3 Parameter Estimation

3.1 Introduction

Over the past 50 years, statisticians, computer scientists and engineers have developed an amazing array of algorithms for extracting interesting features from data. The current vogue name for this huge array of algorithms is machine learning. However, it is worth bearing in mind, that the data is collected through "experiments" (either physical experiment, sample surveys etc). If we redo the experiment and collect a new set of data, the numbers in the new data set are unlikely to match the numbers in the old data set. Further the estimates from both data sets are unlikely to be the same. So what exactly are the numbers that we have extracted from the data?

Given that multiple experiments give rise to different values, we can treat our observations as random and the estimator as a random variable (Rice, nicely describes this on page 257), this is the same as the multiple trajectories for the sample mean seen in Figure 1.4. Once we understand that for each sample, we obtain an estimator, and these are random variables with a sampling distribution (see Definition 1.8). Then we can start to understand what we are estimating. The estimator is an estimate of a parameter in the sampling distribution, usually its mean. But this is not very informative, without understanding what underpins this distribution. To do this, we make assumptions about how the data is collected. In this course, we will usually assume that the data \( \{X_i\}_{i=1}^n \) are iid random variables from a certain distribution, that is a function of an unknown parameter (this is often called the data generating process). The sampling distribution and the data generating process are closely related. And the estimator is an estimate of a parameter in the data generating process. We often call this the population parameter and treat \( \{X_i\}_{i=1}^n \) as iid sample from the population.

**Definition 3.1** (Statistical inference). *Statistical inference is the mapping of what we estimate from the data onto the entire population, which is unobserved. Formally, we say we drawing conclusions or making inference, about the underlying population based on the observed sample.*

Recall, that in Section 1.5 we made inference on the population mean based on the sample mean.

In summary, point estimation involves two main steps:

- Find methods and algorithms which allow us to evaluate features in the data. This is called a point
Parameter Estimation

- Underlying all algorithms is a model and an unobserved population from which the data is collected. The task of a statistician is to make inference about parameters in the population based on the estimator at hand.

Often to construct an estimator we assume that the observations come from a certain family of distributions. In this chapter, we will assume that the \( \{X_i\} \) is an iid sample from a known family of distributions \( \{f(x; \theta)\} \) where \( \theta = (\theta_1, \ldots, \theta_p) \) are a small number of unknown parameters, which we aim to estimate. The family of distributions is determined either from scientific evidence (how the data was collected) or empirical evidence (for example making a histogram of the data).

In most real life situation, the true distribution is unlikely to be known. We can only make intelligent guesses on what it should be. Keep in mind the famous quotes made by various statistician over the past hundred years: "All models are wrong but some models are useful." Therefore, it is also important to understand what the estimator is actually estimating when the model has not been correctly specified (for example, the data comes from a beta-distribution but we estimate the parameters as if it came from an exponential). However, this analysis is beyond this course.

Let us recall the set-up. We assume that we observe the iid random variables \( \{X_i\}_{i=1}^n \) which come from the known family of distributions \( \{f(x; \theta)\} \) where \( \theta = (\theta_1, \ldots, \theta_p) \) and \( \theta \in \Theta \). \( \Theta \) is called the parameter space, it is the set of all parameters where \( f(x; \theta) \) makes sense as a density (it is positive and integrates to one). Our objective is to estimate \( \theta \), this estimator is called a point estimator.

### 3.2 Estimation: Method of moments

The method of moments was developed over a hundred twenty years ago, and is a precursor of the maximum likelihood estimator, described in a later section. It may not be the most efficient estimator (defined in Section 3.6) and is prone to finite sample bias. But it is extremely simple to evaluate and is conceptionally easy to understand. We start with a few motivating examples. A nice description is given in Section 8.4 of Rice.

#### 3.2.1 Motivation

Often use the notation

\[
\mu_r = E(X_r^r)
\]

for the \( r \)th moment.
3 Parameter Estimation

(i) Suppose the random variable $X$ is exponentially distributed with density $f(x; \lambda) = \lambda \exp(-\lambda x) \ (x \geq 0)$, then

$$E(X) = \frac{1}{\lambda}.$$ 

Thus

$$\lambda = \frac{1}{\mu_1}.$$ 

(ii) Suppose the random variables $X$ has a Poisson distribution with probability mass function $f(k; \lambda) = \frac{\lambda^k \exp(-\lambda)}{k!}$. Then

$$E(X) = \lambda.$$ 

Thus $\lambda = \mu_1$

(iii) Suppose the random variables $X$ has a normal distribution with density $f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. Then

$$E(X) = \mu \quad \text{and} \quad E(X^2) = \sigma^2 + \mu^2.$$ 

Thus $\mu = \mu_1$ and $\sigma^2 = \mu_2 - \mu_1^2$

(iv) Suppose the random variable $X$ has Gamma distribution distribution with density $f(x; \alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1} \exp(-\beta x)}{\Gamma(\alpha)}$. Then

$$E(X) = \frac{\alpha}{\beta} \quad \text{and} \quad E(X^2) = \frac{\alpha(\alpha + 1)}{\beta^2}.$$ 

Thus

$$\mu_1 = \frac{\alpha}{\beta} \quad \text{and} \quad \mu_2 = \mu_1^2 + \frac{\mu_1}{\beta}.$$ 

Hence

$$\beta = \frac{\mu_1}{\mu_2 - \mu_1^2} \quad \text{and} \quad \alpha = \frac{\mu_1^2}{\mu_2 - \mu_1^2}.$$ 

Thus the parameters in the distribution are embedded within the moments of the estimators. And we can rewrite the parameters as a function of the moments. Thus by estimating the moments, we can also estimate the parameters by simply substituting the moment estimators into the formula for parameters in terms of the moment. Moments are like means, they can easily be estimated by taking the average. For example, an estimator of $\mu_r = E[X_r^r]$ based on the iid random variables $\{X_i\}$ is

$$\hat{\mu}_r = \frac{1}{n} \sum_{i=1}^{n} X_i^r.$$ 

Therefore if $\theta = g(\mu_1, \ldots, \mu_K)$, then the method of moments estimator $\hat{\theta}$ is

$$\hat{\theta}_n = g(\hat{\mu}_1, \ldots, \hat{\mu}_K).$$
3.2.2 Examples

Example 3.1. (i) Suppose the random variable $X$ is exponentially distributed with density $f(x; \lambda) = \lambda \exp(-\lambda x)$, then a moments estimator of $\lambda$ is

$$\hat{\lambda}_n = \frac{1}{\hat{\mu}_1}.$$  

(ii) Suppose the random variables $X$ has a Poisson distribution with probability mass function $f(k; \lambda) = \frac{\lambda^k \exp(-\lambda)}{k!}$. Then $\hat{\lambda}_n = \hat{\mu}_1$.

(iii) Suppose the random variables $X$ has a normal distribution with density $f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. Then $\hat{\mu}_1 = \hat{\mu}$ and $\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2$.

(iv) Suppose the random variable $X$ has Gamma distribution distribution with density $f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x)$. Then

$$\hat{\beta}_n = \frac{\hat{\mu}_1}{\hat{\mu}_2 - \hat{\mu}_1^2}, \quad \hat{\alpha}_n = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2}.$$  

Method of moment estimators are usually not unique, an example is given below.

Example 3.2. Consider the exponential distribution $f(x; \lambda) = \lambda \exp(-\lambda x)$. Basic algebra gives

$$E(X^r) = \lambda \int_0^\infty x^r \exp(-\lambda x) dx$$

$$= \frac{\lambda}{\lambda^r} \int_0^\infty (\lambda x)^r \exp(-\lambda x) dx \text{ (change variables } y = \lambda x)$$

$$= \frac{1}{\lambda^r} \int_0^\infty y^r \exp(-y) dy$$

$$= \frac{1}{\lambda^r} \Gamma(r + 1),$$

where $\Gamma(r + 1)$ is the gamma function (and does not depend on $\lambda$). Thus for all $r \geq 1$ we have

$$\lambda = \left( \frac{\Gamma(r + 1)}{\hat{\mu}} \right)^{1/r}$$

Based on the above, for any $r$ we can use

$$\hat{\lambda}_{r,n} = \left( \frac{\Gamma(r + 1)}{\hat{\mu}_r} \right)^{1/r}$$

as an estimator of $\lambda$. Question: which moments estimator should one use?

The take home message from the above example is that moments estimator do not have to depend on the first few moments. They can depend on higher order moments, the can also depend on moments of the transformed data, for example even $\{\log X_i\}$.
3.2.3 Sampling properties of method of moments estimators

In this section we derive the sampling properties of the method of moments estimators. We start by considering the estimator of the rate \( \lambda \) in the exponential distribution. From the previous section we observe that the moments estimator is

\[
\hat{\lambda}_n = \frac{1}{\bar{X}_n}.
\]

We simulate the estimator in the case that \( \lambda = 1.5 \) and a plot of 5 trajectories \((n = 1, \ldots, 100)\) is given in Figure 3.1. We observe that they all appear to converge to the truth \( \lambda = 1.5 \) as \( n \to \infty \). Indeed this should be the case. Since \( \bar{X}_n \xrightarrow{p} \mu \), by using Lemma 1.2 we have that \( \hat{\lambda}_n \xrightarrow{p} 1/\mu \). This means the method of moments estimator is asymptotically consistent. In general it will converge to the true parameter as the sample size grows. This is true for all method of moment estimators.

![Figure 3.1: 5 trajectories of \( \hat{\lambda}_n \). The pink dashed line is the truth = 1.5.](image)

To obtain the sampling distribution and variance of \( \hat{\lambda}_n \) we use the results in Section 1.5.4. We observe that \( \hat{\lambda}_n = g(\bar{X}_n) \), thus by using (1.7), on transformations of sample means we can obtain the asymptotic distribution of \( \hat{\lambda}_n \). Since \( g'(\mu) = -1/\mu^2 \neq 0 \) we have

\[
\sqrt{n}(\hat{\lambda}_n - \lambda) \xrightarrow{D} N \left( 0, \frac{\text{var}(X)}{\mu^4} \right) = N(0, \lambda^2)
\]

as \( n \to \infty \). Thus the asymptotic sampling variance of the estimator is

\[
\sigma^2_{\hat{\lambda}_n} = \frac{\lambda^2}{n}.
\]

The corresponding standard error for \( \hat{\lambda}_n \)

\[
\frac{\lambda}{\sqrt{n}}
\]
3 Parameter Estimation

Of course this standard error is quite useless if we want to use it to construct confidence intervals, since \( \lambda \) is unknown. However, we overcome this issue by replacing \( \lambda \) with its estimator \( \hat{\lambda}_n \) to yield the estimated (asymptotic) sampling variance

\[
\hat{\sigma}^2_{\hat{\lambda}_n} = \frac{\hat{\lambda}_n^2}{n}.
\]

Since we know by Lemma 1.2 that \( \hat{\lambda}_n \xrightarrow{p} \lambda \), then we have \( \hat{\lambda}^2_n \xrightarrow{p} \lambda^2 \), thus for sufficiently large \( n \), \( \hat{\sigma}^2_{\hat{\lambda}_n} \) is a good approximation of \( \sigma^2_{\hat{\lambda}_n} \).

To see how good the above approximations of the sampling distribution of \( \hat{\lambda}_n \) and its true finite sample variance we conduct some simulations. For \( n = 3, \ldots, 100 \) we simulate from an exponential with \( \lambda = 1.5 \) and evaluate \( \hat{\lambda}_n \) we do this 500 times.

Comparing the asymptotic and empirical variance

We evaluate the empirical variance:

\[
\hat{\sigma}^2_n = \frac{1}{500} \sum_{i=1}^{500} (\tilde{\lambda}_{i,n} - \hat{\lambda}_n)^2 \quad \text{with} \quad \hat{\lambda}_n = \frac{1}{500} \sum_{i=1}^{500} \tilde{\lambda}_{i,n} \quad (3.2)
\]

this should be close to the true variance (since it was done over 500 replications). In Figure 3.2 we plot \( n\hat{\sigma}^2_n \) against \( n \). Recall that asymptotic sampling variance is \( \lambda^2/n \). It is not the “true” finite sample variance (which can usually only be evaluated through simulations). However, we would expect that \( n\hat{\sigma}^2_n \) (defined in (3.2)) to be “close” to \( \lambda^2 \) for “large” \( n \). We do see that this is the case, for \( n > 40 \) they are closely aligned, but for smaller \( n \) the match is not so close.

Comparing the asymptotic and finite sample sampling distributions

We observe a similar effect for the sampling distribution of the estimators. We recall that the result in (3.1) is an asymptotic approximation of the true finite sample distribution \( \tilde{\lambda}_n \). In other words, for a sufficiently large \( n \)

\[
\tilde{\lambda}_n \sim N\left(\lambda, \frac{\lambda^3}{n}\right).
\]

To see how close this approximation is to the truth, in Figures 3.3-3.5 we make a histogram of the estimates (evaluated over 500 replication) together with a histogram of the corresponding normal approximation. We observe that for sample size \( n = 5 \) the true sampling distribution of \( \tilde{\lambda}_5 \) is right skewed and normal approximation is really quite bad. For sample size \( n = 20 \), the true sampling distribution has less of a right skew, and the sampling distribution is better. And for sample size \( n = 100 \) there appears to be a close match in the true sampling distribution and the normal approximation. In summary, for relatively large sample sizes the asymptotic variance \( \lambda^2/n \) and normal approximation are quite good approximations of the true
Figure 3.2: A plot of the sampling variance times the sample size for $n=3,\ldots,100$. The blue line is $n\overline{\sigma_n^2}$, the red line is the $\lambda^2 = 1.5^2 = 2.25$.

Figure 3.3: Exponential distribution. Left: True sampling distribution when $n = 5$. Asymptotic normal distribution $N(1.5, 2.25/5)$

variance and sampling distribution of the estimator. But for small sample sizes some caution is required when applying these approximations (to constructing confidence intervals and testing). As can be clearly seen the quality of the normal approximation as asymptotic variance is not so good when the sample size is small.

The above results concern the method of moments estimator corresponding to the exponential distribution. However, similar results also hold for general method of moments estimators. Some examples be given in
Figure 3.4: Exponential distribution. Left: True sampling distribution when n = 20. Asymptotic normal distribution \( N(1.5, 2.25/20) \)

Figure 3.5: Exponential distribution. Left: True sampling distribution when n = 100. Asymptotic normal distribution \( N(1.5, 2.25/100) \)

your homework. However, to get more practice, in the section below we consider the asymptotic sampling properties of the moments estimator of the gamma distribution.
Sampling properties of the Gamma distribution

Suppose that \( \{X_i\} \) are iid random variables with Gamma distribution with density
\[
f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x).
\]
Suppose (for simplicity) that \( \beta \) is known. Then the method of moments estimator can simply be constructed using the first moment since
\[
E[X] = \frac{\alpha}{\beta} \quad \alpha = \beta E[X].
\]
This yields the moment estimator
\[
\hat{\alpha}_n = \beta \bar{X}_n
\]
of \( \alpha \). Using Theorem 1.1 we have
\[
\sqrt{n}(\hat{\alpha}_n - \alpha) \xrightarrow{D} N \left( 0, \beta^2 \operatorname{var}(X) \right),
\]
where we note that \( \operatorname{var}(X) = \alpha / \beta^2 \). Thus the asymptotic variance of \( \hat{\alpha}_n \) is \( \alpha \).

3.2.4 Application of asymptotic results to the construction of confidence intervals

From a practical perspective the reason the asymptotic sampling properties of the moments estimators are of interest is to construct confidence intervals for the parameter we are estimating. We recall from elementary statistics that if \( \{X_i\} \) are iid random variables with mean \( \mu \) and variance \( \sigma^2 \), then the sample mean \( \bar{X}_n \) satisfies the following distributional results
\[
\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} N(0, \sigma^2).
\]
Thus for sufficiently large \( n \) we roughly have
\[
\bar{X}_n \approx N \left( \mu, \frac{\sigma^2}{n} \right).
\]
This result is used to construct the \((1 - \alpha)100\%\) confidence interval for the mean \( \mu \)
\[
\left[ \bar{X}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right].
\]
The idea being that if \( \bar{X} \) is normally distributed, then for every 100 intervals constructed on average \((1 - \alpha)100\%\) of them would contain the population mean \( \mu \). To understand this, in Figure 3.6, left plot, we construct 100 95\% confidence intervals (for \( n = 3 \) based on iid normal observations). We observe that in this simulation, 3 out of 100 do not contain the population mean. If the population variance is unknown we replace \( \sigma \) with the estimator \( \hat{s}_n \), this induces additionally variability. We observe in Figure 3.6 (right plot) that by simply replacing \( \sigma \) with \( s_3 \) (but still using iid normal observations) and still using the normal distribution to construct the CI means that the CI has less confidence than the 95\% that is stated (18 confidence intervals...
3 Parameter Estimation

Figure 3.6: Confidence intervals \( \mu \) using \( \bar{X} \), where \( X_i \sim N(\mu = 5, \sigma^2 = 10^2) \) and \( n = 3 \) (over 100 replications).

Left: Confidence interval constructed using \( z \)-values and \( \sigma = 10 \) is used. Right: Confidence interval constructed using \( z \)-values and \( s_3 \) is used.

out of 100 do not contain the mean). Since we know that for normal data \( T_n = \sqrt{n}(\bar{X} - \mu)/s_n \sim t_{n-1} \), when constructing the CI we replace the normal distribution with the t-distribution with \( (n-1) \)-df:

\[
\left[ \bar{X}_n - t_{\alpha/2, n-1} \frac{s_n}{\sqrt{n}}, \bar{X}_n + t_{\alpha/2, n-1} \frac{s_n}{\sqrt{n}} \right].
\]

To see how effective using the t-distribution is, in the plot on the right, we simulate 100 confidence intervals (using iid normal observations, sample size \( n = 3 \)). We use \( s_n \) and the t-distribution to construct the CI. We observe that out 3 out of 100 do not contain the confidence interval. Hence the t-distribution does “make” a confidence interval.

For non-normal data and small sample sizes neither

\[
\left[ \bar{X}_n - z_{\alpha/2} \frac{s_n}{\sqrt{n}}, \bar{X}_n + z_{\alpha/2} \frac{s_n}{\sqrt{n}} \right] \quad \text{nor} \quad \left[ \bar{X}_n - t_{\alpha/2, n-1} \frac{s_n}{\sqrt{n}}, \bar{X}_n + t_{\alpha/2, n-1} \frac{s_n}{\sqrt{n}} \right]
\]

is truly a \((1 - \alpha)100\%\) for the mean \( \mu \).

We now apply the same ideas described above to obtain confidence intervals parameter estimators. We observe the iid random variables \( \{X_i\}_{i=1}^n \) which has density \( f(x; \theta) \). Suppose the method of moments estimator of \( \theta \) is \( \hat{\theta}_n \) and it can be shown that

\[
\sqrt{n}(\hat{\theta}_n - \theta) \overset{D}{\to} N(0, \sigma_\theta^2).
\]
For a sufficiently large \( n \) we roughly have

\[
\hat{\theta}_n \approx N\left( \theta, \frac{\sigma^2}{n} \right).
\]

This result is used to construct the \((1 - \alpha)100\%\) confidence interval for \( \theta \):

\[
\left[ \hat{\theta}_n - z_{\alpha/2} \frac{\sigma_\theta}{\sqrt{n}}, \hat{\theta}_n + z_{\alpha/2} \frac{\sigma_\theta}{\sqrt{n}} \right].
\]

Note that often we have to estimate \( \tilde{\sigma}_\theta \), in which case we replace \( \sigma_\theta \) with its estimator \( \hat{\sigma}_\theta \) to yield the interval

\[
\left[ \hat{\theta}_n - z_{\alpha/2} \frac{\hat{\sigma}_\theta}{\sqrt{n}}, \hat{\theta}_n + z_{\alpha/2} \frac{\hat{\sigma}_\theta}{\sqrt{n}} \right].
\]

Observe that we did not replace \( z_{\alpha/2} \) with the corresponding \( t \)-distribution, because it is not really clear if the \( t \)-distribution is able to model correctly the additional uncertainty caused by replacing \( \sigma_\theta \) with its estimator \( \hat{\sigma}_\theta \).

### 3.3 Monte Carlo methods

We now investigate how to estimate the finite sample distribution of estimators. We focus on the method of moment estimators, but the methods described below also apply to many other estimators.

#### 3.3.1 The parametric Bootstrap

For many estimators asymptotic normality can be shown. But as demonstrated in Figures 3.3-3.5 (and the variance estimator in Figure 3.2) this approximation is not very reliable when the sample size is small. The plots given on the left hand side of Figures 3.3-3.5 are the histograms corresponding to the true sampling distribution and the blue line in Figure 3.2 is the true variance. In an ideal world we would use this distribution and variances for inference (constructing confidence intervals etc, we cover this in a later chapter). But the simulations are based on the simulating the exponential distribution with true underlying parameter (look at the R code in Chapter3Rcode). In reality the true parameter is unknown (else we would not be estimating it).

To get round this problem we can estimate the finite sampling distribution by simulating from the distribution using the \textit{estimated} parameter. To demonstrate what we mean by this, we return to the exponential distribution.

- The idea is to sample from an exponential distribution with sample size \( n = 20 \).
3 Parameter Estimation

• Suppose we observe

\[ \text{demo} = 0.5475, 0.0089, 1.1269, 0.7519, 0.5628, 0.8547, 1.9941, 0.7383, 0.0529, 0.0243, 
0.5029, 0.4628, 0.3951, 0.7799, 2.3609, 1.3204, 0.1754, 1.5920, 1.0729, 0.7659, \]

which are drawn from an exponential distribution (with \( \lambda = 1.5 \), which is treated as unknown). The sample mean is 0.805 and the method of moments estimator of \( \lambda \) is \( \hat{\lambda}_{20} = \bar{X}^{-1} = 1.243 \).

• The bootstrap step is based on simulating from an exponential distribution with \( \lambda = 1.243 \) for \( n = 20 \) and calculate \( \hat{\lambda}^*_n \) from these simulated values. We use the * notation to denote the fact that \( \hat{\lambda}_{20}^* \) is estimated from simulated or the bootstrap sample.

We repeat this several times (say 500 times), storing \( \hat{\lambda}^*_n \) for each replication.

• A plot of the estimated histogram together with the true distribution is given in Figure 3.7. Further the QQplot of the estimated distribution against true distribution (calculated based on the true parameter \( \lambda = 1.5 \)) is given in Figure 3.8. We observe a relatively close match, though there is a slight shift.

• Based on this simulation, the estimated variance is 0.115, whereas the true sampling variance of \( \hat{\lambda}_{20} \) is 0.135.

• For every sample, the estimated distribution will change a little.

From the demonstration above and Figures 3.7 and 3.8 we observe that the Monte Carlo method appears to capture the right skew in the sampling distribution of \( \hat{\lambda}_{20} \), which the normal approximation given in Figure 3.4 clearly does not.

![Sampling Distribution and Histogram](image)

Figure 3.7: Left: True sampling distribution when \( n = 20 \). Right: Distribution based on sampling from estimated exponential with \( \hat{\lambda} = 1.243 \).
Figure 3.8: QQplot of quantiles of true distribution of $\hat{\lambda}_n$ (calculated through simulations) against the estimated distribution of $\hat{\lambda}_n$ based on the estimated parameter.

The Monte Carlo method described above can be generalised to many distributions. The "take home" message is that we replace the unknown parameter with the estimated parameter, when conducting the replications. Monte Carlo methods are also nicely described in Rice, Section 8.4, page 264-265.

### 3.3.2 The nonparametric Bootstrap

Monte Carlo methods, as described, above are very useful. However, by sampling from an exponential distribution we make the assumption the observed data has really been drawn from an exponential distribution. If this is not the case, then this method will not give a good approximation of the sampling distribution of the estimator. It can be completely wrong.

There are more general methods, which allow for what we called misspecification of the distribution. These are nonparametric methods (where the distribution is not assumed known). This entails sampling not from the conjectured distribution (such as the exponential) but sampling (with replacement) from the data itself. This is often called the nonparametric bootstrap. It is robust to misspecification of the distribution. A bootstrap sample in R can be obtained using the command `sample(data, replace = T)`.

We outline the nonparametric bootstrap for the exponential example described in the previous section.

- Given the data

  ```
  demo = 0.5475, 0.0089, 1.1269, 0.7519, 0.5628, 0.8547, 1.9941, 0.7383, 0.0529, 0.0243
  0.5029, 0.4628, 0.3951, 0.7799, 2.3609, 1.3204, 0.1754, 1.5920, 1.0729, 0.7659,
  ```
we resample from with replacement. For example, one bootstrap sample is

```r
> temp = sample(demo, replace = T)
> temp
[1] 0.3480 0.4494 0.3645 3.5009 0.9539 0.7354 0.8323 0.2294 0.6017 0.8323 2.2859
[13] 0.1917 0.7354 0.4627 0.2294 0.1917 0.0348 0.1917 0.4627
```

The estimator corresponding to the above is $\lambda_{20}^* = 1.446$.

- We repeat the bootstrap procedure described above. For each sample we evaluate $\lambda_{20}^*$ and store it. We do this many times (I did it 500, but the more the better).
- A histogram of the bootstrap estimates of $\lambda_{20}$ is given in Figure 3.9. A QQplot using the bootstrap quantiles against the true quantiles is given in Figure 3.10.
- The bootstrap estimated variance of $\lambda_{20}$ is 0.169, whereas the true sampling variance of $\lambda_{20}$ is 0.135 (calculated using simulation).

A possible reason for the over estimation in the bootstrap standard error is that there is slightly more spread the original data than the spread in the true, underlying exponential distribution. The nonparametric bootstrap mimics the properties of the data from which it samples from.

- For every sample, the estimated bootstrap distribution will change a little (this distribution is random).

Based on this one sample, the nonparametric bootstrap seems to estimate the true sampling distribution quite well. It is comparable to the parametric bootstrap described in the previous section. But, in general, if the distribution is correctly specified (for example, the data really does from an exponential), then the parametric bootstrap should perform a little better than the nonparametric bootstrap (well one would think, though I cannot say for sure).

The bootstrap is conceptionally simple to understand. It can be shown that the bootstrap is able to capture the skewness of the finite sampling distribution of the estimator. But the actually details are quite delicate (using Edgeworth expansion) and is beyond this class.

### 3.3.3 The power transform approach

As mentioned above the bootstrap is able to capture in the skewness in the sampling distribution of the estimator (which is useful in inference). An alternative approach, is transform the estimator in such a way to remove the skewness in the sampling distribution.

We recall from HW2, Question 3, that for skewed random variables, the sampling distribution of the sample mean tends to be skewed for small sample sizes. As seen from the simulations above, bootstrap methods
3 Parameter Estimation

Figure 3.9: Left: True sampling distribution when \( n = 20 \). Right: Distribution based on the bootstrap samples.

Figure 3.10: QQplot of quantiles of true distribution of \( \tilde{\lambda}_n \) (calculated through simulations) against the estimated distribution of \( \tilde{\lambda}_n \) based on the estimated parameter.

tend to capture the skewness. An alternative approach is to make a power transform, for the form \( \tilde{\theta}_n^\alpha \), similiar to that described in HW2, question 3. By using Lemma 1.2 we have

\[
\sqrt{n} \left( \tilde{\theta}_n^\alpha - \theta^\alpha \right) \xrightarrow{D} \mathcal{N} \left( 0, \left[ \alpha \theta^{\alpha-1} \sigma^2 \right] \right).
\]

But we also recall from HW2, question 3, that for \( \alpha < 1 \) the power transform tends to reduce the skewness of the parameter and obtain a better approximate the normal distribution. Of course, it is necessary to
3 Parameter Estimation

select the optimal $\alpha$, Chen and Deo (2004) propose a method for selecting the "best" $\alpha$ based on minimising the skewness of the sampling distribution (beyond this course).

3.4 Estimation: Maximum likelihood (MLE)

3.4.1 Motivation

Empirical evidence suggests that the life time of an incandescent light bulb follows an exponential distribution with density $f(x; \lambda) = \lambda \exp(-\lambda x)$ with mean lifetime $\lambda^{-1}$. It is well known that the light bulb manufacturing industry did their level best to ensure that light bulbs did not last "too long", if it lasted too long it would reduce profits (sounds rather like what manufacturers today and goes contrary to the circular economy, but I digress). An interesting summary is give in wiki and here.

Suppose a manufacturer has three options on their machine for producing light bulbs with a mean life time of 750 hours, 1000 hours or 1250 hours. However, the settings on the machine are stuck and the writing has worn away so noone knows the setting it is stuck at. It can be either 750 hours, 1000 hours or 1250 hours. To figure it out, 20 light bulbs are produced in the stuck setting machine and their lifetimes measure. The data is summarized below.

$x = 202.5, 962.4, 342.4, 596.1, 1331.8, 902.8, 501.7, 1393.1, 620.8, 1604.2, 241.0, 372.6, 143.0, 1420.5, 74.7, 2342.4, 1072.3, 1309.2, 1650.7, 163.7$

The sample mean is 862.4, which is slightly closer to 1000 than 750. But simply comparing the sample means is really not enough. It makes sense to consider the joint distribution of the data, which we believe comes from an exponential distribution. Under the assumption that the lifetime of the light bulbs are independent of each other the joint density is the product of the marginals

$$f_X(x; \lambda) = \prod_{i=1}^{20} \lambda \exp(-\lambda x_i) = \lambda^{20} \exp\left(-\lambda \sum_{i=1}^{n} x_i\right) = \lambda^{20} \exp\left(-\lambda \times 20 \times \text{mean}\right)$$

But unlike applications in probability, we are not using the joint density to calculate the probability of an event. The data is observed, it cannot change. Our aim is to select the $\lambda$ based on the observed data; the $\lambda$ that most likely to give the observed data. The idea in statistics is that a "typical" sample is most likely to be drawn from the maximum of the density (close to the peak). Thus from a statistical perspective $f_X(x; \lambda)$ is a likelihood function, and is a function of $\lambda$ rather than $x$. Often to emphasis the dependence on $\lambda$ we rewrite the $f_X(x; \lambda)$ as

$$L(\lambda; x) = \lambda^{20} \exp\left(-\lambda \sum_{i=1}^{n} x_i\right) = \lambda^{20} \exp\left(-\lambda \times 20 \times \text{mean}\right).$$
A plot of $L(\lambda; \mathbf{x})$ is given in Figure 3.11 (see the left hand plot). However, often to make the plot easier to read the logarithm of the likelihood is used

$$L(\lambda; \mathbf{x}) = 20 \log \lambda - \sum_{i=1}^{n} x_i = 20 \log \lambda - 862.4\lambda.$$ 

Since log is a monotonic transform, it does not change the characteristics in $L(\lambda; \mathbf{x})$. The log-likelihood is given on the right hand side of Figure 3.11. As the potential candidates are mean $\lambda^{-1} = 750, 1000, 1250$, these correspond to the red vertical lines in the plots. From the plot we observe that $\lambda^{-1} = 1250$ is unlikely given the observed data. Visually it is difficult to tell the difference between 750, 1000, but the actual log-likelihoods at these values are given in the table below.

<table>
<thead>
<tr>
<th>$\lambda^{-1}$</th>
<th>750</th>
<th>1000</th>
<th>1250</th>
</tr>
</thead>
<tbody>
<tr>
<td>log-likelihood</td>
<td>-155.399</td>
<td>-155.403</td>
<td>-156.416</td>
</tr>
</tbody>
</table>

Figure 3.11: Left: Likelihood. Right: log-likelihood.

We observe that just by a very small margin, $\lambda^{-1} = 750$ maximises the likelihood over $\lambda^{-1} = 1000$. Thus based on the observed data set, using the likelihood criterion we make the decision that the setting is stuck at 750 hours. In fact in this case, we have made the correct decision, the sample was generated from an exponential with mean $\lambda^{-1} = 750$. But we do observe from the plots in Figure 3.11 that the likelihood (and equivalently log-likelihood) is maximised at $\lambda^{-1} = 862.4$ (the sample mean). Thus if any value of $\lambda^{-1}$ were a potential candidate for the mean, using $\lambda^{-1} = 862.4$ or equivalently $\lambda = 862.4^{-1}$ appears to be the most likely. This is exactly the maximum likelihood estimator of $\lambda$ based on the exponential distribution (interestingly it is the same as the method of moments estimator too).

We now formally define the likelihood for iid random variables. Let us suppose that $\{X_i\}_{i=1}^{n}$ are iid random
variables with density \( f(x; \theta) \), the likelihood is defined as

\[
L_n(\theta; X) = \prod_{i=1}^{n} f(X_i; \theta)
\]

and the log-likelihood is the logarithm of the likelihood

\[
\mathcal{L}_n(\theta; X) = \sum_{i=1}^{n} \log f(X_i; \theta).
\]

The maximum likelihood estimator (denoted from now on as MLE) is defined as parameter which maximises the likelihood. If you like maths notation we say

\[
\hat{\theta}_{\text{mle}} = \arg \max_{\theta \in \Theta} \mathcal{L}_n(\theta; X) = \arg \max_{\theta \in \Theta} L_n(\theta; X),
\]

where \( \Theta \) denotes the parameter space (see Section 2.3, recall it is all parameters where \( f(x, \theta) \) is a density or probability mass function). The MLE is said to be the estimator that is most consistent with the observed data (it is most likely given the data).

Remember the data vector \( X \) is observed, so we are maximising over the unknown parameter \( \theta \).

Often it is easier reparametrize a distribution. For example, for the exponential distribution \( f(x; \lambda) = \lambda \exp(-\lambda x) \) we require that \( \lambda > 0 \), but if we rewrite \( \lambda = \exp(\gamma) \), then \( \gamma \in (-\infty, \infty) \) (the parameter space of \( \gamma \) is not restricted). In this reparametrized world we can write the exponential distribution as

\[
g(x; \gamma) = \exp(\gamma) \exp(-\exp(\gamma)x).
\]

In the following lemma we show that if the mapping \( \lambda \mapsto \exp(\gamma) \) is one-to-one (a bijection), then the MLE of \( \gamma \) and \( \lambda \) yield (after transformation) the same values.

**Lemma 3.1** (The invariance property). The invariance property of the MLE says that it makes no difference which parameterization we use for finding the MLE. If \( \hat{\theta} \) is the MLE of \( \theta \) and \( g(\theta) \) is a one-to-one function, then \( g(\hat{\theta}) \) is the MLE of \( g(\theta) \).

The proof is straightforward. We may not use this property so much in this course, but it is very useful in applied statistics. For example, in generalized linear models etc.

### 3.4.2 Examples

We now give some examples of distributions and their MLE. We note that in general to maximise the likelihood

\[
L(\theta; X) = \prod_{i=1}^{n} f(X_i; \theta) \quad \text{equivalently} \quad \mathcal{L}_n(\theta; X) = \sum_{i=1}^{n} \log f(X_i; \theta),
\]
over the parameter space $\Theta$, we differentiate $L(\theta; X)$ with respect to $\theta$ and set the derivative to zero and solve for $\theta$ (there are some issues which arise if the maximum happens at the boundary of the parameter space, which is beyond this course).

We also evaluate the second derivative at the solution to ensure it is the (local maximum, local because there could be a few maximums). If $\frac{d^2 L(\theta; X)}{d\theta^2}$ is negative (or negative definite), then we can be sure we have “caught” the (local) maximum. Often it will not be possible to obtain an explicit expression for the maximum, and we discuss strategies on dealing with this in a later section.

**The Normal/Gaussian distribution**

Suppose $\{X_i\}$ are iid random variables, then the Gaussian likelihood is

$$L(\theta; X_n) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left( -\frac{(X_i - \mu)^2}{2\sigma^2} \right) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left[ -\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{2\sigma^2} \right].$$

Often (especially for the exponential family of distributions) it is easier to take the logarithm and maximise that

$$L_n(\theta; X_n) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{2\sigma^2}.$$

Evaluating the partial derivative of $L_n(\theta; X)$ with respect to $\mu$ and $\sigma^2$ and setting to zero gives

$$\frac{\partial L_n(\theta; X_n)}{\partial \mu} = \frac{\sum_{i=1}^{n}(X_i - \mu)}{\sigma^2} = 0$$

$$\frac{\partial L_n(\theta; X_n)}{\partial \sigma^2} = -\frac{n}{2}\sigma^{-2} + \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{2\sigma^4}.$$

Solving the above gives

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu}_n)^2,$$

which is the same as the method of moments estimator.

An alternative derivation of the MLE (which on first appearances does not appear to have any advantages) is based on transforming the data vector $X_n = (X_1, \ldots, X_n)$. It is uses the transformation proof described in Section 2.4.2 used to prove Theorem 2.3. We recall, we had defined $n$-orthonormal vectors $\{\xi_j\}_{j=1}^{n}$ where the first vector is $\xi_1 = n^{-1/2}(1, 1, 1 \ldots, 1)$ which transform $X_n$ to a different basis. The coefficients of $X_n$ on this new basis are:

$$Y_n = E_n X_n = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{pmatrix} X_n = \begin{pmatrix} \sqrt{n} \bar{X} \\ \langle \xi_2, X_n \rangle \\ \vdots \\ \langle \xi_n, X_n \rangle \end{pmatrix}.$$
Since \( \{X_i\} \) are iid normal, the joint distribution of \( X_n \) is normal with mean \( E[X_nX_n] = (n^{1/2}\mu, 0, \ldots, 0)' \) and variance \( \text{var}(E_nX_n) = E_n\text{var}(X_n)E_n^n = \sigma^2E_nE_n^n = \sigma^2I_n; \)

\[
Y_n \sim \mathcal{N}\left(\begin{pmatrix} \sqrt{n}\mu \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix} \right)
\]

Observe the first entry of \( Y_n \) contains information on the mean, but the rest do not. We can easily calculate the likelihood of the transformed data which is

\[
\mathcal{L}_n(\theta; Y_n) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{n(\bar{X}_n - \mu)}{2\sigma^2} - \frac{\sum_{i=2}^{n} Y_i^2}{2\sigma^2}.
\]

Differentiating this likelihood with respect to \( \mu \) and \( \sigma^2 \) gives

\[
\frac{\partial \mathcal{L}_n(\theta; Y_n)}{\partial \mu} = \frac{n(\bar{X}_n - \mu)}{\sigma^2} = 0
\]

\[
\frac{\partial \mathcal{L}_n(\theta; X)}{\partial \sigma^2} = -\frac{n}{2} \frac{1}{\sigma^4} + \frac{(\bar{X}_n - \mu)^2}{2\sigma^4} + \frac{\sum_{i=2}^{n} Y_i^2}{2\sigma^4}.
\]

Solving the above gives the MLE

\[
\tilde{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i^2 \quad \tilde{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} Y_i^2.
\]

Now on first impression it would appear that the estimators are different, since \( \frac{1}{n} \sum_{i=2}^{n} Y_i^2 \) looks different to \( n^{-1} \sum_{i=1}^{n} (X_i - \tilde{\mu}_n)^2 \). But using equation (1.3) in Section 1.3.2 we have that

\[
\tilde{\sigma}_n^2 = \frac{1}{n} \sum_{i=2}^{n} Y_i^2 = n^{-1} \sum_{i=1}^{n} (X_i - \tilde{\mu}_n)^2.
\]

Thus the MLE estimators for both transformations of the data are the same. Which it will always be. An advantage of the transformed data is that the mean information is contained in only coefficient, whereas the variance information is encrypted in the other coefficients of the vector.

Take home message: So long as the matrix \( E_n \) is known (and non-singular) no information is lost or gained in the transform \( E_n X_n \). But in terms of estimation sometimes it is easier to deal with the transformed data.

Note that the MLE estimator of \( \sigma^2 \) has a bias (which goes away for large \( n \)). We know from Theorem 2.3 that the unbiased estimator of \( \sigma^2 \) is

\[
s_n^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (X_i - \tilde{\mu}_n)^2.
\]

Observe that the MLE coincides with the moments estimator for the normal distribution.
Exponential distribution

We now derive the MLE of the exponential distribution. Suppose that \{X_i\} are iid exponential, then the likelihood is

\[ L(\theta; X_n) = \prod_{i=1}^{n} [\lambda \exp(-\lambda X_i)] = \lambda^n \exp(-\lambda \sum_{i=1}^{n} X_i). \]

The log likelihood is

\[ \mathcal{L}_n(\theta; X_n) = n \log \lambda - \lambda \sum_{i=1}^{n} X_i. \] (3.4)

Differentiating the above wrt \( \lambda \) gives

\[ \frac{d \mathcal{L}_n(\theta; X_n)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i = 0. \]

Thus the MLE is

\[ \hat{\lambda}_n = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}_n}. \]

Again, the MLE coincides with the moments estimator for the normal distribution.

The Poisson distribution

We now derive the MLE of the Poisson distribution. Suppose that \{X_i\} are iid poisson, then the likelihood is

\[ L(\theta; X_n) = \prod_{i=1}^{n} \frac{\lambda^{X_i} \exp(-\lambda)}{X_i!}. \]

The log likelihood is

\[ \mathcal{L}_n(\theta; X_n) = \sum_{i=1}^{n} \left[ X_i \log \lambda - \lambda - \log X_i \right] = 0. \]

Differentiating the above wrt \( \lambda \) gives

\[ \frac{d \mathcal{L}_n(\theta; X_n)}{d\lambda} = \lambda^{-1} \sum_{i=1}^{n} X_i - n. \]

This gives the MLE is

\[ \hat{\lambda}_n = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}_n. \]

Again, the MLE coincides with the moments estimator for the normal distribution.
The Gamma distribution

We now derive the MLE of the exponential distribution. Suppose that \( \{X_i\} \) are iid exponential, then the likelihood is

\[
L(\theta; X_n) = \prod_{i=1}^{n} \frac{\beta^\alpha}{\Gamma(\alpha)} X_i^{\alpha-1} \exp(-\beta X_i)
\]

The log-likelihood is

\[
\mathcal{L}_n(\theta; X_n) = \sum_{i=1}^{n} \left[ (\alpha - 1) \log X_i - \beta X_i + \alpha \log \beta - \log \Gamma(\alpha) \right].
\]

Differentiating the above wrt \( \beta \) and \( \alpha \) gives

\[
\frac{\partial \mathcal{L}_n}{\partial \beta} = \sum_{i=1}^{n} \left[ -X_i + \frac{\alpha}{\beta} \right] = 0
\]

\[
\frac{\partial \mathcal{L}_n}{\partial \alpha} = \sum_{i=1}^{n} \left[ \log X_i + \log \beta - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \right] = 0.
\]

Thus, solving the above we can rewrite the solution of \( \beta \) in terms of \( \alpha \) (this is often called profiling)

\[
\hat{\beta}(\alpha) = \frac{\alpha}{X_n}.
\]

Substituting this into the next derivative gives

\[
\sum_{i=1}^{n} \left[ \log X_i + \log \frac{\alpha}{X_n} - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \right] = 0.
\]

But to solve the above we need to use a numerical routine.

We note that the above MLE does not coincide (in an obvious way) to the method of moments estimator described in Section 3.2.1.

**Research 3.** In many of the examples above we showed that the MLE in the exponential family coincides with the method of moments estimators (the exception is the Gamma). This is not a coincidence, and there are reasons for it.

The curious amongst you may want to investigate why. Hint: As a start you may want to evaluate the expectation of \( X \) and \( \log X \) for the Gamma distribution. Then study the properties of moments of the exponential (natural) family.

The Uniform distribution

Let us suppose \( \{X_i\} \) are iid random variables with a uniform distribution with density

\[
f(x; \theta) = \begin{cases} 
\frac{1}{\theta} & x \in [0, \theta] \\
0 & x \notin [0, \theta].
\end{cases}
\]
We can write the density as \( f(x, \theta) = \theta^{-1} I_{[0, \theta]}(x) \), where \( I_{[0, \theta]}(x) \) is an indicator function which is one between \([0, \theta]\) and zero elsewhere (please draw it). Based on this the likelihood is

\[
L(\theta, X) = \prod_{i=1}^{n} \theta^{-1} I_{[0, \theta]}(X_i).
\]

Now recall that in a likelihood we treat the data as fixed and the parameter \( \theta \) as variable. Thus \( L(\theta, X) \) is a function of \( \theta \). The log-likelihood has no meaning. Instead, it is easier to just directly maximise the likelihood. Since the data is kept fixed and \( L(\theta, X) \) is a function of \( \theta \), we observe (by making a plot) that \( L(\theta, X) \) is zero for any \( X_i \leq \theta \leq X_{i+1} \). Thus \( L(\theta, X) \) is zero for \( \theta \leq \max_i X_i \) (see below).

Since the data is kept fixed and \( L(\theta, X) \) is a function of \( \theta \), we observe (see plot on the left) that \( L(\theta, X) \) is zero for any \( X_i \leq \theta \leq X_{i+1} \). Thus \( L(\theta, X) \) is zero for \( \theta \leq \max_i X_i \). But for \( \theta > \max_i X_i \), it drops at the rate \( \theta^{-n} \). The way to visualize this is to think of a simple data set \( X_1 = 2, X_2 = 2.5 \) and \( X_3 = 3.2 \). \( L(\theta, X) = 0 \) if \( \theta < 3.2 \) (since this event cannot happen), thus \( \theta \) must be 3.2 or greater. And the MLE if 3.2.

Thus, \( L(\theta, X) \) is maximised at \( \theta = \max_i X_i \), and the maximum likelihood estimator for the uniform distribution is

\[
\hat{\theta}_n = \max_{1 \leq i \leq n} X_i.
\]

**Inflated Poisson distribution**

The Poisson distribution is often used to model count data. A plot of a Poisson mass function for different values of \( \lambda \) is given in Figure 3.12. We observe that when \( \lambda \) is small, there is a large mass at zero, then it rapidly drops. On the other hand if \( \lambda \) is large, then the mass tends to be very small at zero. This gives a dichotomy in modelling. Either the probabilities will be large at zero and small elsewhere, or small at zero and large elsewhere. Such data can arise, but often we require more flexibility in the distribution.

There arises many real life situations where the chance of observing zero outcomes is very high, but also the chance of observing 5, 6, 7, can also be high. The regular Poisson cannot model both these behaviour simultaneously. But by mixing two distributions more flexible characteristics can be achieved. For example, suppose

\[
U \in \{0, 1\} \text{ a Bernoulli random variable}
\]

\[
V \in \{0, 1, 2, \ldots\} \text{ a Poisson random variable}
\]
3 Parameter Estimation

Figure 3.12: Left: Mass function of the Poisson distribution (stolen from Wiki)

where

\[
P(U = 0) = p \quad P(U = 1) = 1 - p
\]

\[
P(V = k) = \frac{\lambda^k \exp(-\lambda)}{k!}
\]

We define a new random variable which mixes the Bernoulli and the Poisson:

\[X = (1 - U) + UV.\]

We observe that

\[
P(X = 0) = P(X = 0|U = 0)P(U = 0) + P(X = 0|U = 1)P(U = 1)
= p + \exp(-\lambda)(1 - p)
\]

\[
P(X = k) = \frac{P(X = k|U = 0)}{P(U = 0)} P(U = 0) + \frac{P(X = k|U = 1)}{P(U = 1)} P(U = 1)
= \frac{\lambda^k \exp(-\lambda)(1 - p)}{k!} \quad k = 1, 2, \ldots
\]

Thus combining the two sets of probabilities we can write

\[
P(X = k) = [p + \exp(-\lambda)(1 - p)]^{I(k=0)} \left[ \frac{\lambda^k \exp(-\lambda)(1 - p)}{k!} \right]^{1-I(k=0)} \quad k = 0, 1, 2, \ldots
\]
where \( I(\cdot) \) is an indicator variable: \( I(k = 0) = 1 \) of \( k = 0 \) and \( I(k = 0) = 0 \) if \( k \neq 0 \) (it is like the \texttt{if} and \texttt{else} function when we code). A plot for different \( \lambda \) is given in Figure 3.13. Observe that even for large \( \lambda \), the probability at zero can be large as well as “far” from zero.

![Figure 3.13: Inflated Poisson distribution for \( p = 0.3, \lambda = 1 \) (red), \( \lambda = 4 \) (purple) and \( \lambda = 10 \) (red).](image)

The above model is called the Inflated zero Poisson model and it was first proposed by Diane Lambert for modelling manufacturing defects. Observe that the inclusion of the Bernoulli random variables allows one to model a large number of zeros without the need to use the Poisson distribution. This distribution does not belong to the exponential family.

Given the iid observations \( \{X_i\} \) the likelihood function is

\[
L(\theta, X_n) = \prod_{i=1}^{n} \left( [p + \exp(-\lambda)(1 - p)]^{I(X_i = 0)} \left[ \frac{\lambda^{X_i} \exp(-\lambda)(1 - p)}{X_i!} \right]^{1 - I(X_i = 0)} \right)
\]

and the log-likelihood is

\[
\mathcal{L}_n(\theta, X_n) = \sum_{i=1}^{n} I(X_i = 0) \log [p + \exp(-\lambda)(1 - p)] + \sum_{i=1}^{n} [1 - I(X_i = 0)] \left[ X_i \log \lambda - \lambda - \log X_i! + \log(1 - p) \right].
\]
The derivatives are
\[
\frac{\partial L_n(\theta, X_n)}{\partial p} = \sum_{i=1}^{n} I(X_i = 0) \frac{1 - \exp(-\lambda)}{p + \exp(-\lambda)(1 - p)} - \sum_{i=1}^{n} [1 - I(X_i = 0)] \frac{1}{1 - p} = 0
\]
\[
\frac{\partial L_n(\theta, X_n)}{\partial \lambda} = \sum_{i=1}^{n} I(X_i = 0) \frac{- \exp(-\lambda)(1 - p)}{p + \exp(-\lambda)(1 - p)} + \sum_{i=1}^{n} [1 - I(X_i = 0)] \left[ \frac{X_i}{\lambda} - 1 \right] = 0.
\]

An analytic solution does not exist for the above and a numerical routine needs to be used (note that a moment type of estimator is possible).

Include in the HW the Weibull distribution.

**Using constraints: Lagrange multipliers**

This section goes beyond the syllabus of this course, but it is important when constructing likelihoods. In some examples, one needs to place constraints on the parameters. The simplest case is the multinomial distribution. The multinomial distribution is a generalisation of the binomial distribution. To construct a binomial distribution, we use that for each trial, \( X_i \), there are two options “success” and “failure”, where the probability of a success is \( \pi \). However, for many data sets, each trial may have more than two options, for examples in a survey, one may give a person several fruit \( K \) different options and ask which is their favourite fruit. To simplify notation, we label each fruit \( r \) for \( 1 \leq r \leq K \). The person can only answer one fruit, where the probability of fruit \( i \) being selected is \( \pi_r \). In this set-up, suppose \( N \) people are random selected (with replacement) asked the fruit question. Let \( M = (M_1, \ldots, M_K) \) denote total number of responses for each fruit. For example \( M_r \) denotes the number of people out of \( N \) who like fruit \( r \). The sample space is \( \{(m_1, \ldots, m_K); 0 \leq m_i \leq N, \sum_{r=1}^{K} m_r = K\} \). It can be shown that
\[
P(M_1 = m_1, M_2 = m_2, \ldots, M_K = m_K) = \binom{N}{m_1, \ldots, m_K} \pi_1^{m_1} \cdots \pi_K^{m_K}.
\]

This is called a multinomial distribution. Suppose that \( n \) surveys are conducted and in each survey \( N_i \) people are asked which was their favourite fruit, we observe the random vector \( \{M = (M_{i,1}, \ldots, M_{i,K})\}_{i=1}^{n} \). The log-likelihood is
\[
L_n(\pi_1, \ldots, \pi_K) = \sum_{i=1}^{n} \sum_{r=1}^{K} M_{i,r} \log \pi_r + \sum_{i=1}^{n} \log \binom{N_i}{M_{i,1}, \ldots, M_{i,K}}.
\]

However, we observe that we have an additional condition on the parameters, that is the probabilities \( \sum_{r=1}^{K} \pi_r = 1 \) (thus in the end we only estimate \( (K - 1) \) parameters not \( K \), since the \( \pi_K = 1 - \pi_1 - \ldots - \pi_{K-1} \)). We can either place this condition into the likelihood itself:
\[
L_n(\pi_1, \ldots, \pi_{K-1}) = \sum_{i=1}^{n} \left[ \sum_{r=1}^{K-1} M_{i,r} \log \pi_r + M_{i,K} \log(1 - \pi_1 - \ldots - \pi_{K-1}) \right] \sum_{i=1}^{n} \log \binom{N_i}{M_{i,1}, \ldots, M_{i,K}}
\]
or we include an Lagrange muliplier. This is done by including an extra "dummy" variable into the likelihood

\[ L_n(\pi_1, \ldots, \pi_K, \lambda) = \sum_{i=1}^{n} \sum_{r=1}^{K} M_{i,r} \log \pi_r + \sum_{i=1}^{n} \log \left( \frac{N_i}{M_{i,1}, \ldots, M_{i,K}} \right) - \lambda \left( \sum_{r=1}^{K} \pi_r - 1 \right). \]

the last term forces the \( \sum_{r=1}^{K} \pi_r = 1 \). This can be seen when we differentiate \( L_n(\pi_1, \ldots, \pi_K, \lambda) \) with respect to \( \{\pi_r\}_{r=1}^{K} \) and \( \lambda \):

\[
\frac{\partial}{\partial \pi_s} L_n(\pi_1, \ldots, \pi_K, \lambda) = \sum_{i=1}^{n} \frac{M_{i,s}}{\pi_s} - \lambda \quad 1 \leq s \leq K
\]

\[
\frac{\partial}{\partial \lambda} L_n(\pi_1, \ldots, \pi_K, \lambda) = \sum_{r=1}^{K} \pi_r - 1.
\]

Thus when we solve the \((K+1)\) equations above (by setting them to zero), the very last one constrains \( \sum_{r=1}^{K} \pi_r = 1 \). In this example, the there is not much advantage of adding the additional term \( \lambda \left( \sum_{r=1}^{K} \pi_r - 1 \right) \).

But often placing restrictions to the parameters in a likelihood using Lagrange multipliers can be extremely useful.

### 3.4.3 Evaluation of the MLE for more complicated distributions

Most methods are designed for minimisation of a criterion. Therefore to use these methods we define the negative of the likelihood (which simply turns the likelihood upside down). Therefore, the parameter which maximises the likelihood is the same as the parameter which minimises the negative likelihood. Brute force or minimisation of a function (such as a likelihood) in R can be achieved using the function `optim` or `nlm`.

The type of algorithm that one uses is very important. If the likelihood function is concave (equivalent to the negative of the likelihood being convex), the maximising can be achieved using the Newton-Raphson algorithm or related convex optimisation schemes.

We illustrate the above routine in R for the MLE of the gamma distribution.

```r
# Simulate from a gamma function alpha = 9, beta = 2
# often alpha is called the shape and beta the rate.

xdata = rgamma(n=100,shape =9,rate =2)

# Below is the negative-likelihood based on
# gamma distribution using the generated data.

LikeGamma = function(par){
  n = length(xdata) # xdata is the data vector we input into likelihood
  alpha = par[1]; beta = par[2]

  # Your code here...
}
```


3 Parameter Estimation

\[
\loglik = (\alpha - 1) \sum \log(xdata) - \beta \sum xdata + n\alpha \log(\beta) - n \log(\gamma(\alpha))
\]

\[
n\loglik = -\loglik
\]

\[
\text{return}(n\loglik)
\]

## "minimization" using optim function
## par: initial value; fn = function to minimize
# As initial value we give the vector alpha = 3 and
# beta = 3.
# A better initial value is to give the Moments estimator
# this would speed up the algorithm and if the
# likelihood was not concave (which for this example it is)
# it is more likely to converge to its global minimum.

```r
fit.optim = optim(par= c(3,3), fn=LikeGamma, hessian = T)
```

```r
fit.optim$par
fit.optim$convergence
```

```r
solve(fit.optim$hessian) #inverse hessian
# solve(fit.optim$hessian) gives the asymptotic variance
# of the estimator. See Section 3.5.2 below.
```

Many of the algorithms require an initial value for the estimator \( \hat{\theta} \), to start the algorithm off. If you can provide an initial values which is based on a crude estimator of \( \theta \) (such as a method of moments estimator), this would be great.

Optimisation for a concave distribution is relatively straightforward. Given any initial value you are usually guaranteed to get to the global maximum of the distribution. Under quite general conditions most of the distributions in the exponential are concave, so are ideal for maximising. However, if the distribution is not concave (which often happens in the case of mixtures of distributions as studied in HW2 or the inflated Poisson), then the routine can run into problems. For example, it may converge to a local maximum rather than a global maximum. Typically, runs the routine with several initial values in the hope of finding a global maximum, but this is far from a simple task.

**Remark** (The EM-algorithm). The Expectation-Maximisation algorithm is an algorithm designed for specifically maximising the likelihood. It works by constructing a likelihood based on the observed data and unobserved data. By using a "clever" choice of unobserved data the combined likelihood (often complete likelihood) has a simple form. The EM-algorithm is based on maximisation of the conditional expectation of this complete
3 Parameter Estimation

likelihood. The precise details are beyond this course.

Remark (Profiling). So called profiling the likelihood is another strategy for maximising the likelihood.

3.5 Sampling properties of the MLE

3.5.1 Consistency

It can be shown that the MLE (under certain conditions) is asymptotically consistent. This means that if the distribution is correctly specified then $\hat{\theta}_n \xrightarrow{p} \theta_0$ (where $\theta_0$ is the true parameter in the distribution) as the sample size grows. The proof of this result is quite technical and beyond this class, but we give a heuristic (some ideas) as to why the Maximum likelihood estimator actually works. Rice discusses this in Section 8.5.2.

Let us suppose that $X_i$ are iid random variables with density $f(x; \theta_0)$ the true parameter $\theta_0$ but it is known it belongs to the parameter space $\Theta$. The log-likelihood

$$L_n(\theta; X) = \frac{1}{n} \sum_{i=1}^{n} \log f(X_i; \theta).$$

We observe that we divide by $n$ in the above definition to turn the sum into an average (which we like), in terms of estimation it does not change anything. Since $L_n(\theta; X)$ is an average of iid random variables (in this case $\{\log f(X_i; \theta)\}$) as $n$ grows large it limits to its expectation (see Section 1.4.1)

$$L(\theta) = E[\log f(X; \theta)].$$

We recall that the expectation is defined as

$$E[\log f(X; \theta)] = \int [\log f(x, \theta)]f(x, \theta_0)dx,$$

observe this is a function of $\theta$. To show that the likelihood "works" we need to show that the true parameter $\theta_0$ maximises $L(\theta) = E[\log f(X; \theta)];$ this is like saying in the ideal situation we can observe all the data that we want, then the likelihood is practically $L(\theta) = E[\log f(X; \theta)]$ and $\theta_0$ must maximise it. If at the very minimum, $\theta_0$ did not maximise $E[\log f(X; \theta)]$ then the likelihood method cannot work. Its the bare minimum requirement.

To show that $L(\theta) = E[\log f(X; \theta)]$ is maximised at $\theta_0$ we take the derivative with respect to $\theta$ and show that the derivative is zero at $\theta_0$. Thus

$$\frac{dL(\theta)}{d\theta} = \frac{d}{d\theta} \int [\log f(x, \theta)]f(x, \theta_0)dx.$$
3 Parameter Estimation

We make the assumption that the derivative can be put inside the integral (which for many distributions does hold, but not all):

$$\frac{dL(\theta)}{d\theta} = \int \frac{1}{f(x, \theta)} f(x, \theta_0) dx = \int \frac{df(x, \theta)}{d\theta} \log f(x, \theta) f(x, \theta_0) dx f(x, \theta_0) dx.$$

Thus at $\theta_0$ we have

$$\frac{dL(\theta)}{d\theta} |_{\theta=\theta_0} = \int \frac{df(x, \theta)}{d\theta} |_{\theta=\theta_0} f(x, \theta_0) dx = \int \frac{df(x, \theta)}{d\theta} f(x, \theta_0) dx |_{\theta=\theta_0} = d\frac{1}{d\theta} |_{\theta=\theta_0} = 0.$$

The take home message in the above proof is that the density of pmf integrates to one, it does not depend on a parameter.

Using the above result as a starting point, we obtain the following result.

**Lemma 3.1**

Suppose that $\{X_i\}_{i=1}^n$ are iid random variables with density $f(x; \theta_0)$. Let $\hat{\theta}_n$ be the MLE of $\theta$ based on $\{X_i\}_{i=1}^n$. Then

$$\hat{\theta}_n \xrightarrow{p} \theta_0$$

we $n \to \infty$.

3.5.2 The distributional properties of the MLE

The aim in this section study the sampling properties of the MLE. These results will be used in later chapter for testing and constructing confidence intervals. As we mentioned above many MLE estimators are the same as the method of moments estimator. And we we showed in Section 3.2.3 that the methods of moments estimator is usually asymptotically normal in distribution. In this section we show that (under certain conditions) all MLE estimators are asymptotically normal.

We first illustrate this result with an example. We consider the MLE of the parameters in the Gamma distribution (which is not, in an obvious way, a method of moments estimator). We simulate from a Gamma distribution with $\alpha = 9$ and $\beta = 2$. For simplicity, we treat $\beta = 2$ as known and only estimate $\alpha^1$. The simulations were done over 400 realisations for $n = 1, \ldots, 100$. We estimate $\alpha$ using the routine outlined above and denote the MLE for sample size $n$ as $\hat{\alpha}_n$.

---

1Keep in mind the distribution will change a little if we estimate both. Further the variances will be different if we estimate only one parameters rather than both.
In Figure 3.14 we observe that each realisation does appear to "converge" to the true parameter as the sample size grows. In Figure 3.15 we observe that the estimator has a "finite" sample bias, but the bias decreases as the sample size grows. The $n$ times the variance fluctuates quite a lot but it around $8.5 - 9$ (8.5 is the asymptotic variance times $n$) for a $n$ larger than 30. Finally, in Figure 3.16 we make a plot of the
3 Parameter Estimation

Figure 3.16: Top: Distribution of the standardized sample mean $Z_n = \sqrt{nI(\alpha)(\hat{\alpha}_n - 9)}$ ($n = 3$ and $n = 20$).
Bottom: QQplot against standard normal quantiles.

histogram and QQplot against the standard normal of

$$Z_n = \sqrt{nI(\alpha)(\hat{\alpha}_n - 9)},$$

we discuss what $I(\alpha)$ is below. But we observe that for $n = 3$, the estimator appears to have a small right skew, which is reduced $n = 20$.

These results allude to the result that as the sample size grows, like the method of moment estimators, the distribution of the maximum likelihood estimator asymptotically becomes normal.
Theorem 3.2

Suppose \( \{X_i\} \) are iid random variables with parametric distribution \( f(x; \theta_0) \), where \( \theta_0 \) is unknown. The density \( f(x, \theta) \) is smooth over \( \theta \) (such that we can evaluate its derivative) and the support of the density does not depend on the parameter \( \theta \). Then asymptotically the MLE is

\[
\sqrt{nI(\theta_0)} \left( \hat{\theta}_n - \theta_0 \right) \xrightarrow{D} N(0, 1),
\]

as \( n \to \infty \), where

\[
I(\theta) = -\int_{-\infty}^{\infty} \left( \frac{d^2 \log f(x; \theta)}{d\theta^2} \right) f(x; \theta_0) dx
\]

From the above result for a large enough \( n \) we say that

\[
\hat{\theta}_n \sim N \left( \theta_0, \frac{1}{nI(\theta_0)} \right).
\]

The above theorem is stated for univariate MLE estimators. However, a similar results hold for the MLE of several parameters (for example the MLE of both \( \alpha \) and \( \beta \) in the geometric distribution). In this case we replace \( I(\theta) \) with a matrix, which is the Hessian of the log-likelihood.

Outline of proof of Theorem 3.2

The precise proof of Theorem 3.2 is quite technical. But we give the basic ideas here. In many respects the proof resembles the proof of Lemma 1.2. Even though \( \hat{\theta}_n \) does not appear to be an average, at the estimation method is an average; the log-likelihood which is the sum (or average) of \( \log f(X_i; \theta) \). As the log-likelihood is an average it will asymptotically be normal, which implies (in a way described below) that \( \hat{\theta}_n \) can also be written as an average and will be asymptotically normal.

We first recall that since

\[
\hat{\theta}_n = \arg \max L_n(\theta),
\]

then in general (if \( \hat{\theta}_n \) lies inside the parameter space and not on the boundary), \( \hat{\theta}_n \) is the solution of

\[
\frac{dL_n(\theta)}{d\theta} \bigg|_{\theta = \hat{\theta}_n} = 0.
\]

We recall that \( \frac{dL_n(\theta)}{d\theta} \) is

\[
\frac{dL_n(\theta)}{d\theta} = n^{-1} \sum_{i=1}^{n} \frac{d\log f(X_i; \theta)}{d\theta},
\]
we divide by \( n \) to turn the derivative of the likelihood into an average, which does not change the estimator. Now by using the mean value theorem and expanding \( \frac{dL_n(\theta)}{d\theta} \bigg|_{\theta = \hat{\theta}_n} \) about \( \theta_0 \) (the true parameter) gives

\[
\frac{dL_n(\theta)}{d\theta} \bigg|_{\theta = \hat{\theta}_n} \approx \frac{dL_n(\theta)}{d\theta} \bigg|_{\theta = \theta_0} + (\hat{\theta}_n - \theta_0) \frac{d^2L_n(\theta)}{d\theta^2} \bigg|_{\theta = \hat{\theta}_n},
\]

see the illustration in Figure 3.17 to understand why. Since \( \frac{dL_n(\theta)}{d\theta} \bigg|_{\theta = \hat{\theta}_n} \) this gives

![Illustration of mean value theorem.](image)

Figure 3.17: Illustration of mean value theorem.

\[
\frac{dL_n(\theta)}{d\theta} \bigg|_{\theta = \theta_0} \approx -(\hat{\theta}_n - \theta_0) \frac{d^2L_n(\theta)}{d\theta^2} \bigg|_{\theta = \hat{\theta}_n},
\]

observe that the above involves the Fisher information before taking expectation (often called the observed Fisher information). Going back to the definitions of the likelihood we have

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{df(X_i, \theta_0)}{d\theta} \approx (\hat{\theta}_n - \theta_0) \frac{1}{n} \sum_{i=1}^{n} \frac{df(X_i, \theta_0)}{d\theta^2}.
\]

Since \( \frac{1}{n} \sum_{i=1}^{n} \frac{df(X_i, \theta_0)}{d\theta^2} \) is an average we will replace it by its expectation, which is the Fisher information \( I(\theta_0) \). This gives

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{df(X_i, \theta_0)}{d\theta} \approx (\hat{\theta}_n - \theta_0) I(\theta_0).
\]
Thus

\[
(\hat{\theta}_n - \theta_0) \approx I(\theta_0)^{-1} \frac{1}{n} \sum_{i=1}^{n} \frac{df(X_i, \theta_0)}{d\theta}.
\]

Observe that \(I(\theta)^{-1}\) is a constant, so the limiting distribution of \((\hat{\theta}_n - \theta_0)\) is determined by

\[
\bar{Y}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{df(X_i, \theta_0)}{d\theta},
\]

which is an average of iid random variables. Further, we have shown in Section 3.5.1, that \(E[\log f(X; \theta)]\) is maximised at \(\theta = \theta_0\), which is equivalent (under certain conditions) to \(dE[\log f(X; \theta)]/d\theta\big|_{\theta=\theta_0} = 0\). Thus \(E[\bar{Y}_n] = 0\). Now by the CLT in Theorem 1.1 we have that

\[
\sqrt{n}f(\theta_0)\bar{Y}_n \overset{D}{\to} N(0, 1) \text{ and } \sqrt{n}\bar{Y}_n \overset{D}{\to} N(0, J(\theta_0))
\]

where

\[
J(\theta_0) = \text{var} \left( \frac{df(X_i, \theta_0)}{d\theta} \right)_{\theta=\theta_0}.
\]

Thus altogether we have

\[
(\hat{\theta}_n - \theta_0) \approx I(\theta_0)^{-1} \bar{Y}_n \overset{D}{\to} N \left( 0, I(\theta_0)^{-1}J(\theta_0)I(\theta_0)^{-1} \right) \quad n \to \infty.
\]

Finally, in Lemma 3.3 we prove that \(J(\theta_0) = I(\theta_0)\). Thus the above reduces to

\[
(\hat{\theta}_n - \theta_0) \overset{D}{\to} N \left( 0, I(\theta_0)^{-1} \right) \quad n \to \infty,
\]

which is what we want to show.

**Example: Gamma distribution**

Consider the Gamma example considered above where \(\alpha = 2\) in the distribution is assumed known. The log-likelihood is

\[
L_n(\theta; X_n) = \sum_{i=1}^{n} [(\alpha - 1) \log X_i - 2X_i + \alpha \log 2 - \log \Gamma(\alpha)].
\]

Differentiating the above wrt \(\alpha\) gives

\[
\frac{\partial L_n}{\partial \alpha} = \sum_{i=1}^{n} \left[ \log X_i + \log \beta - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \right].
\]

The second derivative is

\[
-\frac{\partial^2 L_n}{\partial \alpha^2} = \sum_{i=1}^{n} \left[ \frac{\Gamma''(\alpha)}{\Gamma(\alpha)} - \frac{\Gamma'(\alpha)^2}{\Gamma(\alpha)^2} \right].
\]
Notice that the second derivative does not depend on the data. This is quite common for distributions in the exponential family (under certain parameterisations of the parameter), but is not the rule for all distributions. Often the second derivative will depend on the observed data. Based on the above, the Fisher information is

$$I(\alpha) = \left[ \frac{\Gamma''(\alpha)}{\Gamma(\alpha)} - \frac{\Gamma'(\alpha)^2}{\Gamma(\alpha)^2} \right].$$

Therefore the limiting distribution of $\alpha$ is

$$\sqrt{n} \left( \frac{\Gamma''(\alpha)}{\Gamma(\alpha)} - \frac{\Gamma'(\alpha)^2}{\Gamma(\alpha)^2} \right) (\hat{\alpha}_n - \alpha) \xrightarrow{D} N(0, 1). \tag{3.5}$$

This is exactly what we observed in Figures 3.15 and 3.16; the asymptotic variance is $1/(nI(\alpha))$ and the standardized estimator is close to a standard normal for a large enough sample size.

**Example: Exponential distribution**

The exponential data the log-likelihood is

$$L_n(\theta; X_n) = n \log \lambda - \lambda \sum_{i=1}^{n} X_i.$$

Differentiating the above wrt $\lambda$ gives

$$\frac{\partial L_n(\theta; X_n)}{\partial \lambda} = n \lambda - \sum_{i=1}^{n} X_i.$$

The second derivative is

$$\frac{d^2 L_n(\theta; X_n)}{d\lambda^2} = -\frac{n}{\lambda^2}.$$

Thus the Fisher information is

$$I(\lambda) = \frac{1}{\lambda^2}.$$

Therefore the limiting distribution of $\alpha$ is

$$\sqrt{n \frac{1}{\lambda^2}} \left( \hat{\alpha}_n - \lambda \right) \xrightarrow{D} N(0, 1).$$

### 3.5.3 The Fisher information matrix

The Fisher information is an important component in statistics. However, why it involves the second derivative of the likelihood and what it actually means can be quite difficult to understand. We make an attempt in this section.
The Fisher information matrix $I(\theta)$ is really quite a cryptic object, even its name appears nonsensical! The word Fisher comes from Ronald Fisher, so no need to explain that part. But the word information appears strange; roughly speaking $nI(\theta)$ describes how much information the data contains about the true parameter. Below, we try to explain why.

Background: First we recall from our calculus days, that the size of the second derivative at the maximum of a function corresponds to how peaky or curvey it is at the maximum. The "larger" the second derivative the more the curvature (peaky) the likelihood is at the maximum. The "smaller" the second derivative the flatter the function is about the maximum. Recall, that when the second derivative is zero, we have a saddle point. These insights are useful in the discussion below.

For simplicity we focus on the exponential distribution (though the discussion below applies to most distributions). We recall from equation (3.4) that the log-density is
\[
\log f(X, \lambda) = \log \lambda - \lambda X.
\]
The second derivative is
\[
\frac{d^2}{d\lambda^2} \log f(X, \lambda) = -\frac{1}{\lambda^2}.
\]
Thus the information matrix is
\[
I(\lambda) = -\mathbb{E} \left[ \frac{d^2}{d\lambda^2} \log f(X, \lambda) \right] = \frac{1}{\lambda^2}.
\]
Thus for a the data set $\{X_i\}_{i=1}^n$ the information matrix is
\[
nI(\lambda) = \frac{n}{\lambda^2}.
\]
Observe that for the exponential distribution $\frac{d^2}{d\lambda^2} \log f(X, \lambda)$ does not depend on the data. So the curvature of the likelihood is the same for all data sets. However, for other distributions $\frac{d^2}{d\lambda^2} \log f(X, \lambda)$ will depend on the data, which is why we consider the mean curvature. The larger $nI(\lambda)$, the greater the information in the data about $\lambda$. This is easily understood, by studying the likelihood. Since $nI(\lambda)$ corresponds to the (expected) negative second derivative of the likelihood, it measures the steepness of the likelihood about the true, population parameter. The large $nI(\lambda)$ the more pronounced the likelihood about the peak, and the "easier" it is to find the maximum. To see this, we simulate from two different exponential distributions, with $\lambda = 1$ and $\lambda = 100$ and sample size $n = 10$. For $\lambda = 1$, we generate the numbers:

2.0546 0.9861 0.3977 1.9480 0.8082 0.8082 0.0491 2.5444 0.4528 0.9950.

The mean is 1.101 and the log-likelihood is given in Figure 3.18. For $\lambda = 100$, we generate the numbers:

0.0008424 0.0321545 0.0009847 0.0070063 0.0044617 0.0470954 0.0076897 0.0055562 0.0000079 0.0050735.
Figure 3.18: The likelihood for one simulation for an exponential distribution with $\lambda = 1$ and $n = 10$. $nI(\lambda) = 10/\lambda = 10$.

Figure 3.19: The likelihood for one simulation from an exponential distribution with $\lambda = 100$ and $n = 10$. $nI(\lambda) = 10/\lambda = 1$ (left and right plots are the same likelihood, just on different ranges).

The mean is 0.01108 and the log-likelihood is given in Figure 3.19. What we observe is when $\lambda$ is small, the likelihood is more peaked about the maximum. We are better able to find the maximum; this corresponds to a smaller variance and “more” information in the data about the underlying parameter (thus $nI_n(\theta_0)$ is large or equivalently the asymptotic variance $(nI_n(\theta_0))^{-1}$ is small). On the other hand, when $\lambda$ is large, the likelihood is more “flat” about the maximum. Making it much harder to “find” the maximum. It is difficult to distinguish the maximum from other values in the neighbourhood. This means there is “less” information
in the data about the underlying parameter (thus \( nI_n(\theta_0) \) is small or equivalently the asymptotic variance \((nI_n(\theta_0))^{-1}\) is large). If you run the same simulation, but with a larger \( n \), you will see that the likelihood gets increasingly steeper about its maximum.

HW: Ask the same question for the Poisson distribution (using `rpois`).

**Lemma 3.3**

\[
E \left( \frac{d \log L_n(\theta; X)}{d \theta} \right)_{\theta=\theta_0}^2 = -E \left( \frac{d^2 \log L_n(\theta; X)}{d \theta^2} \right)_{\theta=\theta_0}.
\]

**PROOF.** To prove this result we use the fact that the likelihood \( L_n \) is a density/distribution, thus it integrates to one:

\[
\int L_n(\theta, x) dx = 1.
\]

Now by differentiating the above with respect to \( \theta \) gives

\[
\frac{\partial}{\partial \theta} \int L_n(\theta, x) dx = 0.
\]

Thus

\[
\int \frac{\partial L_n(\theta, x)}{\partial \theta} dx = 0 \Rightarrow \int \frac{\partial \log L_n(\theta, x)}{\partial \theta} L_n(\theta, x) dx = 0.
\]

Differentiating again with respect to \( \theta \) and taking the derivative inside gives

\[
\begin{align*}
\int \frac{\partial^2 \log L_n(\theta, x)}{\partial \theta^2} L_n(\theta, x) dx &+ \int \frac{\partial \log L_n(\theta, x)}{\partial \theta} \frac{\partial^2 L_n(\theta, x)}{\partial \theta^2} dx = 0 \\
\Rightarrow \int \frac{\partial^2 \log L_n(\theta, x)}{\partial \theta^2} L_n(\theta, x) dx &+ \int \frac{\partial \log L_n(\theta, x)}{\partial \theta} L_n(\theta, x) \frac{1}{\frac{\partial \log L_n(\theta, x)}{\partial \theta}} \frac{\partial^2 L_n(\theta, x)}{\partial \theta^2} dx = 0 \\
\Rightarrow \int \frac{\partial^2 \log L_n(\theta, x)}{\partial \theta^2} L_n(\theta, x) dx &+ \int \left( \frac{\partial \log L_n(\theta, x)}{\partial \theta} \right)^2 L_n(\theta, x) dx = 0.
\end{align*}
\]

Thus

\[
-E \left( \frac{\partial^2 \log L_n(\theta, X)}{\partial \theta^2} \right) = E \left( \frac{\partial \log L_n(\theta, X)}{\partial \theta} \right)^2.
\]

Note in all the derivations we are evaluating the second derivative of the likelihood at the true parameter \( \theta \) of the underlying distribution.
3.5.4 The curious case of the uniform distribution

We recall that if \( \{X_i\} \) are iid random variables with uniform density \( f(x, \theta) = \theta^{-1}I_{[0, \theta]}(x) \). The uniform distribution is one of those distributions that do not satisfy the so called “regularity” conditions mentioned above (this is because the support of the distribution involves the parameter, thus Lemma 3.3 does not hold). We do not maximise the likelihood (or log-likelihood) by differentiating it. We would expect that its sampling properties are also strange, we now show that they are.

We recall from Section 3.4.2, that the MLE is
\[
\hat{\theta}_n = \max_{1 \leq i \leq n} (X_i)
\]

The sampling distribution of \( \hat{\theta}_n \) actually quite straightforward to derive (as it is not normal, observe the MLE does not involve an average). We start with the distribution function;
\[
P(\hat{\theta}_n \leq x) = P\left( \max_{1 \leq i \leq n} (X_i) \leq x \right)
= P( X_1 \leq x \text{ and } X_2 \leq x \text{ and } \ldots \text{ and } X_n \leq x ) = \prod_{i=1}^{n} P(X_i \leq x) = \frac{x^n}{\theta^n}.
\]

This gives the distribution function. The density is the derivative of the distribution function with respect to \( x \)
\[
f_{\hat{\theta}_n}(x) = \frac{dP(\hat{\theta}_n \leq x)}{dx} = \frac{n x^{n-1}}{\theta^n} \quad x \in [0, \theta]
\]
Thus the distribution is the MLE is \( \frac{n x^{n-1}}{\theta^n} \), which is even close to normal, even for large \( n \). This is one strange property. Now we obtain another. We start with the expectation:
\[
E[\hat{\theta}_n] = n \int_0^{\theta} x \frac{x^{n-1}}{\theta^n} \frac{dx}{\theta^n} = \frac{n}{\theta} \left[ \frac{x^{n+1}}{n+1} \right]_{x=0}^{\theta} = \frac{n}{n+1} \theta.
\]
Thus there is a finite sample bias
\[
\frac{n}{n+1} \theta - \theta = -\frac{1}{n} \theta,
\]
which tends to zero as \( n \to \infty \). Next we consider the variance. We start with the second moment:
\[
E[\hat{\theta}_n^2] = n \int_0^{\theta} x^2 \frac{x^{n-1}}{\theta^n} \frac{dx}{\theta^n} = \frac{n}{\theta} \left[ \frac{x^{n+2}}{n+2} \right]_{x=0}^{\theta} = \frac{n}{n+2} \theta^2.
\]
Thus
\[
\text{var}[\hat{\theta}_n] = \frac{n}{n+2} \theta^2 - \frac{n^2}{(n+1)^2} \theta^2
= n \left( \frac{1}{n+2} - \frac{n}{(n+1)^2} \right) \theta^2
= n \left( \frac{n+1-n}{(n+2)(n+1)^2} \right) \theta^2 = \frac{n}{(n+1)^2(n+2)} \theta^2.
\]
Now observe what happens to the variance as \( n \) gets large:

\[
\text{var}[\tilde{\theta}_n] = \frac{n}{(n+1)^2(n+2)} \theta^2 = \frac{1}{(n+1)^2(n/n + 2/n)} \theta^2 \sim \frac{1}{(n+1)^2} \theta^2.
\]

The standard error for the MLE is:

\[
\sqrt{\text{var}[\tilde{\theta}_n]} = \frac{1}{(n+1)\sqrt{(n/n + 2/n)}} \theta \sim \frac{1}{(n+1)} \theta.
\]

This is (much, much) faster than the usual standard error of the sample mean which is \( \sigma / \sqrt{n} \). Thus the variance of the MLE for the uniform distribution is in general “very good” for small sample sizes (indeed the bias and the variance are of the same “size”). But we do observe that the standard error for the MLE depends on the parameter \( \theta \). Thus the larger \( \theta \), the more spread out the data and the larger that standard error of the MLE.

Figure 3.20 offers an intuitive explanation as to why the MLE for the uniform distribution is so “good”.

3.6 What is the best estimator?

3.6.1 Measuring efficiency

For many different procedures there is not unique estimation method. For example, in the previous sections we considered the method of moments estimator of the Gamma distribution and the MLE for the Gamma
distribution. Given the array of estimation methods how would one choose a particular method. It seems sensible to use the estimator which has a greater change of concentrating about the mean. There are various ways to measure this, but one method is the mean squared error $E[(\hat{\theta}_n - \theta)^2]$, and to select the method with the smallest MSE for a given $n$. We recall (see Section 1.5.1) that

$$E[(\hat{\theta}_n - \theta)^2] = \text{var}(\hat{\theta}_n) + \left( E[\hat{\theta}_n] - \theta \right)^2.$$  

Assuming that the estimator in unbiased (or nearly unbiased), then one would compare the variances of the estimator. Often this is not easy to derive (especially for finite sample sizes), but the asymptotic variance can be derived. For example, the asymptotic variance of the method of moments estimator based on the first moment (as given in (3.3)) is approx

$$\text{var}(\hat{\theta}_{\text{MoM},n}) = n^{-1} \left( \frac{\Gamma''(\alpha)}{\Gamma(\alpha)} - \frac{\Gamma'(\alpha)^2}{\Gamma(\alpha)^2} \right).$$

Thus the asymptotically most efficient estimator is the one with the smallest variance

$$n^{-1} \left( \frac{\Gamma''(\alpha)}{\Gamma(\alpha)} - \frac{\Gamma'(\alpha)^2}{\Gamma(\alpha)^2} \right)^{-1} \times \frac{\alpha}{n}.$$ 

**Definition 3.2 (Efficiency).** Given the two estimators $\hat{\theta}_{1,n}$ and $\hat{\theta}_{2,n}$ which are unbiased (or it is very small), we measure the efficiency of $\hat{\theta}_{1,n}$ relative to $\hat{\theta}_{2,n}$ using the following measure:

$$\text{eff}(\hat{\theta}_{1,n}, \hat{\theta}_{2,n}) = \frac{\text{var}(\hat{\theta}_{1,n})}{\text{var}(\hat{\theta}_{2,n})}.$$  

Thus for the gamma distribution example, the (asymptotic) relative efficiency (since we do not have an analytic expression for the variance of $\hat{\theta}_{\text{MLE},n}$) is

$$\text{as-eff}(\hat{\theta}_{\text{MoM},n}, \hat{\theta}_{\text{MLE},n}) = n^{-1} \left( \frac{\Gamma''(\alpha)}{\Gamma(\alpha)} - \frac{\Gamma'(\alpha)^2}{\Gamma(\alpha)^2} \right).$$

A plot of $\text{as-eff}(\hat{\theta}_{\text{MoM},n}, \hat{\theta}_{\text{MLE},n})$ is given in Figure 3.21. We observe that $\text{as-eff}(\hat{\theta}_{\text{MoM},n}, \hat{\theta}_{\text{MLE},n}) > 1$, thus (asymptotically) the MLE estimator tends to be more efficient than the method of moments estimator (an explanation as to why is given in the next section). However, as $\alpha$ grows, the relative efficiency converges to one. This means that for large $\alpha$ (asymptotically) there is very little difference in performance of the estimators. In the following two sections we explain why the MLE (under uncertain regularity conditions) performs so well. NEED TO DO: RUN SOME SIMULATIONS COMPARING THE TWO ESTIMATORS.

HW Question: Compare different method of moment estimators of exponential distribution.
3 Parameter Estimation

Figure 3.21: A plot of \( \text{asyeff}(\hat{\theta}_{\text{MoM},n}, \hat{\theta}_{\text{MLE},n}) \) for the Gamma distribution.

3.6.2 The Cramer-Rao Bound

As we mentioned in the previous section, different estimators may have different mean squared errors. Any one family of distributions may have an infinite number of different estimator to select. How do we know if we have the best estimator or if it can be improved. If there existed a lower bound for mean squared error, we can use this as a benchmark to compare the estimator with. If the variance of our estimator attains the lower bound or it is only slightly higher than it, then there is no need to search for a better estimator.

We now state a classical result where such a lower bound is obtained. This result only applies to the class of estimators which are unbiased, nevertheless it is extremely powerful and useful. It is interesting to note that was independently derived by C.R. Rao when he was only 24 years old and Harold Cramer.

**Theorem 3.4**

Suppose \( \{X_i\}_{i=1}^n \) are iid random variables with density \( f(x; \theta) \). Let \( T(X_1, \ldots, X_n) \) be an unbiased estimator of \( \theta \) i.e. \( E[T(X_1, \ldots, X_n)] = \theta \). Then under certain regularity conditions on the density we have

\[
\text{var}[T(X_1, \ldots, X_n)] \geq \frac{1}{nI(\theta)},
\]

where \( I(\theta) \) denotes the Fisher information matrix corresponding to \( X \).

This result gives a lower bound on how small the variance of an unbiased estimator can actually be.
Interestingly, we previously showed that asymptotic variance of the MLE is \((nI/n)^{-1}\). This means that for finite samples the MLE may not be best estimator of \(\theta\), but at least asymptotically (this means for large samples), using the MLE will usually yield an estimator that is close to the best.

### 3.7 Sufficiency

Suppose you are an astronaut/cosmonaut exploring Pluto (the unfortunate x-planet). You are collecting rock samples and want to transmit their weights to earth. However, it takes several minutes to transmit one piece of information (the weight of one rock) and you have 1000 rocks. It would take several days to transmit all the information about the rocks to earth. The people on earth are getting quite impatient and want the information as fast as possible. What do you do? Is it really necessary to give the people on earth the exact data (data specific information). May be it is sufficient to produce a synthetic data (based on the information you send then), which reproduces the main features in true data as closely as possible. If such information were sufficient, we should transmit the maximum information about the distribution of the data.

We now make this precise, by defining the notion of a sufficient statistic.

**Definition 3.3 (Sufficiency).** Suppose \(T(X)\) is a function of the data \(X\). Then \(T(X)\) is called a sufficient statistic for the parameter \(\theta\), if the conditional distribution of \(X\) given \(T(X)\) does not depend on the parameter \(\theta\). Formally we write this as

\[
\text{The above above definition tells us the following. Suppose the sample mean } \bar{X}_n \text{ is a sufficient statistic for the population mean } \mu (\text{for some distribution). If we observe the the sample mean } \bar{X}_n = 3. \text{ Then the conditional distribution of } X_1, \ldots, X_n \text{ conditioned on } n^{-1} \sum_{i=1}^n X_i = 3 \text{ does not depend on the population mean } \mu. \text{ Returning to our the example of you collecting rocks samples from Pluto. If it is known that the distribution of the weight of rocks comes form a known parametric family, with unknown parameter } \theta, \text{ and } T(X) \text{ is a sufficient statistic for } \theta. \text{ Rather than transmitting the weights of 1000 rocks to earth you can simply evaluate and send } t = T(X) \text{ to earth (which only takes a few minutes). And the people back on earth can produce a synthetic data set by drawing samples from numbers from the conditional density}
\]

\[
f(X_1 = x_1, \ldots, X_n = x_n | T(X) = t),
\]

which is knowns (since the parametric density is known and the above conditional density does not depend on unknown parameters).

We first give an example of a sufficient statistic and then state necessary and sufficient conditions for a sufficiency.

**Example 3.3 (Bernoulli random variables).** Suppose \(\{X_i\}_{i=1}^n \) are Bernoulli random variables where \(P(X_i = 0) = 1 - \pi \) and \(P(X_i = 1) = \pi\). We now show that \(T_n(\bar{X}) = \sum_{i=1}^n X_i\) is a sufficient statistic for \(\pi\). Our aim is to
3 Parameter Estimation

show that $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n|T_n(X) = t)$ does not depend on $\pi$. There are two ways this can be shown. We first use the brute force method. Using the classical $P(A|B) = P(A \cap B)/P(B)$ we have

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n|T_n(X) = t) = \frac{P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n, T_n(X) = t)}{P(T_n(X) = t)}.$$  

To evaluate the joint probability $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n, T_n(X) = t)$ we first consider a few examples. Suppose $n = 2$ and $T_n(X) = 1$, then

$$P(X_1 = 1, X_2 = 1, T_n(X) = 1) = 0 \text{ but } P(X_1 = 1, X_2 = 0, T_n(X) = 1) = P(X_1 = 1, X_2 = 0) = \pi(1 - \pi).$$

Using the above, in general we observe that

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n, T_n(X) = t) = \begin{cases} 
\pi^t(1 - \pi)^{n-t} & \sum_{i=1}^n x_i = t \\
0 & \sum_{i=1}^n x_i \neq t
\end{cases}$$

This deals with the numerator. Now we consider the denominator, since $T_n = \sum_{i=1}^n X_i \sim \text{Bin}(n, \pi)$, then $P(T_n = t) = \binom{n}{t} \pi^t(1 - \pi)^{n-t}$. Putting the numerator and denominator together gives the conditional probability

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n|T_n(X) = t) = \begin{cases} 
\frac{\binom{n}{t} \pi^t(1 - \pi)^{n-t}}{\sum_{i=1}^n x_i = t} & \sum_{i=1}^n x_i = t \\
0 & \sum_{i=1}^n x_i \neq t
\end{cases}$$

In other words, if $\sum_{i=1}^n x_i = t$ then

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n|T_n(X) = t) = \frac{1}{\binom{n}{t}}.$$  

The above could also be deduced using basic combinatorial arguments.

However, what we do observe is that if to total number of successes is known, than $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n|T_n(X) = t)$ does not depend on the underlying parameter $\pi$. Thus $T_n(X) = \sum_{i=1}^n X_i$ is a sufficient statistic for $\pi$.

In general it is difficult to “come up” with sufficient statistics. The following result gives necessary and sufficient conditions for a sufficient statistic.

**Theorem 3.5**

A necessary and sufficient condition for $T(X)$ to be a sufficient statistic for $\theta$ is that the joint probability function of $X$ (either density or probability mass function) can be factorised as follows

$$f_X(x_1, \ldots, x_n; \theta) = h(x_1, \ldots, x_n)g[T(x_1, \ldots, x_n); \theta].$$

Remark: The important aspect of this result, is that the function $h(\cdot)$ does not depend on $\theta$. All the
information on the function \( \theta \) is contained in the function \( g(\cdot; \theta) \).

It is often (but not always) easier to consider the log of the joint probability function. By the factorisation theorem the log probability is

\[
\log f_X(x_1, \ldots, x_n; \theta) = \log h(x_1, \ldots, x_n) + \log g[T(x_1, \ldots, x_n); \theta].
\]

**Example: The exponential distribution**

Suppose \( \{X_i\} \) are iid random variables with exponential density \( f(x; \lambda) = \lambda \exp(-\lambda x) \). Then the log of the joint density is

\[
\log f_X(x_1, \ldots, x_n; \lambda) = \sum_{i=1}^{n} \log f(x; \lambda) = -\lambda \sum_{i=1}^{n} x_i.
\]

All the data is summarized in \( T(x) = \sum_{i=1}^{n} x_i \). Thus by the factorisation theorem \( T(x) \) is a sufficient statistic for \( \lambda \).

**Example: The normal distribution**

We use the log-expansion given in Example 2.3. The log of the normal density is

\[
\log f_X(x_1, \ldots, x_n; \mu, \sigma^2) = \sum_{i=1}^{n} \log f(x; \mu, \sigma^2) = \frac{\mu}{\sigma^2} \sum_{i=1}^{n} x_i - \frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 + \frac{n\mu^2}{2\sigma^2} + \frac{n}{2} \log(2\pi).
\]

Again we see that the data is “clustered together” in two terms \( \sum_{i=1}^{n} x_i \) and \( \sum_{i=1}^{n} x_i^2 \). Thus we observe that \( T_1(x) = \sum_{i=1}^{n} x_i \) and \( T_2(x) = \sum_{i=1}^{n} x_i^2 \) are sufficient statistics for \( \mu \) and \( \sigma^2 \). Further, \( T_1(x) \) is a sufficient statistic for \( \mu \).

**Example: All distributions in the exponential family**

The above two examples both come from the exponential family, with good reason. All distributions which come from the exponential family can be summarized in a few sufficient statistics (that do not depend on the sample size). To understand why, we recall from Section 2.3 that a family of distributions belong to the exponential family if they can be written as

\[
f(x; \theta) = \exp \left[ \sum_{j=1}^{k} s_j(x)T_j(\theta) + b(\theta) + c(x) \right] \quad x \in A,
\]
where $A$ does not depend on the parameter $\theta$ and $\theta = (\theta_1, \ldots, \theta_K)$. Thus if $\{X_i\}$ are iid random variables with distribution $f(x; \theta)$, then their joint density is

$$
\log f_{X}(x_1, \ldots, x_n; \lambda) = \sum_{i=1}^{n} \left[ \sum_{j=1}^{K} s_j(x_i) T_j(\theta) + b(\theta) + c(x_i) \right]
$$

$$
= \sum_{j=1}^{K} \sum_{i=1}^{n} s_j(x_i) T_j(\theta) + n b(\theta) + \sum_{i=1}^{n} c(x_i).
$$

Thus $T_1(x), \ldots, T_K(x)$ are collectively sufficient statistics for $\theta$. Observe that since $T_1(x), \ldots, T_K(x)$ does not depend on the sample size, we are able to summarize important aspects of the data into just $K$ terms.

**Example: The uniform distribution (not in exponential family)**

The uniform distribution is an example of a distribution that does not belong to the exponential family but whose number of sufficient statistics does not depend on the sample size. We recall that if $\{X_i\}$ are iid random variables with uniform density, then their joint density (see Section 3.4.2) is

$$
f_{X}(x_1, \ldots, x_n; \theta) = \theta^{-n} \prod_{i=1}^{n} I_{(0, \theta]}(x_i),
$$

where $I_{(0, \theta]}(x)$ is the indicator variable, which is one for $x \in [0, \theta]$ and zero elsewhere. With some thought we observe that this can be written as

$$
f_{X}(x_1, \ldots, x_n; \theta) = \theta^{-n} I_{(0, \theta]}(\max_i x_i).
$$

Thus $T(x) = \max_i x_i$ is a sufficient statistic for $\theta$.

**Example: The Weibull distribution; sufficient statistic cannot be reduced**

We now show that the sufficient statistics of the Weibull distribution cannot be reduced to just a few (which do not depend on the sample size).

Suppose that $\{X_i\}$ are iid random variables with the Weibull distribution;

$$
f(x; \lambda, k) = \left( \frac{k}{\lambda} \right) \left( \frac{x}{\lambda} \right)^{k-1} \exp \left( -\left( \frac{x}{\lambda} \right)^k \right) \quad x \geq 0.
$$

The log of the joint density is

$$
\log f_{X}(x_1, \ldots, x_n; \lambda, k) = -\frac{1}{\lambda^k} \sum_{i=1}^{n} x_i^k + (k-1) \sum_{i=1}^{n} \log x_i - n(k-1) \log \lambda + n \log k - n \log \lambda.
$$
Now suppose that $k$ is assumed known. Then $\sum_{i=1}^{n} x_i^k$ is a sufficient statistic for $\lambda$; since
\[
\log f_X(x_1, \ldots, x_n; \lambda, k) = -\frac{1}{\lambda^k} \sum_{i=1}^{n} x_i^k - n(k-1) \log \lambda - n \log (k-1) \sum_{i=1}^{n} \log x_i.
\]
However, no such sufficient statistic exists for $k$ (except for the original data).

### 3.7.1 Application of sufficiency to estimation: Rao-Blackwellisation

Sufficiency has various applications in statistics. But we conclude this section with a very elegant application to estimation. The theorem below is called the Rao-Blackwell theorem.

**Theorem 3.6**

Suppose that $\hat{\theta} = \hat{\theta}(X)$ is an estimator of $\theta$ based on the random variables $X$ (and $E[\hat{\theta}^2] < \infty$) and $T(X)$ is a sufficient statistic for $\theta$. Define the "new estimator" $\tilde{\theta} = E[\hat{\theta}(X)|T(X)]$.

Then
\[
E \left[ \tilde{\theta} - \theta \right]^2 \leq E \left[ \hat{\theta} - \theta \right]^2.
\]

**PROOF.** The proof is straightforward (but will not be tested). We first note that by iterated expectation we have
\[
E \left( \tilde{\theta} \right) = E \left( E[\hat{\theta}|T(X)] \right) = E \left( \hat{\theta} \right)
\]
Thus the bias of both $\hat{\theta}$ and $\tilde{\theta}$ are the same. Next we focus on the variance. By using the well known conditional variance identity (see HW2) we have
\[
\text{var}[\tilde{\theta}] = \text{var}[E[\hat{\theta}|T(X)]] + E \left( \text{var}[\hat{\theta}|T(X)] \right).
\]
Since $E \left( \text{var}[\hat{\theta}|T(X)] \right) \geq 0$, this immediately implies that
\[
\text{var}[\tilde{\theta}] \geq \text{var}[E[\hat{\theta}|T(X)]].
\]
Thus giving the required result. \( \square \)

The Rao-Blackwell theorem allows one to improve an estimator by conditioning on a sufficient statistic. It also makes it understand that "good estimators" should be functions of the sufficient statistic. If you return to the MLE, in particular the MLE of the exponential family, you will observe that the maximum likelihood
3 Parameter Estimation

estimator is a function of the sufficient statistics (this will take some algebraic manipulation). This gives credence to the claim that the MLE yield good estimators.

Under stronger conditions on the sufficient statistic, one can show that if an (unbiased) estimator is a function of the sufficient statistic it cannot be improved (this result is beyond this class but is related to minimal sufficiency and completeness).

3.8 What happens if we get the assumptions wrong

NEED TO DO.

3.9 A historical perspective

Mention Jackknife (bias reduction) proposed by Maurice Quenouille. Discuss Blackwell one of the early Afro-American statistician who made fundamental contributions to statistics.

Also discuss C.R.Rao.