Maximum Likelihood Estimation as an Application of Optimization in Calculus: Resources for Instructors

David Brown
Department of Mathematics and Computer Science
Colorado College
dbrown@coloradocollege.edu

July 2025

1 Introduction

Maximum likelihood estimation (MLE) is a commonly used method in statistics and data science that allows one to find the best fit of a model to some data. Its prevalence is due in part to its flexibility, as it can be used to analyze either qualitative or quantitative data, with a wide range of model structures. In our paper (Brown, Shea, Harrington 2025), we showed that MLE can be introduced as an interesting application of optimization methods in single- and multi- variable calculus classes, since a key step in the process consists of finding the maximum value of a function.

In this document we provide additional resources for instructors who wish to explore the use of MLE in calculus classes, but who may not have prior experience with this method. This is meant to supplement our paper, so we encourage readers to begin there. The rest of document is organized as follows:

The sections with solved problems present detailed step-by-step solutions and interpretations, suitable for use in class to introduce the methods to students. These include several of the examples from our paper. The sections with additional problems do not have solutions included, so they are suitable

for assigned homework or projects. A full set of solutions to these additional problems is available from the author upon request.

In some cases we have incorporated real data in order to highlight the real-world applicability of MLE. However, for more complicated problems we resorted to using "synthetic" (i.e. fake) data in order to make the calculations tractable. In calculus classes we want students to be able to write down the likelihood function and differentiate it by hand, but industrial strength applications of MLE to large data sets are usually done by computer. Our examples that use fake data are identifiable because the data is introduced using the phrasing "suppose that..."

We would love to hear from instructors (and students) about their experiences incorporating MLE into math classes, and we would love to hear ideas for additional models or data sets to incorporate.

2 Some Useful Distributions

MLE involves fitting a model to some data, and often the model is constructed using one or more commonly encountered probability distributions. Here we give a brief introduction to the distributions that we use in our examples. An instructor can choose a subset of these distributions to introduce in class, depending on which MLE examples they want to use. We provide slightly more information about each distribution than is really needed in a calculus class, where one just needs to know the probability mass or density function $p(x)$ and a sense of what kinds of data might be reasonably modeled by the distribution.

2.1 Binomial

The binomial distribution arises when we are counting the number of "successes" that occur in a fixed number of trials, each of which has the same probability of success. It is important that the trials are independent of each other, i.e. whether success or failure occurs on any one trial is not affected by the successes or failures on any other trials. Note that the term "success" is traditional in this field, but we often prefer the term "event" since the occurrences may or may not be cause for celebration.

Let N be the number of trials, and p be the probability of success on each trial. Then the probability of k successes is given by:

$$p(k) = \binom{N}{k} p^k (1-p)^{(N-k)} \tag{1}$$

The set of possible values is: $0 \leq k \leq N$. A function like this that assigns probabilities to a discrete list of possible values is called a "probability mass function".

Example 2.1. Suppose that 11% of people are left-handed. If you sample 40 people at random, the probability that exactly k of them are left-handed is given by $p(k) = \binom{40}{k}(0.11)^k(0.89)^{(40-k)}$. This distribution is plotted in Figure 1.

The binomial distribution has a mean of $\mu = Np$, which is also where the mode (highest probability) occurs. The standard deviation is given by $\sigma = \sqrt{Np(1-p)}$.

2.2 Geometric

The geometric distribution arises most naturally as a model of the number of trials required until a "success" happens, although it can arise in other situations as well. For example, it describes the number of times that you need to flip a coin until Heads first appears. Thinking of the trials as being analogous to the passage of time, we can describe this as the "waiting time" until an event of interest occurs. The geometric distribution is "memoryless", which means that the number of trials from "now" until the next success doesn't depend on how

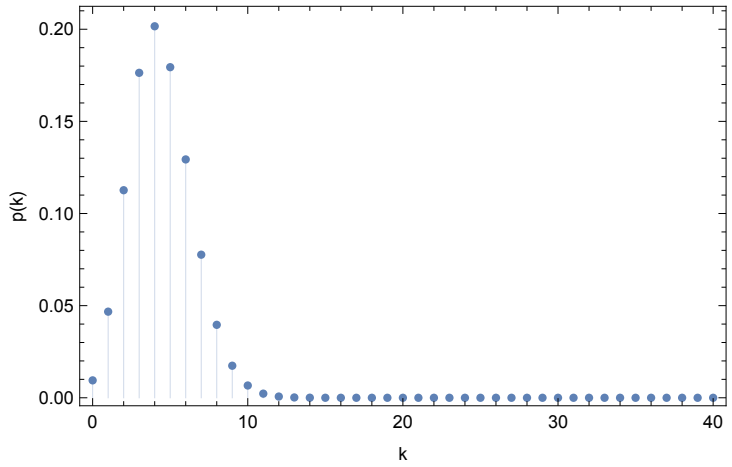

Figure 1: An example of the binomial distribution with $N = 40, p = 0.11$

we define "now". We often use it to model the number of trials from one success until the next one, but we don't have to start our counting at the most recent success. If we perform independent trials, each of which has probability p of success, then the probability that the first success occurs on the kth trial is given by:

$$p(k) = (1-p)^{k-1}p \tag{2}$$

The set of possible values is $1 \leq k < \infty$. The geometric distribution has mean $\mu = 1/p$ and standard deviation $\sigma = \frac{\sqrt{1-p}}{p}$.

Example 2.2. Suppose that a laboratory rat has a panel with 5 levers on it. One of the levers, when pressed, rewards the rat with a piece of vegan cheese-substitute (this particular rat is lactose intolerant). If we assume that the rat presses the levers randomly and with equal probability, then the probability that it first presses the pseudo-cheese lever on the kth attempt is:

$$p(k) = (.8)^{k-1}(.2) \tag{3}$$

This distribution is plotted in Figure 2.

2.3 Poisson

The Poisson distribution is closely related to the binomial, in the sense that it is useful for counting the number of events. However, instead of discrete trials that have a fixed probability of success, the scenario is that events occur randomly at a fixed average "rate", continuously in time. As with the binomial, it is essential that the occurrences of events are independent of each other. Let r be the rate at which events happen. Then the probability that k events occur

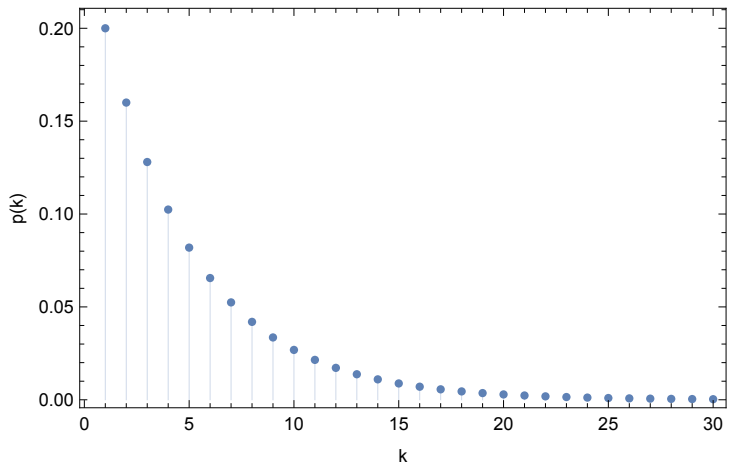

Figure 2: An example of the Geometric distribution with $p = 0.2$

in a time interval of length T is given by:

$$p(k) = \frac{(rT)^k e^{-(rT)}}{k!} \tag{4}$$

The set of possible values is: $0 \leq k < \infty$.

Example 2.3. Suppose that on a pleasant fall afternoon, apples are falling off of a tree at an average rate of 5 per hour. Assuming that the apples fall independently of each other, the probability that exactly k apples fall in the next three hours is given by:

$$p(k) = \frac{(3 \times 5)^k e^{-(3\times 5)}}{k!} \tag{5}$$

This distribution is plotted in Figure 3.

The Poisson distribution has its mean and mode at $\mu = rT$ and the standard deviation $\sigma = \sqrt{rT}$.

While many examples involving the Poisson distribution are expressed in terms of events occurring over time, T can also stand for something like distance or volume. For example we might use the Poisson distribution to model the number of potholes along a length of road, or the number of chocolate chips in a cookie. In this kind of case the rate parameter r would be interpreted as events per unit length, or events per unit volume.

2.4 Exponential

The exponential distribution arises most naturally as the waiting time until an event occurs. Assume the same scenario as above with the Poisson distribution: events occurring at a fixed average rate in continuous time. Just as the

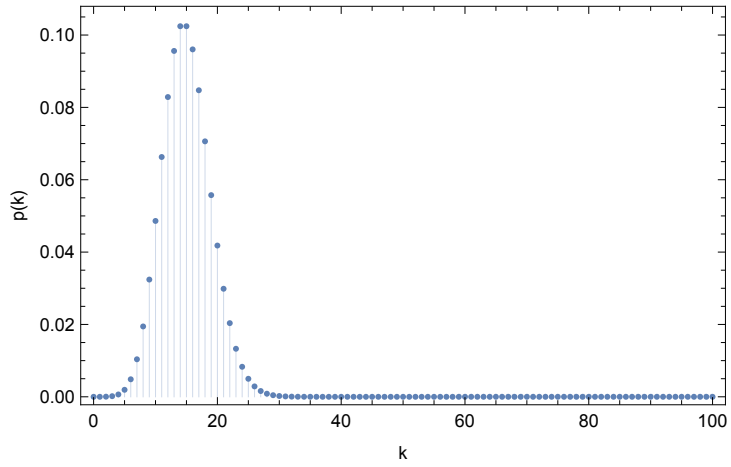

Figure 3: An example of the Poisson distribution with $T = 3, r = 5$

geometric distribution models the waiting time until the next success with discrete trials, the exponential distribution models the waiting time until the next event occurs when the events occur randomly in continuous time. Unlike the binomial, geometric, and Poisson distributions, the exponential distribution is continuous, with possible values $0 \leq t < \infty$. Like the geometric distribution, the exponential distribution is said to be "memoryless". This means that the probability of waiting a certain amount of time until the next event is independent of when we start our clock. In other words, the distribution of the time between two events is the same as the distribution of the time from "now" until the next event, no matter how we define "now". The time until the next event has probability density function:

$$p(t) = re^{-rt} \tag{6}$$

The exponential distribution has mean $\mu = \frac{1}{r}$ and standard deviation $\sigma = \frac{1}{r}$. Since this is a continuous distribution, the probability of any specifically chosen time is 0. A probability density function is used to compute probabilities via integration: $P(a < T < b) = \int_a^b p(t)dt$. However, this can be skipped over in a Calculus 1 class where integration has not been introduced. In the context of MLE, a probability density function is used in exactly the same way as a probability mass function.

Example 2.4. Suppose that on a pleasant fall afternoon, apples are falling off of a tree at an average rate of 5 per hour. Starting our clock at any time we choose, and assuming that the apples fall independently of each other, the probability density for the time until the next apple falls is given by:

$$p(t) = 5e^{-5t} \tag{7}$$

This distribution is plotted in Figure 4.

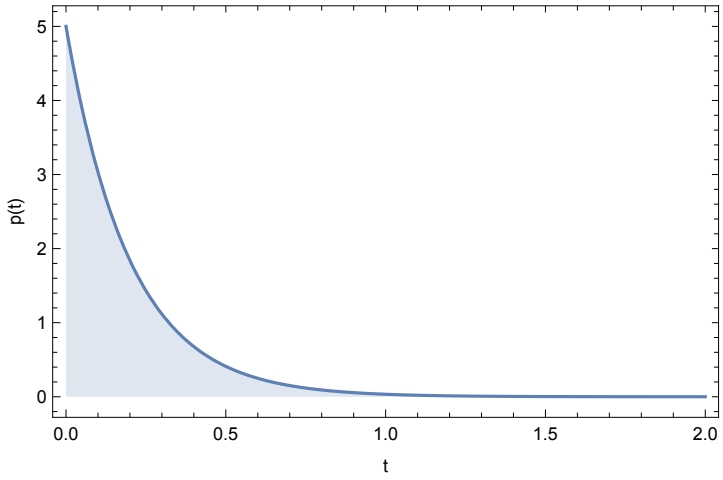

Figure 4: An example of the exponential distribution with $r = 5$

2.5 Normal

The normal distribution is the famous "bell curve" that arises in many different contexts. In a calculus class, one would simply need to state that a normal distribution is a reasonable model for a given data set, perhaps supported by a histogram of the data. Like the exponential distribution, the normal distribution is continuous so we are working with the probability density instead of the actual probability of a value. The normal distribution has two parameters, the mean μ and standard deviation σ. Then the probability density function is given by:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{8}$$

The set of possible values is $-\infty < x < \infty$, but the probability density drops quickly towards 0 as we move away from the mean, so this is still used as a model for many quantities that have a restricted set of possible values. It is often convenient in MLE problems to replace σ^2 with a parameter for the variance, $v = \sigma^2$.

Example 2.5. Suppose that adult gray squirrels have body masses that are normally distributed with a mean of $\mu = 850$g and standard deviation $\sigma = 100$g. Then the probability density function for the mass of an adult gray squirrel is:

$$p(x) = \frac{1}{\sqrt{2\pi 100^2}} e^{-\frac{(x-850)^2}{2(100^2)}} \tag{9}$$

This distribution is plotted in Figure 5.

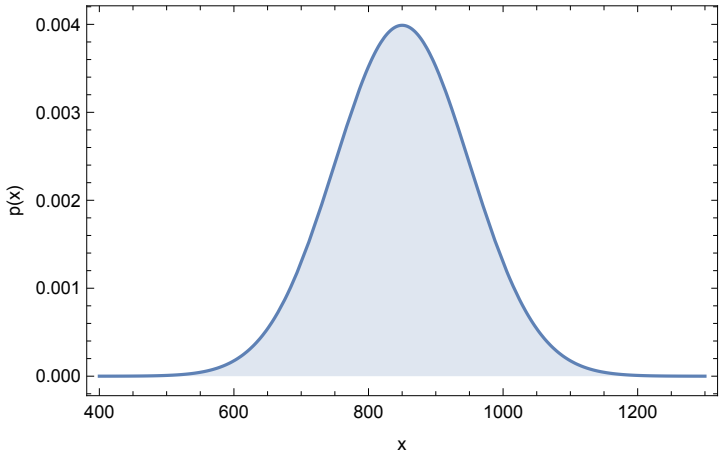

Figure 5: An example of the normal distribution with $\mu = 850$ and $\sigma = 100$

2.6 Pareto, aka Power Law

The Pareto distribution, also known as a power law, is commonly encountered as a model for situations where extremely large values, while rare, are not as rare as would be predicted by something like the normal or exponential distribution. Because of this, the Pareto distribution is often called "fat tailed." For example, extremely wealthy individuals are rare, but they are more common that we would predict using a normal distribution. Indeed, the Pareto distribution was originally used to describe the distribution of wealth, but it has been found to be a good fit to many other kinds of data including populations of towns, size of meteorites, insurance losses after natural disasters, number of papers published by individual researchers, etc. One should be aware that claims of a Pareto distribution rely on the behavior of the distribution for extremely large values, which are by definition rare. As a result, some statisticians urge caution when making the claim that data indicate the presence of a Pareto distribution as opposed to some other candidate model.

The special case of the Pareto distribution that we will consider takes on the continuum of values $1 \leq x < \infty$. For data that has a different minimum value, we can simply rescale the data set by dividing each value by the smallest one. This distribution has a single "shape" parameter $a > 0$. The probability density function is:

$$p(x) = \frac{a}{x^{a+1}} \tag{10}$$

Interestingly, the distribution has a finite mean μ and standard deviation σ only if a is large enough:

$$\mu = \begin{cases} \frac{a}{a-1}, & a > 1 \\ \infty, & 0 < a \leq 1 \end{cases}$$

$$\sigma = \begin{cases} \frac{1}{a-1}\sqrt{\frac{a}{a-2}}, & a > 2 \\ \infty, & 0 < a \leq 2 \end{cases}$$

Example 2.6. Contact-tracing for an infectious disease can reveal how many new infections, x, were caused by each infected person. Let us consider only individuals who infected an least one other person. If most individuals cause few infections but there are some "super-spreaders", a Pareto distribution might be appropriate. Suppose that the number of infections caused by an individual follows a Pareto distribution with $a = 1.5$. Then the probability density function is given by:

$$p(x) = \frac{1.5}{x^{2.5}} \tag{11}$$

for $x \geq 1$. To illustrate what this implies here, we can compute that $P(1 \leq x \leq 2) = 0.65$ and $P(x \geq 10) = 0.03$ .This distribution is plotted in Figure 6.

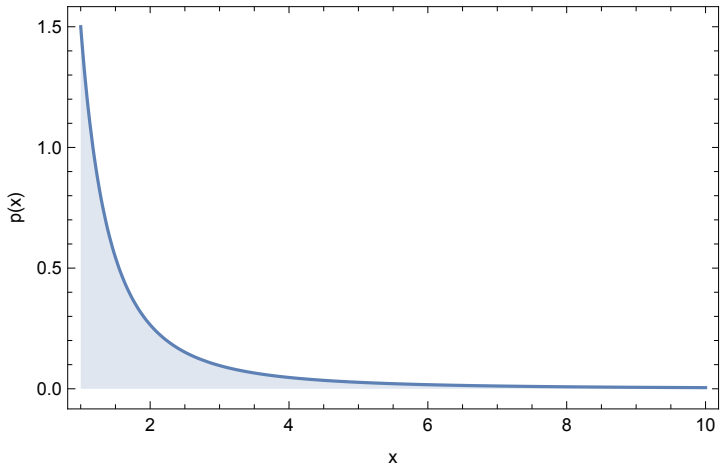

Figure 6: An example of the Pareto distribution with $a = 1.5$

2.7 Gompertz

Researchers from many fields study the distribution of lifespans in humans and other organisms. (This is relevant in biology, ecology, medicine, demography, and life insurance.) There are two related functions that are sometimes confused with each other:

- $p(t)$, the probability that an individual lives from age 0 until exactly age t (and no longer). This is called the lifespan distribution.
- $h(t)$, the risk of dying now for an individual who is currently at age t. This is called the hazard function.

Since living until exactly age t and then dying means that an individual must first survive all ages up to t and then die at time t, these functions are related by:

$$p(t) = \left(1 - \int_0^t p(s)ds\right) h(t) \tag{12}$$

In the simplest scenario, the hazard rate is constant: an individual has an age-independent risk of dying, r. Then one can show that the equation above (along with the requirement that the integral of $p(t)$ over all possible values comes out to be 1) results in an exponential distribution of lifespans: $p(t) = re^{-rt}$.

While a constant hazard rate is in fact found in some species, it is also common for the hazard rate to change with age. For example, a hazard rate that increases with age is called "senescence". One commonly used model incorporates a hazard rate that increases exponentially: $h(t) = ace^{at}$. In other words, the risk of dying increases exponentially with age at a rate a, from an initial level of ac. This results in a lifespan distribution known as the "Gompertz" distribution, named after a 19th century mathematician and actuary who introduced it:

$$p(t) = ace^{at}e^{-c(e^{at}-1)} \tag{13}$$

While we have introduced this in the framework of organisms' lifespans, many mechanical components can be modeled by the same distribution if their risk of failure increases over time due to wear.

Example 2.7. Suppose that a newly hatched fluffy-backed tit-babbler (which is the real name of an actual bird species) has a death rate of 0.2 per year, which increases exponentially at a rate of 0.4 per year. Then the probability of living until age t is given by the Gompertz distribution with $ac = 0.2$ and $a = 0.4$, so $c = 0.5$. This is plotted in Figure 7.

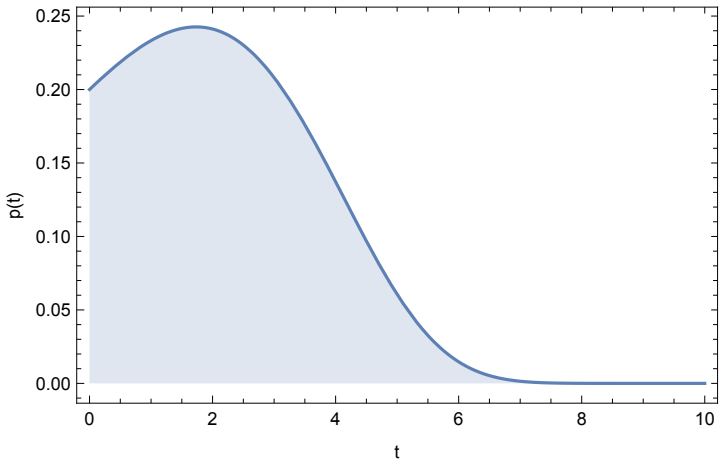

Figure 7: An example of the Gompertz distribution with $a = 0.4, c = 0.5$

2.8 Logistic Regression

Many students are familiar with linear regression, in which we try to fit a straight line to a set of (x, y) data points. In linear regression, both variables are quantitative. In logistic regression, the predictor variable x is quantitative, but the response variable y is qualitative. For simplicity, we will restrict our focus to scenarios in which y only has two possible values, which are labeled as "yes" and "no". For example, y might record whether or not a person likes grunge rock, and x could be their age. The responses are typically encoded as $y = 0$ for "no" and $y = 1$ for "yes".

Logistic regression is useful for scenarios where the y responses switch from mostly "yeses" to mostly "nos" (or vice versa) at some value of x. The transition can be gradual or abrupt. Figure 8 shows an example of data that we might analyze using logistic regression. It appears that y switches from being mostly "no" to mostly "yes" when x is between 2 and 3. There is a region of overlap, where both responses have a reasonable chance of happening. (Some people in their 40s like grunge rock, and some don't.)

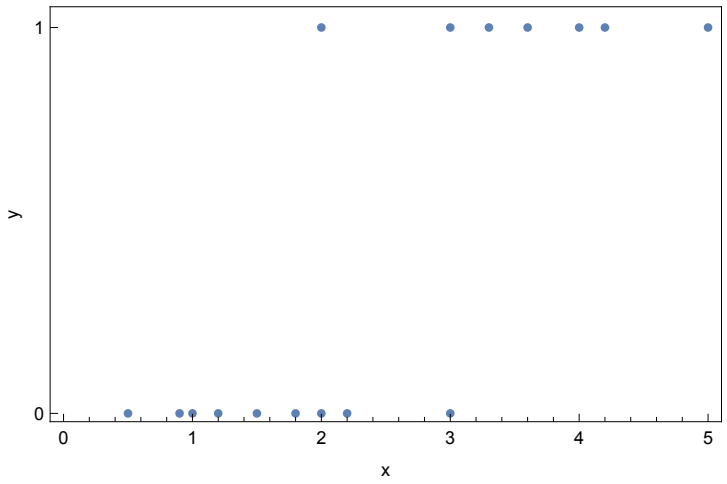

Figure 8: Typical data that we might try to fit using logistic regression.

Looking at Figure 8, we might be tempted to fit a sigmoidal curve to the data. This is the approach of logistic regression. Specifically, we define a function $p(x)$ that is the probability of a data point being a "yes" as a function of x:

$$probability(y = 1) = p(x) = \frac{e^{bx}}{e^{bx} + e^{a}} \tag{14}$$

Then the probability of "no" is:

$$probability(y = 0) = 1 - p(x) = \frac{e^{a}}{e^{bx} + e^{a}} \tag{15}$$

This function $p(x)$ is called a "logistic" function, from which our technique inherits the name. Figure 9 shows the result of fitting $p(x)$ to our sample data by choosing appropriate values of the parameters a and b.

Be alert for a common point of confusion. Here, $p(x)$ is not defining a probability distribution over all possible values of x. In other words, we don't require $\int p(x)dx = 1$. Rather, for any given value of x, it is a probability mass function over the (two) possible values of y.

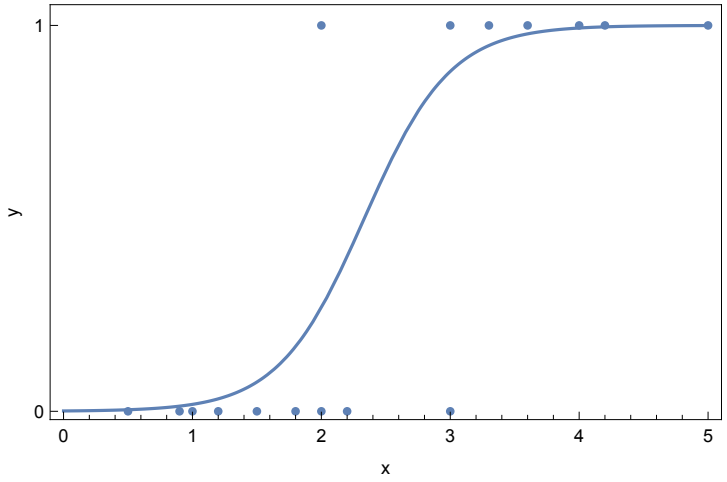

Figure 9: Fit of a logistic regression model to some data.

At this point, students should develop some intuition and facility with the logistic function by doing the following:

- Plot $p(x)$ with different choices of parameters a and b, including negative ones. Based on this exploration, they should describe the role of each parameter in controlling the shape and "location" of the logistic function.
- Compute $\lim_{x\to\infty} p(x)$ and $\lim_{x\to-\infty} p(x)$ in the cases $b > 0$ and $b < 0$.
- Develop a further understanding of the roles of a and b by evaluating $p(0)$ and by finding the value of x at which $p(x) = \frac{1}{2}$.

3 Solved Single-Variable MLE Problems

3.1 Binomial Distribution

Example 3.1. A jar contains an unknown number of red marbles and blue marbles. One at a time, ten marbles are sampled with replacement; seven of them are red. Find the maximum likelihood estimate for the proportion of red marbles in the jar.

Solution: Let p be the proportion of red marbles in the jar. This is the parameter that we want to estimate. The probability of getting any particular sequence of 7 red and 3 blue marbles is found by multiplying the probabilities for each individual outcome, yielding: $p^7(1-p)^3$. Let c be the number of different sequences that can be written with 7 copies of "r" and 3 copies of "b". It turns out that there are $c = \binom{10}{7} = 120$ of these. However, the important fact is that all of these sequences are equally likely, and c doesn't depend on the parameter that we are interested in, p

The likelihood function for this data is: $L(p) = cp^7(1-p)^3$. Since we are trying to find the location p of the maximum, the constant c will not affect our answer so we will drop it from the problem and just use $L(p) = p^7(1-p)^3$.

To find the maximum, we set the first derivative equal to zero:

$$\begin{aligned} L'(p) &= 7p^6(1-p)^3 + p^7 3(1-p)^2 \\ &= p^6(1-p)^2(7(1-p)+3p) \\ &= 0 \end{aligned}$$

From this we find three possibilities:

$$\begin{aligned} p^6 &= 0 \\ (1-p)^2 &= 0 \\ 7(1-p)+3p &= 0 \end{aligned}$$

The first and second possibilities lead to $p = 0$ and $p = 1$, respectively. However, these imply that the jar contains no red marbles or only red marbles, both of which are impossible. So we drop these potential solutions. Solving the last equation for p yields $p = \frac{7}{10}$. This critical point is confirmed to be a maximum via the second derivative test, since $L''(0.7) < 0$.

The likelihood function is plotted in Figure 10 (normalized with $c = 1$ for simplicity).

Instead of working directly with the likelihood function, we can work with the log-likelihood function, $f(p) = \ln L(p)$. Since the natural logarithm is monotone increasing, a value of p that maximizes $f(p)$ will maximize $L(p)$ and vice versa.

Here the log-likelihood function is

$$\begin{aligned} f(p) &= \ln\left(p^7(1-p)^3\right) \\ &= 7\ln p + 3\ln(1-p) \end{aligned}$$

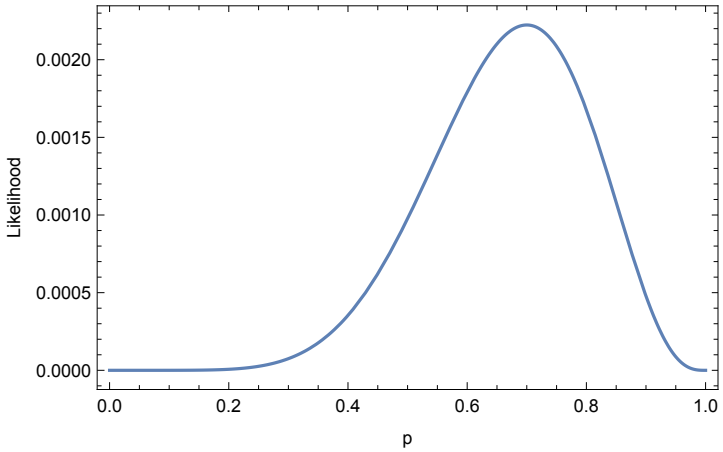

Figure 10: The likelihood function for the marbles problem.

We then find critical points via:

$$\begin{aligned} f'(p) &= \frac{7}{p} - \frac{3}{1-p} \\ &= 0 \end{aligned}$$

Solving this yields the unique critical point $p = \frac{7}{10}$. The log-likelihood function for this example is plotted in Figure 11.

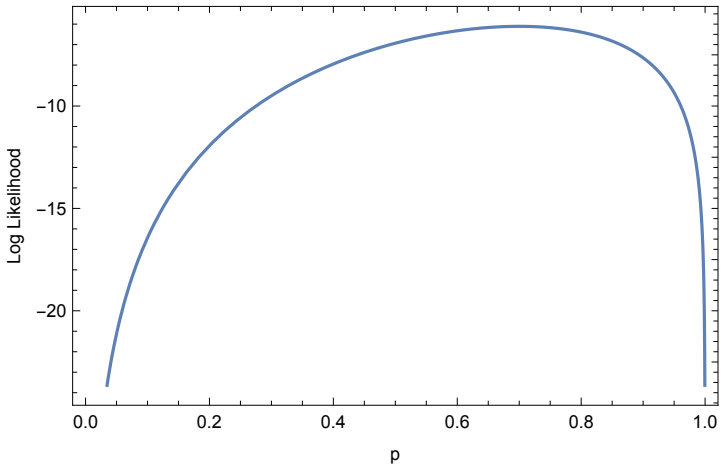

Figure 11: The log-likelihood function for the marbles problem.

Follow-up Questions:

Q1: We claimed that dropping the constant c from the likelihood function

didn't change our parameter estimate. Verify that this is so by re-doing the problem using the likelihood function $L(p) = 120p^7(1-p)^3$.

A1: Using the likelihood function, we can cancel c from the equation when we set $L'(p) = 0$. Using the log-likelihood function, we find that the constant $\ln(c)$ is added to $f(p)$, and it disappears when we differentiate.

Q2: Suppose that we drew 100 marbles with replacement and found that 70 of them were red. What do you think that the MLE estimate for p should be? Verify that this is so. Then plot the likelihood function for this data and compare it to the likelihood function for the original data.

A2: Intuitively, we expect to find the maximum again at $p = 0.7$. This is easily verified by using either the likelihood function $L(p) = p^{70}(1-p)^{30}$ or the log-likelihood function $f(p) = 70\ln p + 30\ln(1-p)$. The graph of $L(p)$ has a peak at $p = 0.7$, and it is more "concentrated" at this maximum than the original likelihood function.

Q3: Suppose that you draw 10 marbles with replacement and find that all 10 of them are red. What is the likelihood function in this case, and where is its maximum?

A3: In this scenario $L(p) = p^{10}$. This has no local maximum. The global maximum on the feasible interval $0 \le p \le 1$ occurs at $p = 1$. Our best estimate based on the available data is that 100% of the marbles in the jar are red.

Example 3.2. In the general population, around 11% of people are left-handed. Is this true among professional musicians? Suppose that a random sample of 50 musicians finds that 8 of them are left-handed. Graph the likelihood function for this data, and interpret what it tells us. Then find the MLE estimate for the proportion of all professional musicians who are left-handed.

Solution: Since this is a random sample, we can assume that the individuals are independent of each other so the probability of distribution of the number of left-handed people in our sample is given by the binomial distribution with $N = 50$. We want to estimate the unknown parameter p, the proportion of left-handed people among all professional musicians. Setting aside the constant c in the binomial distribution, we have the likelihood function:

$$L(p) = p^8(1-p)^{42}$$

This is plotted in Figure 12. This graph shows the relative likelihood of getting the data in the sample, assuming different values of p. It appears that the value of p that makes our data most likely is around $0.16 = 8/50$, which is what we would intuitively expect.

Taking the natural logarithm gives the log-likelihood function:

$$\begin{aligned} f(p) &= \ln\left(p^8(1-p)^{42}\right) \\ &= 8\ln p + 42\ln(1-p) \end{aligned}$$

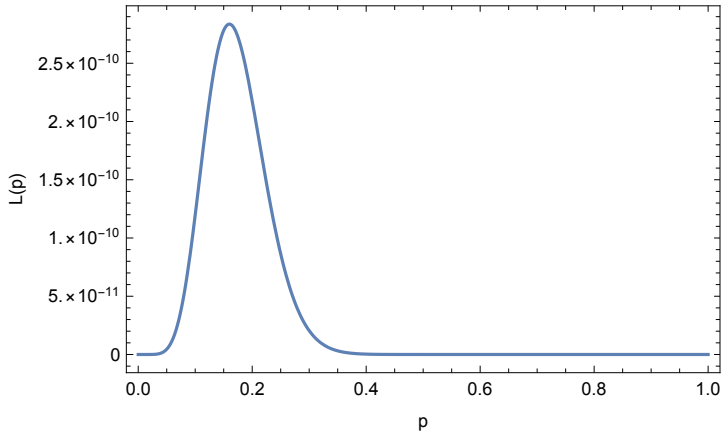

Figure 12: The likelihood function for the left-handed musicians problem.

We any find critical points via:

$$\begin{aligned} f'(p) &= \frac{8}{p} - \frac{42}{1-p} \\ &= 0 \end{aligned}$$

Solving this yields the unique critical point $p = \frac{8}{50} = 16\%$. This is our MLE estimate.

Follow-up Questions:

Q1: In this problem, the constant $c = 536,878,650$. Use this in the likelihood function to find the probability of our sample data occurring if $p = .16$ and if $p = .11$. What do these results mean?

A1: $L(0.16) \approx 0.152$ and $L(0.11) \approx 0.086$. This means that our data were almost twice as likely to occur if $p = .16$ than if $p = .11$.

Q2: In your own words, explain what it means that the MLE estimate is $p = 0.16$. For example, do we know for sure that 16% of professional musicians are left-handed?

A2: The data that we got were more likely to occur if 16% of musicians are left-handed than if any other percent are left-handed. Thus, 16% represents our best estimate based on this data, but it is still possible that the true percent is something else.

Q3: Where in our calculations did we use the assumption of independent data points? Is there some way that the data could have been collected that would make you concerned about whether the assumption of independence was satisfied?

A3: We assumed that the individual in our sample were independent of each other when we formed the likelihood function by multiplying the probabilities of 50 individuals. I would be concerned about the assumption of independence if we sampled musicians who perform together. Maybe left-handed musicians tend to join bands with other lefties, for example!

Example 3.3. Conservation biologists use a method called mark-recapture to estimate the size of a population of wild animals. For example, suppose that biologists are interested in estimating the number of black-footed ferrets in eastern Wyoming. They set out a number of traps and capture 50 adult ferrets. They attach a small ear tag to each individual, then release them. The following year, they set out traps again. This time they catch 65 adult ferrets, of which 10 have ear tags from being caught the first time.

a) Letting N be the unknown total population of adult black-footed ferrets in the study region, justify the likelihood function: $L(N) = (\frac{50}{N})^{10}(1-\frac{50}{N}))^{55}$. What is the range of possible values for N?

b) List several assumptions about the ferrets that we had to make in order to write down the likelihood function.

c) Graph the likelihood function. Estimate the MLE value of N.

d) Use the log-likelihood function $f(N)$ to find the MLE value of N.

Solution:

a) Note that $\frac{50}{N}$ is the fraction of ferrets that have ear tags from the previous year. If we think of tagged ferrets like red marbles and untagged ferrets like blue marbles, we get the likelihood function here the same way we did with marbles, replacing p with $\frac{50}{N}$. We know that there are at least 65 ferrets in the population, so $65 \leq N < \infty$.

b) We need to assume that the ferrets are all independent of each other, i.e. whether an individual is captured is unaffected by the capture of any other individuals. We need to assume that all 50 tagged ferrets are still present during the second year; none of them have died or left the area. We also need to assume that ferrets don't learn to avoid traps after being captured, so that the tagged and untagged individuals are equally likely to be caught.

c) A plot of the $L(N)$ is shown in Figure 13. It appears that the maximum occurs between $N = 300$ and $N = 400$.

d) Taking the natural logarithm of $L(N)$ yields:

$$\begin{aligned} f(N) &= \ln\left(\frac{50}{N}\right)^{10} + \ln\left(1-\frac{50}{N}\right)^{55} \\ &= 10\ln 50 - 10\ln N + 55\ln(N-50) - 55\ln N \\ &= 10\ln 50 - 65\ln N + 55\ln(N-50) \end{aligned}$$

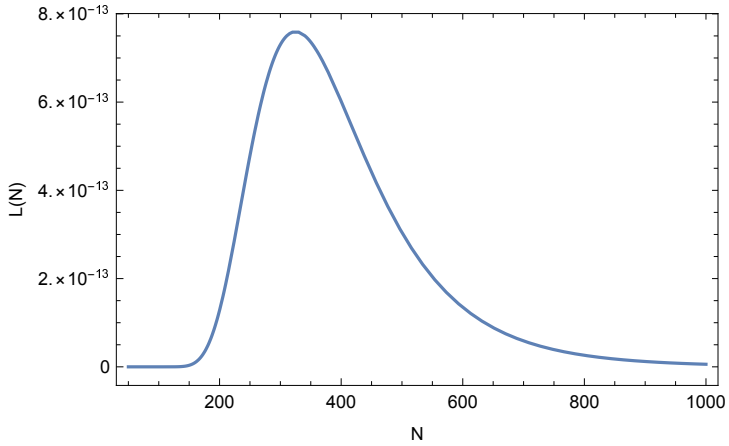

Figure 13: The likelihood function for the ferret population.

Then:

$$\begin{aligned} f'(N) &= -\frac{65}{N} + \frac{55}{N-50} \\ &= 0 \end{aligned}$$

We can simplify this to:

$$-65(N-50) + 55N = 0$$

Solving this yields $N = \frac{65 \times 50}{10} = 325$.

Follow-up Questions:

Q1: How confident would you be in this population estimate? How should conservation biologists report a result like this?

A1: We made several questionable assumptions in constructing our model. In addition, the value of the MLE estimate changes substantially if the number of tagged ferrets captured in the second year changes by just one or two, which seems like it could easily occur. Thus, the precise estimate of $N = 325$ does not seem very reliable. Biologists should probably report a range of plausible values, or a rough estimate like "several hundred".

Q2: Our MLE value is an estimate of the true population size, which may be different. Since the number of data points is fairly large, an important result from statistics says that we can obtain a 95% confidence interval for the true population size from the formula: $N \pm 1.96 \frac{1}{\sqrt{-f''(N)}}$, where N is our MLE value. Find the 95% confidence interval for the true population of ferrets in this region, and interpret what it means.

A2: From our log-likelihood function, $f''(N) = \frac{65}{N^2} - \frac{55}{(N-50)^2}$. Plugging in our estimate of $N = 325$ yields $f''(325) \approx -0.00011$. Then our approximate confidence interval is: $325 \pm 1.96\frac{1}{\sqrt{0.00011}} \approx 325 \pm 185$. We are 95% certain based on our data that the true population is within this interval.

3.2 Geometric Distribution

Example 3.4. You and nine friends are having a disagreement about what proportion of people believe that aliens have visited Earth. You decide to collect some data as follows: each of you will go out and ask random people if they believe. As soon as you find a believer, you will stop and report the number of people you talked to. For example, if the first person you survey believes that aliens have visited Earth, you will report the value 1. If the first three people you survey don't believe and the fourth one does, you will report the value 4.

Suppose that after the ten of you have conducted your surveys, the number of people that you each talked to is given by the following sorted list:

$$\{1,1,1,2,2,3,3,4,5,7\}$$

a) Let g be the proportion of all people who believe that aliens have visited earth. Explain why the probability of finding the first believer on the kth person that you talk to is given by: $p(g) = (1-g)^{k-1}g$.

b) Find the likelihood function for this data $L(g)$ and plot it on the interval $0 \leq g \leq 1$.

c) Compute $L(0.2)$ and $L(0.4)$ and write a brief description of what these values mean.

d) Find the log-likelihood function and use it to find the maximum likelihood estimate of g.

Solution:

a) In order for the first believer to occur on the kth trial, you must first talk to $k-1$ non-believers in a row. This occurs with probability $(1-g)^{k-1}$. Then the probability that the first $k-1$ people surveyed are non-believers and the kth person is a believer is $(1-g)^{k-1}g$.

b) We multiply the probabilities of each data point using our previous formula for $p(k)$:

$$\begin{aligned} L(g) &= p(1)p(1)\cdots p(7) \\ &= (g)(g)\cdots(1-g)^6 g \\ &= g^{10}(1-g)^{19} \end{aligned}$$

This is plotted in Figure 14.

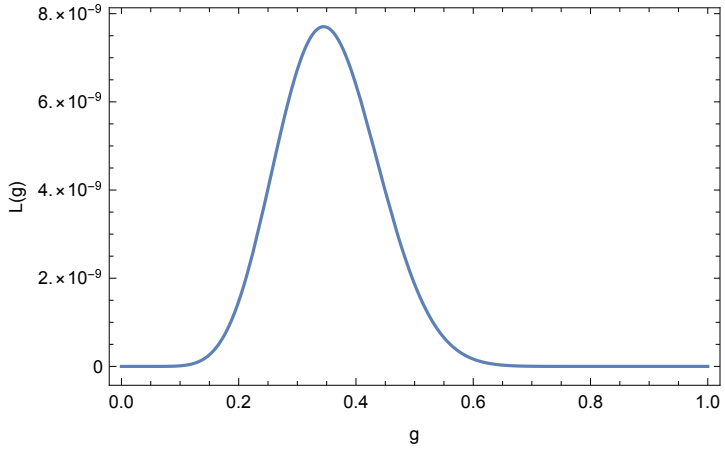

Figure 14: The likelihood function for the proportion of alien believers.

c) $L(0.2) = 0.2^{10}0.8^{19} \approx 1.5 \times 10^{-9}$ and $L(0.4) = 0.4^{10}0.6^{19} \approx 6.4 \times 10^{-9}$. These are the probabilities of obtaining our data set if the true value of g were 0.2 or 0.4, respectively.

d)

$$\begin{aligned} f(g) &= \ln(L(g)) \\ &= 10\ln(g) + 19\ln(1-g) \end{aligned}$$

Then we find the maximum by solving:

$$f'(g) = \frac{10}{g} - \frac{19}{1-g} = 0$$

This has the unique solution $g = \frac{10}{29}$.

3.3 Poisson Distribution

Example 3.5. The following table lists the number of business bankruptcies (in thousands) each year in the state of Colorado. Consider the time periods 2008-2015 and 2016-2023 separately, and assume that within each time period the bankruptcies followed a Poisson distribution. For each of these two time periods, find the MLE estimate of the annual rate of bankruptcies.

Solution:

We consider each year to be a time interval of length $T = 1$. Let b_1 be the bankruptcy rate during 2008-2015 and b_2 be the bankruptcy rate during 2016-2023. Using the Poisson distribution, the probability of 14 (thousand) bankruptcies in 2008 is given by $p(14) = cb_1^{14}e^{-b_1}$, where $c = \frac{1}{14!}$. Assuming

Year	Bankruptcies	Year	Bankruptcies
2008	14	2016	11
2009	19	2017	10
2010	25	2018	9
2011	27	2019	9
2012	26	2020	7
2013	21	2021	6
2014	18	2022	5
2015	14	2023	4

Table 1: Annual business bankruptcies in CO (thousands)

that each year is independent of each other year, then the likelihood function for the 2008-2015 data is found by multiplying the probabilities for each year and dropping the constants c:

$$\begin{aligned} L(b_1) &= b_1^{14} e^{-b_1} b_1^{19} e^{-b_1} ... b_1^{14} e^{-b_1} \\ &= b_1^{164} e^{-8b_1} \end{aligned}$$

The log-likelihood function is then:

$$f(b_1) = 164 \ln(b_1) - 8b_1$$

We find the maximum by solving:

$$f'(b_1) = \frac{164}{b_1} - 8 = 0$$

The maximum occurs at $b_1 = 164/8. = 21.5$ thousand per year.

Similarly, the log-likelihood function for the for the 2016-2023 data is:

$$f(b_2) = 61 \ln(b_2) - 8b_2$$

which has its maximum at $b_2 = 61/8 = 7.625$ thousand per year.

Follow-up Questions:

Q1: Someone who is skeptical about the usefulness of MLE might point out that the estimates of b_1 and b_2 could have been computed much more easily by just dividing the number of bankruptcies by the number of years. What additional value or insight is provided by the MLE approach?

A1: The MLE approach doesn't just produce the single "best" value of a parameter; it allows us to compare how much the data supports different possible values by examining the likelihood (or log-likelihood) function. For example, statisticians can use the likelihood function to produce a confidence interval that quantifies how much uncertainty we have about our estimate.

Q2: What have we assumed about how businesses going bankrupt affect each other? Does this seem reasonable?

A2: The MLE approach assumes that the annual data points are independent, and the Poisson distribution assumes that the bankruptcies (events) in any given year are independent of each other. So, we have assumed that a business going bankrupt doesn't affect any other businesses. This is not particularly realistic! Thus, the results of our analysis should be treated with some caution.

Q3: Look carefully at the data from 2016-2023. Does it seem reasonable to assume that there is a constant rate of bankruptcies (b_2) during this interval? If not, how might you think about improving the model?

A3: There seems to be a downward trend from 2016-2023 that isn't due just to random fluctuations. An improved model might replace the constant (b_2) with a decreasing function of time. This would require at least two parameters, which would turn this into a multivariable problem.

3.4 Exponential Distribution

Example 3.6. A mathematician named Beth was hard at work trying to prove a new theorem, but she kept getting interrupted by her puppy Dot invading her space to ask for attention or treats. She decided to start recording the times at which Dot interrupted her, in minutes from an arbitrary $t = 0$. Here is the data:

$$\{3, 8, 10, 17, 18, 20, 36, 42, 43, 44, 53, 58\}$$

a) Convert the data to time intervals between interruptions, then make a histogram of these intervals. Does it seem reasonable to model this data with an exponential distribution?

b) Assume that time intervals between interruptions follow an exponential distribution with rate parameter r. Find and graph the likelihood function for this data set.

c) Find the MLE estimate of r.

Solution:

a) We find the time intervals between interruptions by taking the difference between subsequent interruptions. This yields:

$$\{3, 5, 2, 7, 1, 2, 16, 6, 1, 1, 9, 5\}$$

The histogram is presented in Figure 15. While the data doesn't look exactly like an exponential distribution, it is reasonably close considering the small data set.

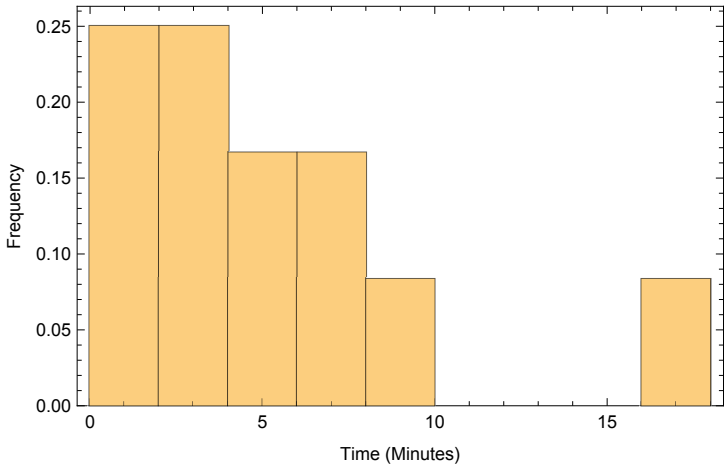

Figure 15: Histogram of puppy interruption time intervals.

b) Let t be the length of a time interval between interruptions. The exponential distribution uses the probability density function $p(t) = re^{-rt}$, where r is the unknown parameter. We get the likelihood function by plugging in all 12 time intervals to the density function and multiplying them together:

$$\begin{aligned} L(r) &= p(3)p(5)p(2)\cdots p(5) \\ &= (re^{-3r})(re^{-5r})(re^{-2r})\cdots(re^{-5r}) \\ &= r^{12}e^{-58r} \end{aligned}$$

This is plotted in Figure 16

c) We compute the log-likelihood function:

$$f(r) = 12\ln(r) - 58r$$

Then we find the maximum by solving

$$f'(r) = \frac{12}{r} - 58 = 0$$

From this we find the MLE solution: $r = 12/58 \approx 0.21$.

Follow-up Questions:

Q1: What are the units of r? How would you explain to a non-expert what this parameter means?

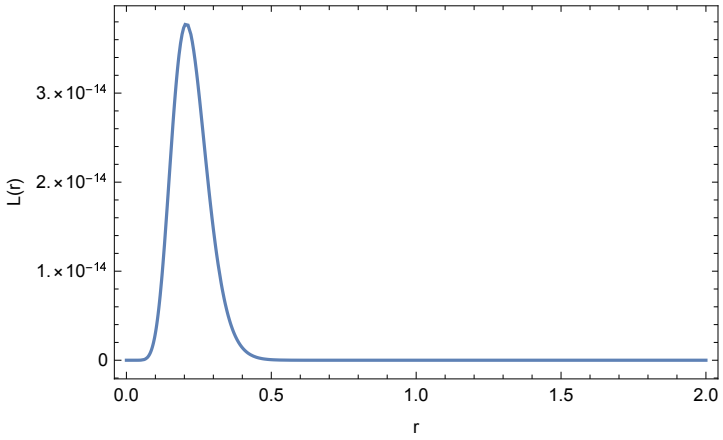

Figure 16: Likelihood function for puppy interruption rate.

A1: Since t is measure in minutes, r has units of 1/minute. One way to see this is that the exponent in the term e^{-rt} must not have any units. The parameter r represents the average rate of Dot's interruptions; according to our MLE result, she interrupts Beth on average 0.21 times per minute.

Q2: Beth has another dog, Archer, who is a little bit older and more mature. Would you predict that the value of r for Archer would be larger, smaller, or about the same as the value for Dot?

A2: Since older dogs are usually a bit more mellow, I would expect that Archer interrupts Beth less frequently, so she would have a smaller value of r.

Q3: An important assumption underlying the exponential distribution of time between events is that the system is "memoryless". This means that the probability of an event occurring is independent of how long it has been since the last event, so the system has no "memory" of previous events. Does this seem like a reasonable assumption in this situation? Can you think of a different situation in which the "memoryless" assumption would clearly be unrealistic?

A3: It is a well-known fact from biology that puppies can't remember anything that happened more than about 3 seconds ago. Thus, it may be reasonable to assume that Dot's next interruption is independent of how long it has been since her last interruption. There are many situations where the probability of an event occurring does depend on how long it has been since the last event. Some examples include earthquakes in a given fault system, economic recessions in a country, and interruptions by a significant other (person).

3.5 Pareto Distribution

Example 3.7. When meteors hit the surface of the earth, they leave an impact crater. Scientists use the size of the impact crater to estimate the energy of the meteor's impact. Table 2 lists the known impact craters with a diameter of at least 100 meters and less than 10,000 years old.

Name	Diameter (100 m)	Age (Years)
Wabar	1	200
Kaali	1	3500
Campo del Cielo	1	4500
Henbury	2	4700
Morasko	1	5000
Boxhole	2	5400
Ilumetsa	1	5600
Luna	16	7000
Macha	3	7300

Table 2: Meteor impact crater diameters (in 100 m).

a) Researchers believe that the size of impact craters may follow a Pareto distribution, which is a specific probability distribution with a peak at the lowest value, and a "tail" that drops off quickly but includes some extremely large values. Does this seem to describe the data in this table? You may want to use a histogram to support your answer.

b) The Pareto distribution has probability density function $p(x) = \frac{a}{x^{a+1}}$. Graph this function on the interval $1 \leq x \leq 5$ for several different values of a (including $a = 1, 2, 3$) and describe what effect the value of a has on the shape of the distribution.

c) Assuming that the impact crater diameters are independent of each other, find the likelihood function for this data set. Then graph it on the interval $0 \leq a \leq 10$. Report a range of values where you are fairly certain that the true value of a lies.

d) Find the log-likelihood function and use it to compute the MLE value of a.

Solution:

a) Five of the nine diameters are at the lowest value (1), and the frequencies drop off quickly as we go up from there. The data set does include one extremely large diameter crater. Thus, a Pareto distribution seems plausible.

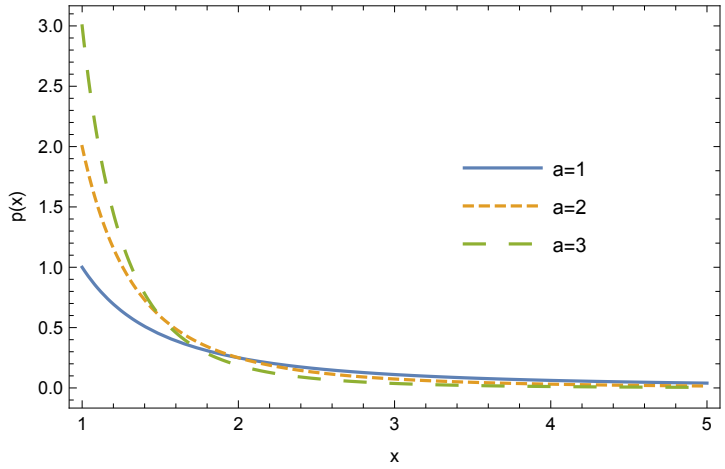

Figure 17: Examples of Pareto distribution for different values of the shape parameter a.

b) The distributions are plotted in Figure 17. We see that the higher the value of a is, the more the distribution is "concentrated" at $x = 1$ and the faster it decreases as x increases.

c) We get $L(a)$ by plugging each crater diameter into $p(x)$ and multiplying these terms:

$$\begin{aligned} L(a) &= p(1)p(1)p(1)\cdots p(16)p(3) \\ &= \left(\frac{a}{1^{a+1}}\right)\left(\frac{a}{1^{a+1}}\right)\left(\frac{a}{1^{a+1}}\right)\cdots\left(\frac{a}{16^{a+1}}\right)\left(\frac{a}{3^{a+1}}\right) \\ &= a^9\left(\frac{1}{192^{a+1}}\right) \end{aligned}$$

This is plotted in Figure 18. The peak of the distribution appears to be slightly below $a = 2$. The range of "fairly certain" values will depend on the individual respondent, but something like $1 \leq a \leq 3$ seems plausible.

d)

$$f(a) = \ln(L(a)) = 9\ln(a) - (a+1)\ln(192)$$

We find the maximum by solving:

$$f'(a) = \frac{9}{a} - \ln(192) = 0$$

This yields the MLE value: $a = \frac{9}{\ln(192)} \approx 1.71$.

Follow-up Questions:

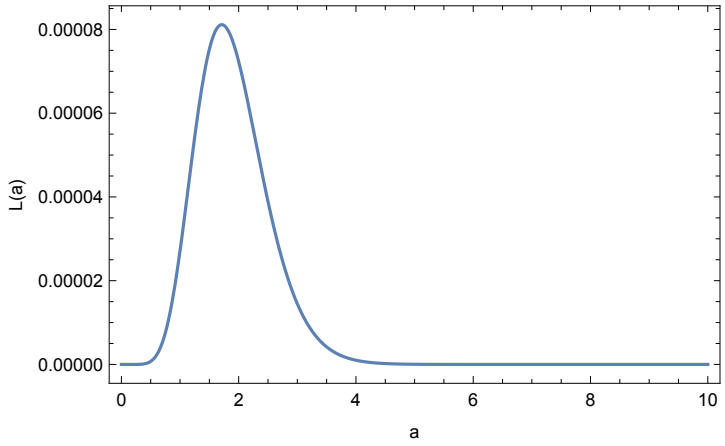

Figure 18: Likelihood function for impact crater diameters.

Q1: Why is it useful to have an estimate of the parameter a?

A1: Among other reasons, the value of a is useful for predicting the distribution of meteor sizes. It could help researchers estimate the probability of large meteors striking Earth in the near future, which would be nice to know (maybe).

Q2: Suppose that a newly discovered crater with a small diameter were added to the data set. How would this affect the shape of $L(a)$ and the MLE estimate? What if a newly discovered crater with a large diameter were added to the data set?

A2: Adding a crater with a small diameter would mean that the Pareto distribution is more concentrated near $x = 1$, which means that the MLE value of a would increase, and the peak of $L(a)$ would shift to the right. Similarly, a large crater added to the data set would lead to a lower estimate of a and a shift of $L(a)$ to the left.

4 Additional Single-Variable Problems

4.1 Binomial Distribution

Example 4.1. In 2015, the US Transgender Survey asked 21,065 working age transgender individuals about their employment status. Of the respondents, 3,049 were unemployed. Find the maximum likelihood estimate for the proportion of transgender individuals in the US who were unemployed in 2015.

Follow-up Questions:

Q1: What are some possible sources of error in the data?

Q2: Our MLE value is an estimate of the true transgender unemployment rate in 2015, which may have been different. Since the number of data points is fairly large, an important result from statistics says that we can obtain a 95% confidence interval for the true proportion unemployed from the formula: $p \pm 1.96\frac{1}{\sqrt{-f''(p)}}$, where p is our MLE value. Find the 95% confidence interval for the true transgender employment rate in 2015, and interpret what it means.

Q3: According to the Bureau of Labor Statistics, the overall national unemployment rate in the U.S. in 2015 was about 5%. Based on our confidence interval, explain why it is not plausible that the actual unemployment rate among transgender adults was also 5%.

Example 4.2. The Colorado Pikeminnow was once an abundant fish species native to much of the southwest United States. The species was declared endangered in 1973. In 2000, the Colorado Pikeminnow's population in the Green River Basin was estimated using a process known as mark-recapture. In mark-recapture, a sample of the population is captured and marked or tagged (for example, by giving a bird a small anklet cuff or giving a fish a microchip), and released back into the wild. Then, another sample of the population is captured after a period of time, and data are recorded on how many of the second sample had been tagged in the first sample.

Researchers tagged 1470 Pikeminnow in 2000. The following year they trapped 1540 individuals, of which 257 had been tagged in 2000. Use MLE to estimate N, the total population of Pikeminnow in the region being studied.

Follow-up Questions:

Q1: What are some possible sources of error in our model?

Q2: Suppose that the second year's sample of 1540 individuals had 10 more or 10 fewer tagged individuals (i.e. 247 or 267). How much would this change the estimated total population size?

Q3: How should researchers convey to policy-makers or the general public the appropriate level of confidence that they should have in this result?

Q4: Our MLE value is an estimate of the true population size, which may be different. Since the number of data points is fairly large, an important result from statistics says that we can obtain a 95% confidence interval for the true population size from the formula: $N \pm 1.96 \frac{1}{\sqrt{-f''(N)}}$, where N is our MLE value. Find the 95% confidence interval for the true population of pikeminnow in this region, and interpret what it means.

Example 4.3. Tay-Sachs disease is a rare genetic disorder passed from parents to children. Its incidence rate is particularly high among people of Eastern European and Ashkenazi Jewish descent. It is caused by the absence of an enzyme that helps break down fatty substances. These fatty substances, called gangliosides, build up to toxic levels in the brain and spinal cord and affect the function of the nerve cells. A gene responsible for producing the enzyme comes in two different forms, called alleles. The two different alleles are denoted T and t. Allele T contributes normal production of the enzyme, while allele t contributes little or no production of the enzyme.

Every person carries two copies of the gene, one inherited from their biological mother and one from their biological father. Thus, there are three possible "genotypes": TT, Tt, and tt. Individuals with genotype tt have Tay-Sachs disease because they do not produce the required enzyme. Individuals with genotype Tt are called "carriers" of Tay-Sachs; they usually do not have any disease symptoms but may pass on the t allele to offspring. Individuals with genotype TT are unaffected by Tay-Sachs.

In a population, the frequency of a specific allele refers to what fraction of all alleles for the gene (pooled from all individuals) are of the specific kind. For example, consider a group of four individuals, whose genotypes are (TT, TT, Tt, tt). Within this group, the frequency of allele t is 3/8.

Volunteers from a high-risk population of Ashkenazi Jewish descent were screened for their Tay-Sachs related genotypes. The results are present in Table 3.

Genotype	Number of Individuals
TT	3477
Tt	122
tt	1

Table 3: Tay-Sachs genotypes in a population

a) Let θ be the frequency of allele t in this population; this is what we will soon estimate. Explain what the following quantities represent: θ^2, $(1-\theta)^2$, $2\theta(1-\theta)$.

b) Write down the likelihood function for θ, then find the MLE estimate of θ.

Follow-Up Questions:

Q1: What would you want to know about how the data were collected in order to reason about whether the assumption of independent data points is reasonable here?

Q2: Graph either the likelihood function or the log-likelihood function. Briefly explain what kind of additional information is conveyed by the graph, beyond the location of the MLE value.

Q3: Suppose that the data set had 10 times as many individuals, but with the same relative proportions as in this data. What would change about the MLE estimate? What would change about the likelihood function?

Q4: Our MLE value is an estimate of the true allele frequency, which may be different. Since the number of data points is fairly large, an important result from statistics says that we can obtain a 95% confidence interval for the true proportion unemployed from the formula: $\theta \pm 1.96\frac{1}{\sqrt{-f''(\theta)}}$, where θ is our MLE value and $f(\theta)$ is the log-likelihood function. Find the 95% confidence interval for the true allele frequency in this population and interpret what it means.

Q5: Suppose that the data set had more individuals, but with the same relative proportions as in this data. What would happen to the confidence interval according to our formula? Does this agree with your intuition?

4.2 Geometric Distribution

Example 4.4. Epidemiologists use contact tracing to study patterns of how an infectious disease spreads in a population. One kind of data that comes from contact tracing is an estimate of how many new infections (cases) are caused by each infected individual. This is known to be highly variable even for a specific disease in a particular population. It is often the case that many infected individuals don't spread the disease at all, while some infected individuals cause many new cases. This variation is due to the complex interaction of many factors including genetics, age, behavior, and luck.

Suppose that epidemiologists identified 100 individuals who transmitted a particular disease to at least one other person. The number of transmissions and the number of instances are listed in Table 4. For example, the first line of the table means that 37 individuals were found to have transmitted the disease to exactly one other individual.

For reasons that are not fully understood, the number of transmissions sometimes follows a geometric distribution. This means that the probability of causing k new cases is given by: $p(k) = r^{k-1}(1-r)$ for some unknown parameter r.

Use MLE to estimate the value of r for this data. Include a graph of the likelihood function and discuss the validity of any assumptions that you had to make in your analysis.

Transmissions	Number Observed
1	37
2	24
3	17
4	13
5	7
6	2

Table 4: Number of infectious disease transmissions

Follow-up Questions:

Q1: Explain what your estimated value of r means in a way that a non-expert could understand.

Q2: Use your model to predict what fraction of individuals who cause at least one transmission cause 10 transmissions.

Q3: Use your model to predict what fraction of infected individuals cause no transmissions, or explain why you can't do this with the available information.

Q4: Our MLE value is an estimate of the true parameter value, which may be different. Since the number of data points is fairly large, an important result from statistics says that we can obtain a 95% confidence interval for the true proportion unemployed from the formula: $r \pm 1.96\frac{1}{\sqrt{-f''(r)}}$, where r is our MLE value and $f(r)$ is the log-likelihood function. Find the 95% confidence interval for the true value of r in this population and interpret what it means.

Example 4.5. Snow avalanches are an important ecological process in many mountain ranges, and can also be a threat to people. In North America, avalanches are rated on their destructive potential (severity) using a numerical scale between 1 (least severe) to 5 (most severe). During the month of January 2024, the Colorado Avalanche Information Center recorded data for the region around Aspen, CO, presented in Table 5.

Severity	Number Observed
1	170
1.5	115
2	92
2.5	29
3	7

Table 5: Avalanches near Aspen, CO

Suppose that experts believe that a geometric distribution may be a good model for the relationship between avalanche severity and frequency. This means

the following: Suppose that the probability that an avalanche has severity 1 is given by C. Then the probability that an avalanche has severity 1.5 is rC, the probability that an avalanche has severity 2 is r^2C, and so forth. Each time we increase to the next severity value, we multiply the previous probability by r. In this model, there is no maximum possible severity. In order to get the probabilities to sum to 1, it turns out that $C = 1-r$. Then the probability that an avalanche has severity x is given by $p(x) = (1-r)r^{2(x-1)}$. Find the MLE estimate for r using this data.

Follow-up Questions:

Q1: Use the model with your parameter estimate to predict what fraction of avalanches under these conditions would have a severity of 4. What about 5? How reliable do you think these prediction are?

Q2: How reasonable is it to assume that the data points in this data set are independent of each other? You may want to look up some information about what conditions or events lead to snow avalanches.

Q3: Other experts think that a better model of avalanche severity is a linear function: $p(x) = mx + b$. In order for the probabilities from $x = 1$ to $x = 5$ to sum to 1, it must be the case that $b = -3m + \frac{1}{9}$, so that $p(x) = m(x-3) + \frac{1}{9}$. Show that in order for $p(1) \geq 0$ and $p(5) \geq 0$, the values of m are constrained to $-\frac{1}{18} \leq m \leq \frac{1}{18}$. Find the log likelihood function for the data using this model. Show that it has no local maximum on the feasible interval, and the global maximum occurs at $m = -\frac{1}{18}$. Use this model to predict what fraction of avalanches would have severity of 4 or 5. How might researchers decide between the two proposed models?

4.3 Poisson Distribution

Example 4.6. Table 6 lists the number of serious fires (2-, 3-, and 4-alarm fires) responded to by the Chicago Fire Department during each month of 2023. We will assume that probability of k fires in a month is modeled by the Poisson distribution: $p(k) = \frac{r^k e^{-r}}{k!}$, where r is the rate of fires per month.

Month	Fires	Month	Fires
Jan	2	Jul	1
Feb	3	Aug	4
Mar	3	Sep	1
Apr	5	Oct	1
May	4	Nov	1
Jun	3	Dec	1

Table 6: Serious Fires in Chicago, 2023

a) Assuming independence, the likelihood function is given by:

$$L(r) = p(2)p(3)\cdots p(1)$$

However, we can simplify things by omitting the $k!$ denominator terms. Explain why this change will not affect the value MLE value of r.

b) Using the simplification from part a), write down the (modified) likelihood function and graph it. Based on your graph, estimate the MLE value of r and report a range of values that you are fairly confident contains the true value of the monthly rate of serious fires in Chicago.

c) Use the log-likelihood function to find the MLE value of r.

Follow-up Questions:

Q1: Our model assumes that the number of fires in any month fluctuates randomly around the average. In other words, we assume that there is no systematic variation due to the season or other factors. Does this seem intuitively plausible? Does this seem to agree with the data?

Q2: Use your MLE estimate and the Poisson distribution to find the probability of Chicago experiencing 6 serious fires in a month. Suppose that Chicago actually suffers 6 serious fires next month. Would you conclude that this could just be part of the random fluctuations, or would you be quite sure that something nefarious is happening (such as a serial arsonist)? What if there were 10 fires next month?

Example 4.7. No manufacturing process is perfect, and there is always some rate of defects (errors) in the product. Understanding the distribution of defects is a critical step in quality control and process improvement.

Suppose that a factory makes high-quality glass for optical instruments used in scientific research. Quality control engineers collect 100 samples of the glass (each 10 cm^2) and record the number of defects in each one, including scratches, chips, and embedded particulates. The data is presented in Table 7. If we

Defects	Number of Samples
0	61
1	30
2	6
3	1
4	2

Table 7: Manufacturing defects.

assume that defects occur randomly and independently of each other at a fixed rate, the number of defects per sample should follow a Poisson distribution. Use MLE to estimate the rate of defects (per 10 cm^2). Include a graph of the likelihood function.

Follow-up Questions:

Q1: Someone might point out that simply dividing the total number of defects by the total number of samples yields the same estimate, $r = \frac{53}{100}$. That is certainly easier than the MLE approach. What, if any, additional information is gained by going through the whole MLE process?

Q2: Use your estimated rate and the Poisson distribution to compute how many samples "should" have been found with each number of defects. Are there any discrepancies between the data and the theoretical prediction that are serious enough to cast doubt on the model and its assumptions?

4.4 Exponential Distribution

Example 4.8. Recent research has shown that clouds are made out of ice crystals, not cotton candy as previously believed. The size distribution of ice crystals in a cloud is affected by factors including temperature, humidity, wind, etc. Understanding these relationships is important for predicting cloud formation, which in turn affects short term weather forecasting and longer term climate predictions.

Scientists study the size of ice crystals in clouds by flying through them in airplanes equipped with specialized equipment. Typically, the size distribution of ice crystals with a diameter above 500 μm follows an exponential distribution, with probability density function $p(x) = re^{-rx}$, where x is the amount by which the diameter exceeds 500 μm.

Suppose that researchers collect data from a specific location in a cloud and find the ice crystal diameters in Table 8.

510	530	590	590	610
640	650	740	740	780
790	870	900	950	1020
1160	1210	1340	1390	1520

Table 8: Ice crystal diameters in μm

Use maximum likelihood to estimate the value of the size decay rate r from this data set. Be sure to first subtract 500 from every diameter. Include a graph of the likelihood function.

Follow-up Questions:

Q1: What are the units of the parameter r? What is the physical meaning of r?

Q2: What happens to the MLE estimate if the the largest ice crystal in the dataset is 2520 instead of 1520 μm?

Q3: What would happen to the shape of likelihood function and the estimate of r if the dataset consisted of 10 copies of each value listed in our dataset? Explain your reasoning without doing any new calculations.

Example 4.9. Scientists study the distribution of lifespans of different organisms for many reasons; there are implications for evolutionary theory, genetics, and human health. Many species that you are familiar with exhibit "senescence", in which the death rate (i.e. probability of dying in a given time interval) increases as an organism gets older. However, some organisms do not appear to exhibit senescence: the probability of dying in a given time interval remains approximately the same no matter how old the organism is. One example is the naked mole-rat, which can live over 30 years in captivity, far longer than other small rodents. (You should immediately stop what you are doing and look up pictures of naked mole-rats online.)

If the mortality rate (risk of dying per unit of time) is a constant r, then it turns out that lifespan will follow an exponential distribution. The probability density of living t years is: $p(t) = re^{-rt}$. Suppose that researchers collect the data in Table 9 on how long naked mole-rats live in the lab. Each data point represents the time in years that an individual lived.

25.4	14.8	13.2	18.1	3.9
6.0	2.4	36.5	3.7	6.6

Table 9: Naked mole-rat lifespans in years.

Use maximum likelihood to estimate the value of the death rate r. Include a graph of the likelihood function.

Follow-up Questions:

Q1: Find the average lifespan in the data. How is it related to the MLE value of r?

Q2: It may not be obvious why a constant death rate leads to an exponential distribution of lifespans. Let's explore that. Let r be the constant per capita death rate and let $N(t)$ be the number of individuals alive at time t.

- Explain why $N(t) - N(t + \Delta t) \approx rN(t)\Delta t$ when Δt is small.
- Rearrange the previous expression and take the limit $\Delta t \to 0$ to obtain the differential equation: $\frac{dN}{dt} = -rN$.
- Show that $N(t) = N(0)e^{-rt}$ satisfies the differential equation.
- Explain why the previous result implies that the *fraction* of individuals still alive at time t is given by $F(t) = e^{-rt}$.
- Explain why $F(t) - F(t+\Delta t) \approx p(t)\Delta t$ for small Δt, where $p(t)$ is the probability density of having a lifespan of length t. (Your argument should just depend on what $F(t)$ and $p(t)$ represent conceptually, not on any specific formulas.)

– Rearrange the previous expression and take the limit $\Delta t \to 0$ to obtain: $p(t) = -\frac{dF}{dt}$.

– Finally, compute $p(t)$ using this last result and the formula you derived for $F(t)$.

4.5 Pareto Distribution

Example 4.10. As of 2024, there had been 30 hurricanes to hit the United States and cause $10 billion or more in damage (inflation-adjusted to 2024 dollars) according to the National Oceanic and Atmospheric Administration (NOAA). They are listed in order of decreasing cost in Table 10 in units of $10 billion.

Name	Year	Cost	Name	Year	Cost
Katrina	2005	20.1	Rita	2005	2.9
Harvey	2017	16.0	Laura	2020	2.8
Ian	2022	12.0	Charley	2004	2.7
Maria	2017	11.5	Hugo	1989	2.3
Sandy	2012	8.9	Irene	2011	1.9
Ida	2021	8.5	Frances	2004	1.6
Helene	2024	7.9	Agnes	1972	1.6
Irma	2017	6.4	Allison	2001	1.5
Andrew	1992	6.1	Betsy	1965	1.4
Ike	2008	4.3	Matthew	2016	1.3
Milton	2024	3.4	Jeanne	2004	1.2
Ivan	2004	3.4	Floyd	1999	1.2
Michael	2018	3.1	Camille	1969	1.2
Florence	2018	3.0	Georges	1998	1.2
Wilma	2005	3.0	Fran	1996	1.0

Table 10: Costliest hurricanes in U.S. history to 2024, in units of $10 billion.

The Pareto distribution is often used to model quantities that have a highly skewed distribution that includes extremely large values. The Pareto distribution has probability density function $p(x) = \frac{a}{x^{a+1}}$ for $x \geq 1$, where a is a parameter.

a) Make a histogram of the hurricane costs. Comment on any striking aspects of the shape of the distribution.

b) Assuming that all of the hurricane costs are independent of each other, write down the likelihood function for the parameter a. Convert it to the log-likelihood function and find the MLE value of a.

b) Graph $p(x)$ for $1 \leq x \leq 25$ using your estimate of a and comment on the shape of the distribution.

Follow-Up Questions:

Q1: Even after adjusting for inflation, the data indicates that extremely expensive hurricanes were more common after 2000 than before. What are some reasons that this might be the case?

Q2: Explain why $10p(55)$ is an estimate for the predicted fraction of expensive hurricanes that would cause between \$500 billion and \$600 billion of damage.

Q3: Using your MLE estimate of a, compute $10p(55)$. Then decide which of the following is a correct description of the value that you computed:

a) The probability that a hurricane will hit the U.S. next year and cause between \$500 billion and \$600 billion of damage.

b) The probability that a hurricane will ever hit the U.S. and cause between \$500 billion and \$600 billion of damage.

c) The fraction of hurricanes hitting the U.S. that will cause between \$500 billion and \$600 billion of damage.

d) Of hurricanes hitting the U.S. that will cause at least \$10 billion of damage, the fraction that will cause between \$500 billion and \$600 billion of damage.

Example 4.11. The Pareto distribution is named after the Italian engineer and economist Vilfredo Pareto, who introduced it in his study of income distributions in the late 1800s. It has been shown to be a good fit for many socio-economic variables that display a strongly skewed pattern, with some values that are very far above the median.

Suppose that you survey 10 random restaurants in an American city and record their revenue in the last 30 days. Here is the data (in units of \$1000):

$$\{85, 114, 92, 76, 89, 510, 84, 98, 214, 78\}$$

Rescale the data by dividing all of the data points by the smallest value, so that the minimum becomes 1. Then fit a Pareto distribution to the data, finding the MLE value of the parameter a. Report any assumptions that you have made in your analysis, and provide a graph to support the decision to use a Pareto distribution for this data set.

Follow-up Questions:

Q1: The Pareto distribution is sometimes described as following an "80:20" rule. For example, the top 20% wealthiest individuals have about 80% of a country's wealth. However, this rule is only satisfied for the Pareto distribution with a particular value of the parameter "a". Does it hold for the restaurant data? If not, suggest an alternative rule.

Q2: Our model is based on the assumption that the data points are independent of each other. Do you have any concerns about whether this is a reasonable assumption for this data set? Is there anything that could cause restaurants' revenues to affect each other?

5 Solved Multi-Variable Problems

5.1 Multinomial Distribution

Example 5.1. A jar contains an unknown number of red marbles, blue marbles, and yellow marbles. One at a time, twenty marbles are sampled with replacement; seven of them are red, ten of them are blue, and three are yellow. Find the maximum likelihood estimate for the proportion of each color of marbles in the jar.

Solution: Let p_r be the proportion of red marbles in the jar and p_b be the proportion of blue marbles. Then the proportion of yellow marbles is just $1 - p_r - p_b$ so we have two parameters to estimate. The likelihood function for this data is: $L(p_r, p_b) = cp_r^7 p_b^{10}(1 - p_r - p_b)^3$ for some constant c. Here c is the multinomial coefficient $\binom{20}{7,10,3}$ but we do not need to know its value for finding the MLE values, and in fact we can drop it from the rest of the solution. Note that we are constrained to the region $0 \leq p_r + p_b \leq 1$. The contour plot for this function is plotted in Figure 19, normalized with $c = 1$. The log-likelihood function is:

$$f(p_r, p_b) = 7\ln(p_r) + 10\ln(p_b) + 3\ln(1 - p_r - p_b)$$

We find the maximum by solving:

$$\begin{aligned} \frac{\partial f}{\partial p_r} &= \frac{7}{p_r} - \frac{3}{1 - p_r - p_b} &&= 0 \\ \frac{\partial f}{\partial p_b} &= \frac{10}{p_b} - \frac{3}{1 - p_r - p_b} &&= 0 \end{aligned}$$

If we clear the denominators and rearrange, we get the system of equations:

$$\begin{aligned} 10p_r + 7p_b &= 7 \\ 10p_r + 13p_b &= 10 \end{aligned}$$

This has the unique solution $p_r = 0.35$, $p_b = 0.5$ as expected.

Follow-Up Question:

Q: Confirm that the critical point of $f(p_r, p_b)$ is a local maximum using the second derivative test.

A: We need the following second derivatives:

$$\begin{aligned} \frac{\partial^2 f}{\partial p_r^2} &= -\frac{7}{p_r^2} - \frac{3}{(1 - p_r - p_b)^2} \\ \frac{\partial^2 f}{\partial p_b^2} &= -\frac{10}{p_b^2} - \frac{3}{(1 - p_r - p_b)^2} \\ \frac{\partial^2 f}{\partial p_r \partial p_b} &= -\frac{3}{(1 - p_r - p_b)^2} \end{aligned}$$

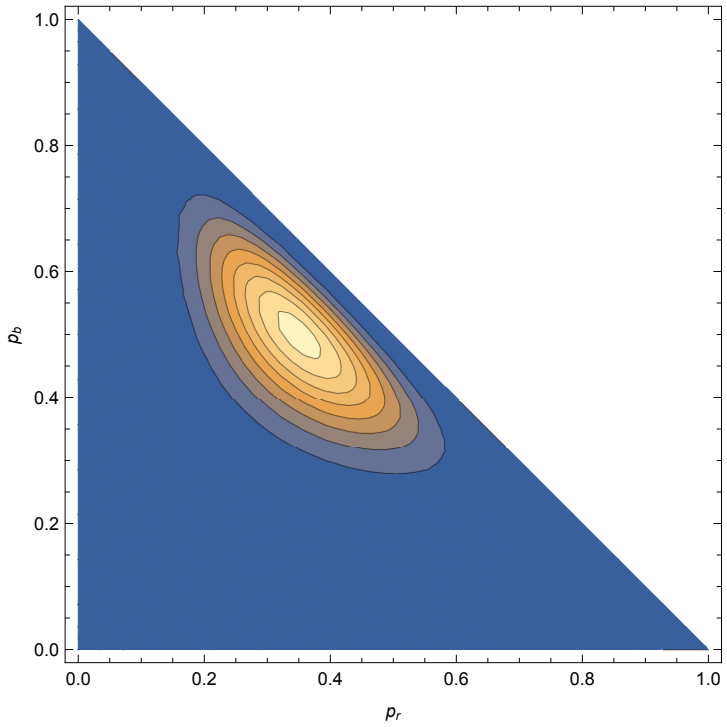

Figure 19: The likelihood function for three colors of marbles.

We plug in $p_r = 0.35$ and $p_b = 0.5$ to get the discriminant (aka the Hessian):

$$\begin{aligned} D(0.35, 0.5) &= \frac{\partial^2 f}{\partial p_r^2}\frac{\partial^2 f}{\partial p_b^2} - \left(\frac{\partial^2 f}{\partial p_r \partial p_b}\right)^2 \\ &\approx (-190)(-173) - (-133)^2 \\ &\approx 15238 \end{aligned}$$

Since this is positive and $\frac{\partial^2 f}{\partial p_r^2}$ is negative, we indeed have a maximum.

5.2 Poisson Distribution

Example 5.2. When concrete is made, adding various materials to the mixture can affect properties like its strength, weight, and workability. One common additive is fly ash, which is an inexpensive way to make concrete easier to shape, but reduces its strength. Suppose that researchers have tested concrete mixtures made with different amounts of fly ash, by placing samples of the concrete in a device that applies a compressive force, then counting the number of cracks

that appear. The data are presented in Table 11. Probability theory tells us

Fly Ash Percent	Number of Cracks
0	3
0	2
0	5
10	6
10	6
10	4
20	8
20	6
20	9

Table 11: Fly ash and concrete strength.

that the number of cracks in a sample is likely to follow a Poisson distribution. Then the probability of observing k cracks in a sample is: $p(k) = \frac{r^k e^{-r}}{k!}$, where r is the average number of cracks per sample. For the rest of this problem, we will dispose of the denominator $k!$ since it acts as a constant and does not affect the parameter estimates. The data suggest that the number of cracks per sample increases with the fly ash percent, perhaps linearly. Develop a model that incorporates this feature, and use MLE to estimate the effect of fly ash on the rate of crack formation.

Solution: We can replace the crack formation rate r with a linear function of fly ash percent (A): $p(k) = (mA + b)^k e^{-(mA+b)}$. For example, the probability of the fourth data point ($A = 10, k = 6$) is proportional to $p(6) = (m10 + b)^6 e^{-(m10+b)}$. We can reasonably assume that the data points here are all independent of each other, so we get the likelihood function by multiplying the probabilities of all of the data points:

$$\begin{aligned} L(m,b) &= p(3)p(2)\cdots p(9) \\ &= (b^3e^{-3})(b^2e^{-2})(b^5e^{-5})((10m+b)^6e^{-(10m+b)})\cdots((20m+b)^9e^{-(20m+b)}) \\ &= b^{10}e^{-3b}(10m+b)^{16}e^{-3(10m+b)}(20m+b)^{23}e^{-3(20m+b)} \end{aligned}$$

A contour plot of this function is presented in Figure 20.

The log-likelihood function is:

$$f(m,b) = 10\ln(b) - 3b + 16\ln(10m+b) - 3(10m+b) + 23\ln(20m+b) - 3(20m+b)$$

We find the maximum by solving:

$$\begin{aligned} \frac{\partial f}{\partial m} &= -90 + \frac{160}{10m+b} + \frac{460}{20m+b} &&= 0 \\ \frac{\partial f}{\partial b} &= -9 + \frac{10}{b} + \frac{16}{10m+b} + \frac{23}{20m+b} &&= 0 \end{aligned}$$

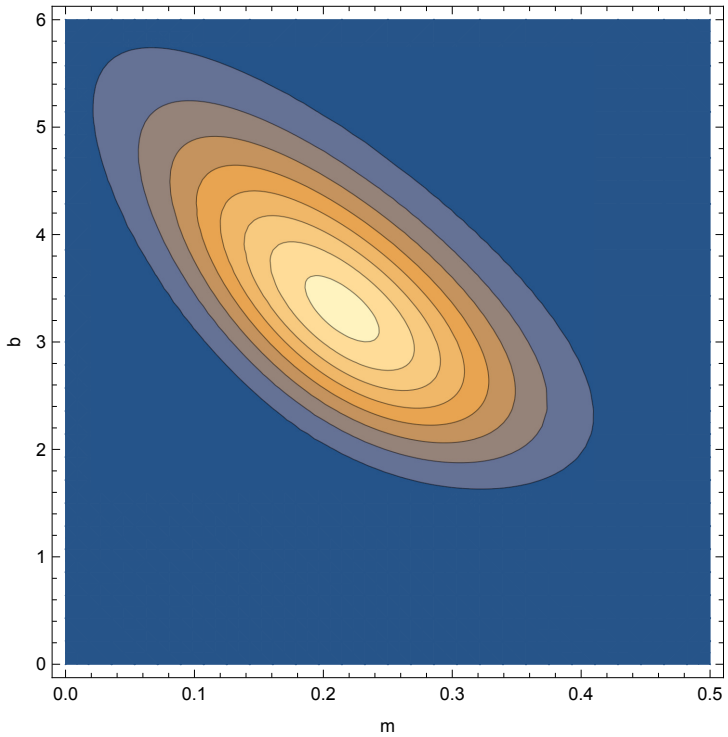

Figure 20: The likelihood function for concrete cracks.

If we clear the denominators this results in a system of two quadratic equations which can be solved exactly with some effort (or use of a computer algebra system). The solution is $m = \frac{637}{2970} \approx 0.214$ and $b = \frac{980}{297} \approx 3.30$. This result means that for every additional one percent of fly ash, we predict an additional 0.214 cracks per sample in the compression test.

Follow-up Questions:

Q1: In the Poisson distribution, the parameter r turns out to be the mean (average). Use your MLE parameter values to compute the model-predicted mean number of cracks in samples with 20% fly ash. Compare this to the mean number of cracks in samples with 20% fly ash in the data set. Do they agree exactly? Why do you think this is?

A1: The mean predicted by the model is $20m + b = 20\frac{637}{2970} + \frac{980}{297} \approx 7.59$. The mean from the three samples with 20% ash in the data set is $\frac{23}{3} \approx 7.67$. These are close, but not exactly the same. The reason is that the estimates of m and b were based on all of the data, not just the samples with 20% ash. The model must "compromise" in trying to fit all of the data, so it can't be expected to match the mean for any one subset perfectly.

Q2: Predict the distribution of cracks in concrete with 15% fly ash, or explain why you shouldn't.

A2: While we don't have data with 15% fly ash, this is well within the range of data to which we fit our model, so interpolation is reasonable. We would predict a Poisson distribution with parameter $r = 15m + b \approx 6.5$.

Q3: Predict the distribution of cracks in concrete with 50% fly ash, or explain why you shouldn't.

A3: We shouldn't use our model to do this. 50% fly ash is far outside the range of data to which the model was fit. We have no compelling reason to expect that the same linear pattern continues indefinitely, so we would be guilty of naive extrapolation, which is a sin.

5.3 Exponential Distribution

Example 5.3. Computer servers at a server farm in the desert occasionally fail. When events like server failures occur at random, the waiting time between failures typically follows an exponential distribution: $p(t) = me^{-mt}$. Here m is the failure rate, and $\frac{1}{m}$ is the average time between failures. Suppose that engineers suspect that the server farm's failures are affected by the outside temperature. In particular, they notice that the time between failures seems to go down if the day's high temperature is above 20° C (Table 12).

Temp. Above 20° C	Time Between Failures (h)
0	4
0	2
0	8
0	10
1	4
3	6
4	4
4	5
7	3
8	1
10	2
10	1

Table 12: Server farm failures.

Incorporate temperature dependence in a model of server failures and use MLE to determine how the failure rate depends on temperature using this data.

Solution: We can build a simple model in which the failure rate increases with temperature (so the time between failures decreases) by replacing the failure rate m in the exponential distribution with a linear function of T, the temperature above 20°. This yields the probability density of waiting t hours between failures: $p(t) = (a + bT)e^{-(a+bT)t}$. We get the likelihood function for the data

by plugging in the twelve (T, t) data points and multiplying the probabilities together:

$$\begin{aligned} L(a,b) &= (ae^{-4a})(ae^{-2a}) \cdots ((a+10b)e^{-2(a+10b)})((a+10b)e^{-(a+10b)}) \\ &= a^4(a+b)(a+3b)(a+4b)^2(a+7b)(a+8b)(a+10b)^2 e^{-(50a+117b)} \end{aligned}$$

A contour plot of the resulting likelihood function is presented in Figure 21.

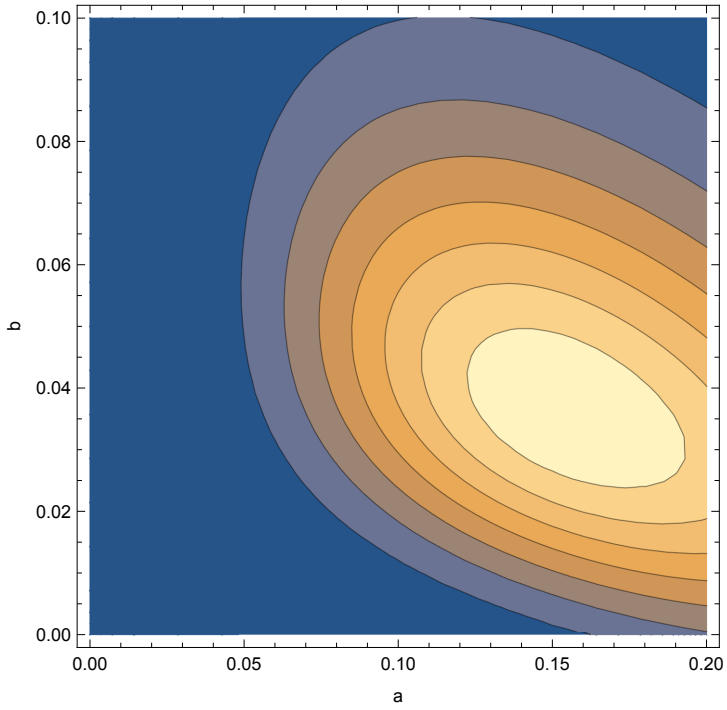

Figure 21: The likelihood function for server failures.

The log-likelihood function is:

$$\begin{aligned} f(a,b) &= 4\ln(a) + \ln(a+b) + \ln(a+3b) + 2\ln(a+4b) \\ &\quad + \ln(a+7b) + \ln(a+8b) + 2\ln(a+10b) - 50a - 117b \end{aligned}$$

We find the maximum by solving:

$$\begin{aligned} \frac{\partial f}{\partial a} &= -50 + \frac{4}{a} + \frac{1}{a+b} + \frac{1}{a+3b} + \frac{2}{a+4b} + \frac{1}{a+7b} + \frac{1}{a+8b} + \frac{2}{a+10b} &= 0 \\ \frac{\partial f}{\partial b} &= -117 + + \frac{1}{a+b} + \frac{3}{a+3b} + \frac{8}{a+4b} + \frac{7}{a+7b} + \frac{8}{a+8b} + \frac{20}{a+10b} &= 0 \end{aligned}$$

Unfortunately, even after clearing the denominators and simplifying we end up with a pair of 6th degree polynomials to solve. At this point we turn to numerical

methods and find that there is a unique solution with $a, b > 0$: $a \approx 0.155$, $b \approx 0.036$. This implies that for every one degree increase in temperature, the server failure rate increases by about 0.036 per hour.

Follow-up Questions:

Q1: Explain in non-mathematical terms what the assumption of independent data points means in this context. Does it seem like a reasonable assumption?

A1: Independence means that the failure of one server has no effect on the failure of another server. They may both be responding to some external factor like temperature, but they don't directly affect each other. This is a reasonable assumption as long as the failure of a server doesn't increase the workload on other servers and thereby increase their chance of failing.

Q2: Why did we choose to model the failure rate as a linear function of temperature, as opposed to some other function? What might cause us to make a different choice?

A2: Our decision was purely for convenience and simplicity. A detailed exploration of how the failure rate depends on temperature, especially if we had more data, might induce us to use a more complicated function, at the cost of more challenging calculations and perhaps more unknown parameters to estimate.

Q3: Use your fitted model to predict the mean time between failures if the outside temperature is 26° C or explain why we can't do that.

A3: 26° C corresponds to an "excess" temperature of 6, which is within the range of data that we used to parameterize the model. Thus, it is reasonable to estimate the failure rate as $a + 6b = 0.155 + 6(0.035) \approx 0.37$. This implies a mean time between failures of $\frac{1}{0.37} \approx 2.7$ hours.

Q4: Use your fitted model to predict the mean time between failures if the outside temperature is 35° C or explain why we can't do that.

A4: 35° C corresponds to an "excess" temperature of 15, which is outside the range of data that we used to parameterize the model. We have no good reason to expect that the linear pattern continues to such a high temperature, so we should not use our model to extrapolate. If you do it anyway, we can't be friends.

5.4 Logistic Regression

Example 5.4. Suppose that a political campaign is interested in the relationship between the number of times they contact someone, and whether the person donates money to the campaign. Table 13 presents data on 10 people identified as potential donors.

Contacts	Donation	Contacts	Donation
1	N	3	N
1	N	3	Y
2	N	4	Y
2	Y	4	Y
2	N	5	Y

Table 13: Political campaign potential donor data.

a) Based on this data, give an intuitive estimate of how many times the campaign should contact someone in order to have a 50% chance of getting a donation from them.

b) Let $p(x)$ be the logistic function. We will interpret $p(x)$ as the probability that someone who has been contacted x times makes a donation. Explain why the first data point in the table has probability $\frac{e^a}{e^b+e^a}$ and the last data point has probability $\frac{e^{5b}}{e^{5b}+e^a}$.

c) Assume that the data points are all independent of each other, and find the likelihood function for this data set, $L(a, b)$. Make a contour plot of this function in the region $0 \leq a \leq 10, 0 \leq b \leq 5$. Based on this plot, estimate the MLE values of a and b, and use them to predict how many times the campaign should contact someone in order to have a 50% chance of getting a donation from them. (Hint: we need $p(x) = 0.5$.)

d) Find the log-likelihood function, $f(a, b)$, as well as its first partial derivatives.

e) Use a computer algebra system or other numerical methods to find the MLE values of a and b. Compare these values to your prior guesses.

f) Plot $p(x)$ using your MLE parameter values, then write a brief nontechnical explanation of what this would mean to the political campaign.

Solution:

a) The shift from mostly "no" to mostly "yes" appears to occur between $x = 2$ and $x = 3$ contacts. I would guess around $x = 2.5$.

b) The first data point has $x = 1$ and y is a "no", so we evaluate $1 - p(1)$. The last data point has $x = 5$ and y is a "yes", so we evaluate $p(5)$.

c) We multiply ten terms of the form $p(x)$ and $1 - p(x)$ as appropriate for each data point:

$$\begin{aligned} L(a, b) &= (1 - p(1))(1 - p(1))(1 - p(2)) \cdots (p(4))(p(5)) \\ &= \left(\frac{e^a}{e^b + e^a}\right)\left(\frac{e^a}{e^b + e^a}\right)\left(\frac{e^a}{e^{2b} + e^a}\right) \cdots \left(\frac{e^{4b}}{e^{4b} + e^a}\right)\left(\frac{e^{5b}}{e^{5b} + e^a}\right) \end{aligned}$$

The contour plot for this function is in Figure 22. It looks like the maximum is at around $a = 5.5, b = 2$. By solving $p(x) = 0.5$, we find that $x = \frac{a}{b}$. Based on the estimated values, this would mean $x \approx \frac{5.5}{2} = 2.75$.

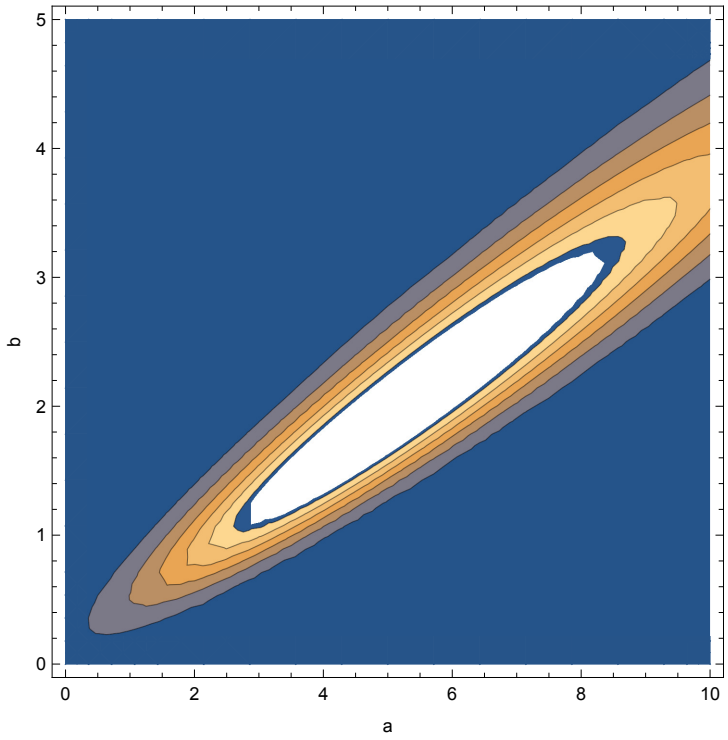

Figure 22: The likelihood function for campaign contributions.

d) Taking the logarithm of $L(a, b)$ and simplifying:

$$\begin{aligned} f(a,b) = & \ln\left(\frac{e^a}{e^b+e^a}\right) + \ln\left(\frac{e^a}{e^b+e^a}\right) + \cdots + \ln\left(\frac{e^{5b}}{e^{5b}+e^a}\right) \\ = & \ a - \ln(e^b+e^a) + a - \ln(e^b+e^a) + \cdots + 5b - \ln(e^{5b}+e^a) \\ = & \ 5a + 18b - 2\ln(e^b+e^a) - 3\ln(e^{2b}+e^a) - 2\ln(e^{3b}+e^a)\cdots \\ & -2\ln(e^{4b}+e^a) - \ln(e^{5b}+e^a) \end{aligned}$$

From this we get:

$$\begin{aligned} \frac{\partial f}{\partial a} = & \ 5 - 2\frac{e^a}{e^b+e^a} - 3\frac{e^a}{e^{2b}+e^a} - 2\frac{e^a}{e^{3b}+e^a} - 2\frac{e^a}{e^{4b}+e^a} - \frac{e^a}{e^{5b}+e^a} \\ \frac{\partial f}{\partial b} = & \ 18 - 2\frac{e^b}{e^b+e^a} - 6\frac{e^{2b}}{e^{2b}+e^a} - 6\frac{e^{3b}}{e^{3b}+e^a} - 8\frac{e^{4b}}{e^{4b}+e^a} - 5\frac{e^{5b}}{e^{5b}+e^a} \end{aligned}$$

e) Using the FindRoot command in Mathematica, I get $a \approx 5.11$ and $b \approx 1.96$. This is pretty close to what I guessed!

f) The result is plotted in Figure 23. The curve shows a prediction, based on our prior data, of how likely someone is to make a donation based on how many times they have been contacted. We see that people contacted once are very unlikely to donate, and people contacted four times are very likely do donate. We have a 50% success rate for people contacted between 2 and 3 times. I would recommend aiming for four contacts per potential donor, to get a high rate of donations without wasted effort.

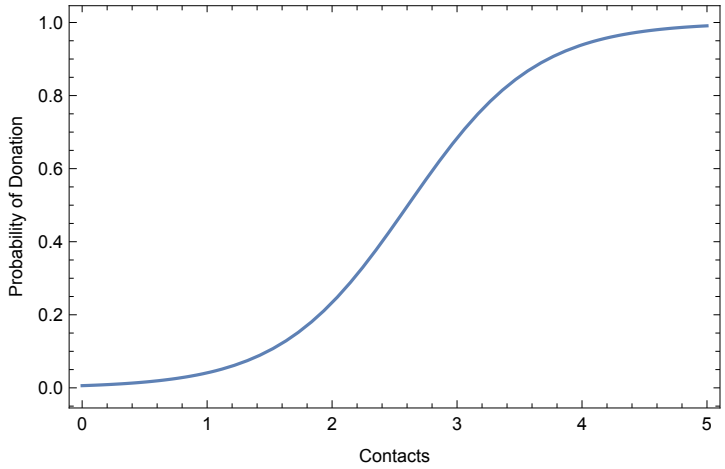

Figure 23: The predicted success rate for campaign contributions.

Follow-up questions:

Q1: The contour plot seems to show a "ridge" of (a, b) values that are (approximately) equally consistent with the data. What is the apparent relationship between a and b along this ridge, and what does this imply about the model's predictions in this case?

A1: The ridge appears to lie along a straight line through the origin, so along it $\frac{a}{b}$ is constant. Since $p(x) = 0.5$ when $x = \frac{a}{b}$, this indicates that we can reliably predict the value of x corresponding to a 50% success rate. However, all (a, b) points along the ridge are similarly consistent with the data, so we can't reliable determine the values of a and b separately. This means that we can't be certain about how steep the logistic curve is, and thus how sudden or gradual the transition between failures and successes may be. This is likely due to the small data set.

Q2: The campaign manager looks at your analysis and concludes that contacting a potential donor 20 times would be good strategy, since it

would essentially guarantee that they make a contribution. Explain what is wrong with this reasoning.

A2: This would be a case of severe extrapolation, which is a sin. Since we fit our model to data with $1 \leq x \leq 5$, there is no reason to expect that the pattern continues out to $x = 20$. Indeed, common sense suggests that potential donors may resent being spammed with too many contacts, and that the probability of a contribution goes back down if x is too high.

6 Additional Multi-Variable Problems

6.1 Poisson Distribution

Example 6.1. A radiation detector is placed near a sample of an unknown radioactive substance. The detector measures the passage of ionizing particles, which are released by the radioactive material as it decays. Table 14 presents the data; for each hour, the table presents the number of particles detected in the previous one-hour interval. If the amount of the radioactive material were

Hour	Particles Detected
1	31
2	26
3	27
4	19
5	17

Table 14: Ionizing particles detected by a radiation detector.

constant, probability theory tells us that the number of particles detected in an hour would be a random variable from the Poisson distribution. Let r be the average rate of particles passing through the detector per hour. Then the probability of detecting k particles in one hour would be given by: $p(k) = \frac{r^k e^{-r}}{k!}$. Since the denominator does not depend on the rate r, we can drop it for the rest of this problem. In our scenario, the amount of the radioactive material does not stay constant because it is breaking down. Specifically, the amount of radioactive material decreases exponentially as time passes. Incorporate this assumption by replacing the rate in the Poisson distribution with an exponentially decreasing function of time. Then write down the likelihood function for the data in the table and use it to estimate the substance's decay rate.

Follow-up questions:

Q1: Make a contour or surface plot of the likelihood function. Based on this, how much certainty do you have in your parameter estimate?

Q2: Convert the decay rate you estimated to a half-life. Then look up a list of half-lives of radioactive materials. Based on this, what are the 5 most likely identities of the unknown substance?

Q3: An alternative way to analyze the data would be more familiar to most scientists. Replace the number of particles detected each hour with the natural logarithm of that number. A plot of the time versus the log of particles detected should show a linear trend. Fit a straight line to this using ordinary least squares. The slope of the line is an alternative estimate of the decay rate. How different is the estimate found this way from our MLE estimate? Is the difference large enough to affect our identification of the mystery material?

Example 6.2. Researchers have found that calls to emergency services (911 calls in the United State) are related to temperature. In many cases, the frequency of emergency calls increases with temperature, due to direct effects (such as heat stroke) and indirect effects (increased risky behavior). Consider a hypothetical smallish town which has data presented in Table 15, with the daily high temperature (measured in degrees Celsius above 30) and the number of emergency calls received for a two-week period. If emergency calls are inde-

Day	Temperature	Calls	Day	Temperature	Calls
1	3	16	8	8	24
2	4	18	9	7	22
3	4	17	10	7	20
4	5	19	11	6	18
5	7	22	12	4	17
6	7	23	13	3	15
7	7	22	14	3	16

Table 15: Daily high temperature (^0C above 30) and number of emergency calls in a town.

pendent of each other, the number of calls in a day should follow the Poisson distribution: $p(k) = \frac{r^k e^{-r}}{k!}$. Here, r is the call rate, which is the average number of calls in a day. Since the denominator does not depend on the rate r, we will drop it for the rest of this problem.

A scatterplot of temperature versus number of calls reveals an approximately linear relationship: $calls \approx -35 + 1.5 \times T$. This suggests replacing the constant rate r in the Poisson distribution with an expression $a + bT$ to model how the call rate depends on daily high temperature. (Here, T is the temperature above 30^0C. Write down the likelihood function for the emergency call data in the table and use it to find the MLE values of a and b. (Note: you may need to use numerical methods to find the maximum.)

Follow-up questions:

Q1: What are the units and meaning of the parameters a and b?

Q2: Make a contour or surface plot of the likelihood function. Based on this, how much certainty do you have in your parameter estimate?

Q3: Using the Poisson distribution was based on the assumption that emergency service calls on a given day are independent of each other. Furthermore, when we wrote down the likelihood function we assumed that the number of calls on one day is independent of the number of calls on another day. How reasonable do you think each of these assumptions is? Explain your reasoning. (Note: replacing these assumptions with something more realistic would greatly increase the complexity of the model and the difficulty of estimating parameter values.)

Q4: Plot the probability distribution for the number of calls that the model would predict for this town on a day with high temperature 37^0C. Then do it again for a day with high temperature 40^0C. Discuss how much confidence an emergency service worker should have in each of these predictions.

6.2 Exponential Distribution

Example 6.3. Hurricanes in the Atlantic are rated on a severity scale of Category 1 (least severe) to Category 5 (most severe). Climate scientists predict that global warming, which causes ocean temperatures to increase, will lead to an increase in the frequency of very severe hurricanes because warm water contributes more thermal energy to hurricane development. The temperature of the ocean is a complex phenomenon, since it varies with season, location, and depth. However, a standardized measure of average ocean temperature has been developed: the Sea Surface Temperature Anomaly (SSTA). This provides a global yearly average surface ocean surface temperature, measured in degrees Celsius above or below a long-term baseline. Table 16 lists the year of each of the thirty Category 5 hurricanes that occurred in the Atlantic from 1961 - 2024, along with the SSTA for that year (data from www.epa.gov). In this example we will explore whether this data is consistent with the prediction of an increasing frequency of major hurricanes as ocean temperatures have increased.

Year	SSTA	Year	SSTA	Year	SSTA
1961	-0.15	1992	0.02	2016	0.52
1966	-0.22	1998	0.24	2017	0.46
1967	-0.23	2003	0.25	2017	0.46
1969	-0.06	2004	0.24	2018	0.41
1971	-0.29	2005	0.24	2019	0.49
1977	-0.04	2005	0.24	2019	0.49
1979	0.02	2005	0.24	2022	0.41
1980	0.04	2005	0.24	2023	0.64
1988	0.08	2007	0.16	2024	0.69
1989	0.04	2007	0.16	2024	0.69

Table 16: Year of each Category 5 hurricane in the Atlantic from 1961 - 2024, along with global mean sea surface temperature anomaly (SSTA) in that year.

Probability theory tells us that if events happen at random times, with an average rate of r events per time unit, the waiting time *between* events should follow an exponential distribution. This means that the probability density of waiting a time interval w between events is given by $p(w) = re^{-rw}$.

a) Convert the data to a list of waiting times between hurricanes, along with the SSTA. This will produce a list of 29 ordered pairs (T_i, w_i) of which the first one is $(-0.22, 5)$. (Do you see why we need to start with

the second hurricane?) Make a histogram of the waiting times (without the SSTA data) and comment on the shape of the distribution.

b) Find the overall average rate of Category 5 hurricane formation by simply dividing the total number of hurricanes by the total time period in the data set.

c) Make a scatterplot of the waiting times versus the SSTA. Can you see a slight downward trend? What does this suggest about the relationship between the rate of hurricane formation and ocean temperature?

d) We now begin the MLE process. Based on our exploration of the data, it seems reasonable to replace the constant rate r in the exponential distribution with linear function of temperature (SSTA). Explain how this leads to the following likelihood function:

$$L(a,b) = \prod_{i=1}^{29} (a + bT_i) e^{-(a+bT_i)w_i}$$

e) What are the meaning and units of the parameters a and b?

f) Find a formula for the log-likelihood function (without plugging in specific data yet). Then compute its partial derivative with respect to each parameter.

g) Plug the specific data points into your partial derivatives and set them equal to zero to yield two (ugly) equations that we need to solve for a and b.

h) Use a computer algebra system or other numerical methods to solve the equations above. If you need to provide a starting guess, recall your previous estimate for the overall rate r. Depending on your method, you may find more than one apparent solution. However, only one candidate solution (a, b) produces real-valued output when plugged into the log-likelihood function. This is our MLE solution!

i) Make a contour plot of the log-likelihood function in a region around the MLE values. Use it to confirm that the solution you found is reasonable.

Follow-up questions:

Q1: Based on this analysis, what will happen to the frequency of Category 5 hurricanes if the ocean warms by another 1^0C? What about 2^0C? Do you have equal confidence in both of these predictions?

Q2: Generally speaking, more data allows us to make more precise estimates of parameter values. Consider several ways in which we could increase the amount of data available to us in this analysis:

– Include other categories of hurricanes.

– Use a longer historical record (prior to 1961).

– Include hurricanes / typhoons from other parts of the world.

Assuming that our goal is to predict the frequency of major hurricanes in the Atlantic, discuss the pros and cons of expanding our potential data set in each of these proposed ways.

Example 6.4. Veterinary researchers are interested in the efficacy of a new medication for reducing the frequency of epileptic seizures in dogs. Suppose that they choose 16 dogs from the same breed, with similar histories of seizures. They give each dog one of 4 doses of the new medication (0, 1, 2, or 3 mg). Then they record the number of days until the dog's next seizure. The data is presented in Table 17.

Dose	Days	Dose	Days	Dose	Days	Dose	Days
0	10	1	20	2	28	3	42
0	11	1	21	2	32	3	43
0	12	1	24	2	35	3	50
0	15	1	30	2	48	3	55

Table 17: Medication dosage (mg) and time until next seizure in dogs.

Probability theory tells us that if random events occur at a constant average rate r, then "waiting time" until the next event follows an exponential distribution: $p(t) = re^{-rt}$. The average waiting time is $\frac{1}{r}$, so the higher the rate is, the less time on average until the event occurs. The data indicates that as the dose increases, the waiting time increases, which means that the rate of seizures decreases.

Assume that the seizure rate depends on the dose x in the form: $r = \frac{a}{1+bx}$. Use MLE to estimate values for a and b from the data. Include a contour plot of the log-likelihood function in the region $0.07 \leq a \leq 0.1$, $0.8 \leq b \leq 1.2$. Write a brief summary of your conclusions, including precise explanations of what your parameter values mean in terms that a medical researcher would use. (Hint: think about the units of a and b, using the fact that $1/r$ is an amount of time.)

Follow-up questions:

Q1: We assumed that the relationship between the seizure rate and dose was $r = \frac{a}{1+bx}$. What does this imply about the relationship between the time until a seizure and the dose? Does the data appear to be consistent with this?

Q2: Based on your results, how much confidence would you have in predicting the effect of a 4 mg dose? What about 10 mg?

Q3: What aspects of your results would you confidently apply to other breeds of dogs?

6.3 Gompertz Distribution

Example 6.5. A brilliant and wise biologist (who happens to be married to the author) carried out a study on the lifespans of painted lady butterflies. She raised them in controlled laboratory conditions, isolated from each other so that they did not have any affect on each other's health or behavior. The length (in days) of the adult stage of 24 individuals is presented in Table 18.

23	45	35	51	90	70
28	35	67	63	117	39
59	37	110	31	68	41
131	9	89	18	121	37

Table 18: Adult stage lifespans (in days) of 24 individual butterflies.

a) Make a histogram of the butterfly lifespan data. Comment on any patterns that you see. Is it clear from this histogram what happens to the hazard rate (i.e. risk of dying) over time?

b) Assume a Gompertz distribution for lifespans. Write down the general form for the likelihood function $L(a, c)$ and log-likelihood function $f(a, c)$ for an arbitrary dataset with n lifespans t_i. Since we are not plugging any data points in yet, your formulas should involve products and/or sums of n terms. Simplifying your expressions will make the next steps easier.

c) Find the pair of equations that must be solved to find the MLE parameter values. Hint:

$$\frac{\partial f}{\partial c} = \frac{n}{c} + n - \sum_{i=1}^{n} e^{at_i}$$

.

d) Now plug in the butterfly lifespan data and use a computer algebra system or other numerical approach to find the MLE parameter estimates. (If you need to provide a starting guess, try $a \approx 0.01, c \approx 0.5$)

e) Make a contour plot of the log-likelihood function in a region that includes your MLE values to verify that your solution is reasonable.

f) Plot the Gompertz distribution with your MLE parameter values and compare it to the histogram of the data.

Follow-up questions:

Q1: Plot the hazard function $h(t)$ with your MLE parameter values. Find $\frac{h(100)}{h(0)}$ and interpret what this means in biological terms.

Q2: I claim that the increasing risk of dying with age is not readily apparent in the histogram of lifespans. Can you explain why there is a peak in deaths followed by a decrease, even though the hazard function is monotone increasing?

Q3: Why was it important for our MLE procedure that the butterflies were raised in isolation from each other?

Example 6.6. Suppose that the human resources (HR) department at a technology company is concerned about their ability to retain software engineers. Table 19 shows the length of time that individual software engineers stayed with the company (in months) before leaving. HR notices that most departures occur at around one year of employment, and they are considering some possible strategies to address this problem. However, you are familiar with the Gompertz distribution so you decide to analyze the data by fitting a Gompertz model to the employment "lifetime" data. Carry out an analysis based on finding the MLE parameter values. Be sure to include a plot of the likelihood function and a plot of the Gompertz distribution and hazard function with the MLE parameter values. Based on your analysis, is HR right to be concerned about the spike in departures around 12 months, or should they focus on something else?

8	10	10	12
12	12	14	14
15	18	20	24
28	32	33	48

Table 19: Duration of employment in months for sixteen software engineers.

Follow-up questions:

Q1: In forming the likelihood function, what assumption did you make about the relationship between different employee's lengths of employment? Discuss how realistic you think this is.

Q2: The Gompertz distribution is just one possible model for this data. An important part of data analysis involves comparing the performance of different possible models. One way to do this is to fit different models to the same data set, and compare the maximum values of each model's likelihood function. However, a model with more parameters can almost always provide a better fit to a given data set than a model with fewer parameters. If we just score models based on the maximum likelihood, we will end up choosing more complicated models, even if they don't really do a better job of capturing the main trends or features of the data. A better approach is to "penalize" a model for having too many parameters, so that we balance the competing goals of simplicity and accuracy.
The most commonly used approach for comparing models was developed by Japanese statistician Hirotugo Akaike in the 1970s. Assume that we have a model with k parameters that we are fitting to a data set via MLE. (For the Gompertz distribution, $k = 2$.) Let f^* be the value of the log-likelihood function at its maximum (i.e. when the MLE parameter values are plugged in). The Akaike Information Criterion is defined as: $AIC = 2(k - f^*)$. Akaike showed that this quantifies the amount of

"information loss" when using the model to represent the data. The less information lost, the better, so when comparing different models the one with the smallest value of AIC is preferred.
Compute the AIC for the Gompertz model fit to the employment data. Then consider a simpler model based on an exponential distribution of time until an employee leaves: $p(t) = re^{-rt}$. Use MLE to fit this model to the same data and compute the resulting AIC. Finally, decide which model would be preferred for this data set.

6.4 Logistic Regression

Example 6.7. This example applies the method of logistic regression to some data. If you have not yet done so, study the introductory material on logistic regression.

Civil engineers are interested in the relationship between a bridge's age and whether it can pass a safety inspection. Suppose that engineers study 12 concrete overpass bridges. Table 20 presents data on their ages (in years) and whether they pass a safety inspection.

Age	Pass?	Age	Pass?
35	Y	58	Y
38	Y	62	Y
42	Y	64	N
46	Y	68	N
53	Y	71	N
55	N	75	N

Table 20: Bridge ages and whether they pass inspection.

a) Based on this data, give an intuitive estimate of how old a bridge of this type will be when it has a 50% chance of passing a safety inspection.

b) Let $p(x)$ be the logistic function. We will interpret $p(x)$ as the probability that a bridge that is x years old will pass a safety inspection. Explain why the first data point in the table has probability $\frac{e^{35b}}{e^{35b}+e^{a}}$ and the last data point has probability $\frac{e^{a}}{e^{75b}+e^{a}}$.

c) Assume that the bridges are all independent of each other, and find the likelihood function for this data set, $L(a, b)$. Make a contour plot of this function in the region $-30 \leq a \leq 0, -0.5 \leq b \leq 0$. Based on this plot, make a guess about possible MLE values of a and b, and use them to predict how old a bridge of this type must be to have a 50% chance of passing inspection.

d) Find the log-likelihood function, $f(a, b)$, as well as its first partial derivatives.

e) Use a computer algebra system or other numerical methods to find the MLE values of a and b. (If you need to provide starting guesses, choose a point near the middle of the contour plot.) Compare these values to your prior guesses.

f) Plot $p(x)$ using your MLE parameter values. Then plot $p(x)$ using another set of parameters from the apparent "ridge" in the contour plot. Discuss the similarities and differences in these curves.

Follow-up questions:

Q1: The contour plot seems to show a "ridge" of (a, b) values that are (approximately) equally consistent with the data. What is the apparent relationship between a and b along this ridge, and what does this imply about the model's predictions in this case?

Q2: Based on your logistic regression, at what age would you recommend that a bridge of this type be subject to more frequent inspections than the standard every two years? Explain your reasoning.

Q3: What would it mean for a logistic regression model to fit the data "perfectly"? What are some possible reasons that the fit is not perfect in the case of the bridge data?

Example 6.8. Astronomers have become quite proficient at detecting and characterizing exoplanets, which are planets that orbits stars other than our sun. Indeed, recent research indicates that most stars have planets orbiting them. If we are interested in planets that might harbor life, however, we need to be more selective. Based on our current understanding of biochemistry, life is most likely to exist on smallish rocky planets that are the right distance from their start (and thus the right temperature) to hold liquid water. Does that sound like a planet you know and love? Such "earth-like" planets in the "habitable zone" are still quite common! A 2013 paper by Petigura et al. estimated that around 22% of "sun-like" stars (similar in size and temperature to our sun) have earth-like planets in orbiting in the habitable zones.

Whether or not a star has an earth-like planet depends on a variety of characteristics, like the star's age, size, presence of a binary star, and so forth. Suppose that researchers select 12 sun-like stars, measure their sizes, and determine whether each has earth-like planet. (Star sizes are measured in units of solar radius, R_0. By definition, our sun has size $1R_0$.) The data is presented in Table 21.

Carry out logistic regression analysis, using MLE to determine parameter values. Include plots of the likelihood function and a plot of the best-fit logistic curve (on the interval $0.8 \leq x \leq 1.1$. Summarize key aspects of your findings in a few sentences. (Hint: carry out your parameter search near $a = 10$.)

Star Size	Earth-Like Planet?	Star Size	Earth-Like Planet?
0.8	N	1.0	N
0.8	N	1.0	N
0.8	N	1.0	Y
0.9	N	1.1	N
0.9	N	1.1	Y
0.9	Y	1.1	Y

Table 21: Star sizes and presence of exoplanets.