STARTed @ last 10 mins of Lecture on Wed 10/20

START Friday 10/22/2021

1 *Probability Models*

In the first four chapters we covered probability theory and learned how to carry out calculations based upon knowing the underlying probability model.

Example 42 Goals Scored in a Soccer Match

Let $X =$ the number of goals scored by a team in matches in an English soccer league. A plausible model is the Poisson (λ) distribution. (For Americans, there is an excruciatingly small probability of a goal in a small time interval.) We will suppose that the mean of the distribution is $\mu_X = \lambda = 2$. Find (i) the probability of two or more goals being scored by a team in a match and (ii) find the probability of two or more goals being scored by a team in a match given that at least one goal is scored.

$$P(X \geq 2) = 1 - P(X \leq 1) = 1 - e^{-2} - 2e^{-2} = 0.594.$$

$$P(X \geq 2 | X \geq 1) = \frac{P(X \geq 2)}{P(X \geq 1)} = \frac{1 - e^{-2} - 2e^{-2}}{1 - e^{-2}} = 0.687$$

2 *Statistical Models*

- Some *data* are collected. These data are only a subset of all the data that could have been collected. All possible data are usually referred to as the *population*.

- We want to know, at least, certain aspects of the population.

- The data collected contain information concerning the population.

- We use probability models to summarize properties of the population.

- *Statistical inference* involves using the collected data to draw conclusions, or *inferences*, about the population.

An important aspect of statistical inference is that it involves *uncertainty*. This is because the collected data are only a subset of the population.

Probability is the tool used to quantify the *amount* of uncertainty in an inference.

In discussing statistical inference, we will limit ourselves, for the most part, to the following situation:

1. We have a random sample $X_1, \ldots, X_n$ from some distribution, call it f.
2. We assume that the form of f is known except for a few parameters, call them $\theta_1, \ldots, \theta_k$.
3. We use the data $X_1, \ldots, X_n$ to draw conclusions concerning the unknown parameters $\theta_1, \ldots, \theta_k$.

In the previous scenario, the pdf f or pmf p is referred to as the *population*. We write f (or p) as follows to indicate its dependence on the unknown parameters:

$$f(x) = f_{\theta_1,\ldots,\theta_k}(x).$$

Examples:

- The population is an *exponential* density, in which case

$$f_\theta(x) = \theta e^{-\theta x} I_{(0,\infty)}(x).$$

- The population is *normal*, i.e.,

$$f_{\theta_1,\theta_2}(x) = \frac{1}{\sqrt{2\pi}\theta_2} \exp\left[-\frac{1}{2}\left(\frac{x-\theta_1}{\theta_2}\right)^2\right].$$

$\theta_1 = \mu$

$\theta_2 = \sigma$

The *parameter space* is the collection of all possible parameter values and is denoted Ω.

In the exponential example, the parameter space might be

$$\Omega = \{\theta : \theta > 0\}$$

and in the normal example

$$\Omega = \{(\theta_1, \theta_2) : -\infty < \theta_1 < \infty,\ \theta_2 > 0\}.$$

In each of these cases the only constraints on the parameters are those required in order for $f(\,\cdot\,|\theta_1, \ldots, \theta_k)$ to be a probability density. For example, in the normal case we know that θ_2 is the standard deviation of the distribution, which can't be negative.

Sometimes, however, the investigator's knowledge of the problem may allow him/her to reduce the parameter space even more.

If the population were normal and one knew that the population mean must be positive, then the parameter space would be

$$\Omega = \{(\theta_1, \theta_2) : \theta_1 > 0, \, \theta_2 > 0\}.$$

The goal is to draw conclusions concerning the unknown parameters using information in the data $X_1, \ldots, X_n$.

- A function of $X_1, \ldots, X_n$ that **does not depend on any unknown parameters** is called a *statistic*.

Any inferences that we draw must depend only on the *statistic* and not on the unknown *parameters*.

We use upper case letters such as $X_1, X_2, \ldots$ to denote random variables and observed values of these random variables using lower case letters $x_1, x_2, \ldots$

Example 42 Goals Scored in a Soccer Match

Suppose that we observe n scores of teams participating in English soccer matches. Let $X_i =$ the number of goals scored in the i^{th} recorded occasion (two scores are observed during each match). We assume that $X_1, \ldots, X_n$ form a random sample from a Poisson (λ) distribution.

We wish to use the observed numbers of goals to gain information about the unknown value of λ. We take the parameter space to be $\Omega = (0, \infty)$.

The pmf of the i^{th} observation X_i is

$$p_\lambda(x_i) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}.$$

The pmf for the sample $X_1, \ldots, X_n$ is

$$p_\lambda(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}.$$

3 *Methods of Data Collection*

There various approaches to collecting data:

- Sampling from a finite population: Suppose there is a finite collection Π of objects with a numerical characteristic $X(\pi)$ for each object $\pi \in \Pi$. We can observe a subset of these objects which is called a sample. The basic approach to sampling from a finite population is simple random sampling. We use the information in the sample to say something about the population Π.

- Observational data: In this case there is not an obvious probabilistic mechanism that generates the data. In this situation, statistical methods are used to analysis the data, but the results cannot be justified on the basis of statistical theory without making further assumptions.

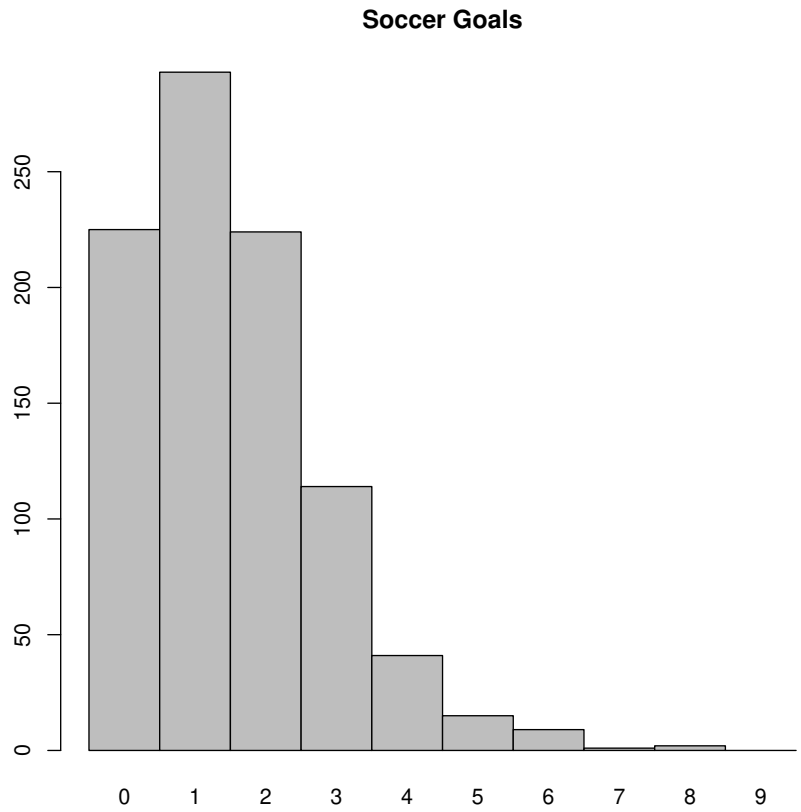
- Experimental data: In this case the researcher controls the experimental conditions so that the assumption of independent responses is reasonable. We will use this as our basic setting for statistical inference.

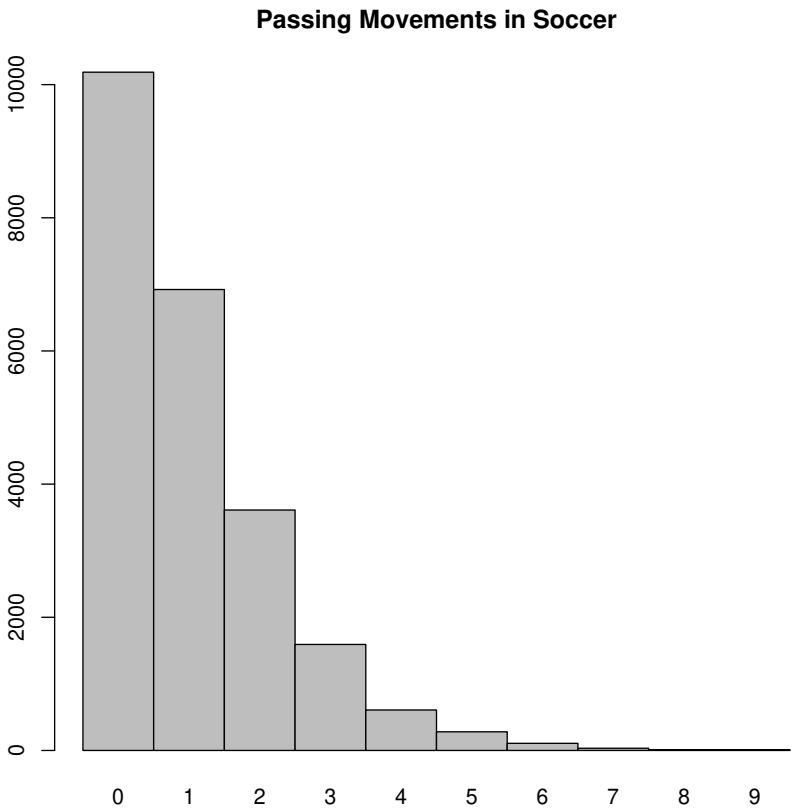
4 Data Description and Summary

We will now illustrate some commonly used statistical approaches to the description and summary of data. We first will present a histogram of the data where the frequencies of values are plotted. Some commonly used summary statistics are also presented. We then present a boxplot which is a simplified view of the data.

Example 42 Goals Scored in a Soccer Match

We consider a data set that consists of the soccer goals scored by each team in 462 games one season in an English soccer league. We also consider the number of passes in passing movements in 42 English soccer matches. Histograms for these data follow:





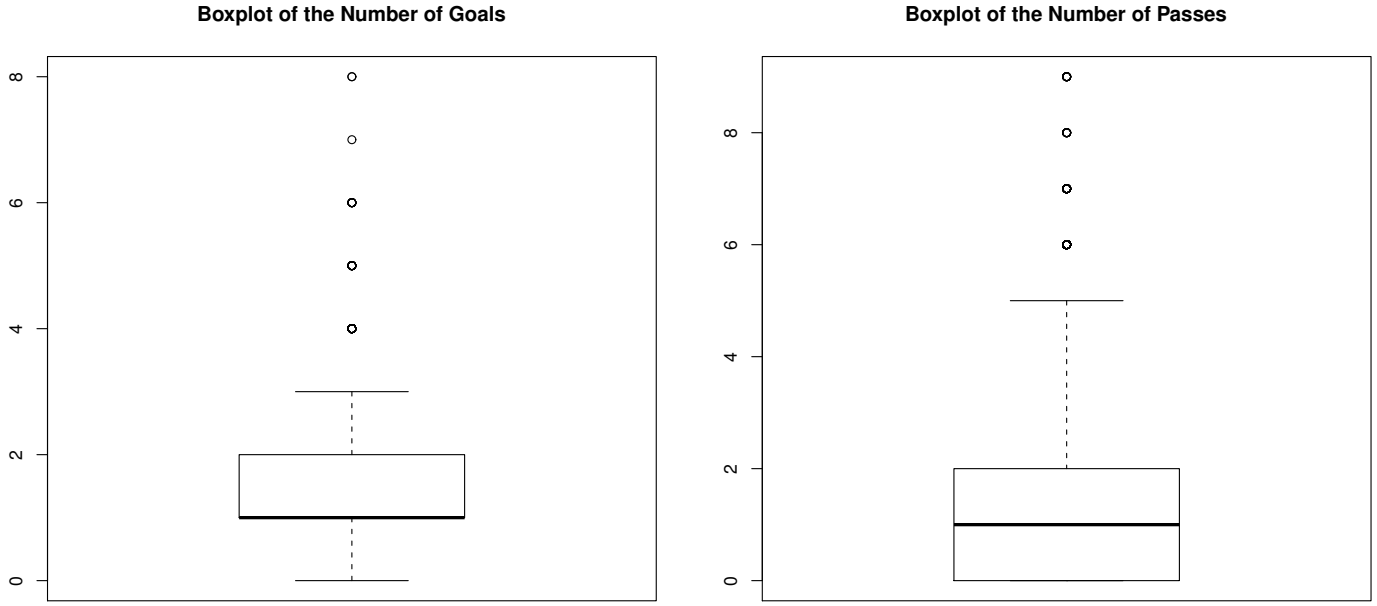
The sample mean, variance, and standard deviation for the goals and passes are

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 1.514,\ s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 1.745,\ s_X = \sqrt{1.745} = 1.328.$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 1.019,\ s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 1.503,\ s_Y = \sqrt{1.503} = 1.226.$$

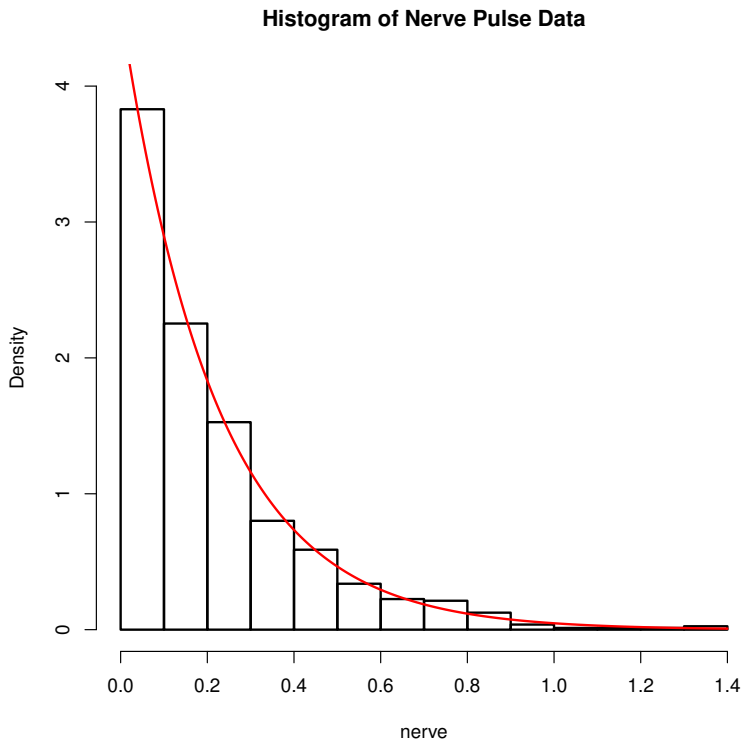
Boxplots are based on the following summary quantiles:

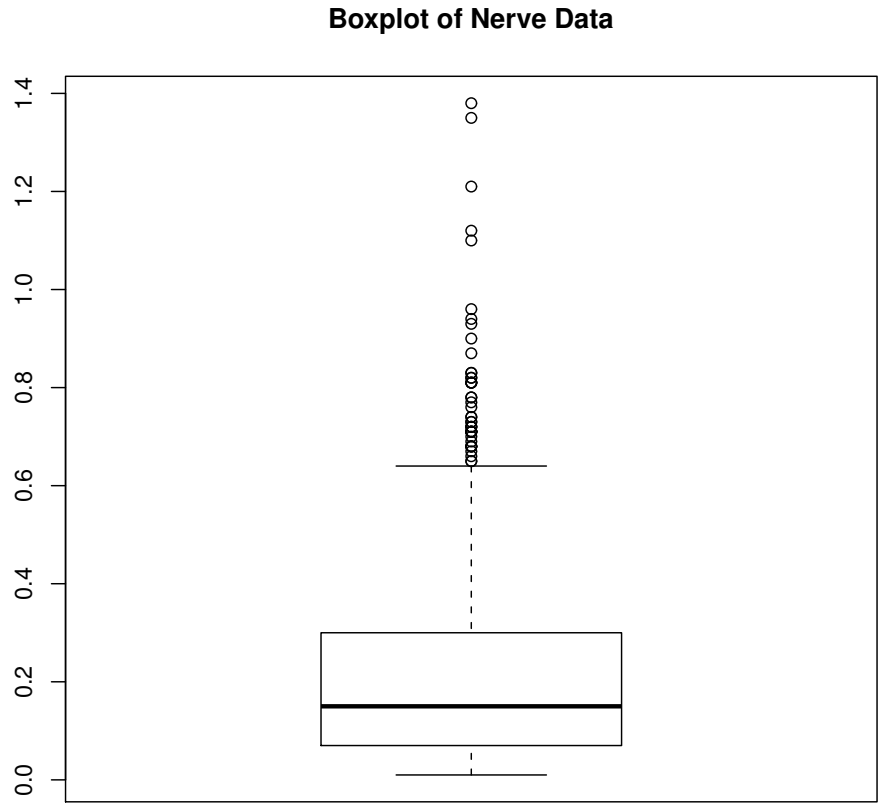
```
> summary(goals)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   1.000   1.514   2.000   8.000
> summary(passes)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   1.000   1.019   2.000   9.000
```



For data with only a few values, the boxplots are not very informative. We will consider another data set where there are many possible values.

Example 43 Cox and Lewis (1966) reported 799 waiting times between successive pulses along a nerve fiber. The data appear in the following histogram:





Summary statistics:

```
> summary(nerve)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    SD     Var
 0.0100  0.0700  0.1500  0.2186  0.3000  1.3800  0.2092 0.04376
```

5 Formulating a Model

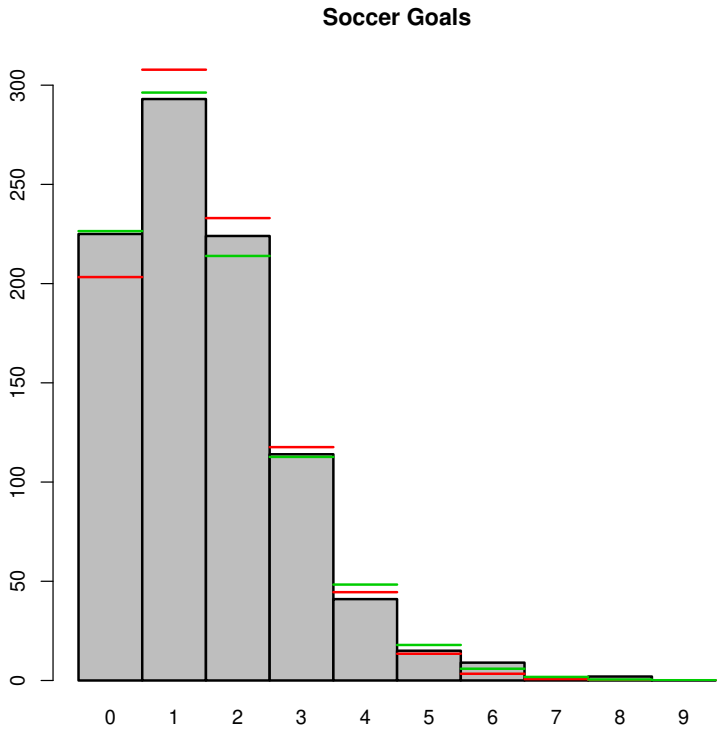
- In a soccer match, a team can score a goal when it has possession of the ball and mounts an attack. Each attack yields a goal with a small probability. If we can assume a constant probability of a goal and independence of attacks, a binomial model for the number of goals is plausible. Since there are many attacks with a small probability of success, a Poisson approximation to the binomial model would be plausible. A model that allows the Poisson mean parameter to vary is the negative binomial model.

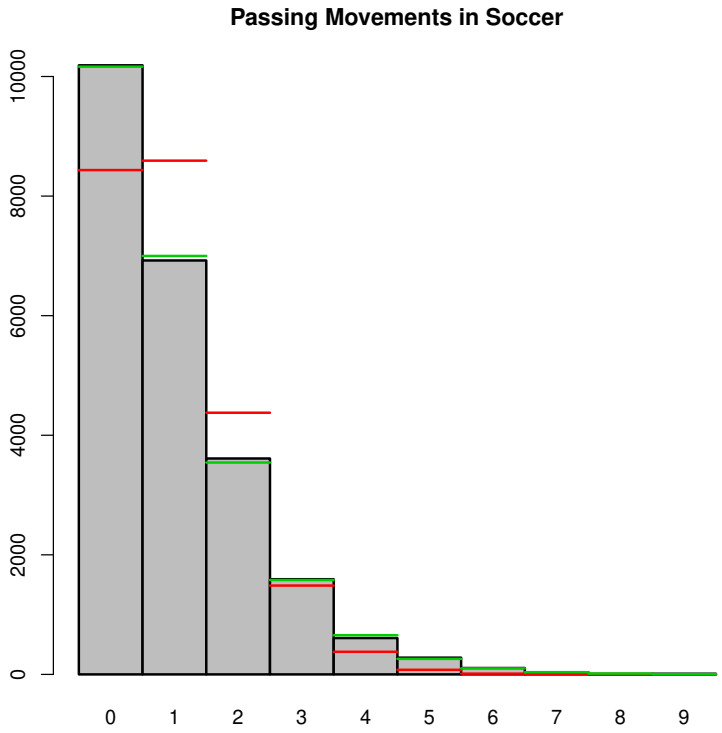
- A passing movement is defined as the number of passes from player to player on the same team until the play ends. Again a Poisson model might be reasonable for the number of passes in a passing movement.

5.1 Fitting the Model to Data

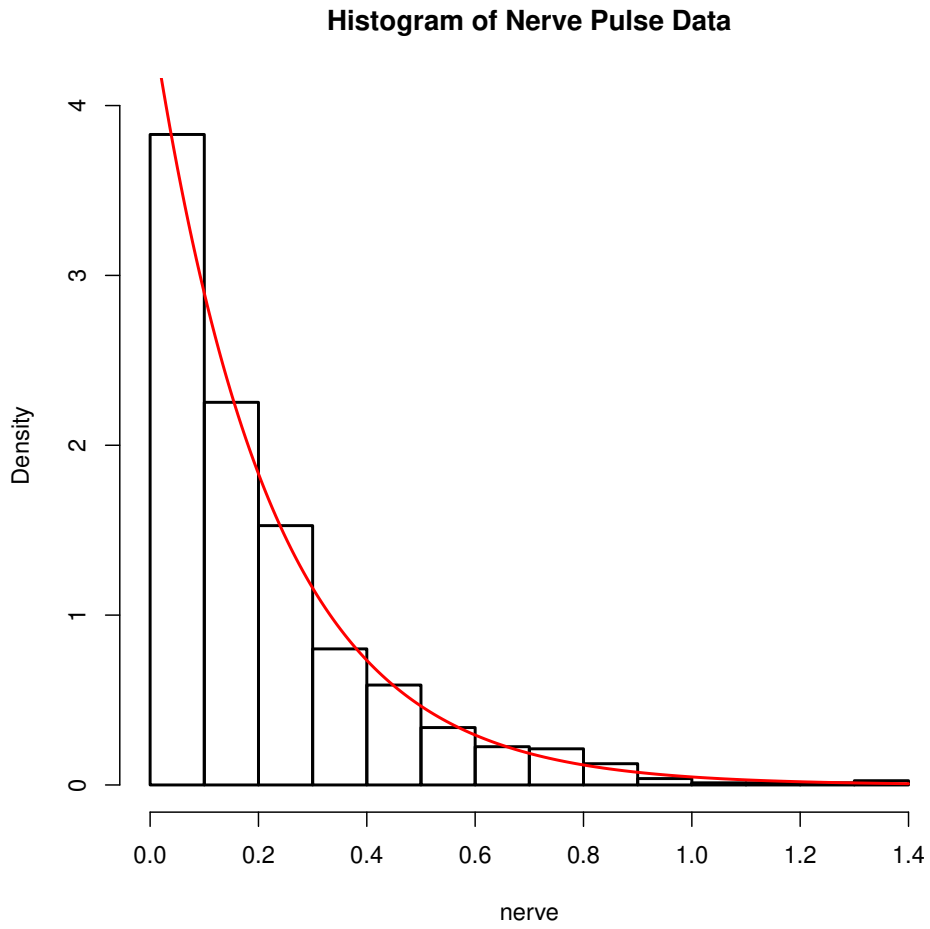
We introduce the idea of parameter estimation as the problem of choosing a value of the parameter that matches the pmf of the distribution to the histogram of the observed data.

- For the Poisson distribution, we will use the Poisson pmf with $\lambda = \bar{x}$, the sample mean.
- For the negative binomial distribution, we will use estimated parameter values obtained by methods we will see later.
- We examine plots that feature the following:
  - A histogram of the observed data in gray
  - Lines in red for the fitted Poisson distribution
  - Lines in green for the fitted negative binomial distribution





Example: Cox and Lewis (1966) reported 799 waiting times between successive pulses along a nerve fiber. The data appear in the following histogram:



The mean is $\bar{x} = 0.2186$ and $\hat{\lambda} = 1/0.2186 = 4.575$. The exponential distribution with $\lambda = 4.575$ is superimposed on the histogram.

5.2 Types of Inference

- *Point Estimation:* Computing a value from the data that is thought to be near the true value of the parameter.
- *Interval Estimation:* Computing from the data a range of plausible values for the unknown parameter.
- *Hypothesis Testing:* Using the data to choose between two statements concerning the unknown parameter.
- *Prediction:* Using the data to predict future values coming from the population.
- *Goodness of Fit:* Using the data to assess whether the assumed distribution is a reasonable model.