DEPARTMENT OF MATHEMATICS
TECHNICAL REPORT

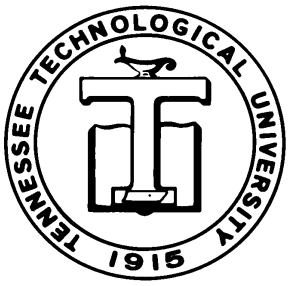
COMPARISON OF CENTRALITY ESTIMATORS FOR SEVERAL DISTRIBUTIONS

JAMIE McCREARY

Under the supervision of
DR. MICHAEL ALLEN

April 2001

No. 2001-3



TENNESSEE TECHNOLOGICAL UNIVERSITY
Cookeville, TN 38505

Comparison Of Centrality Estimators For Several Distributions

Jamie McCreary
Department of Mathematics
Tennessee Technological University
Cookeville, TN 38505

Michael Allen
Department of Mathematics
Tennessee Technological University
Cookeville, TN 38505

1. Introduction

The measure of central tendency is the most commonly used tool in statistical data analysis. The ability to determine an "average" provides a way to locate data centrality. Central tendency is usually determined by one of three methods. One can calculate the mean, median or midrange of a sample set. However, does the best method to determine the central point of a distribution vary with the types of distributions involved? In this paper we attempt to determine which methods are best used for several different distributions. Specifically we will examine the Normal, Uniform, and Cauchy distributions.

2. General method

We will examine the best estimator of centrality using non-rigorous methods. We will accomplish this by use of the Monte Carlo Method. The Monte Carlo procedure involves the generation of a set of pseudo-random numbers. These numbers, when generated according to a specified distribution's algorithm, form the data sets for our analysis. We will use the standard notation of B to indicate the number of Monte Carlo simulation sets. For our analysis we will be calculating the sample mean, sample median, and sample midrange for sets of different sizes. Briefly stated, the sample mean is the familiar arithmetic average, whereas the sample midrange is the arithmetic average of the highest and lowest values of a data set. The sample median is simply the middle number from a group of ordered data. The formulas for these estimators are

Sample mean: $\overline{x} = \frac{1}{n}\sum x_i$ where n is number of data in the set

Sample median: The value of the i^{th} position of the ordered data, where $i = \frac{n+1}{2}$

Sample midrange: $\frac{x_{max}+x_{min}}{2}$

As an example of how these three estimators are calculated, let S=$\{1,2,5,6,8,11,13,15,17,19\}$. The sum of all the numbers in this set is 97, so our mean is $\frac{97}{10}$ =9.7. To find the median in this case, since we have an even number of data points, we take the average of the two middle numbers, 8 and 11. Therefore, our median is 9.5. The midrange is the midpoint of the highest and lowest numbers, 1 and 19. So the midrange is 10.

To determine the best estimator we will use the least squares criteria, which gives us the estimator that produces the lowest variance. Variance is defined as:

$$Var(X) = E(X^2) - (E(X))^2$$

and the formula used to calculate the variance for a sample is given as:

$$\widehat{Var}(X) = \frac{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2}{n-1}$$

Using this formula the variance of the above example set is 39.34.

The criteria we are using to determine the best estimator, namely the minimum variance or least squares criteria, is related to a value known as the Fisher Information, or I($\widehat{\theta}$). This value is, in very basic terms, a score of how easy it is to distinguish between the true parameter θ and any close estimates $\widehat{\theta}$. The higher I($\widehat{\theta}$), the better your estimate. From Leyman (1983), we know that if p_θ is of the exponential family, then I($\widehat{\theta}$)=$\frac{1}{var(\widehat{\theta})}$. Also, from Leyman (1983), we know that if p_θ is of the exponential family and the parameter of interest is mu, the center of the data, then I($\widehat{\theta}$) is at a maximum if $\widehat{\theta}$ is the sample mean. So, we assumed from this theorem that the sample mean would be the best estimator where it could be calculated.

3. The analysis results

For our analysis we generated 1000 of these simulation sets for each of our selected sample sizes. For example, assume we wanted to generate 10 data points from a normal distribution. We would need to specify the central point, μ, and standard deviation, σ, of our population. The program would then execute the algorithm that selected 10 points from this population. We could then calculate the mean, median and midrange of these 10 numbers. The computer would then repeat this process 999 more times. This would then gives us three Monte Carlo samples of size 1000, one for the sample

mean for a sample of size 10, one for the sample median for a sample of size 10 and one for the sample midrange for a sample of size 10.

The Central Limit Theorem tells us that a group of sample means will be approximately normal, with the same central point, μ, as the original population. Therefore, we would expect the sample means of these 1000 sets to be centered on μ. Likewise, because of symmetry in the original population, the sample medians and sample midranges also center themselves around μ. We measure how these measures of central tendency spread out by calculating the variance of each of the Monte Carlo sets. According to our criteria, the one with the lowest variance is the best estimator.

3.1 Summary of the results for the normal distribution

Figure 3.3.1 shows a theoreical distribution of the normal. This distribution is shown with $\mu = 0$ and $\sigma = 1$ which is referred to as the Standard Normal. The mean of the distribution is associated with the peak of the bell curve. Thus, one can see if the mean were changed, this would cause the entire curve to shift to the left or right. Standard deviation affects the width of the bell shape. A change in the standard deviation would cause the curve to either flatten out or become steeper or narrower.

We let the normal sets be simulated as shown in Table 3.1.1.

Table 3.1.1

	Sample size	B	μ	σ
N_1	10	1000	100	5
N_2	20	1000	100	5
N_3	100	1000	100	5
N_4	1000	1000	100	5

In Table 3.1.2 we have the variances for our estimators for the four sets.

Table 3.1.2

	Var(Mean)	Var(Median)	Var(Midrange)
N_1	2.65302	3.540903	4.493488
N_2	1.226015	1.923207	3.632372
N_3	0.2727760	0.4300725	2.355683
N_4	0.02276362	0.03668264	1.503110

As you can see, the mean has the lowest variance in every case. Thus, the results from this test of the normal distribution confirmed our expectations

that the sample mean would be the best estimator. To emphasize this result, Figures 3.1.2, 3.1.3, 3.1.4 and 3.1.5 show the sampling distributions of the three estimators for the four different sample sizes. Again, one can readily see that the sample mean has the smallest variance in each case. In Table 3.1.3 we have the calculated efficiencies (i.e., the ratio of the variances) for the three estimators:

Table 3.1.3

	Mean vs. Median	Mean vs. Midrange	Median vs. Midrange
N_1	.749	.590	.788
N_2	.637	.338	.529
N_3	.634	.116	.183
N_4	.621	.015	.024

3.2 Summary of the results for the Cauchy distribution

For the Cauchy distribution we had different expectations. As can be seen from the graph of the theoretical Cauchy distribution in Figure 3.2.1, this distribution is quite different from the normal. Basically, the Cauchy distribution has what is called heavy tails. In other words, with the Cauchy distribution, there is always a chance of having a very extreme value, either positive or negative. The position and shape of the Cauchy is described by two parameters much in the same way the normal is described. However, in the case of the Cauchy, the parameters are location and scale not mean and standard deviation. The effect of changing these parameters is explained fairly well by their names. With regards to our expectations in this part of the study, the Cauchy distribution violates an assumption of the Central Limit Theorem that requires a finite variance. The heavy-tailed Cauchy has an infinite variance and mean and therefore we should not be able to calculate a meaningful sample mean for our sets. For this distribution we expect the sample mean and likewise the sample midrange to perform poorly (We say this as well for the midrange because it is in its simple form a sample mean.). Hence, that leaves the median, which exists for all distributions, to be the best estimator of centrality for the Cauchy (See Appendix for details). Table 3.2.1 shows the simulations for the Cauchy distribution sets.

Table 3.2.1

	Sample size	B	Location	Scale
C_1	10	1000	5	25
C_2	20	1000	5	25
C_3	100	1000	5	25
C_4	1000	1000	5	25

Looking at the results in Table 3.2.2, we see that our expectations were correct, and that the mean is not the best estimator of centrality for the Cauchy distribution based on our criteria of minimum variance. In fact, we see that the median is the only useful method of determinig centrality in a Cauchy distribution of these three. Again, as a quick summary, Figure 3.2.2 shows the resulting sampling distributions for the three estimators for a sample size of 10. Notice in Figure 3.2.2 that the median is centered at about 5, which is where we centered the original distribution for our simulations.

Table 3.2.2

	Var(Mean)	Var(Median)	Var(Midrange)
C_1	149534.6	218.4611	3644467
C_2	223825.1	93.08229	21756615
C_3	15781623	17.02836	39415852614
C_4	12304283	1.687272	3.078242×10^{12}

3.3 Summary of the results for the uniform distribution

Figure 3.3.1 shows a theoretical uniform distribution. The uniform is best described simply by its endpoints. We assumed the mean would be the best estimator in this case because it could be calculated. We simulated the uniform distribution sets as shown in Table 3.3.1.

Table 3.3.1

	Sample size	B	Minimum = a	Maximum = b
U_1	10	1000	-10	10
U_2	20	1000	-10	10
U_3	100	1000	-10	10
U_4	1000	1000	-10	10

However, as we see in Table 3.3.2 our results showed a departure from our expectations. The midrange has the lowest variance for each sample size.

Table 3.3.2

	Var(Mean)	Var(Median)	Var(Midrange)
U_1	3.36165	7.43223	1.532403
U_2	1.556647	4.163423	0.3856667
U_3	0.3024832	0.8995905	0.01635394
U_4	0.03338836	0.1035157	0.0002013420

Table 3.3.3 shows the calculated efficiencies.

Table 3.3.3

	Mean vs. Median	Mean vs. Midrange	Median vs. Midrange
U_1	.452	2.193	4.854
U_2	.374	4.032	10.753
U_3	.336	18.518	55.556
U_4	.323	166.667	500

Also, Figures 3.3.2, 3.3.3, 3.3.4 and 3.3.5 reiterate this unexpected result that the midrange appears to be the best estimator for the uniform distribution.

3.3.1 A closer look at the uniform distribution

Certain distributions belong to a family of distributions known as the exponential family. This name is refers to the probability density functions (pdf's) that are used to define distributions. Examining the different types of distributions in the exponential family is beyond the scope of this article; it will be sufficient to examine the properties and characteristics of a distribution which make it a part of the exponential family. Distributions in the exponential family can be defined as those having pdf's of the form

$$p_\theta(x) = exp[\sum_{i=1}^{s} n_i(\theta)T_i(x) - B(\theta)]h(x)$$

where, most importantly, $h(x)$ is a function of only x and defines the sample space for the distribution. As detailed by Lehmann (1983) this one assumption of the exponential family is:

> The distributions P_θ have common support, so that without loss of generality, the set $A = \{x \mid p_\theta > 0\}$ is independent of θ.

In other words, the sample space of $p_\theta(x)$ must be independent of θ in the exponential family.

As an example of how to determine if a distribution is part of the exponential family, consider the Normal distribution. Its pdf is given by

$$p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \text{ for } -\infty < x < \infty$$

where $\theta = (\mu, \sigma)$. Note that $p_\theta(x)$ can be rewritten as

$$p_\theta(x) = e^{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \mu^2 - 2\ln(2\pi\sigma^2)}$$

where, using the definition written above for a distribution in the exponential family, $s = 2$, $n_1(\theta) = -\frac{1}{2\sigma^2}$, $T_1(x) = x^2$, $n_2(\theta) = \frac{\mu}{\sigma^2}$, $T_2(x) = x$, $B(\theta) = \mu^2 + 2\ln(2\pi\sigma^2)$ and $h(x) = 1$. Hence, the normal is part of the exponential family because we can write its pdf in the required form and its support is independent of any of the parameters, i.e., $h(x) = 1$.

Upon examination, we realized that the uniform distribution violates this basic assumption of the exponential family. The pdf of the uniform is:

$$f(x) = \begin{cases} \frac{1}{(b-a)},\ a \le x \le b \text{ where } a = \xi - \frac{\lambda}{2},\ b = \xi + \frac{\lambda}{2} \\ 0 \text{ elsewhere} \end{cases}$$

from Hogg and Craig (1995) where ξ is the center of the distribution and λ is simply the distance between the endpoints of the distribution. In exponential form, the pdf is given as

$$p_\theta(x) = e^{-\ln(b-a)} I_{[a,b]}(x)$$

where, from the definition given above, $\theta = \xi$, $s = 1$, $n_1(\theta) = -\ln(b-a)$, $T_1(x) = 1$, $B(\theta) = 0$ and $h(x) = I_{[a,b]}(x)$ and $I_{[a,b]}(x) = 1$ if $a \le x \le b$, 0 otherwise. Note that, though, $h(x)$ is really $h_\theta(x)$ and is based on θ because a and b are both based on θ.

Therefore, it is evident that the uniform distribution's support (i.e., the sample space of x) is dependent upon θ and thus violates the assumption that the sample space be independent of θ. So, how does this affect our study? Previously it was stated that if p_θ is of the exponential family and the parameter of interest is mu, the center of the data, then $\mathrm{I}(\widehat{\theta})$ is at a maximum if $\widehat{\theta}$ is the sample mean. So, we see that the uniform does not

meet the criteria to be in the exponential family and therefore we should not use the result given above.

Next we examine why the midrange has the lowest variance. For this purpose we obtain the variance of the midrange of a uniform distribution by using order statistics. Hence, using the previous notation for the uniform distribution, the result is

$$Var(\text{midrange}) = \frac{b-a}{2(n+1)(n+2)}$$

from David (1970). Note this variance can be rewritten as

$$Var(\text{midrange}) = \frac{\left(\xi + \frac{\lambda}{2}\right) - \left(\xi - \frac{\lambda}{2}\right)}{2(n+1)(n+2)} = \frac{\lambda}{2(n+1)(n+2)}$$

Obviously, this gives us an n^2 term in the denominator, whereas the variance of the sample mean only has an n term in the denominator. To show this fact, note that

$$Var(\text{mean}) = \frac{\sigma^2}{n}$$

from Hogg and Craig(1995). To compare the two variances, we need to put them into similar terms. Hence, it is known that the variance of the uniform distribution is given as

$$Var(x) = \frac{(b-a)^2}{12}$$

Without loss of generality let the above variance be simply σ^2. Now we find $\lambda = b - a$ in terms of σ^2. Thus, with respect to the variance for the midrange,

$$\sigma^2 = \frac{\lambda^2}{12} \Leftrightarrow \lambda = \sqrt{12\sigma^2}$$

and hence

$$Var(\text{midrange}) = \frac{\lambda}{2(n+1)(n+2)} = \frac{\sqrt{12\sigma^2}}{2(n+1)(n+2)} = \frac{\sqrt{3}\sigma}{(n+1)(n+2)}$$

Next, we consider the efficiency of the two estimators. Thus,

$$\frac{Var(\text{midrange})}{Var(\text{mean})} = \frac{\frac{\sqrt{3}\sigma}{(n+1)(n+2)}}{\frac{\sigma^2}{n}} = \frac{n\sqrt{3}}{\sigma(n+1)(n+2)} < \frac{\sqrt{3}}{\sigma(n+2)} < 1 \text{ for } n > \frac{\sqrt{3}}{\sigma} - 2$$

since σ is a fixed value. Therefore, we see that the sample midrange beats the sample mean as an estimator for the mean of the uniform distribution with respect to the least squares criteria as soon as $n > \frac{\sqrt{3}}{\sigma} - 2$.

In a similar manner the sample midrange can as well be shown to outperform the sample median. Hence, we can safely conclude that the median is the best estimator of centrality of the three when the data set comes from a uniform distribution with a relatively large n.

5. Summary

We see now that our assumption of the sample mean being the best estimator of centrality was incorrect. It is obvious that the distribution in question determines the best estimator and we must the fully check all the necessary conditions of a theorem before we apply its result.

6. Appendix

The probability density function of the Cauchy distribution is given by

$$f(x) = \frac{1}{\beta\pi(1 + (\frac{x-\alpha}{\beta})^2)} \text{ for } -\infty < x < \infty$$

where $-\infty < \alpha < \infty$ and $\beta > 0$. Next, the characteristic function of the Cauchy is given by

$$\varphi(t) = e^{it\alpha - \beta|t|}$$

from Hogg and Craig (1995).

By definition, the k^{th} moment of X can calculated by

$$E(X^k) = \frac{\varphi^{(k)}(0)}{i^k}$$

where $\varphi^{(k)}(0)$ is the k^{th} derivative of φ evaluated at 0 (from Chow and Teicher (1988)). For example, if $X \sim N(\mu, \sigma^2)$, then

$$\varphi(t) = e^{i\mu t - \frac{\sigma^2 t^2}{2}}$$

and then

$$E(X) = \frac{\varphi'(0)}{i} = \frac{i\mu e^{i\mu(0) - \frac{\sigma^2(0)^2}{2}}}{i} = \mu$$

Analogously, $E(X)$ for the Cauchy distribution is then

$$E(X) = \frac{\varphi'(0)}{i} = \frac{\frac{d}{dt}e^{it\alpha - \beta|t|}\Big|_{t=0}}{i} = \begin{cases} \lim_{t\to 0^+} \frac{i\alpha e^{it\alpha} - \beta e^{-\beta t}}{i} \\ \lim_{t\to 0^-} \frac{i\alpha e^{it\alpha} + \beta e^{-\beta t}}{i} \end{cases} = \begin{cases} \alpha - \frac{\beta}{i} \\ \alpha + \frac{\beta}{i} \end{cases}$$

Since the two limits are not equal, this implies that the mean (i.e. $E(X)$) of the Cauchy distribution does not exist. To further show this result, consider the direct definition of $E(X)$ which is

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

Hence, in the case of the Cauchy distribution, we have

$$E(X) = \int_{-\infty}^{\infty} \frac{1}{\beta\pi} \frac{x}{(1 + \left(\frac{x-\alpha}{\beta}\right)^2)} dx$$

Using the change of variable technique with $y = \frac{x-\alpha}{\beta}$ we have

$$E(X) = \int_{-\infty}^{\infty} \frac{1}{\pi} \frac{\alpha + \beta y}{(1+y^2)} dx = \int_{-\infty}^{\infty} \frac{1}{\pi} \frac{\alpha}{(1+y^2)} dx + \int_{-\infty}^{\infty} \frac{1}{\pi} \frac{\beta y}{(1+y^2)} dx$$

$$= \alpha + \frac{\beta}{2\pi} \ln(1+x^2)\Big|_{-\infty}^{\infty}$$

which does not exist. In a similar fashion, $E(X^k)$ can be shown not to exist for any $k \geq 1$.

Yet, a measure of the center of the Cauchy distribution can be found by looking at the population median. By definition, the population median is the value m such that $P(X \leq m) = P(X > m) = \frac{1}{2}$. Hence, we simply need to find the value of m that solves the following integral equation:

$$\int_{-\infty}^{m} \frac{1}{\beta\pi} \frac{1}{(1 + \left(\frac{x-\alpha}{\beta}\right)^2)} dx = \frac{1}{2}$$

Again, using the change of variable technique with $y = \frac{x-\alpha}{\beta}$ we have

$$\int_{-\infty}^{\frac{m-\alpha}{\beta}} \frac{1}{\pi(1+y^2)} dy = \frac{1}{2}$$

Integrating, we get

$$\frac{1}{\pi}\left[\tan^{-1}(x)\right]\Big|_{-\infty}^{\frac{m-\alpha}{\beta}} = \frac{1}{2}$$

$$\Rightarrow \frac{1}{\pi}\tan^{-1}\left(\frac{m-\alpha}{\beta}\right) - \lim_{x\to-\infty}\left(\frac{1}{\pi}\tan^{-1}(x)\right) = \frac{1}{2}$$

$$\Rightarrow \frac{1}{\pi}\tan^{-1}\left(\frac{m-\alpha}{\beta}\right) - \left(-\frac{1}{2}\right) = \frac{1}{2}$$

$$\frac{1}{\pi}\tan^{-1}\left(\frac{m-\alpha}{\beta}\right) = 0$$

$$\Rightarrow \frac{m-\alpha}{\beta} = 0$$

$$\Rightarrow m = \alpha$$

Hence, the population median of the standard Cauchy distribution is α. In our simulations we chose $\alpha = 5$ and $\beta = 25$ for variety.

7. References

Chow, Yuan Shin and Teicher, Henry (1988). Probability Theory. Springer-Verlag, New York.

David, H.A. (1970). Order Statistics. John Wiley & Sons, Inc., New York.

Hogg, R.V. and Craig, A.T. (1995). Introduction to Mathematical Statistics (5th Ed.). Prentice Hall, Upper Saddle River, NJ.

Lehmann, E.L. (1983). Theory of Point Estimation. John Wiley & Sons, Inc., New York.

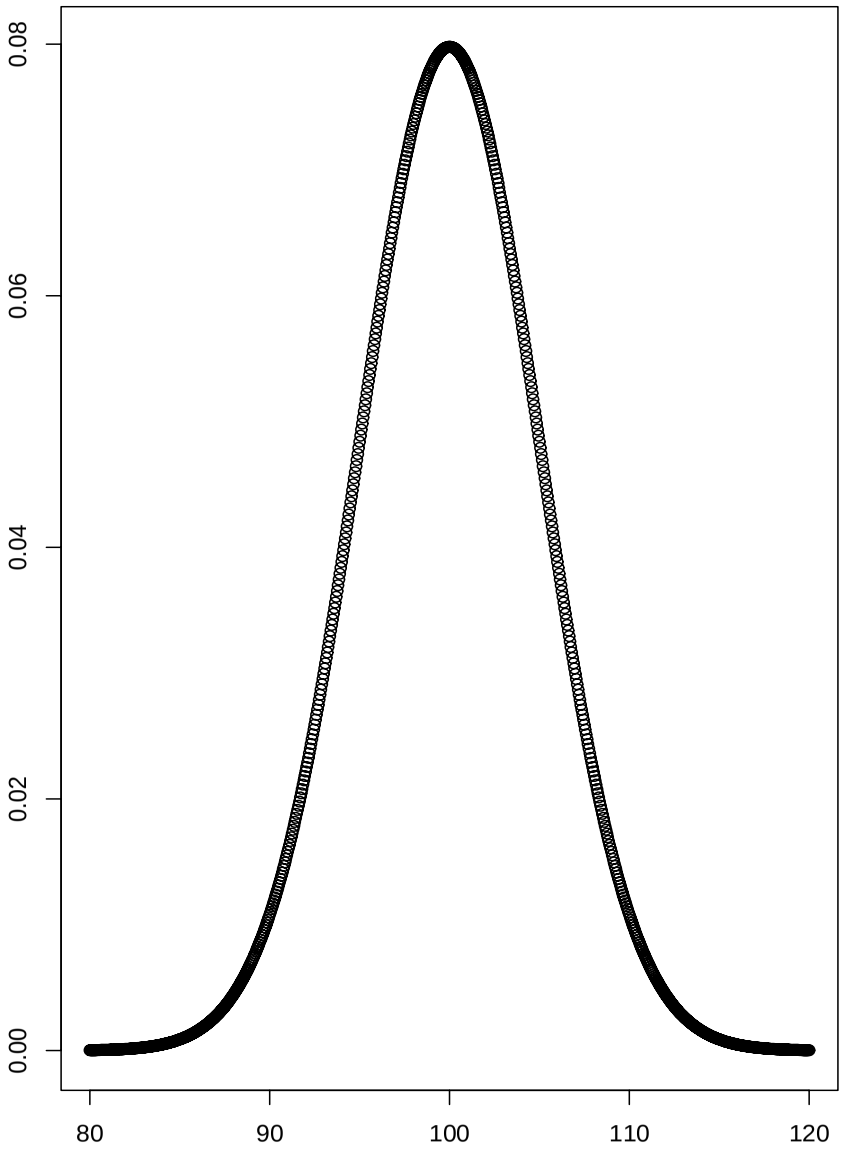


Figure 3.1.1: Theoretical Normal Distribution with $\mu = 100$ and $\sigma = 5$

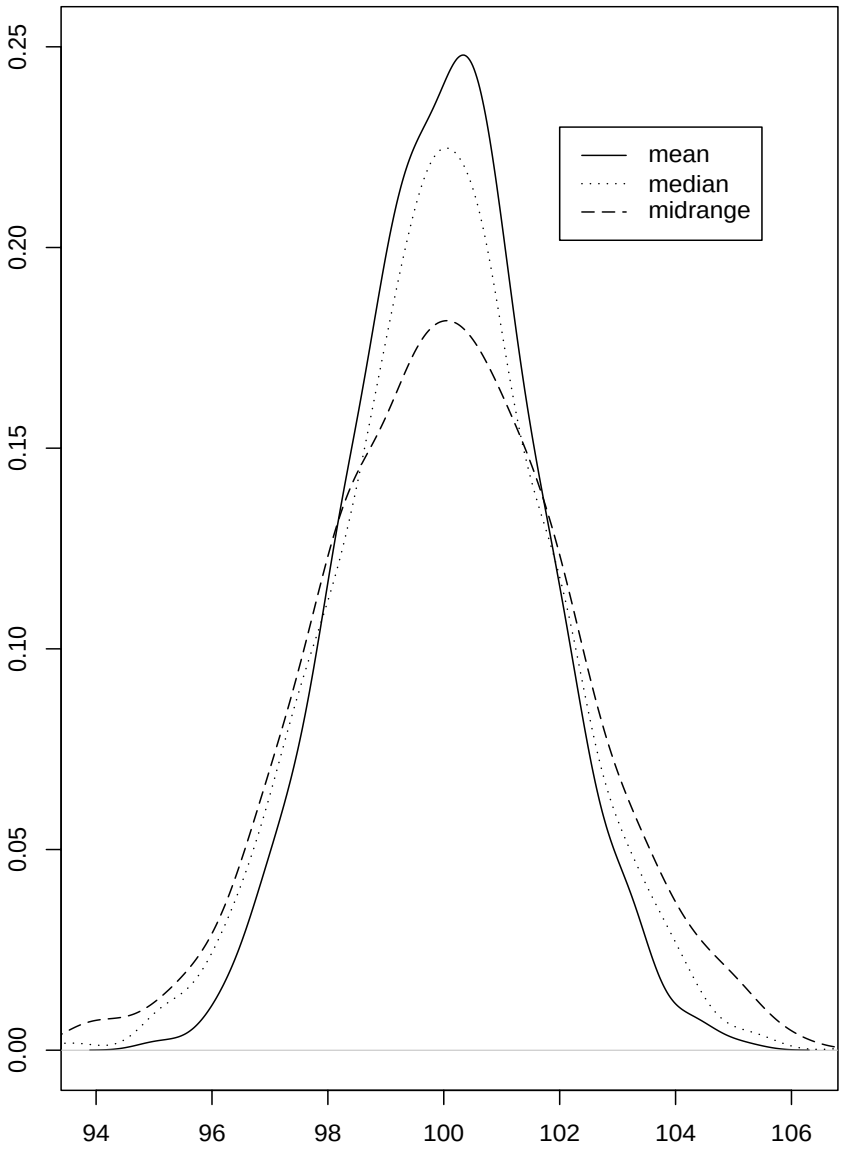


Figure 3.1.2: Normal Distribution 1

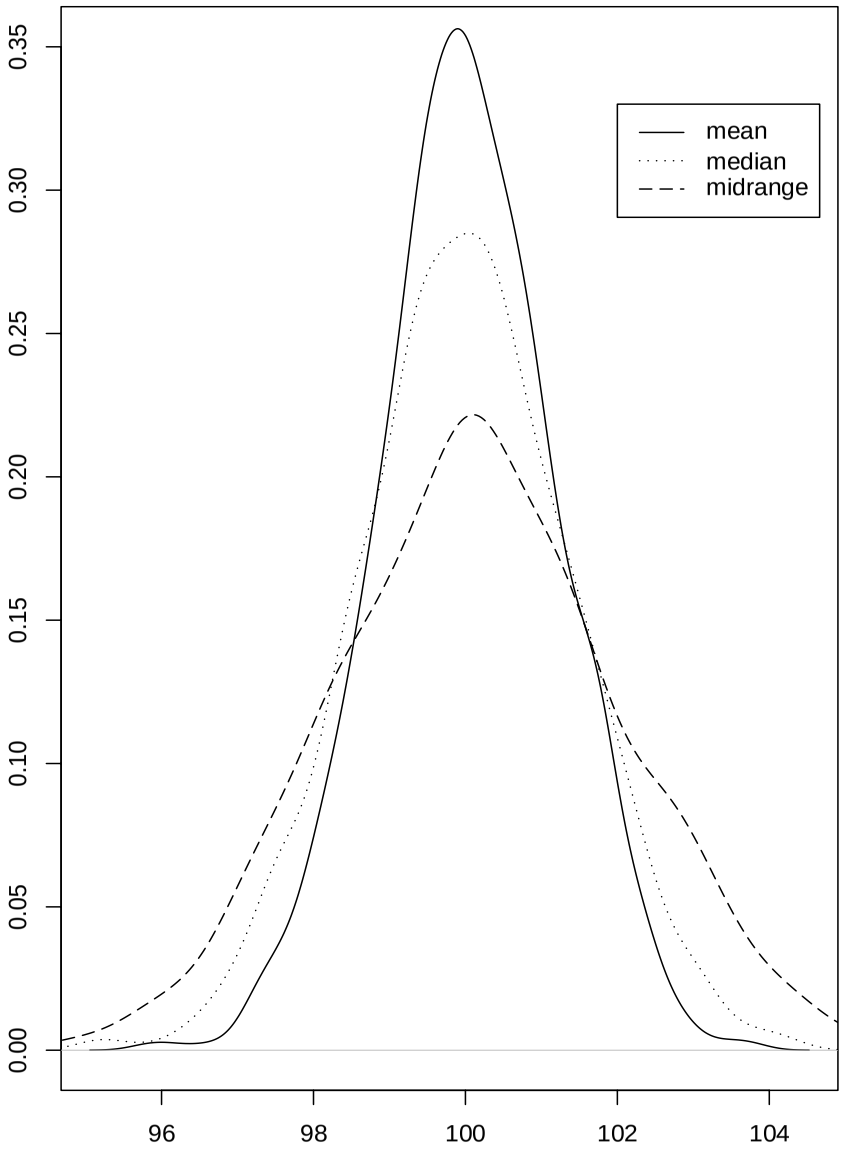


Figure 3.1.3: Normal Distribution 2

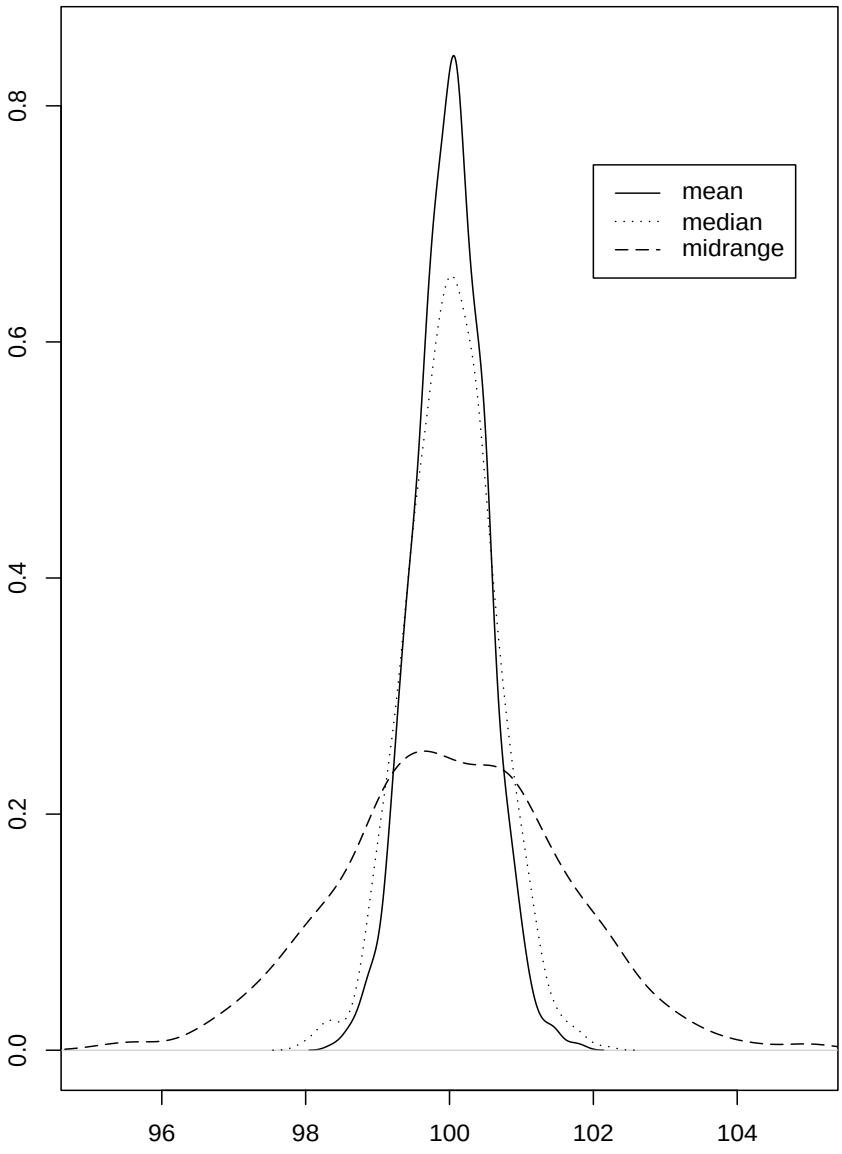


Figure 3.1.4: Normal Distribution 3

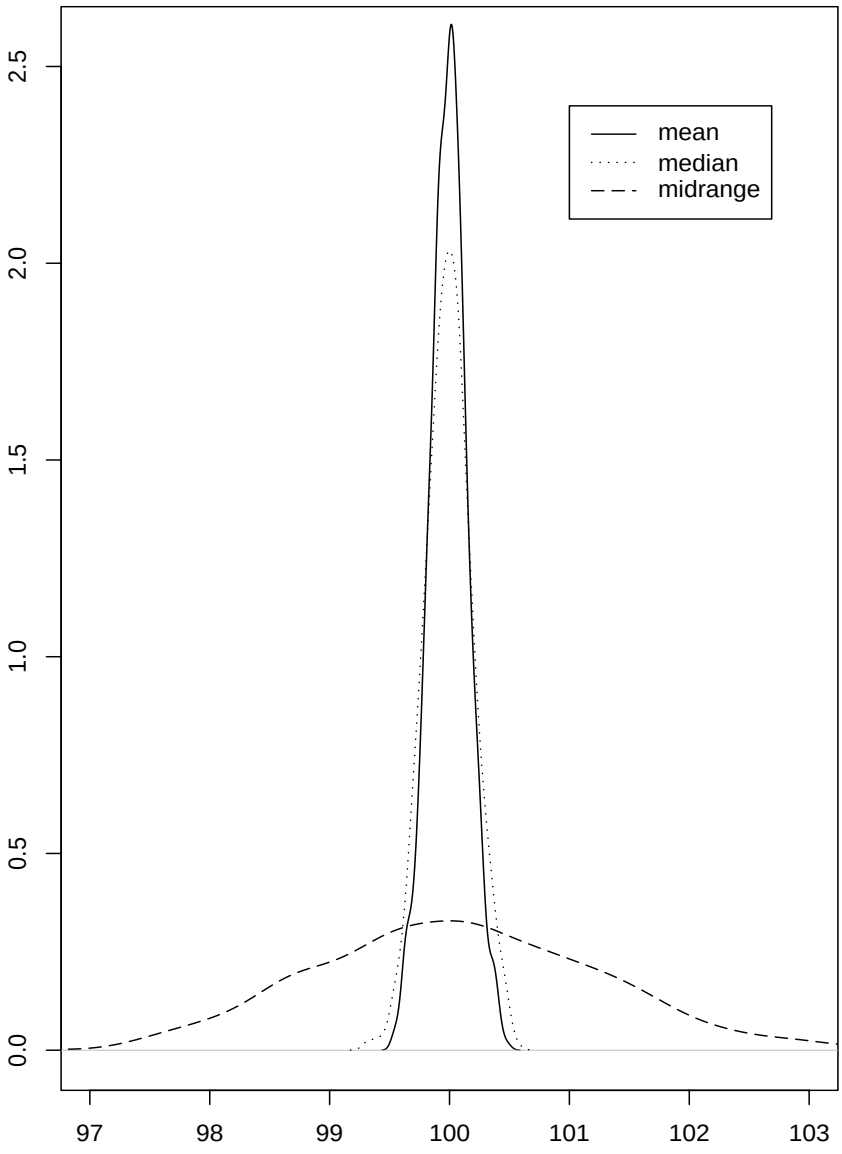


Figure 3.1.5: Normal Distribution 4

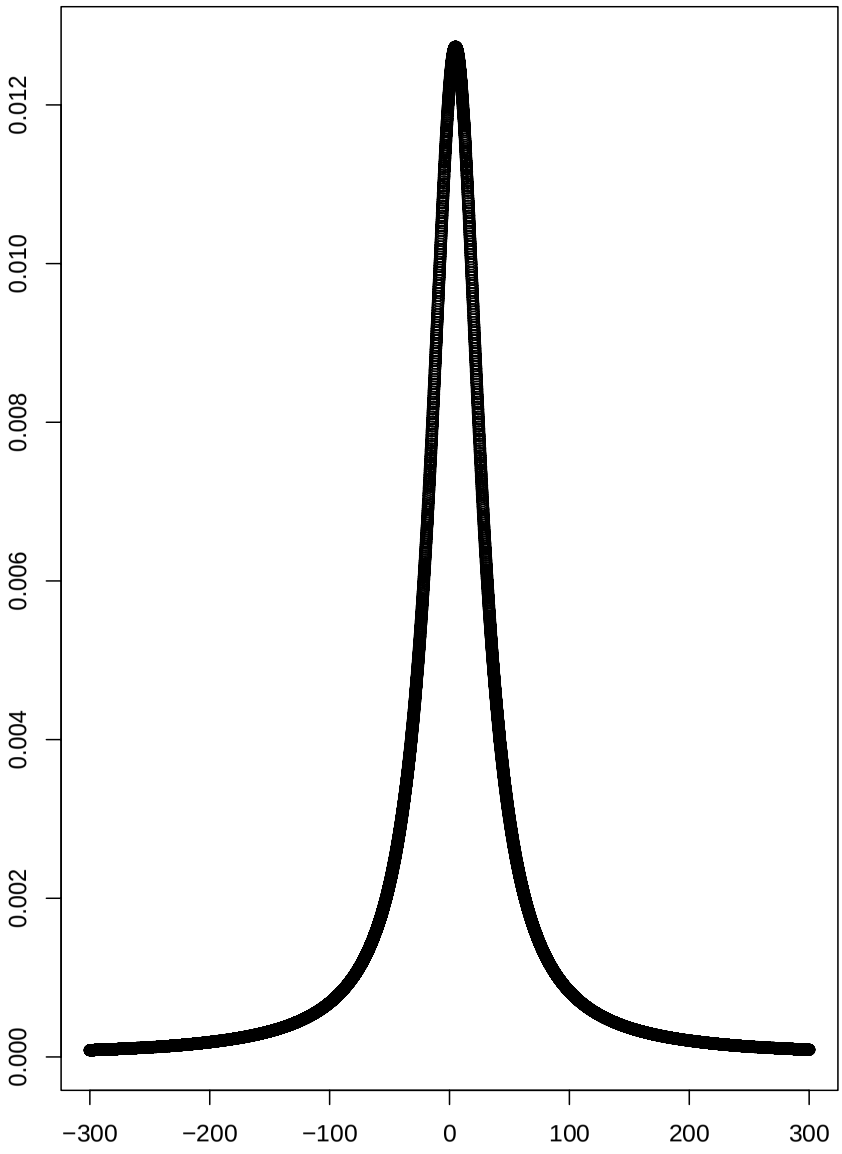


Figure 3.2.1: Theoretical Cauchy Distribution with $\alpha = 5$ and $\beta = 25$

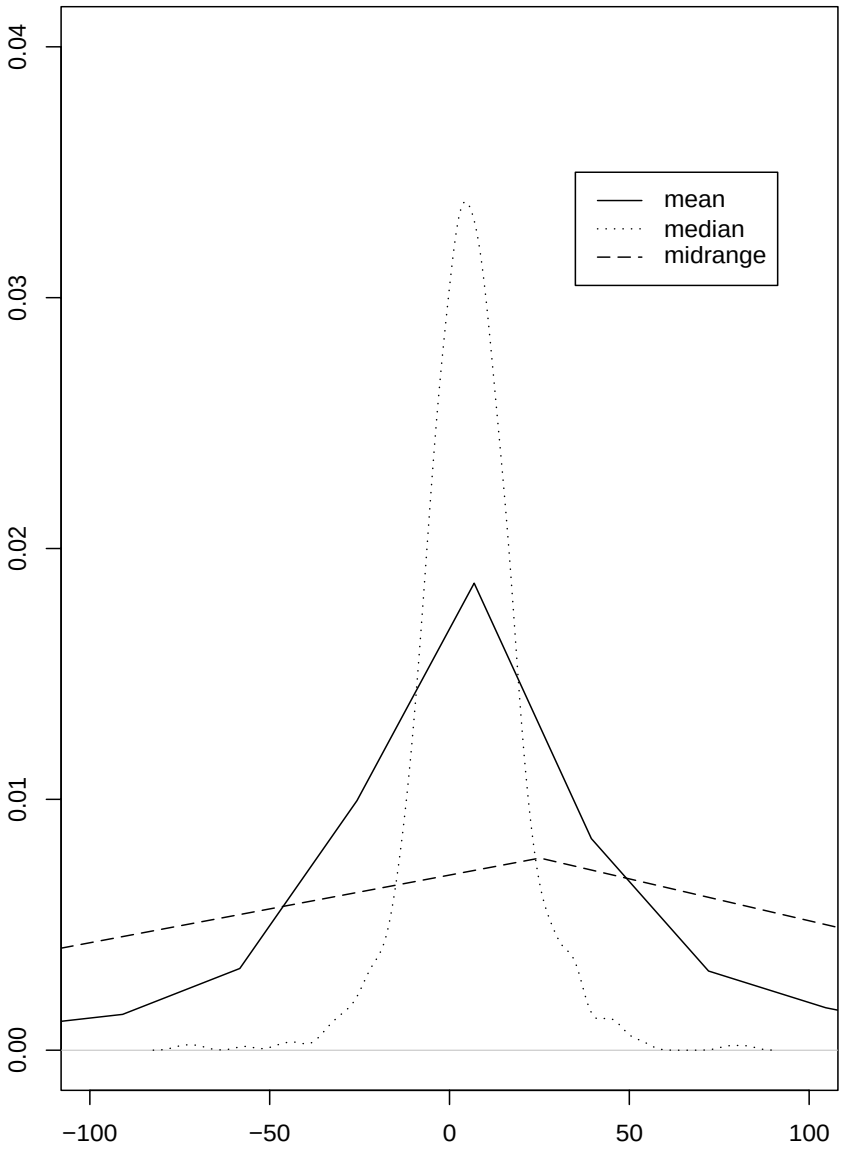


Figure 3.2.2: Cauchy Distribution 1

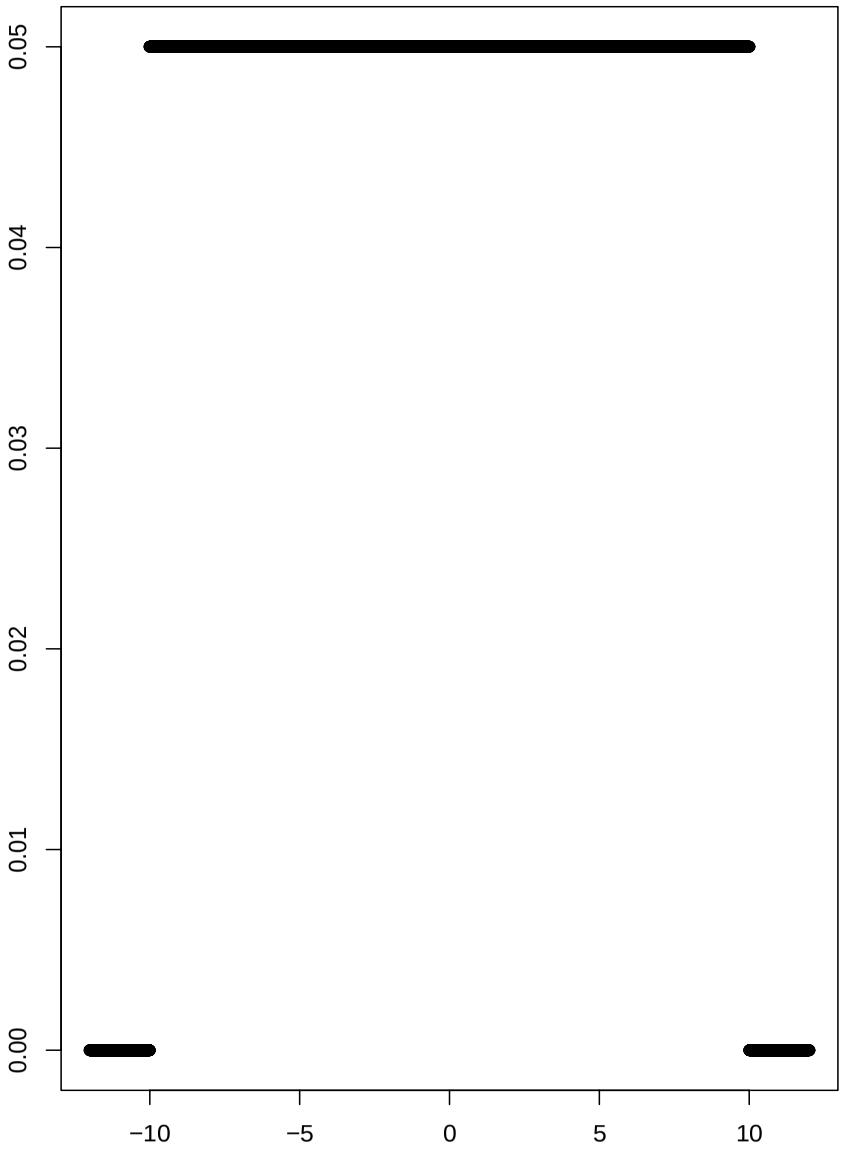


Figure 3.3.1: Theoretical Uniform Distribution with $a = -10$ and $b = 10$

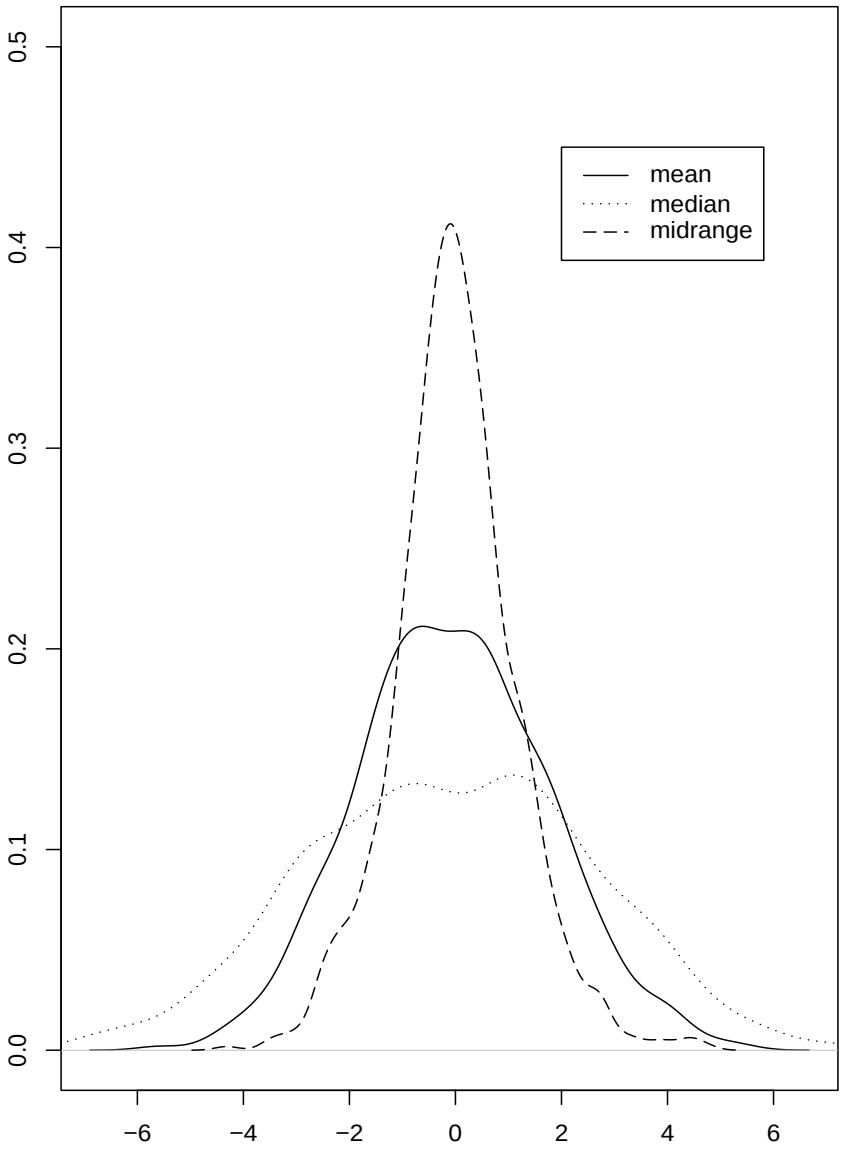


Figure 3.3.2: Uniform Distribution 1

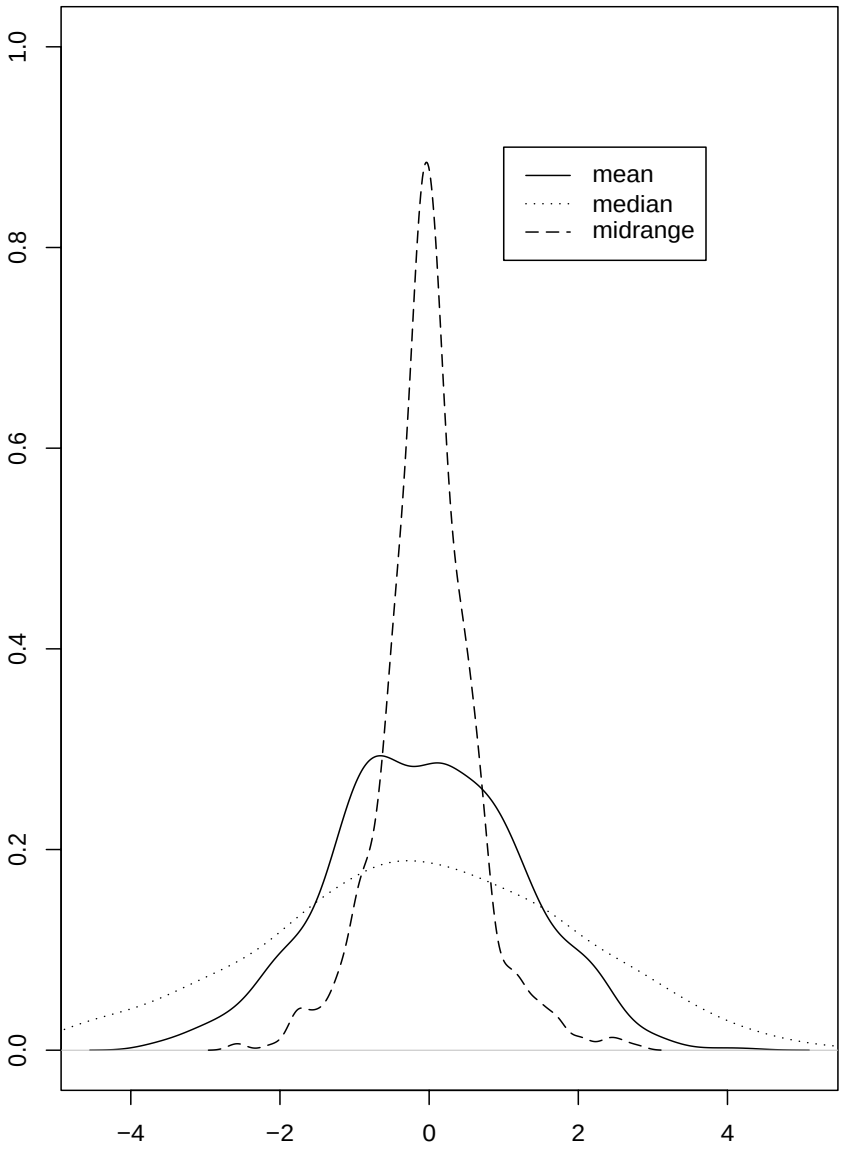


Figure 3.3.3: Uniform Distribution 2

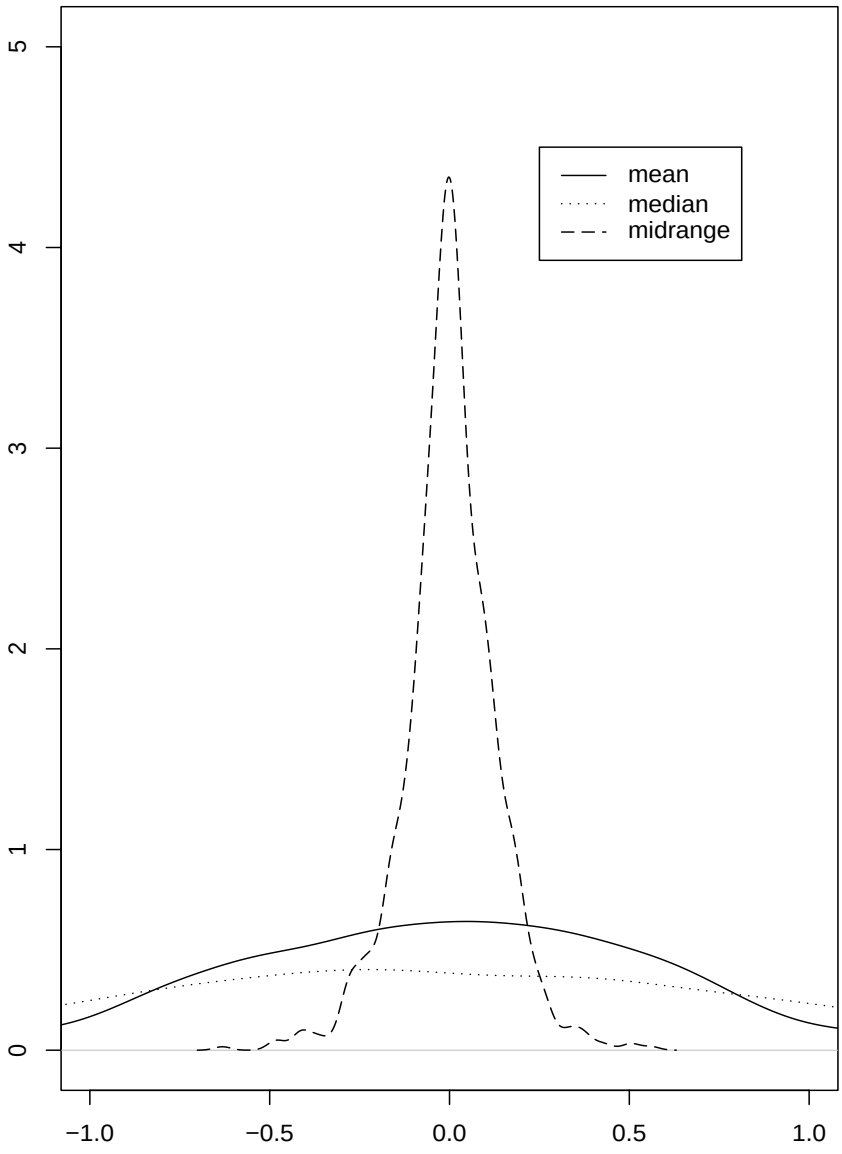


Figure 3.3.4: Uniform Distribution 3

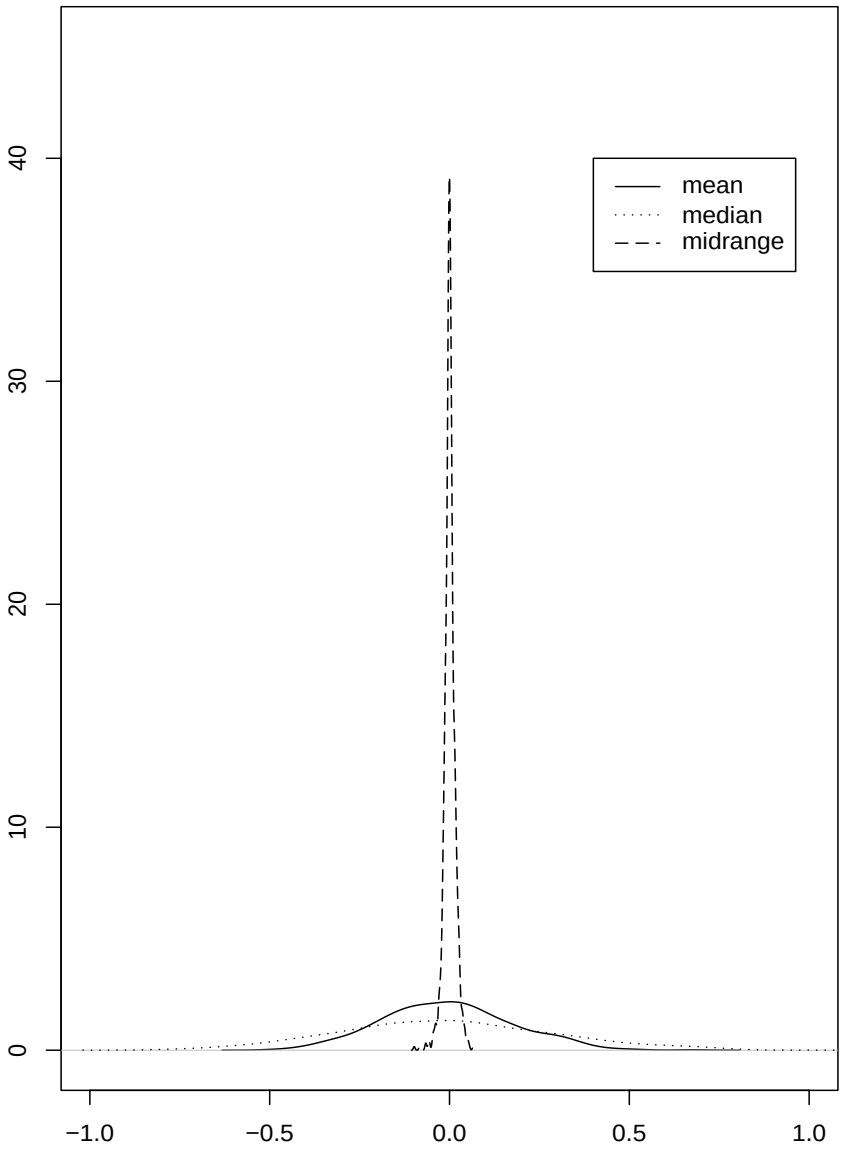


Figure 3.3.5: Uniform Distribution 4