Chapter 1

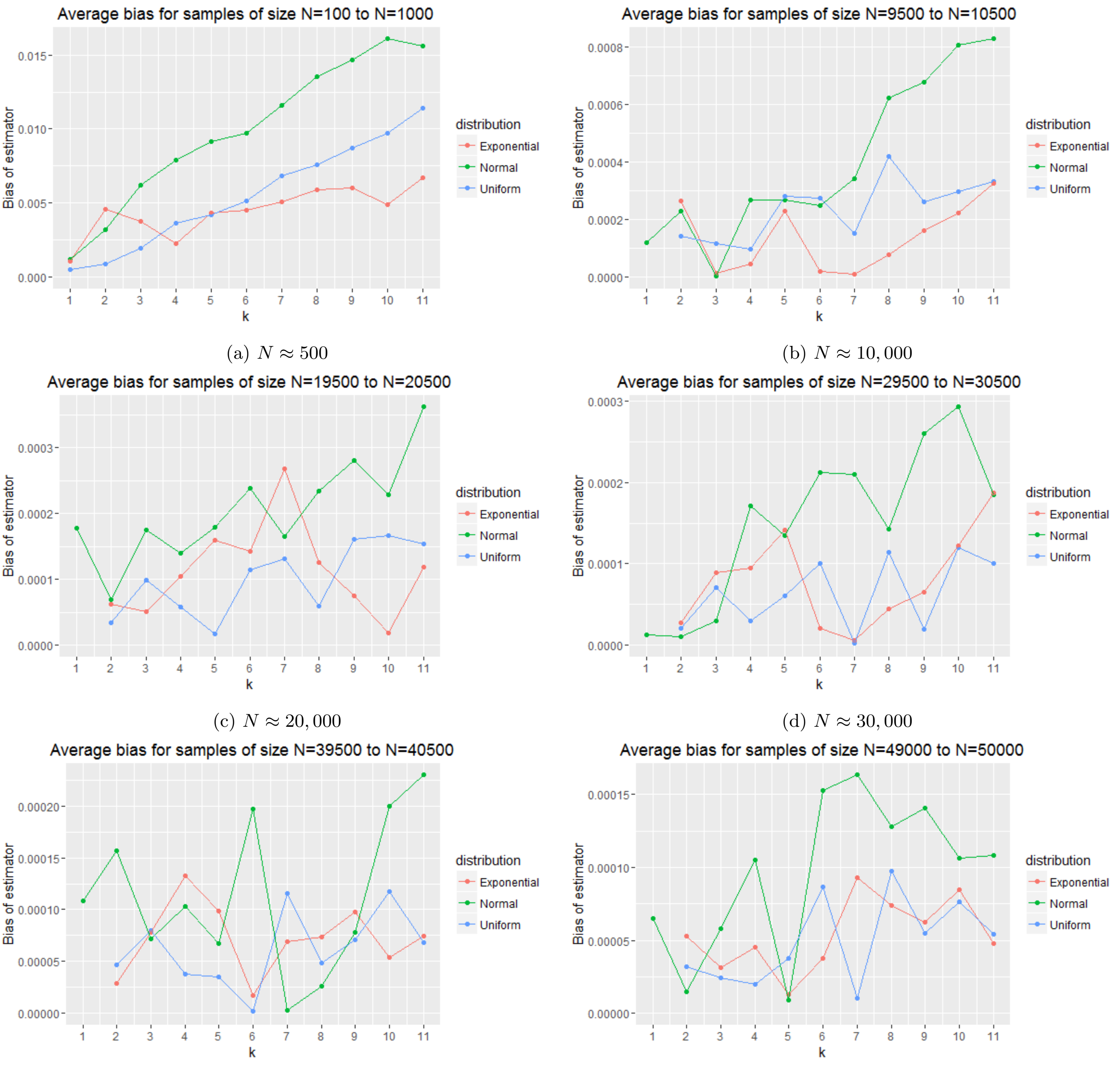
Conclusion

As seen from the analysis undertaken in Chapter **??**, taking an estimation of entropy from samples of the normal, uniform and exponential distributions, appear to act differently. There is no conclusive answer, that agrees across all three distributions, as for the best value of k for the estimator, or if the bias of the estimator is of $O\left(\frac{1}{N^a}\right)$ or $O\left(\left(\frac{k}{N}\right)^a\right)$.

I previously discussed the supposed best value of k as shown by the graphs with regression lines of the bias for each distribution. Where I mean optimal in terms of reducing the bias, so the estimated value is as close to the exact value as possible. To view this in more detail, comparing distributions and sample sizes, I have taken 1000 samples around the sample sizes of 500, $10,000$, $20,000$, $30,000$, $40,000$ and $49,500$, found the average bias of the estimator for each k and each distribution. I have graphically represented this in Figure 1.1.

If we ignore the distribution, there is not an always obvious best value of k shown by the graphs. Moreover, considering the analysis completed in section **??**, we found that the values of k that satisfy Theorems **??** and **??**, depend on N so that $k \in \{k_0^*, ..., k_1^*\}$, where $k_0^* \approx 1$ and $k_1^* = O(N^{\frac{2}{5}})$. Thus for the samples in the graphs above, we can create a table of the values of k that satisfy the conditions to be able to assume consistency and unbias of the estimator, Table 1.1.

Although, for samples of size $N \leq 1,000$, we can see that, quite obviously, the smallest bias occurs with an estimator found using $k = 1$, and this is true for all the distributions, graph (a). If we consider the comparison graphs in Chapter **??**, when $N \leq 1,000$ we approximately have $log(N) \leq 7$, and in Figures **??**, **??** and **??** we have the lowest line, the smallest bias values at $k = 2$, $k = 3$ and $k = 4$ for the normal, uniform and exponential samples respectively. This difference in results could be due to a number of reasons. Firstly, from the current analysis $k = 1$ appears to be the optimal value for the smallest bias; however, when looking at the uniform and exponential samples, the regression line for $k = 1$ was omitted, due to the large number of `Inf` values (reasoning for these values were discussed previously). Secondly, the graphs shown in Chapter



(a) $N \approx 500$

(b) $N \approx 10,000$

(c) $N \approx 20,000$

(d) $N \approx 30,000$

(e) $N \approx 40,000$

(f) $N \approx 49,500$

Figure 1.1: *Comparing the average bias about sample size* N, *for each* k *and each distribution.*

Table 1.1: *Values of k satisfying Theorems* **??** *and* **??** *for sample sizes N*

N	k
500	$\{1,2,3\}$
$1,000$	$\{1,2,3,4\}$
$10,000$	$\{1,2,...,7\}$
$20,000$	$\{1,2,...,9\}$
$30,000$	$\{1,2,...,9\}$
$40,000$	$\{1,2,...,10\}$
$50,000$	$\{1,2,...,11\}$

?? are of regression lines, which we have assumed to be linear as $N \to \infty$. Thus, when looking at the information shown by the lines in a particular region where N is small, we cannot say that these are accurate. Therefore, the difference in results is not a problem; and the graph in Figure 1.1 (a), is more accurate when looking at this N. Moreover, from this graph, we can see a uniform increase in the bias, against the size of k for both the uniform and normal distributions; thus for $N \approx 500$ and $k \in \{1,2,3\}$, the optimal estimator is definitely found using $k = 1$.

Considering graph (b) for sample size $N \approx 10,000$, we have that the smallest values of bias occur when $k = 3$ or 7 for the exponential, $k = 3$ for the normal and $k = 4$ for the uniform distributions. This agrees with the information found when finding k to satisfy condition **??** we have that for this sample size $k \in \{1,2,...,7\}$. This also implies that the most accurate estimator for a sample of this size could be found when using $k = 3$ (or possibly 4), since for all three distributions, at these values of k, we do have some of the smallest sizes of bias. When $N \approx 10,000$ we have $log(N) \approx 9.2$, which is where on Figures **??**, **??** and **??**, the regression lines begin to cross. Thus, looking at these figures makes it difficult to decipher which value of k appears to have the smallest bias, however, if we look at Figures **??**, **??** and **??**, which zoom in around $log(N) \approx 9$, we can see that for the normal distribution, the lowest lines occur at $k = 2$ then 3, for the uniform $k = 4$ then 3 and for the exponential $k = 8$ then 3. The results in these graphs, as well as the graphs in this chapter all give the impression that $k = 3$ is a strong choice for the optimal value of k when finding the estimator for sample size $N \approx 10,000$.

For graph (c), with $N \approx 20,000$, we have a range of values for k with the smallest bias, depending on the distribution; for the exponential $k = 10$, the normal $k = 2$ and uniform $k = 5$. However, considering that for this sample size we must have $k \in \{1,2,...,9\}$, the value of k for the exponential distribution cannot be $k = 10$ for the Theorems **??** and **??** to be satisfied; thus the next smallest bias occurs when $k = 3$. Due to this, nothing conclusive can be said about the optimal value of k, although, we could draw the conclusion that for this sample size, the best k is from the range $k = \{2,3,5\}$. However, when choosing the values of k that satisfy condition **??**, I made some approximations, which depended on the choice of $\alpha > d = 1$, and a value τ which is smaller than a bound. Due to this, the value of tau could be chosen to be slightly larger; for example $\tau = \frac{19}{80} < \frac{1}{4}$, which would give $k_1^* = N^{\frac{19}{80}} = 20000^{\frac{19}{80}} = 10.507$. So $k = 10$ could be included in the values that satisfy the Theorems and cannot be discounted. In fact, if we did use this new value of τ, then the values for the optimum k would be $k = \{2, 5, 10\}$, which makes even less sense.

Moreover, we can also examine the graphs in Chapter **??**, for this sample size $N \approx 20,000$ which gives $log(N) \approx 9.9$. Thus, Figures **??**, **??** and **??**, which zoom around this sample size, show that the optimal choice of k for the normal distribution is $k = 5$, for the uniform is $k = 7$ then 5 and for the exponential $k = 8$ or 10. The exponential agrees on $k = 10$ and the uniform agrees on $k = 5$, but these are both the second lowest lines on the graph, and other that this non of the above results agree across distributions with those from earlier. This could be due to a number of reasons, but the most obvious one being that there is not a fixed value of k for this sample size that coincides for all distributions as the optimal value of k. Thus, for this sample size we cannot draw any conclusions about the optimal value of k.

When $N \approx 30,000$, graph (d), we have the optimal values of k - when the smallest bias occurs - being 2 for the normal distribution and 7 for both the uniform and exponential distributions. Interestingly, for all three distributions $k = 2$ has either the first, second or third smallest bias, out of all k for its own distribution. However, while the normal distribution has a general increase for $k > 2$, the other two distributions later dip to their lowest value of bias at $k = 7$. Thus, while one may say that $k = 7$ is the optimal, since it agrees with two of the three distributions, for the third distribution, the normal distribution, the bias is significantly higher. Henceforth, I would actually assume that a safer option, for the best value of k across any distribution would be when $k = 2$, where the bias for all three distributions is significantly small. Considering the regression lines, found in Chapter **??**, for $log(N) \approx 10.3$, we can see that the optimal choice of k appears to be $k = 5$ for the normal distribution, $k = 8$ for the uniform and $k = 10$ for the exponential. None of these agree with the possibility the $k = 2$ is the optimal value, in fact when considering the graphs of the comparison of the regression lines for each distribution, we can see that the line for $k = 2$ is not nearly the lowest of the group. More specifically for both the exponential and uniform distributions $k = 2$ appears to be the worst possible value of k in decreasing the bias, and since we are now considering a relatively large sample size, one would expect these graphs to have more accurate results. Although all these values of k discussed fit within those found in Table 1.1; no value of k stands out as an optimal value, thus no conclusion can be drawn for this sample size.

Now, for the graph (e), when $N \approx 40,000$ we have the smallest bias occurring at $k = 6$ for the uniform and exponential and $k = 7$ for the normal. If we consider the value for the bias at $k = 6$ for the normal, we can see sudden significant increase in the size of the bias and a massive drop between this value, and that for $k = 7$. Moreover, considering $k = 7$ for the uniform and exponential distributions, we can see a substantial increase from the bias at $k = 6$. Indicating, that the best value of k may be 6 or 7; but this strongly depends on the distribution that the sample is taken from, not just the sample size. Now looking at the regression comparison graphs, around $N \approx 40,000$ we have $log(N) \approx 10.6$, which shows that for the normal distribution the optimal value of $k = 5$ and 7, for the uniform it's $k = 7$ then $6, 8, 9$ (lines are very close) and for the exponential $k = 10$ then $6, 7, 8$ (very close regression lines). This slightly agrees with above, that $k = 6$ or 7 could be the best value of k to choose for the estimator; moreover, it also shows less of a dependence on distribution. Thus, I would suggest that for sample sized $N \approx 40,000$, that $k = \{6, 7\}$ would be the best candidates for the values of k.

Lastly, for the final graph (f) with sample size $N \approx 50,000$, we can see that instead of the uniform and exponential agreeing as before, we now have the normal and exponential agreeing on its minimum value of bias at $k = 5$, and the uniform minimum occurring at $k = 7$. For the uniform distribution, the estimator with $k = 5$ still holds a small bias, but there are four other values of k with smaller bias for this distribution. For the normal distribution, there is a large jump at $k = 7$, where the bias is much larger than before, and the exponential also shows a substantial jump. For this sample size, we have $log(N) \approx 10.8$, so considering the enlarged regression graphs; Figures **??**, **??** and **??**, we can see that for each distribution the best value of k in reducing the bias is given by $k = 7$ then 5 for the normal distribution, $k = 7$ or 8 for the uniform and $k = 10$ for the exponential. Only the normal distribution aligns with above in that $k = 7$ is the optimum value, but the rest does no coincide. However, both $k = 5$ and $k = 7$ appear in both analysis, a number of times so would be inclined to possibly agree that $k = \{5, 7\}$ are the optimal choices for this sample size $N \approx 50,000$, but this result is not certain for all distributions.

By considering all the information found from the graphs in Figure 1.1, for small sample sizes $N \leq 1,000$ we can say that $k = 1$ would possibly be the best choice for k. However, I don't think anything conclusive can be drawn about a general distribution from a larger sample size N, despite some of the rough deductions above, there is no definitive choice for k being obviously shown throughout each distribution considered.

Perhaps an interesting graph to look at is that depicted in Figure 1.2, where I have taken all the data for every sample size $100 \leq N \leq 50,000$ and averaged the bias per distribution. I have done this to see if there is a safer value of k to choose when wanting the estimator of entropy from a random sample to have the smallest bias.

From this, we can see that for a sample distributed like the normal, the safest value of k to choose for any sample size is $k = 1$, since on average, this gives the smallest bias of ≈ 0.000125, regardless of its sample size. Then for all other values of k, the bias of the estimator increases with k; hence, if in doubt, I would recommend choosing $k = 1$ if you suspect your 1-dimensional data to be normally distributed.

Considering the uniform distribution, we can see that the smallest average bias occurs at $k = 2$, with the same behaviour as the normal for $k \geq 3$, whereby

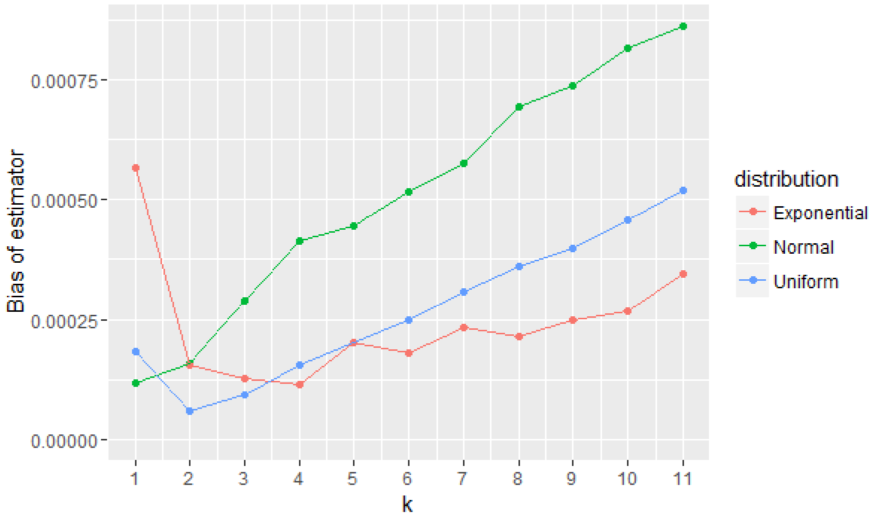


Figure 1.2: *The average bias over all sample sizes N for each k and each distribution.*

the bias increases with k. The average bias at $k = 2$ is ≈ 0.0000625, which is about half the size of that for the normal distribution. However, even though the relative size of the bias within a distribution is important to look at, the size between distributions is not as informative, due to the variability of the samples. Where for the normal distribution we have variance $\sigma^2 = 1$ and for the uniform we have variance $\frac{12}{100^2} = 0.0012$, which is significantly smaller - so we would expect a more accurate estimator due to this. Thus, when uncertain of which k to choose, if the 1-dimensional data appears to be uniformly distributed, then it seems to be that $k = 2$ would be the best choice.

Lastly, looking at the most interesting result shown, those of the exponential distribution, which shows that, on average, the smallest bias occurs when $k = 4$, with the bias generally increasing on both sides as k increases/decreases away from 4. The bias here is ≈ 0.000125 - similar to that of the normal distribution, which is surprising owing to the fact that the samples from the exponential distribution were taken with a variability of $\frac{1}{0.5^2} = 4$, which is four times larger than the variance of the samples from the normal distribution. This is something very interesting about the exponential distribution, and I have noticed throughout my analysis that samples from this distribution do indeed act quite differently to those from the other two.

Another important aim of this paper was to consider whether the distributions imply the bias is of equations **??** or **??**; i.e, whether bias is of $O\left(\frac{1}{N^a}\right)$ or $O\left(\left(\frac{k}{N}\right)^a\right)$. From the previous section, namely Figures **??**, **??** and **??**, implied the behaviours for the distributions considered shown in Table 1.2.

Table 1.2: *Suspected behaviour of Bias*$|\hat{H}_{N,k}|$ *for each distribution*

Normal, $N(0,1)$	Uniform, $U[0,100]$	Exponential, $exp(0.5)$
$O\left(\left(\frac{k}{N}\right)^a\right)$	$O\left(\left(\frac{k}{N}\right)^a\right)$	$O\left(\frac{1}{N^a}\right)$

This table shows that for samples from the normal and uniform distributions, both seemed to have the bias of the estimator showing the relation where c_k depends on k^a, since the graphs of this correspondence, appeared to show an almost exponential relationship between the two variables. However, for the exponential distribution a very different picture occurred; thus, implying for this distribution that there was no obvious dependence on k^a, and in fact c_k is just constant.

Overall, this paper has shown that analytically, the estimator is asymptotically unbiased and that we cannot know how to choose k for a sample by only considering its size N. We can only can find a range of k that satisfies the conditions from the theoretical works, but the analytical results show that there are many discrepancies as to which value of k is optimal, depending on how the sample is distributed. We have also not been able to draw any solid conclusions about the form of the bias of the estimator; however, the data here only showed implications of the connection between c_k and k^a, it has not proved or disproved it either way.

There are large amounts of further research to be done into the estimation of entropy, but time permits that this paper can look at nothing more. With more time, it would be interesting to see how the estimator works analytically on samples from different distributions; for example, Gamma, Chi-Squared or Beta distributions. Also, it would be intriguing to look analytically at the estimator for samples from higher dimensions, with known entropy.