Introduction to Statistics
Maximum Likelihood Estimates
Class 10, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Know the three overlapping "phases" of statistical practice.
2. Know what is meant by the term *statistic*.
3. Be able to define the likelihood function for a parametric model given data.
4. Be able to compute the maximum likelihood estimate of unknown parameter(s).

2 Introduction to statistics

Statistics deals with data. Generally speaking, the goal of statistics is to make inferences based on data. We can divide this the process into three phases: collecting data, describing data and analyzing data. This fits into the paradigm of the scientific method. We make hypotheses about what's true, collect data in experiments, describe the results, and then infer from the results the strength of the evidence concerning our hypotheses.

2.1 Experimental design

The design of an experiment is crucial to making sure the collected data is useful. The adage 'garbage in, garbage out' applies here. A poorly designed experiment will produce poor quality data, from which it may be impossible to draw useful, valid inferences. To quote R.A. Fisher one of the founders of modern statistics,

> To consult a statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of.

2.2 Descriptive statistics

Raw data often takes the form of a massive list, array, or database of labels and numbers. To make sense of the data, we can calculate summary statistics like the mean, median, and interquartile range. We can also visualize the data using graphical devices like histograms, scatterplots, and the empirical cdf. These methods are useful for both communicating and exploring the data to gain insight into its structure, such as whether it might follow a familiar probability distribution.

2.3 Inferential statistics

Ultimately we want to draw inferences about the world. Often this takes the form of specifying a statistical model for the random process by which the data arises. For example, suppose the data takes the form of a series of measurements whose error we believe follows a normal distribution. (Note this is always an approximation since we know the error must have some bound while a normal distribution has range $(-\infty, \infty)$.) We might then use the data to provide evidence for or against this hypothesis. Our focus in 18.05 will be on how to use data to draw inferences about model parameters. For example, assuming gestational length follows a $N(\mu, \sigma)$ distribution, we'll use the data of the gestational lengths of, say, 500 pregnancies to draw inferences about the values of the parameters μ and σ. Similarly, we may model the result of a two-candidate election by a Bernoulli(p) distribution, and use poll data to draw inferences about the value of p.

We can rarely make definitive statements about such parameters because the data itself comes from a random process (such as choosing who to poll). Rather, our statistical evidence will always involve probability statements. Unfortunately, the media and public at large are wont to misunderstand the probabilistic meaning of statistical statements. In fact, researchers themselves often commit the same errors. In this course, we will emphasize the meaning of statistical statements alongside the methods which produce them.

Example 1. To study the effectiveness of new treatment for cancer, patients are recruited and then divided into an experimental group and a control group. The experimental group is given the new treatment and the control group receives the current standard of care. Data collected from the patients might include demographic information, medical history, initial state of cancer, progression of the cancer over time, treatment cost, and the effect of the treatment on tumor size, remission rates, longevity, and quality of life. The data will be used to make inferences about the effectiveness of the new treatment compared to the current standard of care.

Notice that this study will go through all three phases described above. The experimental design must specify the size of the study, who will be eligible to join, how the experimental and control groups will be chosen, how the treatments will be administered, whether or not the subjects or doctors know who is getting which treatment, and precisely what data will be collected, among other things. Once the data is collected it must be described and analyzed to determine whether it supports the hypothesis that the new treatment is more (or less) effective than the current one(s), and by how much. These statistical conclusions will be framed as precise statements involving probabilities.

As noted above, misinterpreting the exact meaning of statistical statements is a common source of error which has led to tragedy on more than one occasion.

Example 2. In 1999 in Great Britain, Sally Clark was convicted of murdering her two children after each child died weeks after birth (the first in 1996, the second in 1998). Her conviction was largely based on a faulty use of statistics to rule out sudden infant death syndrome. Though her conviction was overturned in 2003, she developed serious psychiatric problems during and after her imprisonment and died of alcohol poisoning in 2007. See `http://en.wikipedia.org/wiki/Sally_Clark`

This TED talk discusses the Sally Clark case and other instances of poor statistical intuition: `http://www.youtube.com/watch?v=kLmzxmRcUTo`

2.4 What is *a* statistic?

We give a simple definition whose meaning is best elucidated by examples.

Definition. A statistic is anything that can be computed from the collected data.

Example 3. Consider the data of 1000 rolls of a die. All of the following are statistics: the average of the 1000 rolls; the number of times a 6 was rolled; the sum of the squares of the rolls minus the number of even rolls. It's hard to imagine how we would use the last example, but it is a statistic. On the other hand, the probability of rolling a 6 is *not* a statistic, whether or not the die is truly fair. Rather this probability is a property of the die (and the way we roll it) which we can estimate using the data. Such an estimate is given by the statistic 'proportion of the rolls that were 6'.

Example 4. Suppose we treat a group of cancer patients with a new procedure and collect data on how long they survive post-treatment. From the data we can compute the average survival time of patients in the group. We might employ this statistic as an estimate of the average survival time for future cancer patients following the new procedure. The "expected survival time" for the new procedure (if that even has a meaning) is *not* a statistic.

Example 5. Suppose we ask 1000 residents whether or not they support the proposal to legalize marijuana in Massachusetts. The proportion of the 1000 who support the proposal is a statistic. The proportion of all Massachusetts residents who support the proposal is *not* a statistic since we have not queried every single one (note the word "collected" in the definition). Rather, we hope to draw a statistical conclusion about the state-wide proportion based on the data of our random sample.

The following are two general types of statistics we will use in 18.05.

1. Point statistics: a single value computed from data, such as the sample average $\overline{x}_n$ or the sample standard deviation s_n.
2. Interval statistics: an interval $[a, b]$ computed from the data. This is really just a pair of point statistics, and will often be presented in the form $\overline{x} \pm s$.

3 Review of Bayes' theorem

We cannot stress strongly enough how important Bayes' theorem is to our view of inferential statistics. Recall that Bayes' theorem allows us to 'invert' conditional probabilities. That is, if H and D are events, then Bayes' theorem says

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}.$$

In scientific experiments we start with a hypothesis and collect data to test the hypothesis. We will often let H represent the event 'our hypothesis is true' and let D be the collected data. In these words Bayes' theorem says

$$P(\text{hypothesis is true} \mid \text{data}) = \frac{P(\text{data} \mid \text{hypothesis is true}) \cdot P(\text{hypothesis is true})}{P(\text{data})}$$

The left-hand term is the probability our hypothesis is true given the data we collected. This is precisely what we'd like to know. When all the probabilities on the right are known exactly, we can compute the probability on the left exactly. This will be our focus next week. Unfortunately, in practice we rarely know the exact values of all the terms on the right. Statisticians have developed a number of ways to cope with this lack of knowledge and still make useful inferences. We will be exploring these methods for the rest of the course.

Example 6. Screening for a disease redux
Suppose a screening test for a disease has a 1% false positive rate and a 1% false negative rate. Suppose also that the rate of the disease in the population is 0.002. Finally suppose a randomly selected person tests positive. In the language of hypothesis and data we have:
Hypothesis: $H =$ 'the person has the disease'
Data: $D =$ 'the test was positive.'
What we want to know: $P(H|D) = P(\text{the person has the disease} \mid \text{a positive test})$

In this example all the probabilities on the right are known so we can use Bayes' theorem to compute what we want to know.

$$\begin{aligned} P(\text{hypothesis} \mid \text{data}) &= P(\text{the person has the disease} \mid \text{a positive test}) \\ &= P(H|D) \\ &= \frac{P(D|H)P(H)}{P(D)} \\ &= \frac{.99 \cdot .002}{.99 \cdot .002 + .01 \cdot .998} \\ &= 0.166 \end{aligned}$$

Before the test we would have said the probability the person had the disease was 0.002. After the test we see the probability is 0.166. That is, the positive test provides some evidence that the person has the disease.

4 Introduction to MLE

Suppose we know we have data consisting of values $x_1, \ldots, x_n$ drawn from an exponential distribution. The question remains: which exponential distribution?!

We have casually referred to *the* exponential distribution or *the* binomial distribution or *the* normal distribution. In fact the exponential distribution $\exp(\lambda)$ is not a single distribution but rather a one-parameter family of distributions. Each value of λ defines a different distribution in the family, with pdf $f_\lambda(x) = \lambda e^{-\lambda x}$ on $[0, \infty)$. Similarly, a binomial distribution $\text{bin}(n, p)$ is determined by the two parameters n and p, and a normal distribution $N(\mu, \sigma^2)$ is determined by the two parameters μ and σ^2 (or equivalently, μ and σ). Parameterized families of distributions are often called parametric distributions or parametric models.

We are often faced with the situation of having random data which we know (or believe) is drawn from a parametric model, whose parameters we do not know. For example, in an election between two candidates, polling data constitutes draws from a Bernoulli(p) distribution with unknown parameter p. In this case we would like to use the data to estimate the value of the parameter p, as the latter predicts the result of the election.

Similarly, assuming gestational length follows a normal distribution, we would like to use the data of the gestational lengths from a random sample of pregnancies to draw inferences about the values of the parameters μ and σ^2.

Our focus so far has been on computing the probability of data arising from a parametric model with known parameters. Statistical inference flips this on its head: we will estimate the probability of parameters given a parametric model and observed data drawn from it. In the coming weeks we will see how parameter values are naturally viewed as hypotheses, so we are in fact estimating the probability of various hypotheses given the data.

Probability

Statistics

5 Maximum Likelihood Estimates

There are many methods for estimating unknown parameters from data. We will first consider the maximum likelihood estimate (MLE), which answers the question:

For which parameter value does the observed data have the biggest probability?

The MLE is an example of a point estimate because it gives a single value for the unknown parameter (later our estimates will involve intervals and probabilities). Two advantages of the MLE are that it is often easy to compute and that it agrees with our intuition in simple examples. We will explain the MLE through a series of examples.

Example 7. A coin is flipped 100 times. Given that there were 55 heads, find the maximum likelihood estimate for the probability p of heads on a single toss.

Before actually solving the problem, let's establish some notation and terms.

We can think of counting the number of heads in 100 tosses as an experiment. For a given value of p, the probability of getting 55 heads in this experiment is the binomial probability

$$P(55 \text{ heads}) = \binom{100}{55} p^{55}(1-p)^{45}.$$

The probability of getting 55 heads depends on the value of p, so let's include p in by using the notation of conditional probability:

$$P(55 \text{ heads} \,|\, p) = \binom{100}{55} p^{55}(1-p)^{45}.$$

You should read $P(55 \text{ heads} \,|\, p)$ as:

'the probability of 55 heads given p,'

or more precisely as

'the probability of 55 heads given that the probability of heads on a single toss is p.'

Here are some standard terms we will use as we do statistics.

- Experiment: Flip the coin 100 times and count the number of heads.
- Data: The data is the result of the experiment. In this case it is '55 heads'.
- Parameter(s) of interest: We are interested in the value of the unknown parameter p.

- Likelihood, or likelihood function: this is $P(\text{data}\,|\,p)$. Note it is a function of both the data and the parameter p. In this case the likelihood is

$$P(55 \text{ heads}\,|\,p) = \binom{100}{55} p^{55}(1-p)^{45}.$$

Notes: **1.** The likelihood $P(\text{data}\,|\,p)$ changes as the parameter of interest p changes.

2. Look carefully at the definition. One typical source of confusion is to mistake the likelihood $P(\text{data}\,|\,p)$ for $P(p\,|\,\text{data})$. We know from our earlier work with Bayes' theorem that $P(\text{data}\,|\,p)$ and $P(p\,|\,\text{data})$ are usually very different.

Definition: Given data the maximum likelihood estimate (MLE) for the parameter p is the value of p that maximizes the likelihood $P(\text{data}\,|\,p)$. That is, the MLE is the value of p for which the data is most likely.

answer: For the problem at hand, we saw above that the likelihood

$$P(55 \text{ heads}\,|\,p) = \binom{100}{55} p^{55}(1-p)^{45}.$$

We'll use the notation $\hat{p}$ for the MLE. We use calculus to find it by taking the derivative of the likelihood function and setting it to 0.

$$\frac{d}{dp}P(\text{data}\,|p) = \binom{100}{55}(55p^{54}(1-p)^{45} - 45p^{55}(1-p)^{44}) = 0.$$

Solving this for p we get

$$\begin{aligned}
&55p^{54}(1-p)^{45} = 45p^{55}(1-p)^{44}\\
&55(1-p) = 45p\\
&55 = 100p\\
&\text{the MLE is } \hat{p} = .55
\end{aligned}$$

Note: **1.** The MLE for p turned out to be exactly the fraction of heads we saw in our data.

2. The MLE is computed from the data. That is, it is a statistic.

3. You should check that this critical point is actually the maximum. You could use the second derivative test. An easier way is to notice that we are interested only in $0 \le p \le 1$; that the probability is bigger than zero for $0 < p < 1$; and that the probability is equal to zero for $p = 0$ and for $p = 1$. From these facts it follows that the critical point must be the unique maximum.

5.1 Log likelihood

If is often easier to work with the natural log of the likelihood function. For short this is simply called the log likelihood. Since $\ln(x)$ is an increasing function, the maxima of the likelihood and log likelihood coincide.

Example 8. Redo the previous example using log likelihood.

<u>**answer:**</u> We had the likelihood $P(55 \text{ heads} \,|\, p) = \binom{100}{55} p^{55}(1-p)^{45}$. Therefore the log likelihood is

$$\ln(P(55 \text{ heads} \,|\, p) = \ln\left(\binom{100}{55}\right) + 55\ln(p) + 45\ln(1-p).$$

Maximizing likelihood is the same as maximizing log likelihood. We check that calculus gives us the same answer as before:

$$\begin{aligned}
\frac{d}{dp}(\text{log likelihood}) &= \frac{d}{dp}\left[\ln\left(\binom{100}{55}\right) + 55\ln(p) + 45\ln(1-p)\right] \\
&= \frac{55}{p} - \frac{45}{1-p} = 0 \\
&\Rightarrow 55(1-p) = 45p \\
&\Rightarrow \hat{p} = .55
\end{aligned}$$

5.2 Maximum likelihood for continuous distributions

For continuous distributions, we use the probability density function to define the likelihood. We show this in a few examples. In the next section we explain how this is analogous to what we did in the discrete case.

Example 9. Light bulbs
Suppose that the lifetime of *Badger* brand light bulbs is modeled by an exponential distribution with (unknown) parameter λ. We test 5 bulbs and find they have lifetimes of 2, 3, 1, 3, and 4 years, respectively. What is the MLE for λ?

<u>**answer:**</u> We need to be careful with our notation. With five different values it is best to use subscripts. Let X_j be the lifetime of the i^{th} bulb and let x_i be the value X_i takes. Then each X_i has pdf $f_{X_i}(x_i) = \lambda \mathrm{e}^{-\lambda x_i}$. We assume the lifetimes of the bulbs are independent, so the joint pdf is the product of the individual densities:

$$f(x_1, x_2, x_3, x_4, x_5 \,|\, \lambda) = (\lambda \mathrm{e}^{-\lambda x_1})(\lambda \mathrm{e}^{-\lambda x_2})(\lambda \mathrm{e}^{-\lambda x_3})(\lambda \mathrm{e}^{-\lambda x_4})(\lambda \mathrm{e}^{-\lambda x_5}) = \lambda^5 \mathrm{e}^{-\lambda(x_1+x_2+x_3+x_4+x_5)}.$$

Note that we write this as a conditional density, since it depends on λ. Viewing the data as fixed and λ as variable, this density is the likelihood function. Our data had values

$$x_1 = 2,\; x_2 = 3,\; x_3 = 1,\; x_4 = 3,\; x_5 = 4.$$

So the likelihood and log likelihood functions with this data are

$$f(2,3,1,3,4 \,|\, \lambda) = \lambda^5 \mathrm{e}^{-13\lambda}, \quad \ln(f(2,3,1,3,4 \,|\, \lambda) = 5\ln(\lambda) - 13\lambda$$

Finally we use calculus to find the MLE:

$$\frac{d}{d\lambda}(\text{log likelihood}) = \frac{5}{\lambda} - 13 = 0 \;\Rightarrow\; \boxed{\hat{\lambda} = \frac{5}{13}}.$$

Note: **1.** In this example we used an uppercase letter for a random variable and the corresponding lowercase letter for the value it takes. This will be our usual practice.

2. The MLE for λ turned out to be the reciprocal of the sample mean $\bar{x}$, so $X \sim \exp(\hat{\lambda})$ satisfies $E(X) = \bar{x}$.

The following example illustrates how we can use the method of maximum likelihood to estimate multiple parameters at once.

Example 10. Normal distributions
Suppose the data $x_1, x_2, \ldots, x_n$ is drawn from a $\mathrm{N}(\mu, \sigma^2)$ distribution, where μ and σ are unknown. Find the maximum likelihood estimate for the pair (μ, σ^2).

answer: Let's be precise and phrase this in terms of random variables and densities. Let uppercase $X_1, \ldots, X_n$ be i.i.d. $\mathrm{N}(\mu, \sigma^2)$ random variables, and let lowercase x_i be the value X_i takes. The density for each X_i is

$$f_{X_i}(x_i) = \frac{1}{\sqrt{2\pi}\,\sigma} \mathrm{e}^{-\frac{(x_i-\mu)^2}{2\sigma^2}}.$$

Since the X_i are independent their joint pdf is the product of the individual pdf's:

$$f(x_1, \ldots, x_n \,|\, \mu, \sigma) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \mathrm{e}^{-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}}.$$

For the fixed data $x_1, \ldots, x_n$, the likelihood and log likelihood are

$$f(x_1, \ldots, x_n|\mu, \sigma) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \mathrm{e}^{-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}}, \quad \ln(f(x_1, \ldots, x_n|\mu, \sigma)) = -n\ln(\sqrt{2\pi}) - n\ln(\sigma) - \sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}.$$

Since $\ln(f(x_1, \ldots, x_n|\mu, \sigma))$ is a function of the two variables μ, σ we use partial derivatives to find the MLE. The easy value to find is $\hat{\mu}$:

$$\frac{\partial f(x_1, \ldots, x_n|\mu, \sigma)}{\partial \mu} = \sum_{i=1}^n \frac{(x_i - \mu)}{\sigma^2} = 0 \;\Rightarrow\; \sum_{i=1}^n x_i = n\mu \;\Rightarrow\; \hat{\mu} = \frac{\sum_{i=1}^n x_i}{n} = \overline{x}.$$

To find $\hat{\sigma}$ we differentiate and solve for σ:

$$\frac{\partial f(x_1, \ldots, x_n|\mu, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^n \frac{(x_i - \mu)^2}{\sigma^3} = 0 \;\Rightarrow\; \hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}.$$

We already know $\hat{\mu} = \overline{x}$, so we use that as the value for μ in the formula for $\hat{\sigma}$. We get the maximum likelihood estimates

$$\begin{aligned} \hat{\mu} &= \overline{x} && = \text{the mean of the data} \\ \hat{\sigma}^2 &= \sum_{i=1}^n \frac{1}{n}(x_i - \hat{\mu})^2 = \sum_{i=1}^n \frac{1}{n}(x_i - \overline{x})^2 && = \text{the variance of the data.} \end{aligned}$$

Example 11. Uniform distributions
Suppose our data $x_1, \ldots x_n$ are independently drawn from a uniform distribution $U(a, b)$. Find the MLE estimate for a and b.

answer: This example is different from the previous ones in that we won't use calculus to find the MLE. The density for $U(a, b)$ is $\frac{1}{b-a}$ on $[a, b]$. Therefore our likelihood function is

$$f(x_1, \ldots, x_n \,|\, a, b) = \begin{cases} \left(\frac{1}{b-a}\right)^n & \text{if all } x_i \text{ are in the interval } [a, b] \\ 0 & \text{otherwise.} \end{cases}$$

This is maximized by making $b - a$ as small as possible. The only restriction is that the interval $[a, b]$ must include all the data. Thus the MLE for the pair (a, b) is

$$\hat{a} = \min(x_1, \ldots, x_n) \qquad \hat{b} = \max(x_1, \ldots, x_n).$$

Example 12. Capture/recapture method

The capture/recapture method is a way to estimate the size of a population in the wild. The method assumes that each animal in the population is equally likely to be captured by a trap.

Suppose 10 animals are captured, tagged and released. A few months later, 20 animals are captured, examined, and released. 4 of these 20 are found to be tagged. Estimate the size of the wild population using the MLE for the probability that a wild animal is tagged.

answer: Our unknown parameter n is the number of animals in the wild. Our data is that 4 out of 20 recaptured animals were tagged (and that there are 10 tagged animals). The likelihood function is

$$P(\text{data} \mid n \text{ animals}) = \frac{\binom{n-10}{16}\binom{10}{4}}{\binom{n}{20}}$$

(The numerator is the number of ways to choose 16 animals from among the $n-10$ untagged ones times the number of was to choose 4 out of the 10 tagged animals. The denominator is the number of ways to choose 20 animals from the entire population of n.) We can use R to compute that the likelihood function is maximized when $n = 50$. This should make some sense. It says our best estimate is that the fraction of all animals that are tagged is 10/50 which equals the fraction of recaptured animals which are tagged.

hard to compute

Example 13. Hardy-Weinberg. Suppose that a particular gene occurs as one of two alleles (A and a), where allele A has frequency θ in the population. That is, a random copy of the gene is A with probability θ and a with probability $1 - \theta$. Since a diploid genotype consists of two genes, the probability of each genotype is given by:

genotype	AA	Aa	aa
probability	θ^2	$2\theta(1-\theta)$	$(1-\theta)^2$

Suppose we test a random sample of people and find that k_1 are AA, k_2 are Aa, and k_3 are aa. Find the MLE of θ.

answer: The likelihood function is given by

$$P(k_1, k_2, k_3 \mid \theta) = \binom{k_1 + k_2 + k_3}{k_1}\binom{k_2 + k_3}{k_2}\binom{k_3}{k_3}\theta^{2k_1}(2\theta(1-\theta))^{k_2}(1-\theta)^{2k_3}.$$

So the log likelihood is given by

$$\text{constant} + 2k_1\ln(\theta) + k_2\ln(\theta) + k_2\ln(1-\theta) + 2k_3\ln(1-\theta)$$

We set the derivative equal to zero:

$$\frac{2k_1 + k_2}{\theta} - \frac{k_2 + 2k_3}{1-\theta} = 0$$

Solving for θ, we find the MLE is

$$\hat{\theta} = \frac{2k_1 + k_2}{2k_1 + 2k_2 + 2k_3},$$

which is simply the fraction of A alleles among all the genes in the sampled population.

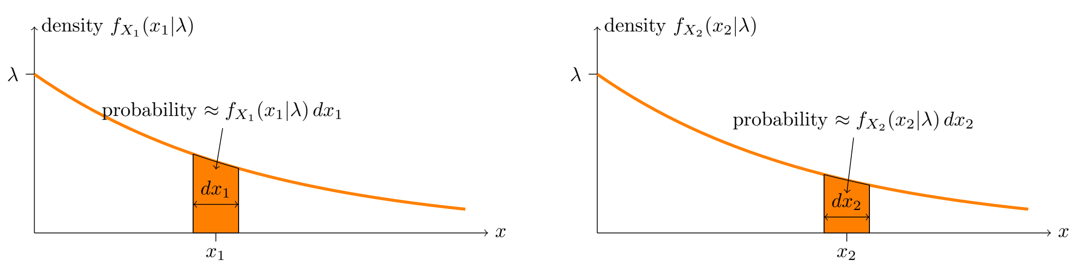
6 Why we use the density to find the MLE for continuous distributions

The idea for the maximum likelihood estimate is to find the value of the parameter(s) for which the data has the highest probability. In this section we 'll see that we're doing this is really what we are doing with the densities. We will do this by considering a smaller version of the light bulb example.

Example 14. Suppose we have two light bulbs whose lifetimes follow an exponential(λ) distribution. Suppose also that we independently measure their lifetimes and get data $x_1 = 2$ years and $x_2 = 3$ years. Find the value of λ that maximizes the probability of this data.

answer: The main paradox to deal with is that for a continuous distribution the probability of a single value, say $x_1 = 2$, is zero. We resolve this paradox by remembering that a single measurement really means a range of values, e.g. in this example we might check the light bulb once a day. So the data $x_1 = 2$ years really means x_1 is somewhere in a range of 1 day around 2 years.

If the range is small we call it dx_1. The probability that X_1 is in the range is approximated by $f_{X_1}(x_1|\lambda)\,dx_1$. This is illustrated in the figure below. The data value x_2 is treated in exactly the same way.



The usual relationship between density and probability for small ranges.

Since the data is collected independently the joint probability is the product of the individual probabilities. Stated carefully

$$P(X_1 \text{ in range, } X_2 \text{ in range}|\lambda) \approx f_{X_1}(x_1|\lambda)\,dx_1 \cdot f_{X_2}(x_2|\lambda)\,dx_2$$

Finally, using the values $x_1 = 2$ and $x_2 = 3$ and the formula for an exponential pdf we have

$$P(X_1 \text{ in range, } X_2 \text{ in range}|\lambda) \approx \lambda e^{-2\lambda}\,dx_1 \cdot \lambda e^{-3\lambda}\,dx_2 = \lambda^2 e^{-5\lambda}\,dx_1\,dx_2.$$

Now that we have a genuine probability we can look for the value of λ that maximizes it. Looking at the formula above we see that the factor $dx_1\,dx_2$ will play no role in finding the maximum. So for the MLE we drop it and simply call the density the likelihood:

$$\text{likelihood } = f(x_1, x_2|\lambda) = \lambda^2 \mathrm{e}^{-5\lambda}.$$

The value of λ that maximizes this is found just like in the example above. It is $\hat{\lambda} = 2/5$.

7 Appendix: Properties of the MLE

For the interested reader, we note several nice features of the MLE. These are quite technical and will not be on any exams.

The MLE behaves well under transformations. That is, if $\hat{p}$ is the MLE for p and g is a one-to-one function, then $g(\hat{p})$ is the MLE for $g(p)$. For example, if $\hat{\sigma}$ is the MLE for the standard deviation σ then $(\hat{\sigma})^2$ is the MLE for the variance σ^2.

Furthermore, the MLE is asymptotically unbiased and has asymptotically minimal variance. To explain these notions, note that the MLE is itself a random variable since the data is random and the MLE is computed from the data. Let $x_1, x_2, \ldots$ be an infinite sequence of samples from a distribution with parameter p. Let $\hat{p}_n$ be the MLE for p based on the data $x_1, \ldots, x_n$.

Asymptotically unbiased means that as the amount of data grows, the mean of the MLE converges to p. In symbols: $E(\hat{p}_n) \to p$ as $n \to \infty$. Of course, we would like the MLE to be close to p with high probability, not just on average, so the smaller the variance of the MLE the better. Asymptotically minimal variance means that as the amount of data grows, the MLE has the minimal variance among all unbiased estimators of p. In symbols: for any unbiased estimator $\tilde{p}_n$ and $\epsilon > 0$ we have that $\mathrm{Var}(\tilde{p}_n) + \epsilon > \mathrm{Var}(\hat{p}_n)$ as $n \to \infty$.