Chapter 8

Sampling and Estimation

We discuss in this chapter two topics that are critical to most statistical analyses. The first is **random sampling**, which is a method for obtaining observations from a statistical population that has many advantages. After obtaining a random sample, the next step of the analysis is the selection of a probability distribution to model the observations, such as the Poisson or normal distributions. One then seeks to **estimate the parameters** of these distributions (λ, μ, σ^2, etc.) using the information contained in the random sample, the second topic of this chapter. We will examine one common method of parameter estimation called maximum likelihood.

8.1 Random samples

A basic assumption of many statistical procedures is that the observations are a **random sample** from a statistical population (see Chapter 3). A sample from a statistical population is a random sample if (1) each element of the population has an equal probability of being sampled, and (2) the observations in the sample are independent (Thompson 2002). This definition has a number of implications. It implies that a random sample will resemble the statistical population from which it is drawn, especially as the sample size n increases, because each element of the population has an equal chance of being in the sample. Random sampling also implies there is no connection or relationship between the observations in the sample, because they are independent of one another.

What are some ways of obtaining a random sample? Suppose we are interested in the distribution of body length for insects of a given species, say in a particular forest. This defines the statistical population of interest. One way to obtain a random sample would be to number all the insects, and then write the numbers on pieces of paper and place them in a hat. After mixing the pieces, one would draw n numbers from the hat (without peeking) and collect only those insects corresponding to these numbers. This method of sampling would yield a random sample, because each individual would have an equal probability of being selected, and the observations would also be independent. For many insect species this method would be impractical, however, because they can be difficult to find and number. It would be more useful for statistical populations where the number of elements is known and they can be uniquely identified, as in surveys of human populations (Thompson 2002).

A more feasible way of sampling insects would be to place traps in the forest and in this way sample the population. If we want to successfully approximate a random sample with our trapping scheme, however, some knowledge of the biology of the organism is essential. For example, suppose that insect size varies in space because of differences in food plants or microclimate. A single trap deployed at only one location could therefore yield insects different in length than those in the overall population. A better sampling scheme would deploy multiple traps at several locations within the forest. The location of the traps could be randomly chosen to avoid conscious or unconscious biases by the trapper, such as deploying the traps close to a road for convenience. There is also the problem that insects susceptible to trapping could differ in length from the general population. This implies that the population actually sampled could differ from the target statistical population, and a careful analyst would consider this possibility. Thus, the biology of the organism plays an integral role in designing an appropriate sampling scheme.

8.2 Parameter estimation

Suppose we have obtained a random sample from some statistical population, say the lengths of insects trapped in a forest, or the counts of the insects in each trap. The first step faced by the analyst is to chose a probability distribution to model the data in the sample. For insect lengths, a normal distribution could be a plausible model, while counts of the insects per trap might have a Poisson distribution. Once a distribution has been selected, the next task is to estimate the parameters of the distribution using the sample data. The dominant method of parameter estimation in modern statistics is **maximum likelihood**. This method has a number of desirable statistical properties although it can also be computationally intensive.

Maximum likelihood obtains estimates of the parameters using a mathematical function (see Chapter 2) known as the likelihood function. The likelihood function gives the probability or density of the observed data as a function of the parameters in the probability distribution. For example, the likelihood function for Poisson data would be a function of the Poisson parameter λ. We then seek the maximum value of the likelihood function (hence the name maximum likelihood) across the potential range of parameter values. The parameter values that maximize the likelihood are the maximum likelihood estimates. In other words, **the maximum likelihood estimates are the parameter values that give the largest probability (or probability density) for the observed data.**

8.2.1 Maximum likelihood for Poisson data

We will first illustrate estimation using maximum likelihood with a random sample drawn from a statistical population where the observations are Poisson. For simplicity, let $n = 3$ and suppose the observed values are $Y_1 = 8$, $Y_2 = 5$, and $Y_3 = 6$. We begin by calculating the probability of observing this sample, which in fact is its likelihood function. Because we have a random sample, the Y_i values are independent of each other, and so this probability is the product of the probability for each Y_i. We have

$$\begin{aligned} L(\lambda) &= P[Y_1 = 8] \times P[Y_2 = 5] \times P[Y_3 = 6] & (8.1) \\ &= \frac{e^{-\lambda}\lambda^8}{8!} \times \frac{e^{-\lambda}\lambda^5}{5!} \times \frac{e^{-\lambda}\lambda^6}{6!} & (8.2) \end{aligned}$$

The notation $L(\lambda)$ is used for likelihood functions and indicates the likelihood is a function of the parameter λ of the Poisson distribution. The method of maximum likelihood estimates λ by finding the value of λ that maximizes this function (Mood *et al.* 1974). Note that the location of the maximum will vary with the data in the sample.

We can find the maximum likelihood estimate graphically by plotting $L(\lambda)$ as function of λ (Fig. 8.1). For these particular data values, the maximum occurs at $\lambda = 6.3$, and so the maximum likelihood estimate (often abbreviated MLE) of λ is this value. This is also the value of $\bar{Y}$ for these data, which suggests that $\bar{Y}$ might be the maximum likelihood estimator of λ in general.

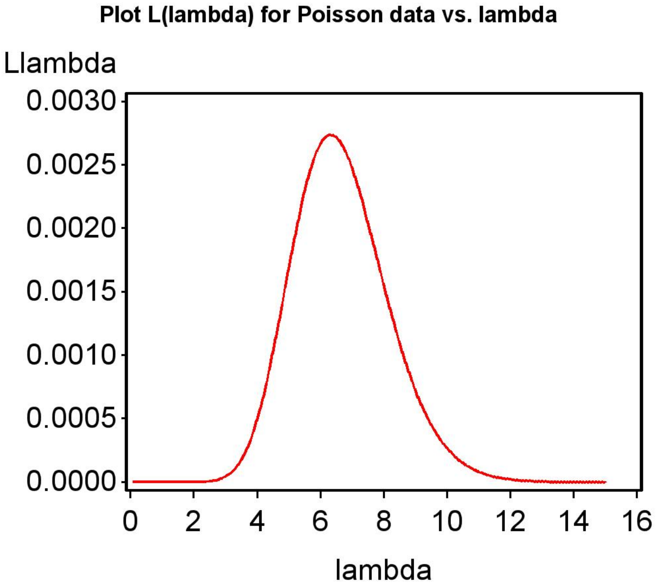


Figure 8.1: Plot of $L(\lambda)$ vs. λ

For readers interested in the mathematics, this also can be shown using derivatives. Let y_1, y_2, and y_3 be the observed values of Y_1, Y_2, and Y_3. The likelihood function can then be written as

$$L(\lambda) = \frac{e^{-\lambda}\lambda^{y_1}}{y_1!} \times \frac{e^{-\lambda}\lambda^{y_2}}{y_2!} \times \frac{e^{-\lambda}\lambda^{y_3}}{y_3!} = \frac{e^{-3\lambda}\lambda^{y_1+y_2+y_3}}{y_1!y_2!y_3!} \tag{8.3}$$

We want to find the maximum of $L(\lambda)$ (Eq. 8.3), which should occur when the derivative of this function with respect to λ equals zero. This follows because the derivative is the slope of a function, and at the maximum the slope is equal to zero. Differentiating $L(\lambda)$ with respect to λ and simplifying, we obtain

$$\frac{dL(\lambda)}{d\lambda} = \frac{e^{-3\lambda}}{y_1!y_2!y_3!}\left[(y_1+y_2+y_3)\lambda^{y_1+y_2+y_3-1} - 3\lambda^{y_1+y_2+y_3}\right]. \tag{8.4}$$

This derivative can only equal zero if the term in square brackets is zero:

$$\left[(y_1 + y_2 + y_3)\lambda^{y_1+y_2+y_3-1} - 3\lambda^{y_1+y_2+y_3}\right] = 0 \tag{8.5}$$

or

$$(y_1 + y_2 + y_3)\lambda^{y_1+y_2+y_3-1} = 3\lambda^{y_1+y_2+y_3}. \tag{8.6}$$

Canceling the quantity $\lambda^{y_1+y_2+y_3}$ from both sides of this equation, we find that

$$(y_1 + y_2 + y_3)\lambda^{-1} = 3, \tag{8.7}$$

or

$$\hat{\lambda} = \frac{y_1 + y_2 + y_3}{3}. \tag{8.8}$$

Note that this is the sample mean $\bar{Y}$ for $n = 3$, and it is can be shown that $\bar{Y}$ is the maximum likelihood estimator of λ for any n. Statisticians often write the estimator of a parameter like λ using the notation $\hat{\lambda}$, pronounced 'λ-hat.' An **estimator** can be thought of as the formula or recipe for obtaining an estimate of a parameter, with the **estimate** itself obtained by plugging actual data values into the estimator.

8.2.2 Poisson likelihood function - SAS demo

We can use a SAS program to further illustrate the behavior of the likelihood function for Poisson data (see program listing below). In particular, we will show how $L(\lambda)$ changes as the observed data and the sample size n changes. The program first generates n random Poisson observations for a specified Poisson parameter value of $\lambda = 6$ (`mu_parameter = 6`). It then plots $L(\lambda)$ across a range of λ values. In this scenario we actually know the underlying value of λ and can see how well maximum likelihood estimates its value. See SAS program below.

The program makes extensive use of loops in the data step, to generate the Poisson data and also values of the likelihood function for different values of λ. One new feature of this program is the use of a SAS macro variable(SAS Institute Inc. 2016). In this case, a macro variable labeled `n` is defined and assigned a value of 3 using the command

```
%let n = 3;
```

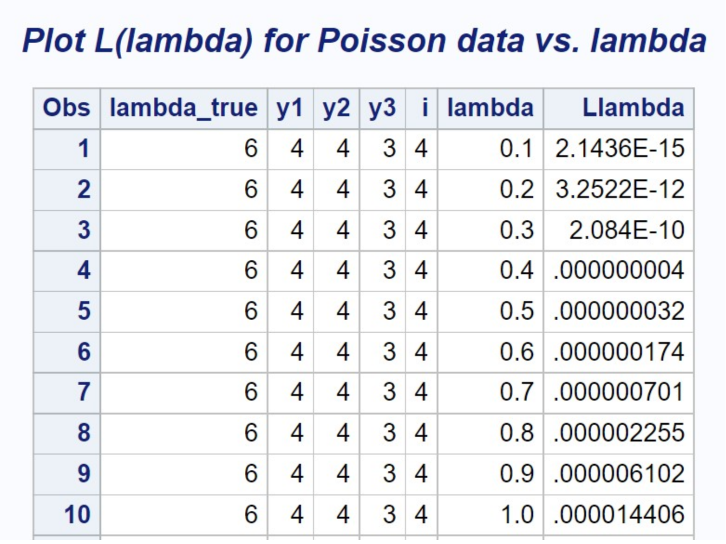
We can then refer to this value throughout the program using the notation `&n`. Otherwise, if we wanted to change the sample size n in the program we would have to type in a new value everywhere sample size is used in the calculations.

SAS program

```
* likepois_random.sas;
title "Plot L(lambda) for Poisson data vs. lambda";
data likepois;
    * Generate n random Poisson observations with parameter lambda;
    %let n = 3;
    lambda_parameter = 6;
    array ydata (&n) y1-y&n;
    do i=1 to &n;
        ydata(i) = ranpoi(0,lambda_parameter);
    end;
    * Find likelihood as function of lambda;
    do lambda=0.1 to 15 by 0.1;
        Llambda = 1;
        do i=1 to &n;
            Llambda = Llambda*pdf(’poisson’,ydata(i),lambda);
        end;
        output;
    end;
run;
* Print data;
proc print data=likepois;
run;
* Plot likelihood as a function of lambda;
proc gplot data=likepois;
    plot Llambda*lambda=1 / vaxis=axis1 haxis=axis1;
    symbol1 i=join v=none c=red width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
quit;
```

Examining the SAS output and graphs from the first two runs of the program (Fig. 8.3, 8.4), we see that the likelihood function is different. This is because the observed data are different for each run. The peak in the likelihood function always occurs at the value of $\bar{Y}$ for each data set, and this is the maximum likelihood estimate of λ.

The last run shows the effect of increasing the sample size in the program, from $n = 3$ to $n = 10$. Note that the peak of the likelihood function lies quite close to the specified value $\lambda = 6$ (Fig. 8.5). This illustrates an important property of maximum likelihood estimators - they converge on the true value as $n \to \infty$. This property is known as consistency in mathematical statistics.



Plot L(lambda) for Poisson data vs. lambda

Obs	lambda_true	y1	y2	y3	i	lambda	Llambda
1	6	4	4	3	4	0.1	2.1436E-15
2	6	4	4	3	4	0.2	3.2522E-12
3	6	4	4	3	4	0.3	2.084E-10
4	6	4	4	3	4	0.4	.000000004
5	6	4	4	3	4	0.5	.000000032
6	6	4	4	3	4	0.6	.000000174
7	6	4	4	3	4	0.7	.000000701
8	6	4	4	3	4	0.8	.000002255
9	6	4	4	3	4	0.9	.000006102
10	6	4	4	3	4	1.0	.000014406

etc.

Figure 8.2: `likepois_random.sas - proc print`

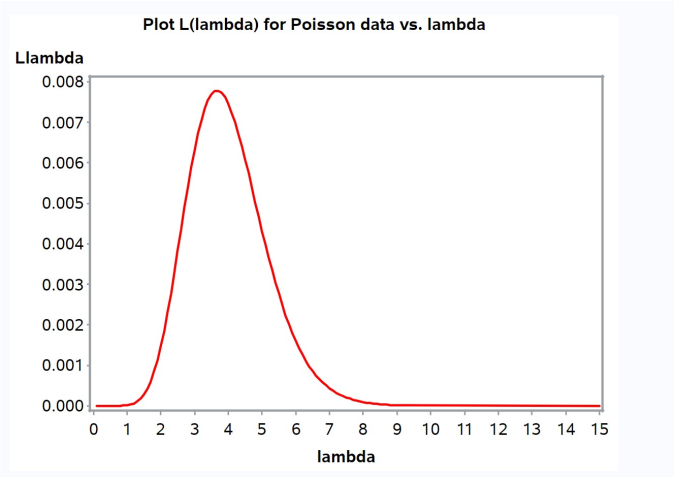


Figure 8.3: `likepois_random.sas - proc gplot` $(n = 3)$

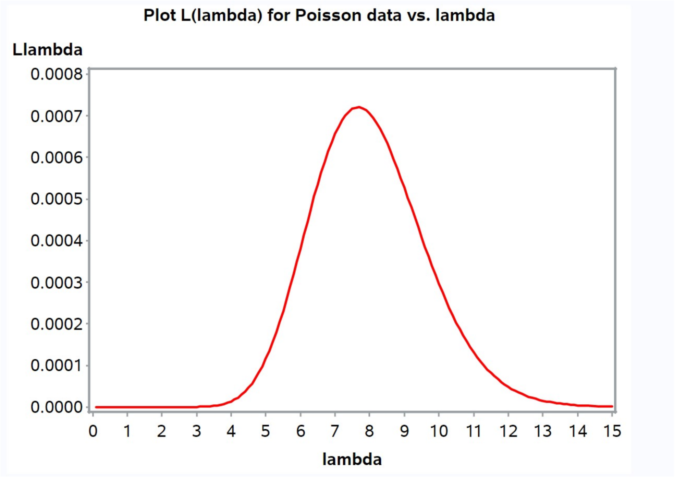


Figure 8.4: `likepois_random.sas - proc gplot` $(n = 3)$

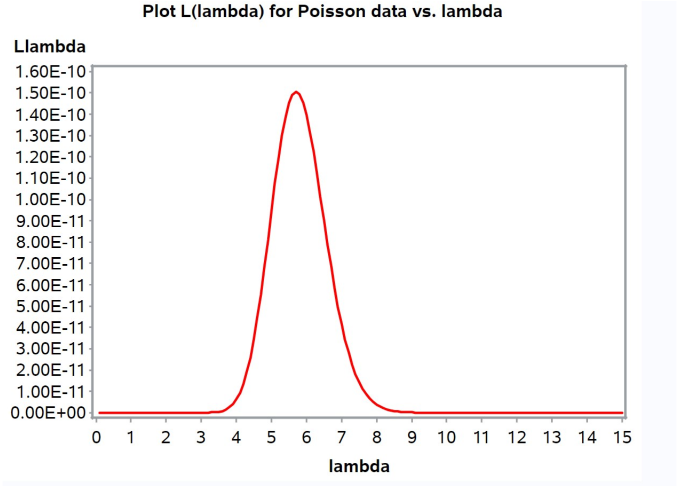


Figure 8.5: `likepois_random.sas - proc gplot` $(n = 10)$

8.2.3 Maximum likelihood for normal data

Now suppose we draw a random sample from a population with a normal distribution, such as body lengths, etc. For simplicity, let $n = 3$ again and the observed values be $Y_1 = 4.5$, $Y_2 = 5.4$, and $Y_3 = 5.3$. The likelihood function in this case is the probability density values for the observed data:

$$L(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(4.5-\mu)^2}{\sigma^2}} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(5.4-\mu)^2}{\sigma^2}} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(5.3-\mu)^2}{\sigma^2}}. \tag{8.9}$$

Note that the terms in the likelihood for normal data are probability densities, instead of probabilities as with Poisson data.

We can find the maximum likelihood estimate graphically by plotting $L(\mu, \sigma^2)$ as function of μ and σ^2. The likelihood function in this case describes a dome-shaped surface (Fig. 8.6). With these particular data, the maximum occurs at about $\mu = 5.07$ and $\sigma^2 = 0.16$, and so these are the maximum likelihood estimates of μ and σ^2.

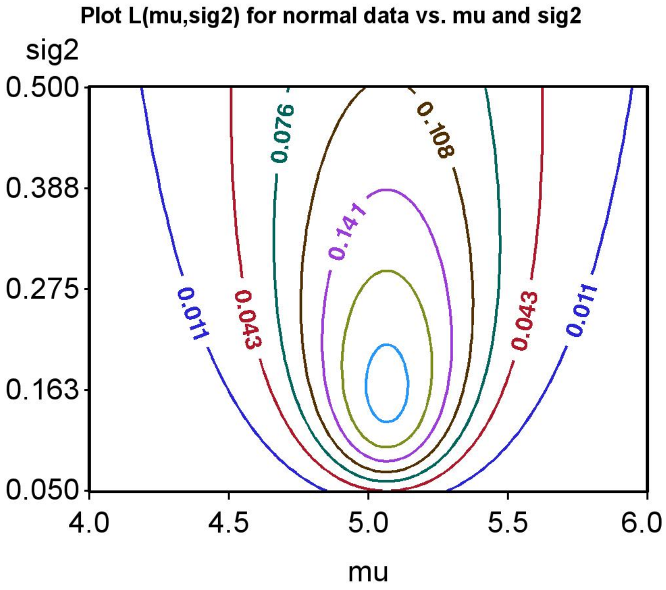


Figure 8.6: Plot of $L(\mu, \sigma^2)$ vs. μ and σ^2

Using a bit of calculus, it can be shown that the maximum likelihood estimators of these parameters are, for any sample size n:

$$\hat{\mu} = \bar{Y} \tag{8.10}$$

and

$$\hat{\sigma}^2 = \frac{\Sigma_{i=1}^{n}(Y_i - \bar{Y})^2}{n}. \tag{8.11}$$

Note that does not quite equal the sample variance s^2, which uses $n - 1$ (rather than n) in the denominator:

$$s^2 = \frac{\Sigma_{i=1}^{n}(Y_i - \bar{Y})^2}{n - 1}. \tag{8.12}$$

Recall that s^2 is an unbiased estimator of σ^2, and so $\hat{\sigma}^2$ derived using maximum likelihood is actually a biased estimator of σ^2. It would consistently generate values that underestimate σ^2 because n is greater than $n - 1$. For cases like this one where bias is known, it is common to use a bias-corrected version of the maximum likelihood estimator (i.e., $n - 1$ rather than n in the denominator).

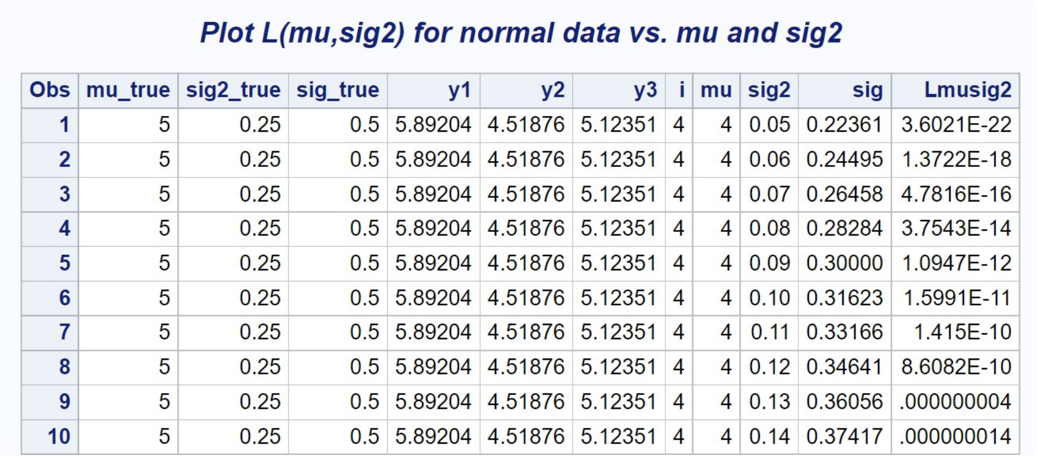
8.2.4 Normal likelihood function - SAS demo

We will use another SAS program to illustrate the behavior of the likelihood function for normal data. The program first generates n random normal observations for a specified, known value of $\mu = 5$ and $\sigma^2 = 0.25$. It then plots the likelihood function across a range of possible μ and σ^2 values. See SAS program below.

Examining the SAS output and graphs from the first two runs of the program (Fig. 8.8, 8.9), we see that the likelihood function changes with the observed data. The peak always occurs at $\hat{\mu}$ and $\hat{\sigma}^2$ for each data set. The last run shows the effect of increasing the sample size from $n = 3$ to $n = 10$. Note that the peak of the likelihood function lies quite close to the specified values of $\mu = 5$ and $\sigma^2 = 0.25$ (Fig. 8.10). This again illustrates the consistency of maximum likelihood estimates.

SAS program

```
* likenorm_random.sas;
title "Plot L(mu,sig2) for normal data vs. mu and sig2";
data likenorm;
    * Generate n random normal observations with parameters mu and sig2;
    %let n = 3;
    mu_parameter = 5; sig2_parameter = 0.25; sig_parameter = sqrt(sig2_parameter);
    array ydata (&n) y1-y&n;
    do i=1 to &n;
        ydata(i) = mu_parameter + sig_parameter*rannor(0);
    end;
    * Find likelihood as a function of mu and sig2;
    do mu=4 to 6 by 0.01;
        do sig2=0.05 to 0.5 by 0.01;
            sig = sqrt(sig2);
            Lmusig2 = 1;
            do i=1 to &n;
                Lmusig2 = Lmusig2*pdf(’normal’,ydata(i),mu,sig);
            end;
            output;
        end;
    end;
run;
* Print data, first 25 observations;
proc print data=likenorm(obs=25);
run;
* Plot likelihood as a function of mu and sig2;
* Contour plot version;
proc gcontour data=likenorm;
    plot sig2*mu=Lmusig2 / autolabel nolegend vaxis=axis1 haxis=axis1;
    symbol1 height=1.5 font=swissb width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
quit;
```



Plot L(mu,sig2) for normal data vs. mu and sig2

Obs	mu_true	sig2_true	sig_true	y1	y2	y3	i	mu	sig2	sig	Lmusig2
1	5	0.25	0.5	5.89204	4.51876	5.12351	4	4	0.05	0.22361	3.6021E-22
2	5	0.25	0.5	5.89204	4.51876	5.12351	4	4	0.06	0.24495	1.3722E-18
3	5	0.25	0.5	5.89204	4.51876	5.12351	4	4	0.07	0.26458	4.7816E-16
4	5	0.25	0.5	5.89204	4.51876	5.12351	4	4	0.08	0.28284	3.7543E-14
5	5	0.25	0.5	5.89204	4.51876	5.12351	4	4	0.09	0.30000	1.0947E-12
6	5	0.25	0.5	5.89204	4.51876	5.12351	4	4	0.10	0.31623	1.5991E-11
7	5	0.25	0.5	5.89204	4.51876	5.12351	4	4	0.11	0.33166	1.415E-10
8	5	0.25	0.5	5.89204	4.51876	5.12351	4	4	0.12	0.34641	8.6082E-10
9	5	0.25	0.5	5.89204	4.51876	5.12351	4	4	0.13	0.36056	.000000004
10	5	0.25	0.5	5.89204	4.51876	5.12351	4	4	0.14	0.37417	.000000014

etc.

Figure 8.7: `likenorm_random.sas - proc print`

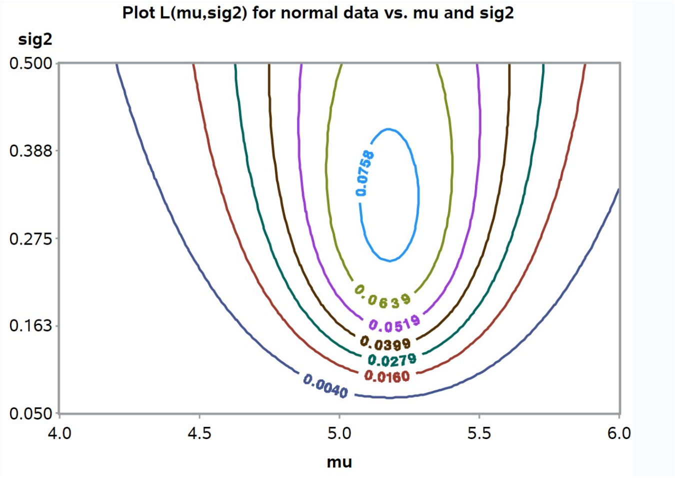


Figure 8.8: `likenorm_random.sas - proc gcontour` ($n = 3$)

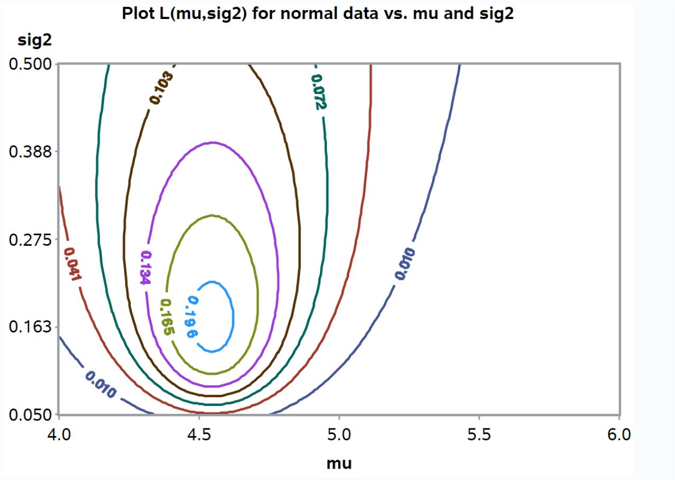


Figure 8.9: `likenorm_random.sas - proc gcontour` ($n = 3$)

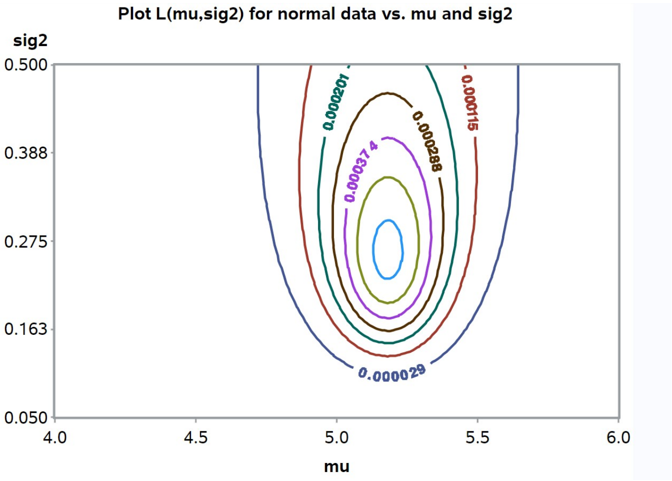


Figure 8.10: `likenorm_random.sas - proc gcontour` $(n = 10)$

8.3 Optimality of maximum likelihood estimates

Why should we use maximum likelihood estimates? There are other methods of parameter estimation, but maximum likelihood estimates are optimal in a number of ways (Mood *et al.* 1974). We have already seen that they are **consistent**, approaching the true parameter values as sample size increases. Increasing the sample size also reduces the variance of these estimators. We can observe this behavior for $\hat{\mu} = \bar{Y}$, the estimator of μ for the normal distribution. Recall that the variance of $\bar{Y}$ is σ^2/n, which decreases for large n. Maximum likelihood estimates are also **asymptotically unbiased**, meaning their expected value approaches the true value of the parameter as the sample size n increases. We can see this in operation for $\hat{\sigma}^2$ (Eq. 8.11), the maximum likelihood estimator of σ^2, vs. s^2 (Eq. 8.12), an unbiased estimator of σ^2. Note that the difference between n vs. $n-1$ in the denominator becomes very small as n increases. Finally, maximum likelihood estimates are **asymptotically normal**, meaning their distribution approaches the normal distribution for large n.

There are other uses for the likelihood function besides parameter estimation. We will later see how the likelihood function can be used to develop statistical tests called likelihood ratio tests. Many of the statistical tests we will study are actually likelihood ratio tests. Likelihood methods provide an essential tool for developing new statistical procedures, provided that we can specify a probability distribution for the data.

8.4 References

Mood, A. M., Graybill, F. A. & Boes, D. C. (1974) *Introduction to the Theory of Statistics.* McGraw-Hill, Inc., New York, NY.

Thompson, S. K. (2002) *Sampling.* John Wiley & Sons, Inc., New York, NY.

SAS Institute Inc. (2016) *SAS 9.4 Macro Language: Reference, Fifth Edition.* SAS Institute Inc., Cary, NC.

8.5 Problems

1. The exponential distribution is a continuous distribution that is used to model the time until a particular event occurs. For example, the time when a radioactive particle decays is often modeled using an exponential distribution. If a variable Y has a exponential distribution, then its probability density is given by the formula

$$f(y) = \frac{e^{-y/\lambda}}{\lambda} \tag{8.13}$$

for $y \geq 0$. The distribution has one parameter, λ, which is the mean decay time ($E[Y] = \lambda$).

   (a) Use SAS and the program `fplot.sas` to plot the exponential probability density with $\lambda = 2$, for $0 \leq y \leq 5$. Attach your SAS program and output.

   (b) Suppose you have a sample of four observations y_1, y_2, y_3 and y_4 from the exponential distribution. What would be the likelihood function for these observations?

   (c) Plot the likelihood function for $y_1 = 1$, $y_2 = 2$, $y_3 = 2$ and $y_4 = 3$ over a range of λ values. Show that the maximum occurs at $\hat{\lambda} = \bar{Y}$, the maximum likelihood estimator of λ. Attach your SAS program and output.

2. The geometric distribution is a discrete distribution that is used to model the time until a particular event occurs. Consider tossing a coin – the number of tosses before a head appears would have a geometric distribution. If a variable Y has a geometric distribution, then the probability that Y takes a particular value y is given by the formula

$$P[Y = y] = f(y) = p(1 - p)^y \tag{8.14}$$

where p is the probability of observing the event on a particular trial, and $y = 0, 1, 2, \ldots, \infty$. The distribution has only one parameter, p.

   (a) Use SAS and the program `fplot.sas` to plot this probability distribution for $p = 0.5$, for $y = 0, 1, \ldots, 10$. Attach your SAS program and output.

(b) Suppose you have a sample of three observations y_1, y_2, and y_3 from the geometric distribution. What would be the likelihood function for these observations?

(c) Plot the likelihood function for $y_1 = 1$, $y_2 = 2$, and $y_3 = 3$ over a range of p values. Show that the maximum occurs at $\hat{p} = 1/(\bar{Y} + 1)$, the maximum likelihood estimator of p. Attach your SAS program and output.