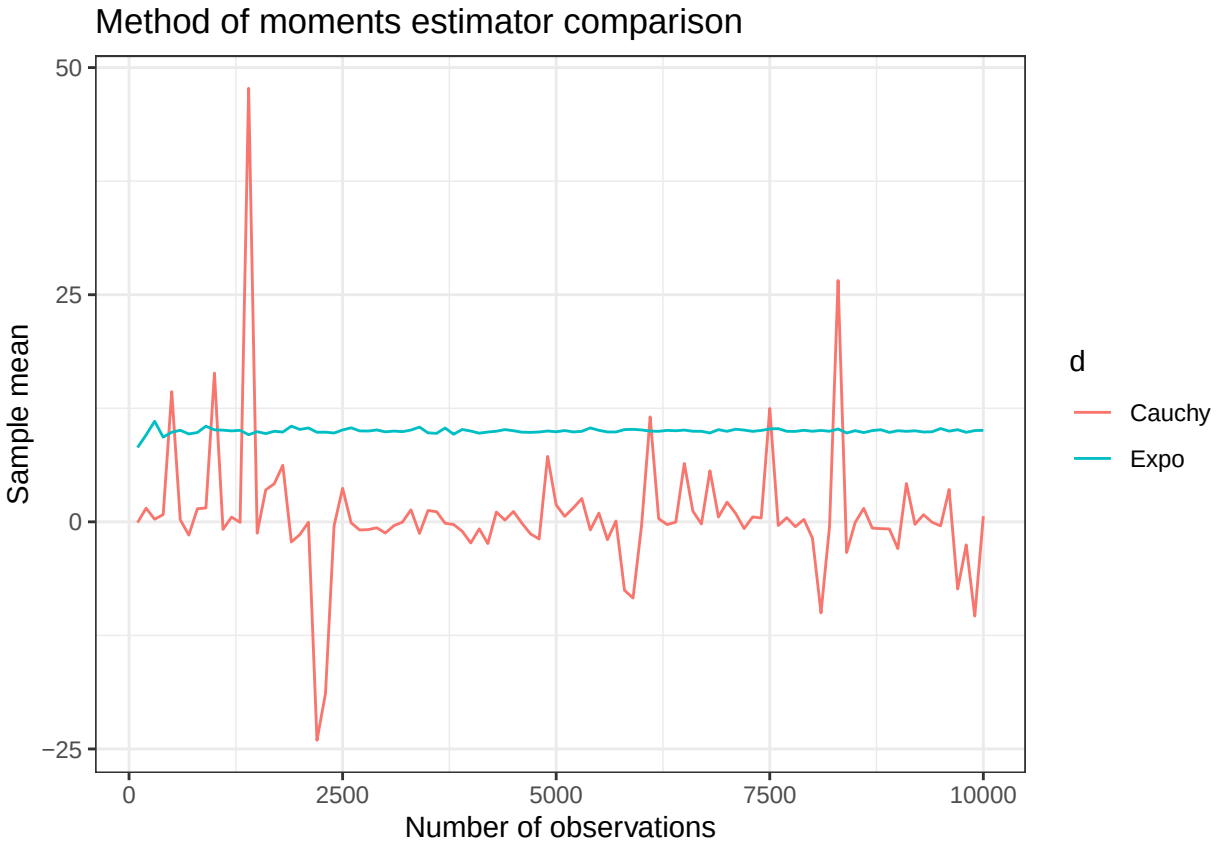
Generate samples of size $100, 200, \ldots, 10000$. *Plot the sample mean of the samples against the sample size in a line plot.*

```
set.seed(11101)
n_obs <- seq(100, 10000, 100)
cauchy_mom <- data.frame(
  n_obs = n_obs,
  samp_mean = sapply(n_obs, function(x) mean(rcchy(x)))
)
ggplot(cauchy_mom, aes(x = n_obs, y = samp_mean)) +
  geom_line() +
  labs(
    x = "Number of observations",
    y = "Sample mean",
    title = "Method of moments estimator for Cauchy distribution"
  ) +
  theme_bw()
```



(d) Method of moments comparison *Using the same sampling scheme, compare the behavior of the method of moments estimator for an Exponential distribution with rate parameter* `0.1`*. (Review: How does this compare to a standard Exponential distribution? The notation in R can be unintuitive.)*

```
set.seed(11101)
expo_mom <- data.frame(
  n_obs = n_obs,
  samp_mean = sapply(n_obs, function(x) mean(rexp(x, 0.1)))
) %>%
  mutate(d = "Expo")
cauchy_mom <- mutate(cauchy_mom, d = "Cauchy")
dist_mom <- rbind(expo_mom, cauchy_mom)

ggplot(dist_mom, aes(x = n_obs, y = samp_mean, color = d)) +
  geom_line() +
  labs(
    x = "Number of observations",
    y = "Sample mean",
    title = "Method of moments estimator comparison"
  ) +
  theme_bw()
```



(e) Challenge: Multiple iterations *Repeat (c), but run multiple simulations of the procedure and plot multiple lines on the graph.*

Problem 2 (Medical diagnosis)

This problem will be a review of conditioning

We are trying to estimate what proportion of the Harvard body (staff, grad students, and undergraduates) have COVID-19. The administration conducts frequent tests to determine the prevalence and incidence of infection.

From Jan 26-28, there have been 8,750 tests. There has been a total of 27 positive cases.

(a) Perfect test *Assume that the test is perfect and that whether a person tests positive is distributed i.i.d. Bernoulli with proportion p.*

Determine the MLE and produce an estimate of p given the observations. Assume the data is in the vector $\vec{x}$ of length n.

Likelihood function:

$$L(p|\vec{x}) = \prod_{j=1}^{n} p^{x_j}(1-p)^{1-x_j} = p^k(1-p)^{n-k},$$

where k is $\sum_{j=1}^{n} x_j$.

Log-likelihood:

$$\ell(p|\vec{x}) = k\ln p + (n-k)\ln(1-p)$$

MLE after setting the derivative of the above to 0:

$$\frac{\partial \ell(p|\vec{X})}{\partial p} = \frac{k}{p} + \frac{k-n}{1-p} = 0$$
$$-k + pk = pk - np$$

$$\hat{p} = \frac{k}{n} = \frac{27}{8750} \approx 0.00309.$$

Note that this is the likelihood function of a Binomial distribution.

Note: Why is i.i.d. Bernoulli a poor choice of model here?

Whether someone tests positive for an infectious disease is not independent of the probability that someone in close proximity also tests positive.

(b) Imperfect test *The sensitivity of a test is the probability that someone who has the condition tests positive.*

The specificity of a test is the probability that someone who does not have the condition tests negative.

Let sensitivity be a and specificity by b. Correct the likelihood function to accomodate for error in the test.

Calculate the probability someone tests positive $P(T)$ with the probability that someone has COVID-19, $P(D)$.

$$P(T) = P(T|D)P(D) + P(T|D^c)P(D^c) = ap + (1-b)(1-p) = 1 + p(a+b-1) - b.$$

Replace p in the likelihood function with $P(T)$.

$$L(p|\vec{x}) = (1 + p(a+b-1) - b)^k (b - p(a+b-1))^{n-k}.$$

Problem 3 (Sunspots)

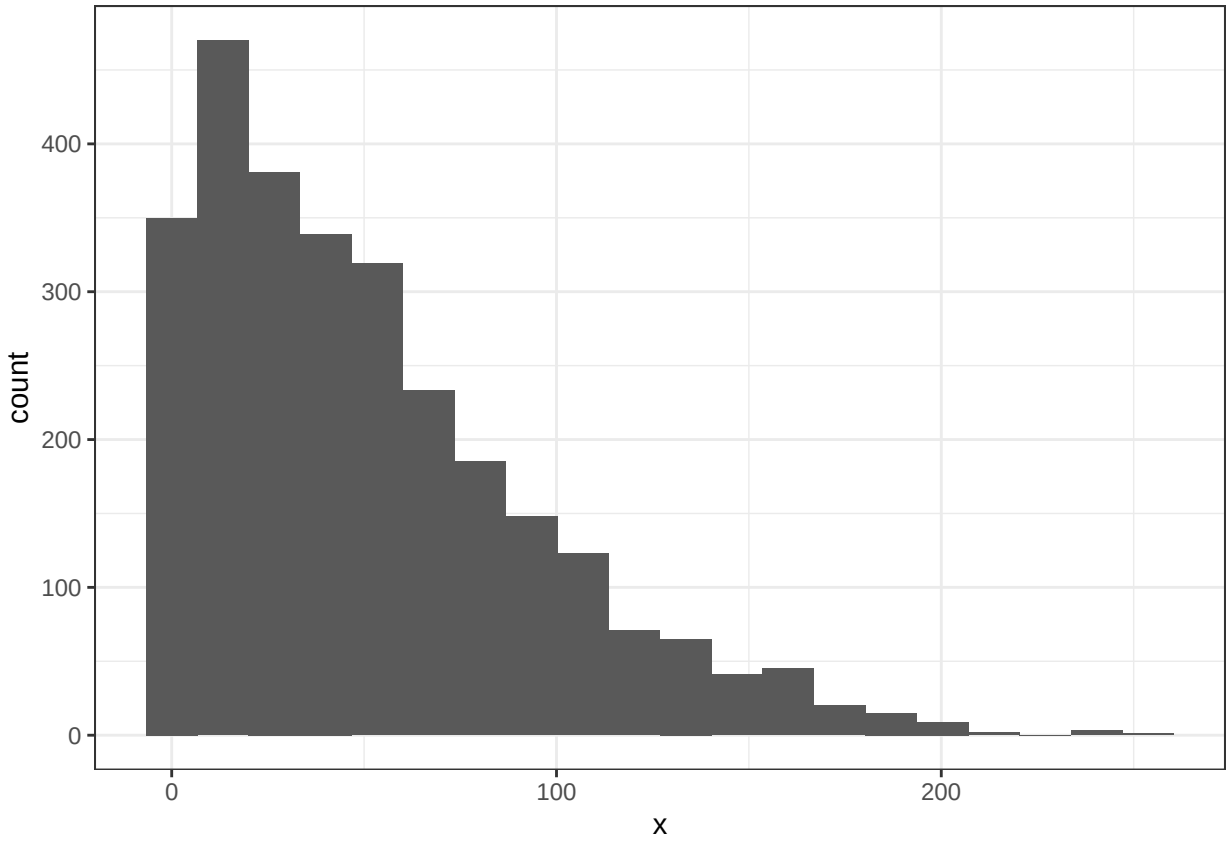
You can see what datasets R has available using the function `data()`

(a) Histogram and ECDF *Plot a histogram and the ECDF of the data* `sunspots` *from R.*

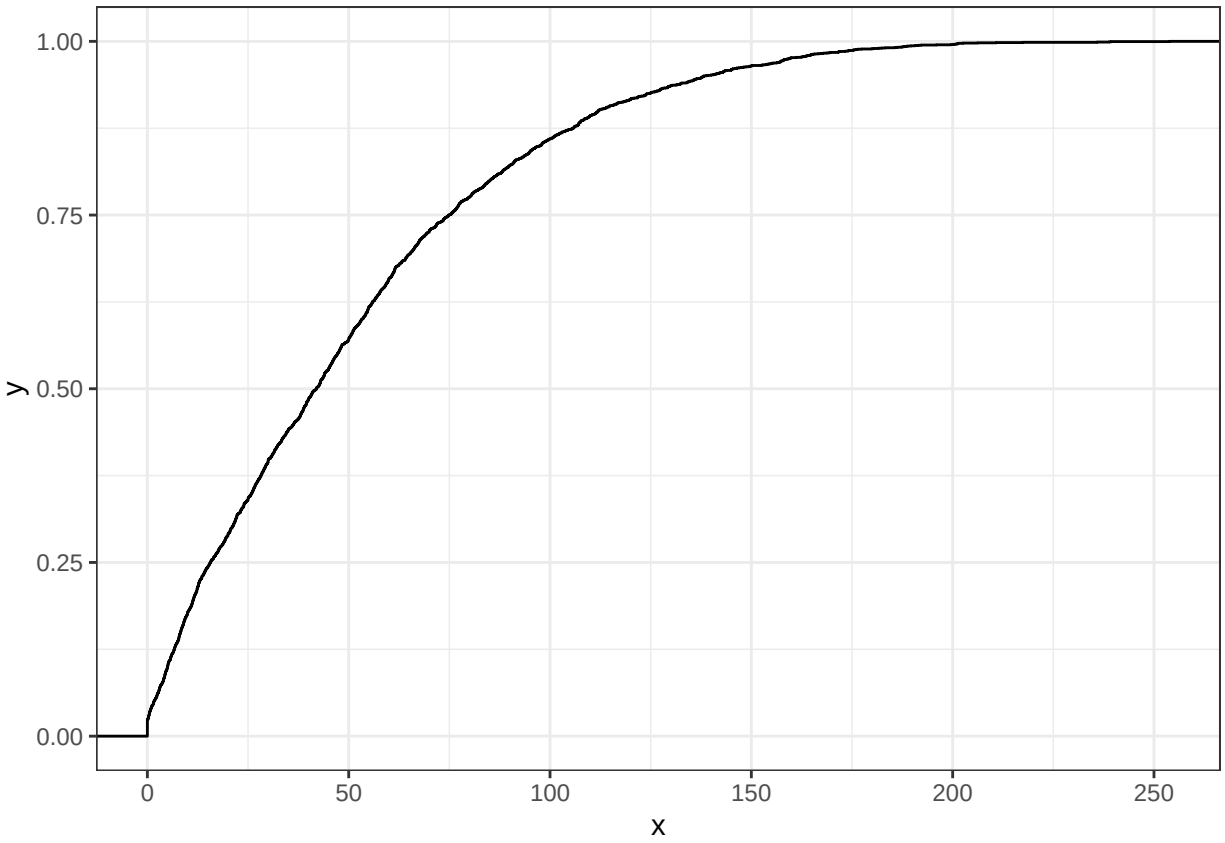
```
# Get an idea of the data
# Histogram
class(sunspots)
```

```
## [1] "ts"
```

```
sspots <- data.frame(x = as.numeric(sunspots))
ggplot(sspots, aes(x = x)) +
  geom_histogram(bins = 20)
```

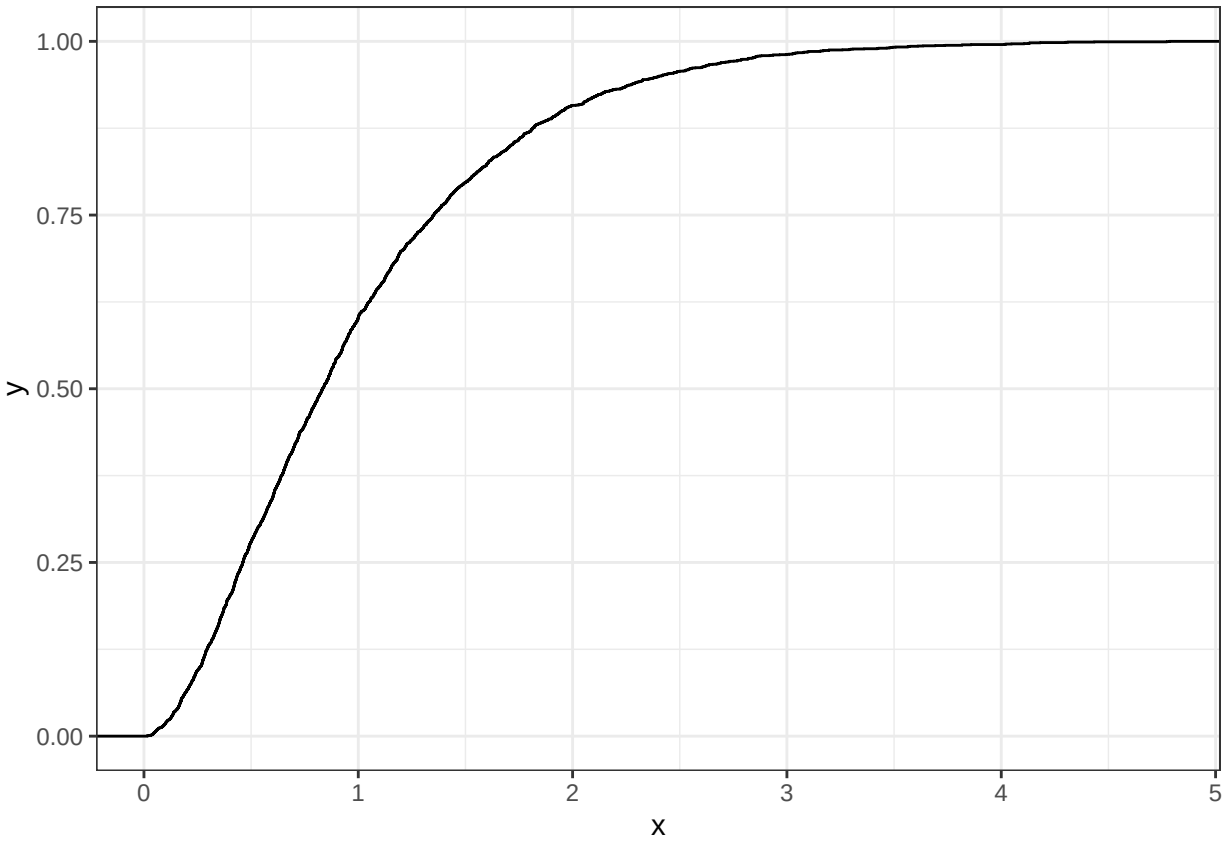



```
# ECDF
ggplot(sspots, aes(x = x)) +
  stat_ecdf(geom = "step")
```

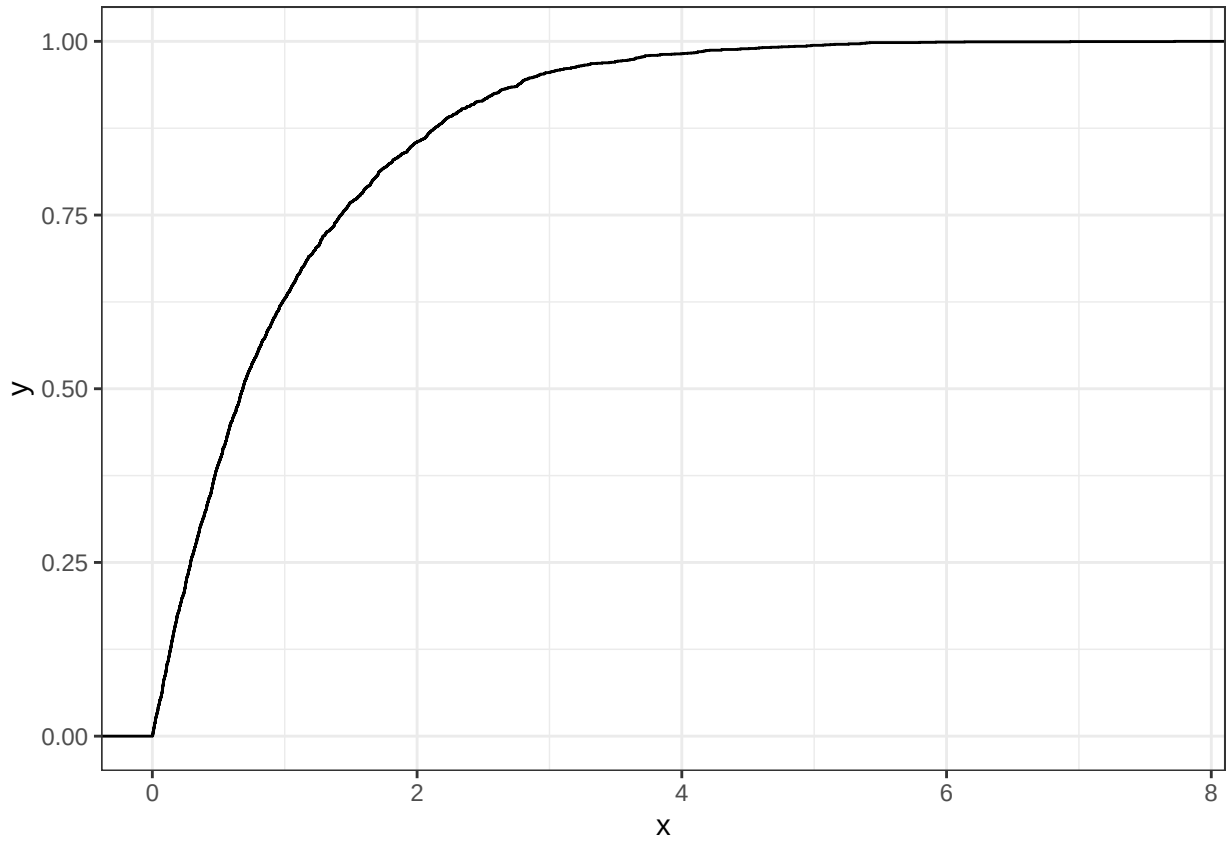


(b) Modeling on monthly sunspot distribution *What distribution does this remind you of? Pick one or more reasonable distributions and compare the ECDFs (number of observations in the sample should equal the number of observations of sunspots).*

```
gamma_df <- data.frame(x = rgamma(nrow(sspots), 2, 2))
ggplot(gamma_df, aes(x = x)) +
  stat_ecdf(geom = "step")
```




```
expo_df <- data.frame(x = rexp(nrow(sspots)))
ggplot(expo_df, aes(x = x)) +
  stat_ecdf(geom = "step")
```



(c) MLE *Assume that the data is distributed Exponential with an unknown rate parameter λ. Write the likelihood function and find the MLE.*

Assume that the likelihood and log-likelihood are conditioned on the observations $x_1, \ldots, x_n$.

The PDF of the Exponential distribution is $f(x) = \lambda e^{\lambda x}$

The likelihood function is

$$L(\lambda) = \prod_{j=1}^{n} \lambda e^{\lambda x_j} = \lambda^n \exp\left(-\lambda \sum_{j=1}^{n} x_j\right).$$

The log-likelihood is given by

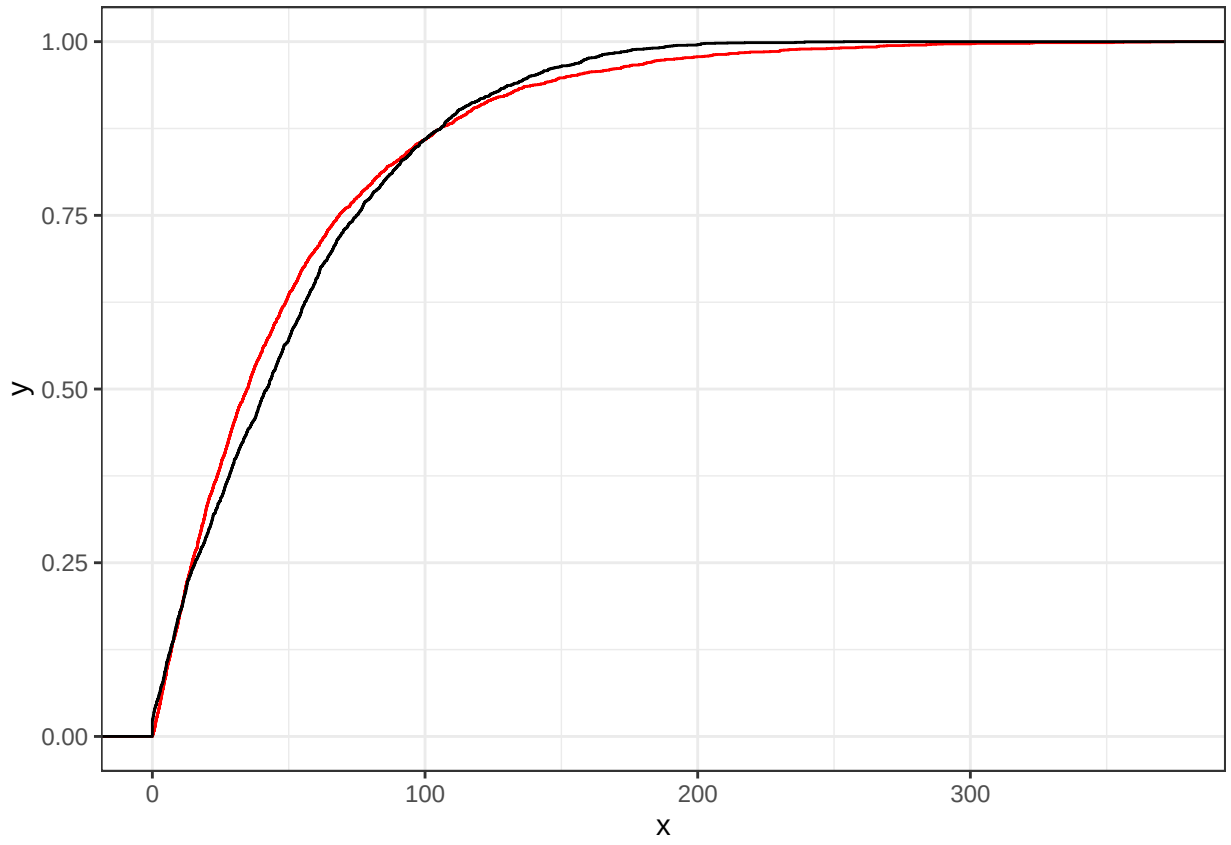
$$\ell(\lambda) = n \ln(\lambda) - \lambda \sum_{j=1}^{n} x_j.$$

Setting the derivative of the above equal to zero yields the MLE

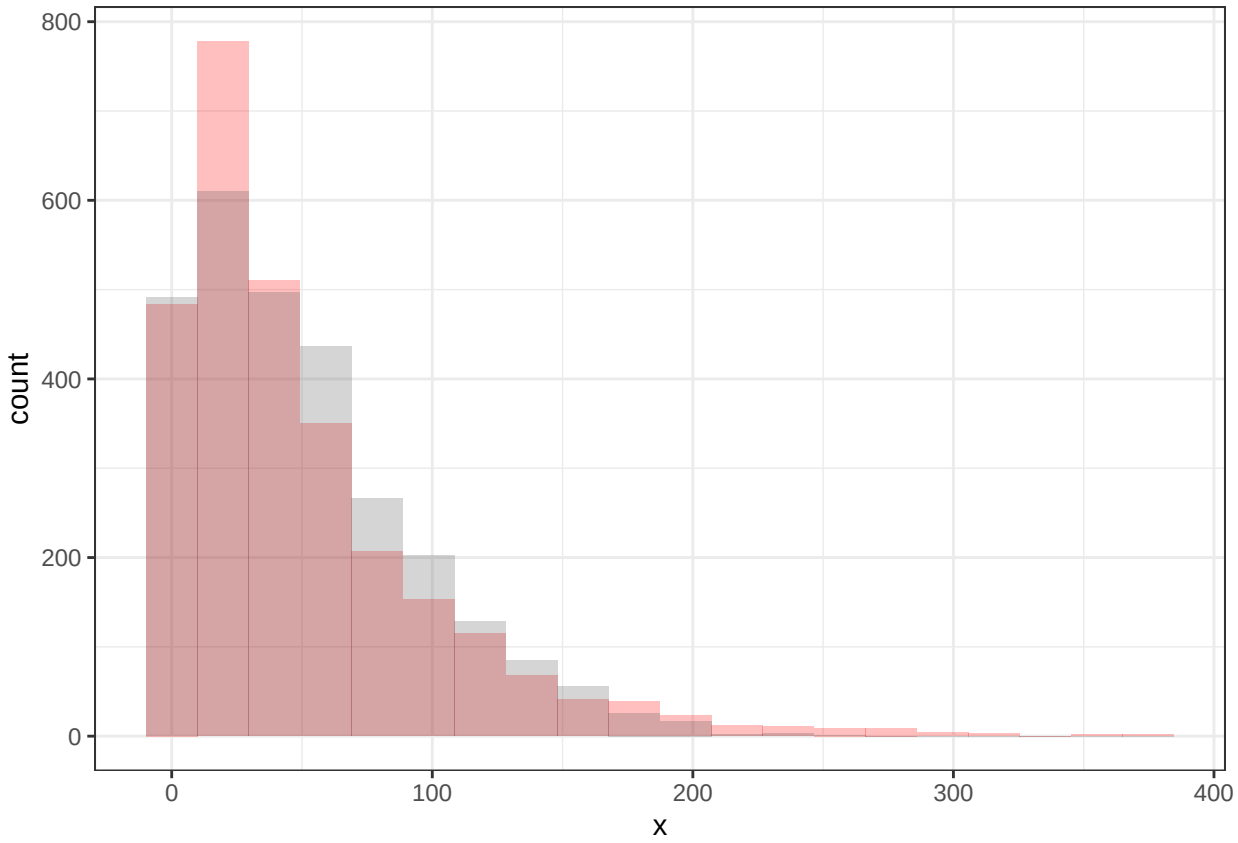
$$\hat{\lambda} = \frac{n}{\sum_{j=1}^{n} x_j}$$

(d) MLE of the data *Calculate the estimate of the rate parameter based on the data. Compare the graphs of the histogram and ECDF of an Exponential model and the observations.*

```
lambda_mle <- nrow(sspots)/sum(sspots$x)
expo_df <- data.frame(x = rexp(nrow(sspots), lambda_mle))
ggplot(expo_df, aes(x = x)) +
  stat_ecdf(geom = "step", color = "red") +
  stat_ecdf(data = sspots, aes(x = x), geom = "step")
```




```
ggplot(expo_df, aes(x = x)) +
  geom_histogram(alpha = 0.25, fill = "red", bins = 20) +
  geom_histogram(data = sspots, aes(x = x), alpha = 0.25, bins = 20)
```


800
600
400
200
0
count
0
100
200
300
400
x

Problem 4 (DNA sequence)

A DNA sequence can be composed of four different possible base pairs. For example, consider the following sequence:

CTACCTTCAATTGCTGGAACG

(a) Multinomial model *For simplicity, assume that DNA base pairs are selected from a Multinomial distribution (ignore properties of DNA base pair-matching).*

Let the probabilities corresponding to the base pair selection be represented as $\theta = (p_a, p_c, p_g, p_t)$.

Write the log-likelihood of θ

Let W_x be the count of base-pairs with outcome x. We have that

$$(W_a, W_c, W_g, W_t) \sim \text{Multinomial}(n, \theta).$$

We have the following likelihood and log-likelihood expressions given observations w_x (and $x \in \{a, c, g, t\}$):

$$L(\theta) = \prod_{i=1}^{n} \prod_{x} p_x^{I_{x_i = x}} = \prod_{x} p_x^{w_x},$$

$$\ell(\theta) = \sum_{x} w_x \log p_x.$$

(b) Markov chain *We can use a Markov chain model to accommodate for possible violations of independence in the genome. Let us model the sequence as a lag-1 Markov model.*

In a lag-1 Markov model, we have $Pr(X_s | X_{s-1}, \ldots, X_1) = Pr(X_s | X_{s-1})$.

This model can be represented with a 4×4 *transition matrix:*

$$T = \begin{pmatrix} \tau_{aa} & \tau_{ac} & \tau_{ag} & \tau_{at} \\ \tau_{ca} & \tau_{cc} & \tau_{cg} & \tau_{ct} \\ \tau_{ga} & \tau_{gc} & \tau_{gg} & \tau_{gt} \\ \tau_{ta} & \tau_{tc} & \tau_{tg} & \tau_{tt} \end{pmatrix}$$

How many parameters does this matrix have?

12. The rows and columns must sum to 1.

Assume that the marginal distribution $Pr(X_1 = x), x \in \{t, c, g, a\}$ *is known. Write the log-likelihood of this model.*

The joint likelihood of the data (stored in $\vec{x}$)

$$f_{\vec{X}}(\vec{x}|T) = f(x_1|T) \prod_{i=2}^{10} f(x_t | x_{t-1}, T) = P(X_1 = x_1|T) \prod_{i=2}^{10} P(X_i = x_i | X_{i-1} = x_{i-1}, T)$$

where we have

$$P(X_i = y | X_{i-1} = x, T) = \tau_{xy}.$$

Recall that we assume we know $f(x_1|T) = P(X_1 = x_1|T)$.

Let W_{xy} be the number of observations where the base pair x is followed by y.

We then have the log-likelihood

$$\ell(T|\vec{x}) = \sum_{x,y} w_{xy} \log \tau_{xy} : x, y = a, c, g, t.$$

This is constrained so that each row of T sums to 1.

The summarizing statistic is the number of each type of base pair change.