prepared by Professor Jenny Baglivo

Statistics can be defined as the art and science of analyzing data:

- *Art:* The art is in asking the right questions, gathering the right information, and summarizing the information in the right form.
- *Science:* The science, based on probability theory, is in analyzing the data formally.

This course focuses on the science of statistics, building on your knowledge of probability theory, and includes substantial applications to many disciplines.

In probability theory we study families of probability distributions. Two families of particular interest are the Poisson and exponential families.

Example 1: Poisson distribution. Let λ ("lambda") be a positive real number. The discrete random variable X is said to have a Poisson distribution with parameter λ when its probability density function (PDF) is as follows:

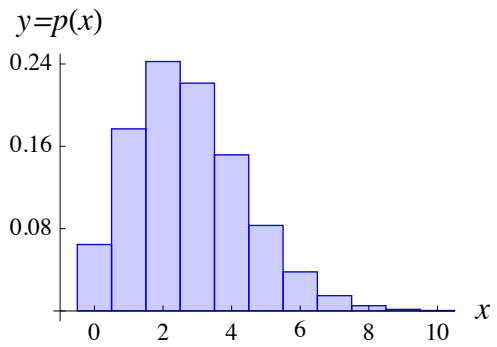
$$p(x) = P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!} \quad \text{when } x = 0, 1, 2, \ldots; \text{ and 0 otherwise.}$$

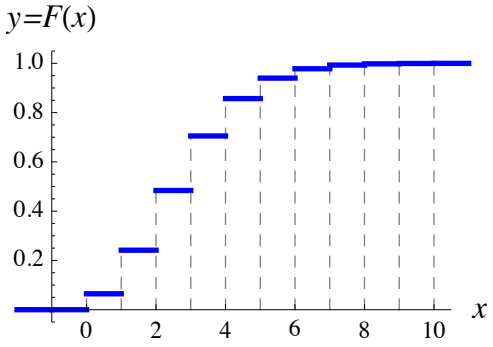
The cumulative distribution function (CDF) of X,

$$F(x) = P(X \leq x) \quad \text{for all } x,$$

is a step function since X is discrete.

Let $\lambda = 2.74$. Then the PDF of X, represented using a probability histogram, is shown in the *left plot* below, and the CDF of X is shown in the *right plot*.

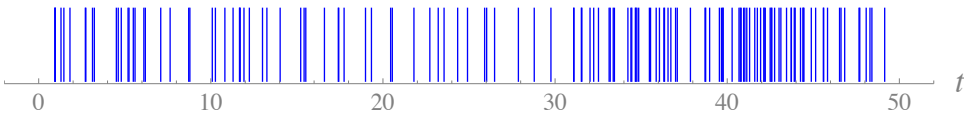




In a probability histogram, area is used to represent probability. For each nonnegative integer x, the area of the rectangle whose base is centered at x is $p(x)$; the sum of the areas is 1.

Application to earthquake analysis. To illustrate how the Poisson distribution can be used to analyze data, we consider data on occurrences of minor-to-light earthquakes (magnitudes 3.5 to 4.4) in the northeastern United States and eastern Canada between 1947 and 1996. (*Source:* Prof. J. Ebel, Weston Observatory, Boston College.)

There were 137 minor-to-light earthquakes during this period, where $t = 0$ corresponds to the beginning of 1947 and $t = 50$ corresponds to the end of 1996. The time of each event is represented by a vertical line in the graph below.



Geophysicists often use Poisson distributions to model the number of earthquakes occurring in fixed time periods. If we divide the observation period $[0, 50]$ into 50 1-year subintervals, and count the number of earthquakes in each subinterval, we get the following summary table:

Number of Events:	0	1	2	3	4	5	6	7
Number of Intervals:	2	12	11	12	5	4	1	3

There were no earthquakes in 2 subintervals, exactly 1 earthquake in 12 subintervals, and so forth. The average was 2.74 (137/50) events per year over the 50-year observation period.

Assuming the list giving the number of earthquakes in each subinterval can be thought of as the values of a random sample from a Poisson distribution, then it is reasonable to use 2.74 to estimate the mean of that distribution. Further, the numbers on the second row of the table above should be close to values we would predict from this distribution:

Event:	$X = 0$	$X = 1$	$X = 2$	$X = 3$	$X = 4$	$X = 5$	$X = 6$	$X \geq 7$
$50P_{2.74}(Event)$:	3.229	8.846	12.119	11.069	7.582	4.155	1.897	1.103

The model predicts that no earthquakes will occur on average 3.229 times in 50 years, that exactly 1 earthquake will occur on average 8.846 times in 50 years, and so forth.

Two questions we might ask are the following:

1. Is the observed average number of events per unit time the "best" estimate of the parameter of a Poisson distribution?
2. Is the Poisson distribution itself a "good" model for these data?

We will learn the fundamental principles needed to answer these and similar questions.

Example 2: Exponential distribution. Let λ be a positive real number. The continuous random variable X is said to have exponential distribution with parameter λ when its probability density function (PDF) is as follows:

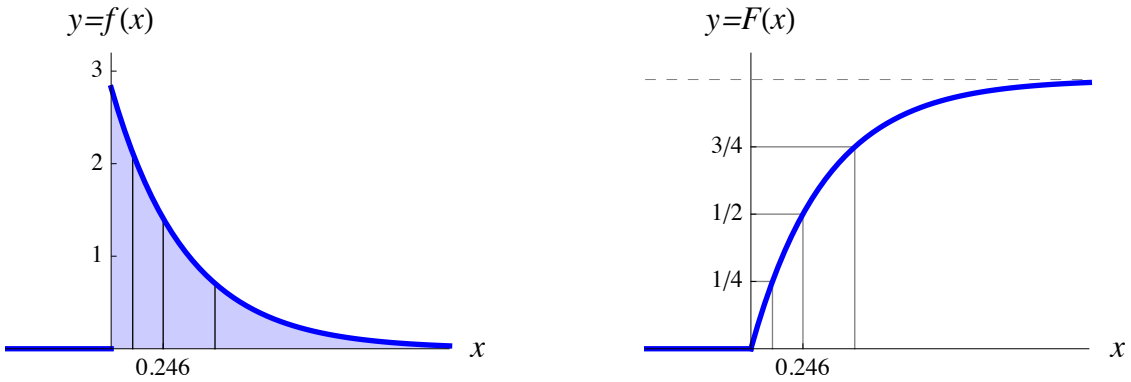
$$f(x) = \lambda e^{-\lambda x} \quad \text{when } x > 0; \text{ and 0 otherwise.}$$

The cumulative distribution function (CDF) of X,

$$F(x) = P(X \leq x) = 1 - e^{-\lambda x} \quad \text{when } x > 0; \text{ and 0 otherwise.}$$

Note that $f(x) = \frac{d}{dx}F(x)$ for all x except 0.

Let $\lambda = 2.821$. Then the PDF of X is shown in the *left plot* below, and the CDF of X is shown in the *right plot*.



The locations of the 25^{th}, 50^{th} and 75^{th} percentiles of the X distribution are marked.

Application to earthquake analysis, continued. An alternative way to analyze earthquakes is to use an exponential distribution to model the time between successive events. The data pictured on the previous page,

$$t_1 = 0.94,\ t_2 = 0.951,\ \ldots,\ t_{136} = 48.4054,\ t_{137} = 49.1497,$$

yield 136 time differences,

$$t_2 - t_1,\ t_3 - t_2,\ \ldots,\ t_{137} - t_{136}$$

over the observation period $[t_1, t_{137}]$.

The time differences data can be summarized as follows:

Interval:	$[0, 0.2)$	$[0.2, 0.4)$	$[0.4, 0.6)$	$[0.6, 0.8)$	$[0.8, 1.0)$	$[1.0, 1.2)$	$[1.2, 1.4)$
Number in Interval:	63	31	14	9	9	4	6

A time difference in the interval $[0, 0.2)$ occurred 63 times, a time difference in the interval $[0.2, 0.4)$ occurred 31 times, and so forth. The average time difference was ≈ 0.354 years.

The reciprocal of the average, $2.821 \approx \frac{1}{0.354}$, is an estimate of the yearly rate earthquakes occurred during the interval from the first to the last observed earthquake.

Assuming the time differences data can be thought of as the values of a random sample from an exponential distribution, then it is reasonable to use 2.821 to estimate the parameter of the distribution. Further, the numbers on the second row of the table above should be close to values we would predict from this distribution:

Interval:	$[0, 0.2)$	$[0.2, 0.4)$	$[0.4, 0.6)$	$[0.6, 0.8)$	$[0.8, 1.0)$	$[1.0, 1.2)$	$[1.2, \infty)$
$136P_{2.821}(X \in Interval)$:	58.643	33.356	18.973	10.792	6.138	3.492	4.606

The suitability of using

- The reciprocal of the mean time difference to estimate the rate λ, and
- The exponential distribution with the time differences data,

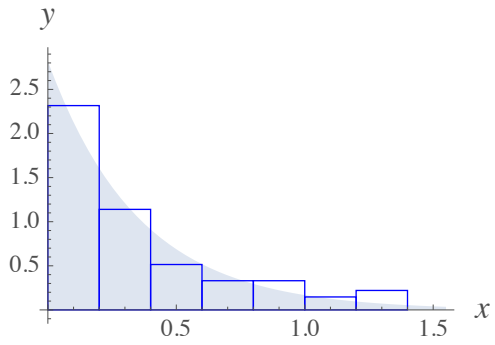
can be examined formally using the techniques we will learn this semester.

Graphical comparison of model with data. Practitioners often use an empirical histogram superimposed on the density function of a continuous model as a graphical comparison of a proposed model with sample data.

To illustrate the technique, we use the time differences data with the 7 intervals

$$[0, 0.2),\ [0.2, 0.4),\ \ldots,\ [1.2, 1.4).$$

For each interval, we plot a rectangle whose base is that interval and whose *area* is the proportion of observations lying in that interval. The sum of the areas of the rectangles is 1.



The plot on the right above shows the histogram for the time differences data superimposed on the density function for an exponential distribution with parameter 2.821.