TRIED RNRF TP1 Report

Carl Robinson

29th Sept 2017

1.1 - Throws of a die

A fair six-sided die was thrown n times. The frequency of the values that occurred are represented as percentages of the total number of throws.

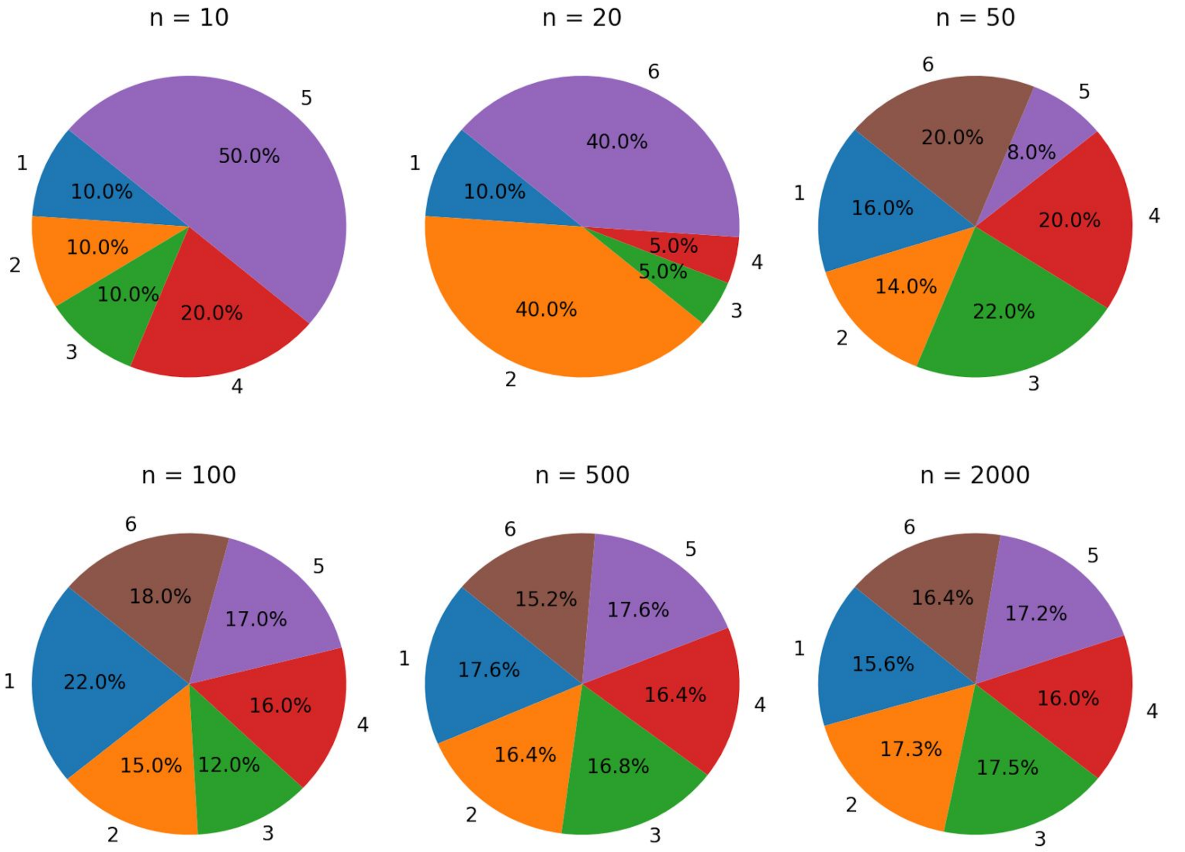

Figure: Pie charts illustrating distribution of values obtained by throwing a fair six-sided die n times

- When n is small, the variance is high. When n=10, some classes are represented frequently (e.g. 5 was drawn 50% of the time), and other classes are not represented at all (e.g. 6 was never drawn).
- Conversely, when n is large the variance is low, so the frequency of each class are very similar. When n=2000, all six classes are represented fairly equally, having frequencies in a narrow range of 15.6% - 17.5%. The class frequencies appear to all tend towards the same value, in this case 1/6 = 16.67%

Table: Distribution of values obtained by throwing a fair six-sided die n times

Number of throws, n	Minimum class percentage, %	Maximum class percentage, %	Difference between min and max class percentage, %	Mean of throw values	Standard deviation
10	10	40	30	2.8	1.720
20	10	30	20	3.95	1.658
50	14	22	8	3.18	1.740
100	12	23	11	3.47	1.615
500	15	18.4	3.4	3.56	1.726
2000	15.6	17.8	2.2	3.529	1.716

- From the table we can clearly see the minimum class percentage start very low at 10% when n=10, and increase to a value of 15.6% when n=2000.
- The same is true for the maximum class percentage, which starts very high at 40% when n=10, and steadily decreases to 17.8% when n=2000.
- In both cases, as n tends towards infinity, the class percentage values tend towards the equal distribution value of 16.67%. This can be seen in the difference between the minimum and maximum class percentages; there is a 30% difference when n=10, but only a 2.2% difference when n=2000.
- The mean throw value tends towards 3.5. This is the expected value for a six-sided die where the probability of each side occurring is equal (a uniform distribution). This is calculated as (a + b)/2, where a and b represent the minimum and maximum values on the sides of the die (1 and 6).
- The standard deviation fluctuates as n increases, but does not actually tend in any one direction. It is 1.720 when n=10, and 1.716 when n=2000. This is because the theoretical variance for a uniform distribution is calculated as (1/12)*(b - a)^2. The variables a and b represent the minimum and maximum values on the sides of the die (1 and 6). As they remain the same regardless of the number of throws, the variance is a constant. The standard deviation is the square root of the variance, so this remains constant too.

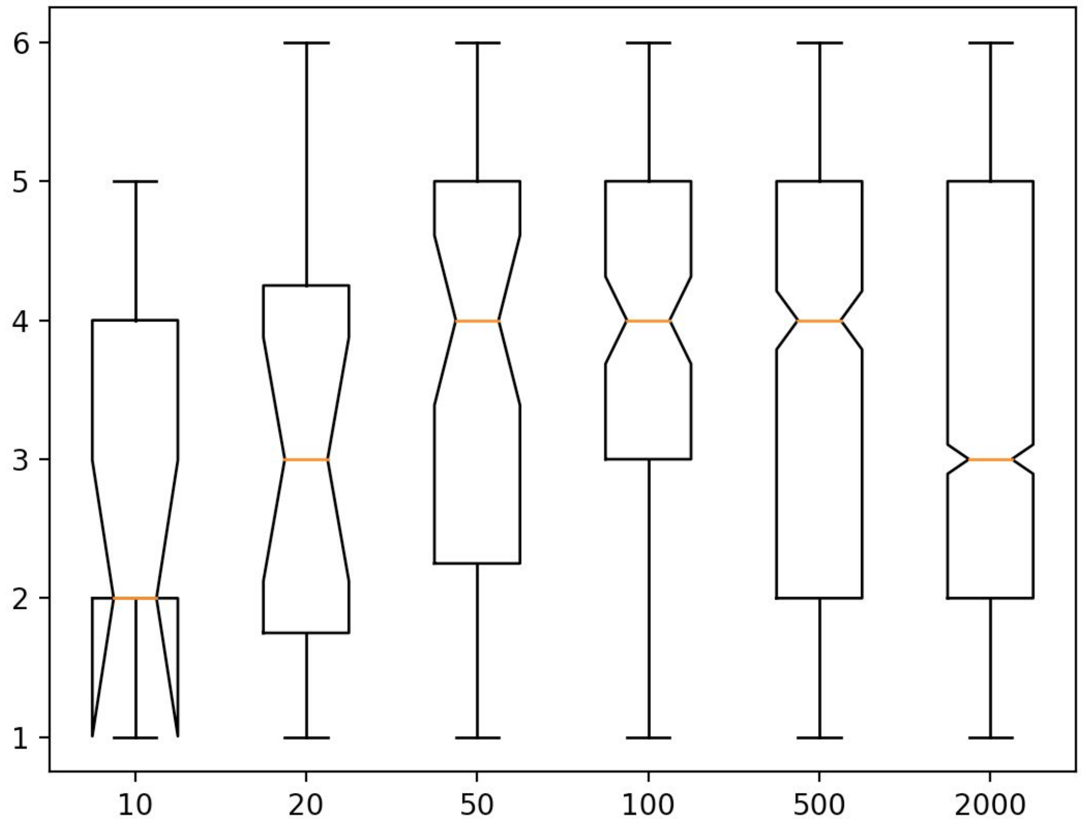

Figure: Boxplot of distribution of values obtained by throwing a fair six-sided die 10 to 2000 times

- The whiskers of the boxplot (moustache) extend to the most extreme data points not considered outliers. They show that the range of values thrown when n=10 was 1 to 5, whereas it was 1 to 6 when n >= 20.
- The orange line represents the median value, which as expected is either 3 or 4 in all sets of throws where n >= 50. The median is only 2 when n=10, as the low number of throws contains a high number of 1s and 2s.
- The bottom edge of the black box represents the lower hinge (25th percentile), defined as the point below which 25% of all values obtained are at or below. When n=500 and n=2000, this has the value of 2, meaning 25% of all throws were either a 2 or below.
- The top edge of the black box represents the upper hinge (75th percentile), defined as the point above which 25% of all values obtained are at or above. When n=500 and n=2000, this has the value of 5, meaning 25% of all throws were either a 2 or above.
- The notches that touch the orange line represent the 95% confidence interval (CI) around the median. The width of the notch is proportional to the interquartile range of the sample, and inversely proportional to the square root of the size of the sample. This boxplot shows that as n increases, the notch size decreases, indicating a narrower confidence interval, with n=2000 producing the narrowest CI around the value 3.

2.1 - Normalised histogram and density function

2.1.1 - Use of simulated unidimensional data

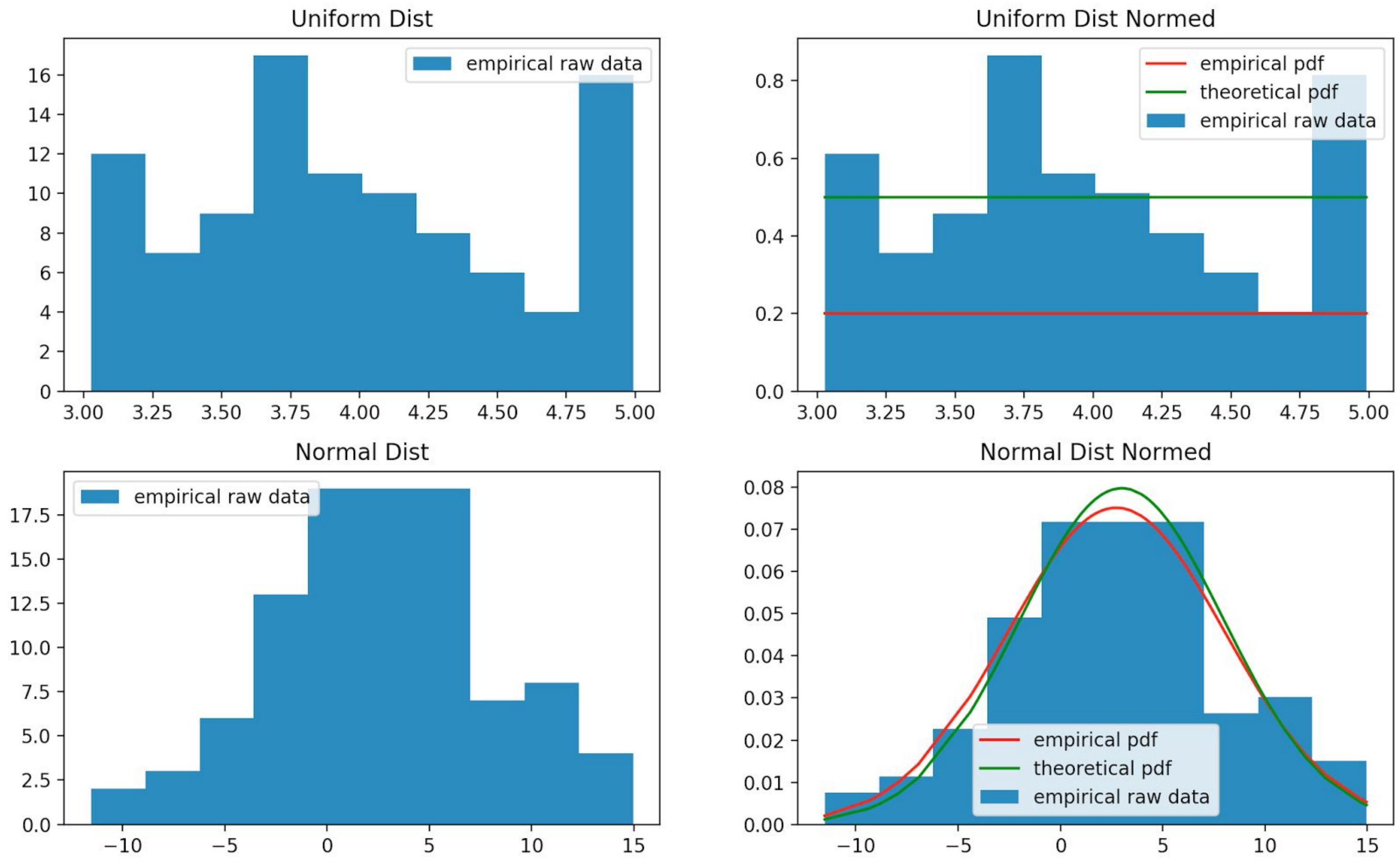

Figure: Raw and normalised histograms of uniform distribution of random numbers between 3 and 5, and normal distribution of random numbers with mean=3 and std=5. Probability density functions shown as curves.

- The normalised uniform distribution's theoretical pdf is calculated as 1 / (b - a). The resulting value is 0.5, which makes sense as the integral equals 1 when a=3.0 and b=5.0.
- The uniform distribution's empirical pdf is calculated as 0.2, by calling the following Python function: stats.uniform.pdf(unif_array, loc=min(unif_array), scale=max(unif_array)). My expectation was that this would also equal 0.5, given that the minimum and maximum values of the generated data (a and b) were close to the a and b values used to calculate the theoretical pdf. I cannot explain this result.
- The normalised normal distribution's theoretical pdf is calculated as 1 / (b - a). The resulting value is 0.5, which makes sense as the integral of the histogram equals 1.0 when a=3.0 and b=5.0.
- The normal distribution's empirical pdf closely matches the theoretical pdf, with a slightly larger variance when n=500. The empirical variance would tighten up around the expected value as n increases.

2.1.2 - Illustration of the Central Limit Theorem (CLT)

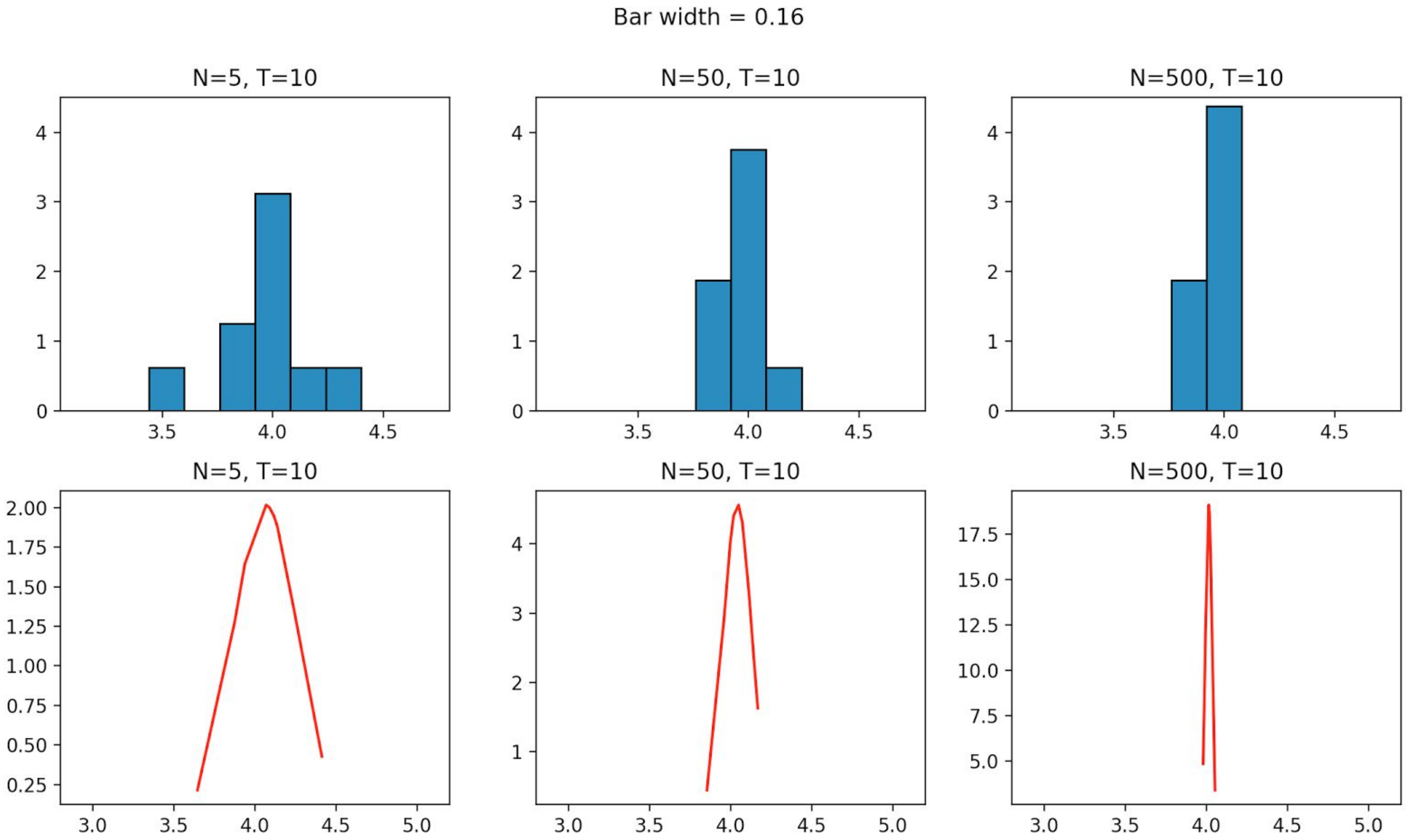

Figure: Normalised histograms of X_bar mean values of T=10 arrays of increasing N random uniformly distributed numbers between 3.0 and 5.0. Probability density function plots illustrate theoretical distribution of means.

- Due to the low number of T=10 arrays, the histograms have an uneven asymmetric appearance, with some bins not having values at all. For the same reason, the pdf plots are clearly segmented, ersatz representations of the normal distribution.
- Despite this, the histograms do show that as n in increases, the values of X_bar tend towards the expected value of 4.0, which is the expected value of its N uniformly distributed random numbers between 3.0 and 5.0.
- The pdf plots also clearly illustrate the trend that as N increases, the variance in the means decreases, as can be seen by a significant narrowing of the normalised curve's width, and a large increase in height (maintaining an surface area of 1 beneath the curve).
- On the pdf plots, the 'peak' of the normal distribution when N=5 is to the right of the theoretical mean of 4.0, but as N increases to 500 this peak is squarely centered over 4.0, more accurately representing the normal distribution.

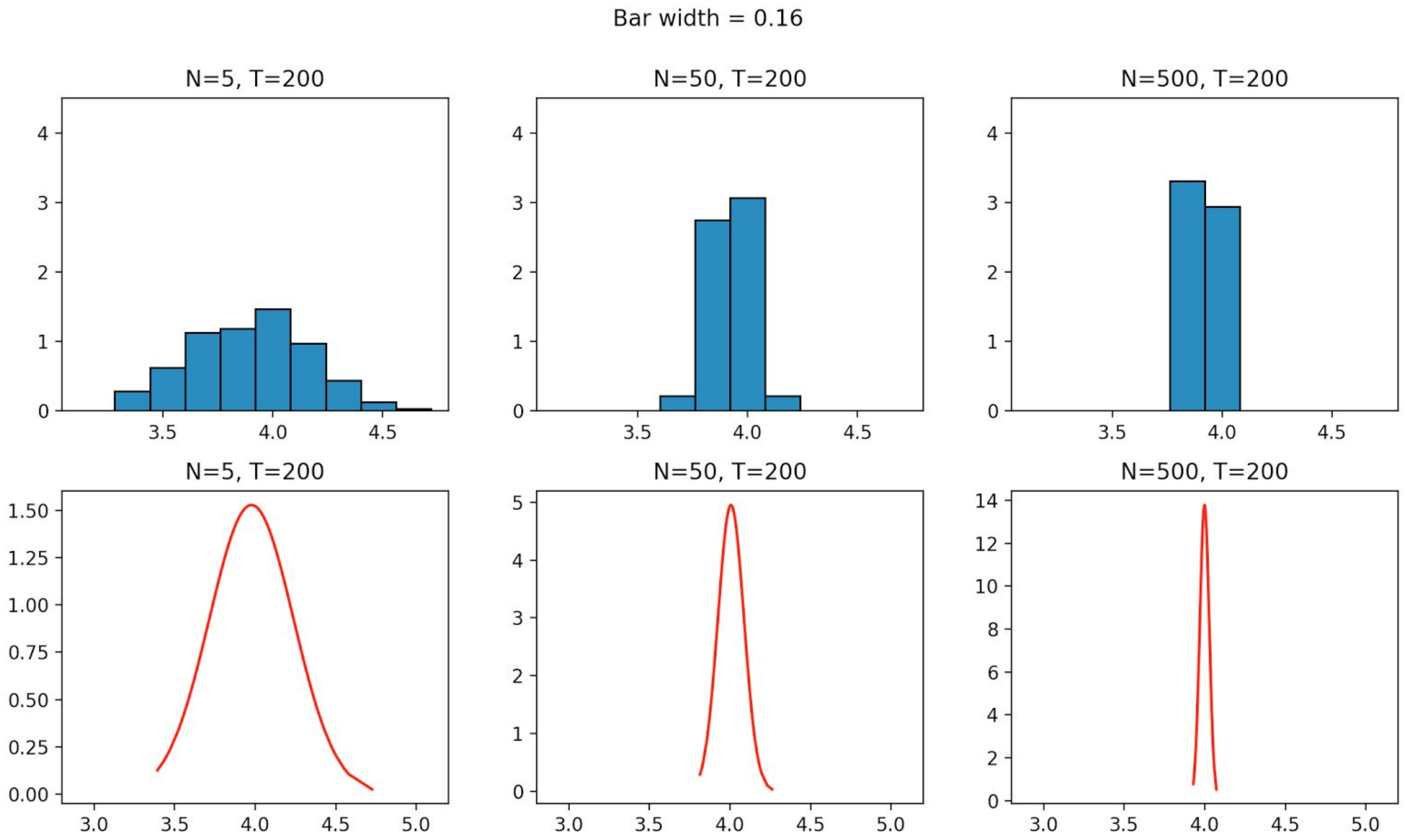

Figure: Normalised histograms of X_bar mean values of T=200 arrays of increasing N random uniformly distributed numbers between 3.0 and 5.0. Probability density function plots illustrate theoretical distribution of means.

- As with the previous figure, the heights of the histograms points are normalised by the value [bin_width * T], so that the surface area of each graph always sums 1. This enables comparison between the graphs.
- With T=200 there are many more elements in each bin, so the histograms more accurately represent the normal distribution that the mean values are tending towards.
- For the same reason, the pdf plots also resemble much smoother Gaussian curves when T=200 than when T=10. This is due to the Central Limit Theorem, which states that as the number of T means of uniformly distributed numbers increases, the more the means tend towards the normal distribution.
- As before, an increasing N value causes the variance to decrease, as the means of each of the T arrays centre more closely around the expected value of 4.0.

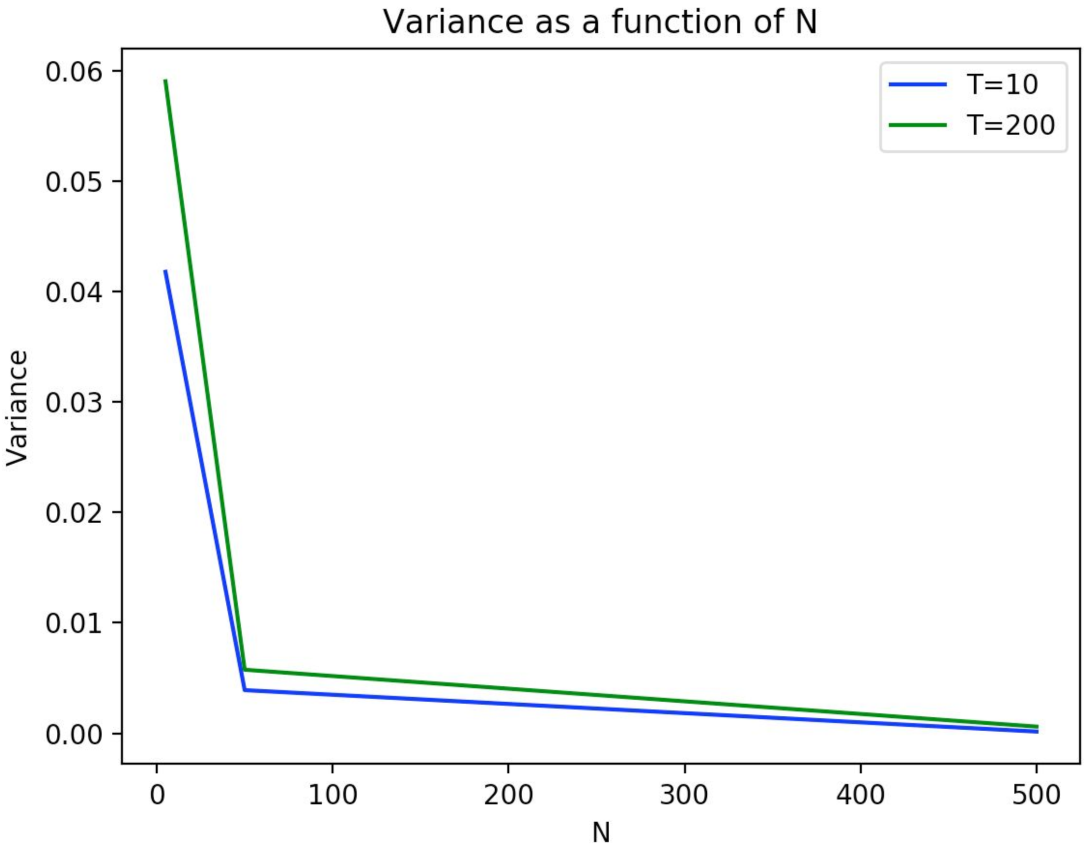

Figure: Variance of T=10 and T=200 means as N increases

- Plotting the variance against N shows how the variance converges towards zero, regardless of the size of T.
- As with the pdf plots, an increasing N value causes the variance to decrease, as the means of each of the T arrays centre more closely around their expected value of 4.0. This is an example of the law of large numbers.
- It can be noted that the variance when T=200 is slightly higher than when T=10, especially when N is low. The increased number of means in the set make it more likely for mean values at the low/high extremes of the normal distribution to appear, which cause the variance to increase. However, the effect of T on variance is negligible compared with the effect N has on it.

2.1.3 - Convergence of empirical density towards theoretical density

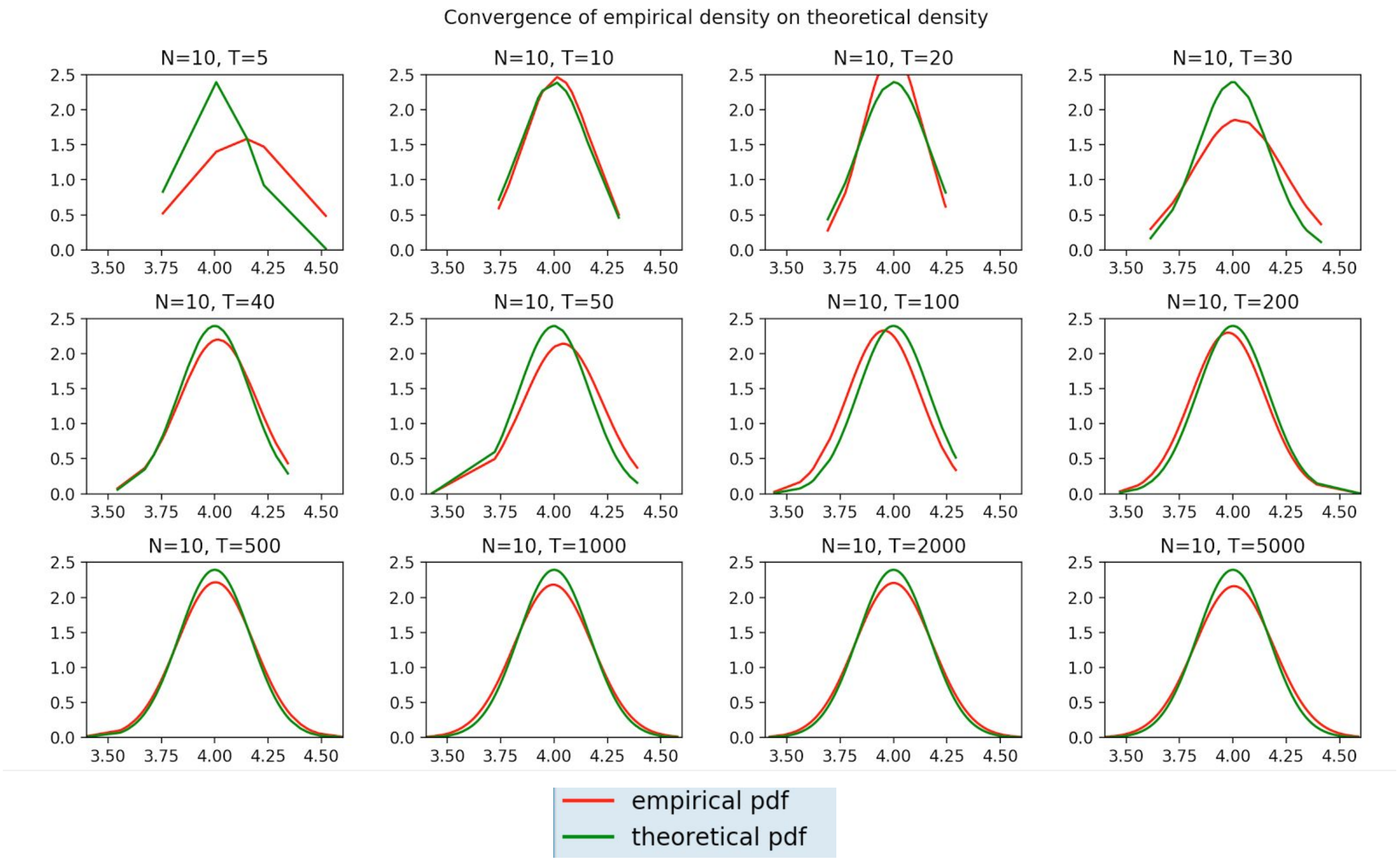

Figure: Convergence of empirical density on theoretical density as a function of the number of repetitions, T

- For the empirical pdf, the mean and standard deviation are calculated using the set of T mean values.
- For the theoretical pdf, the theoretical mean is calculated using the formula (a + b)/2, where a and b are the low and high intervals used in generating the random uniformly distributed numbers. The theoretical standard deviation is calculated here as (b-a)^2 / 24.
- As we will see, the properties of the theoretical normal distribution of the means are determined by the mean and variance of the uniform distribution of the random numbers from which the means are derived.
- When T is low, the theoretical and empirical pdfs do not match well. The variance of the empirical pdf varies significantly between the first few graphs in the series (e.g. compare T=20 with T=30), due to the low numbers of X_bar values contributing to the distribution.
- As T increases, the empirical pdf widens, indicating a larger variance caused by the greater number of X_bar means. The plot also centres around the theoretical mean of 4.0. The two graphs converge, with the fit improving as T increases.
- This series of graphs clearly shows that with a fixed N value, increasing the number of repetitions will create a distribution of X_bar means that tends towards a normal distribution with a mean and standard deviation defined by the properties of its constituent uniform distributions.

3.1 Descriptive statistics

	Mean average, mm	Standard Deviation, mm	Minimum Value, mm	Maximum Value, mm	Range, mm
Zone 1 (NW)	513.04	151.69	271.47	889.43	617.96
Zone 2 (NE)	546.45	142.13	330.63	840.26	509.62
Zone 3 (SW)	1146.15	247.68	734.67	1817.30	1082.63
Zone 4 (SE)	1007.95	174.35	682.45	1432.00	749.55

Table: Rainfall in mm for Senegal, from 1950 to 2000, across 4 geographic zones

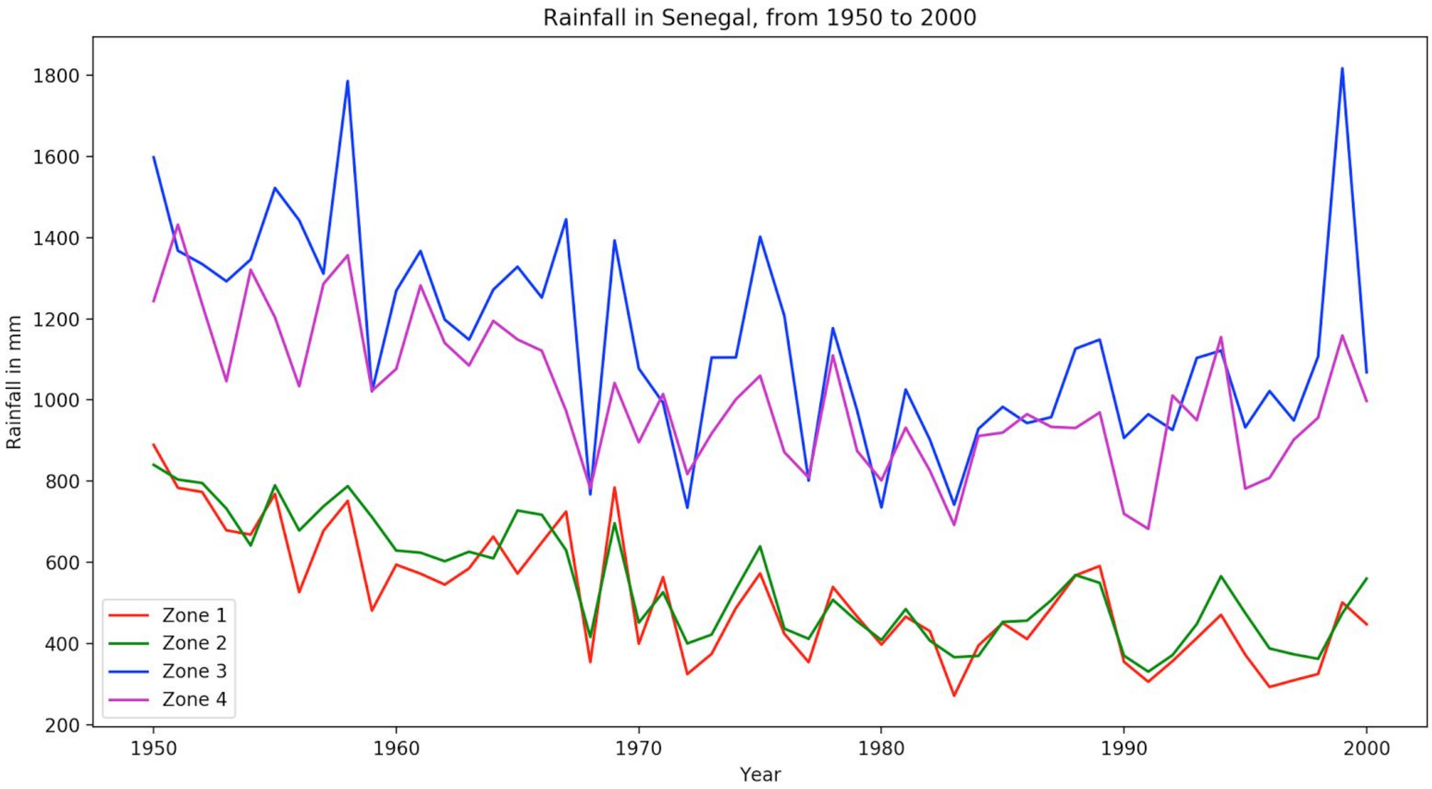

Figure: Rainfall in mm for Senegal, from 1950 to 2000, across 4 geographic zones

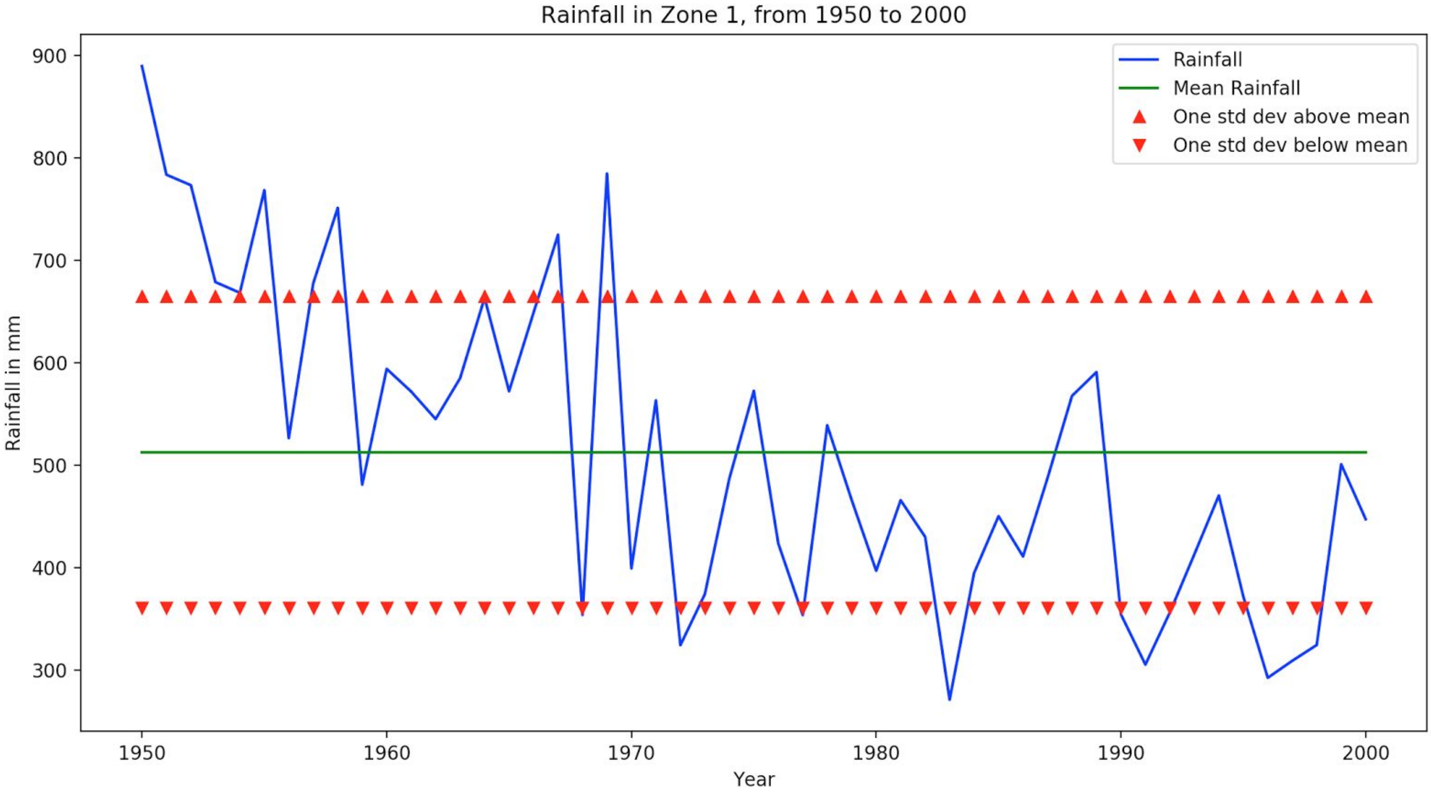

Figure: Rainfall in mm for Zone 1, displaying mean rainfall and one standard deviation from the mean

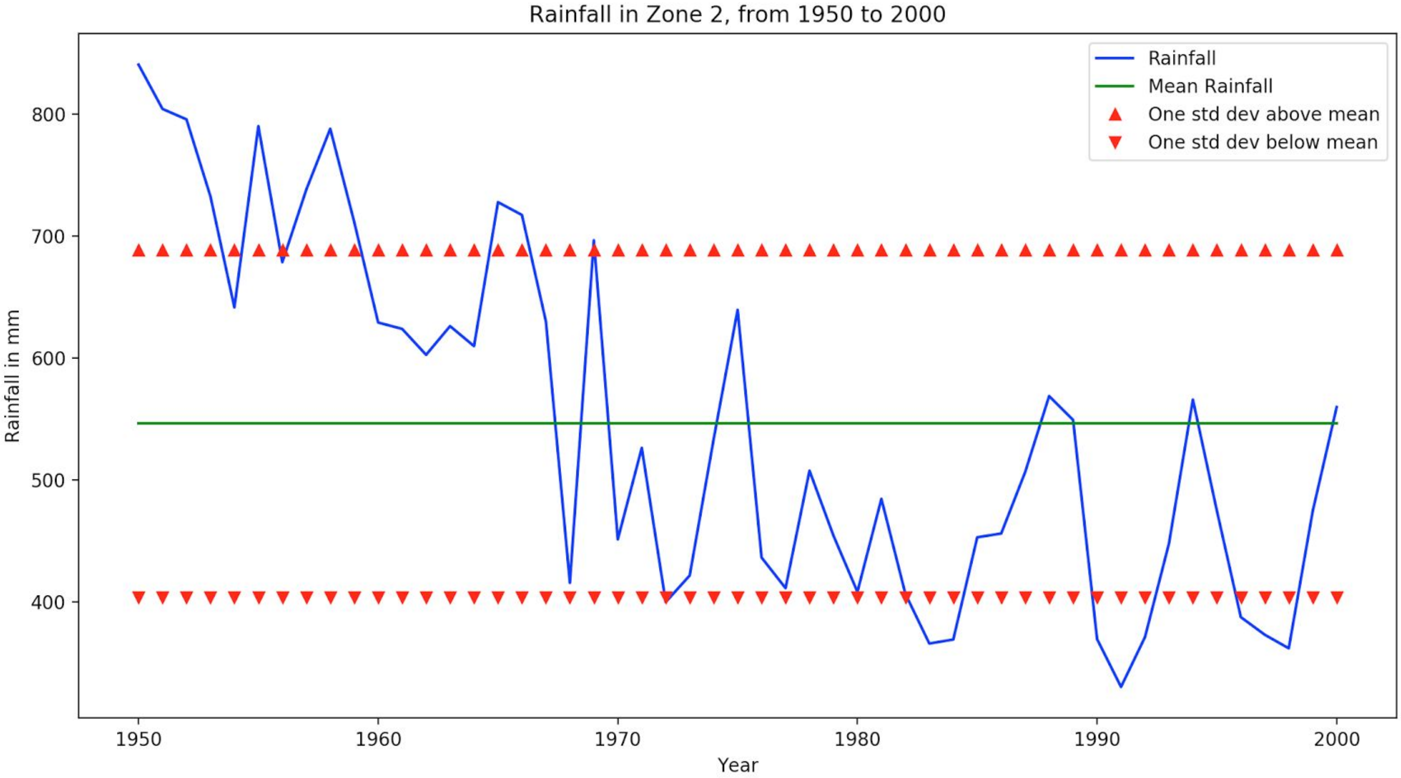

Figure: Rainfall in mm for Zone 2, displaying mean rainfall and one standard deviation from the mean

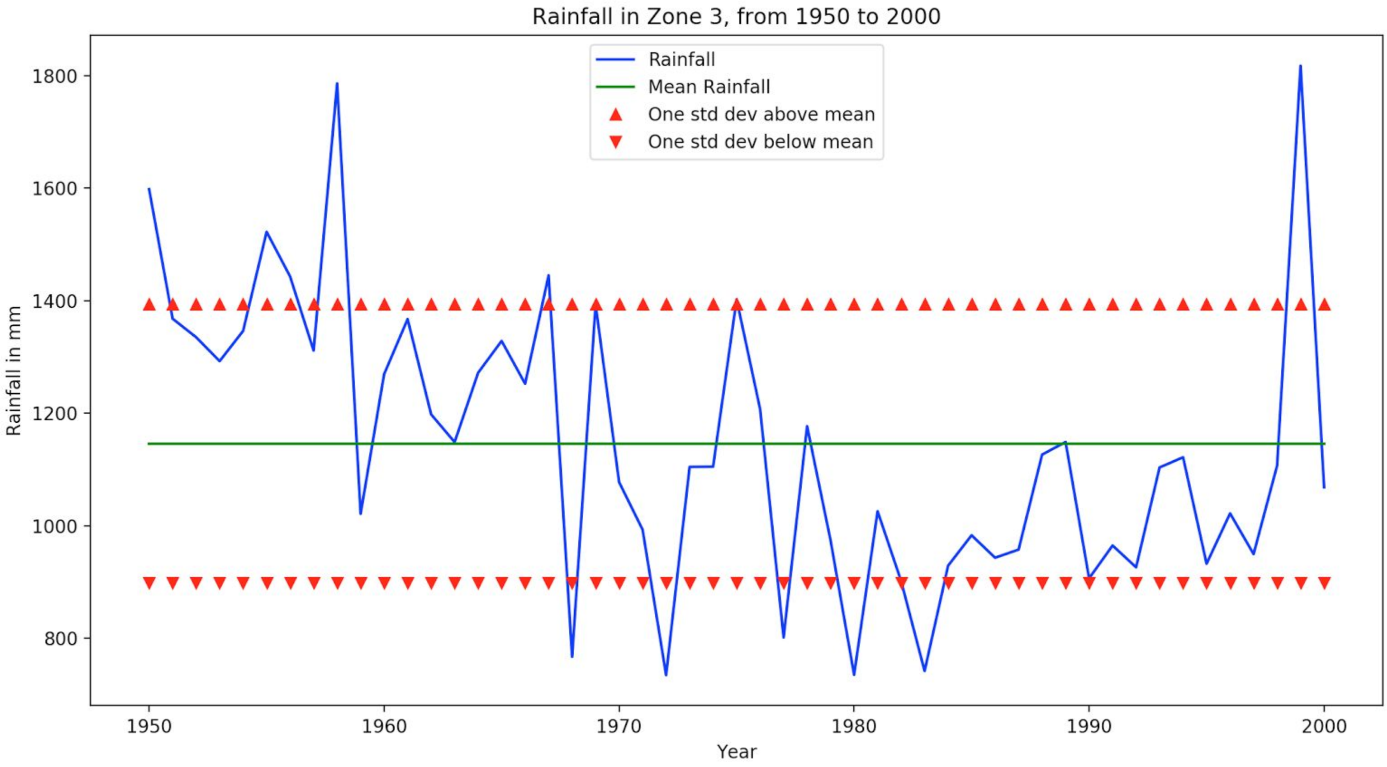

Figure: Rainfall in mm for Zone 3, displaying mean rainfall and one standard deviation from the mean

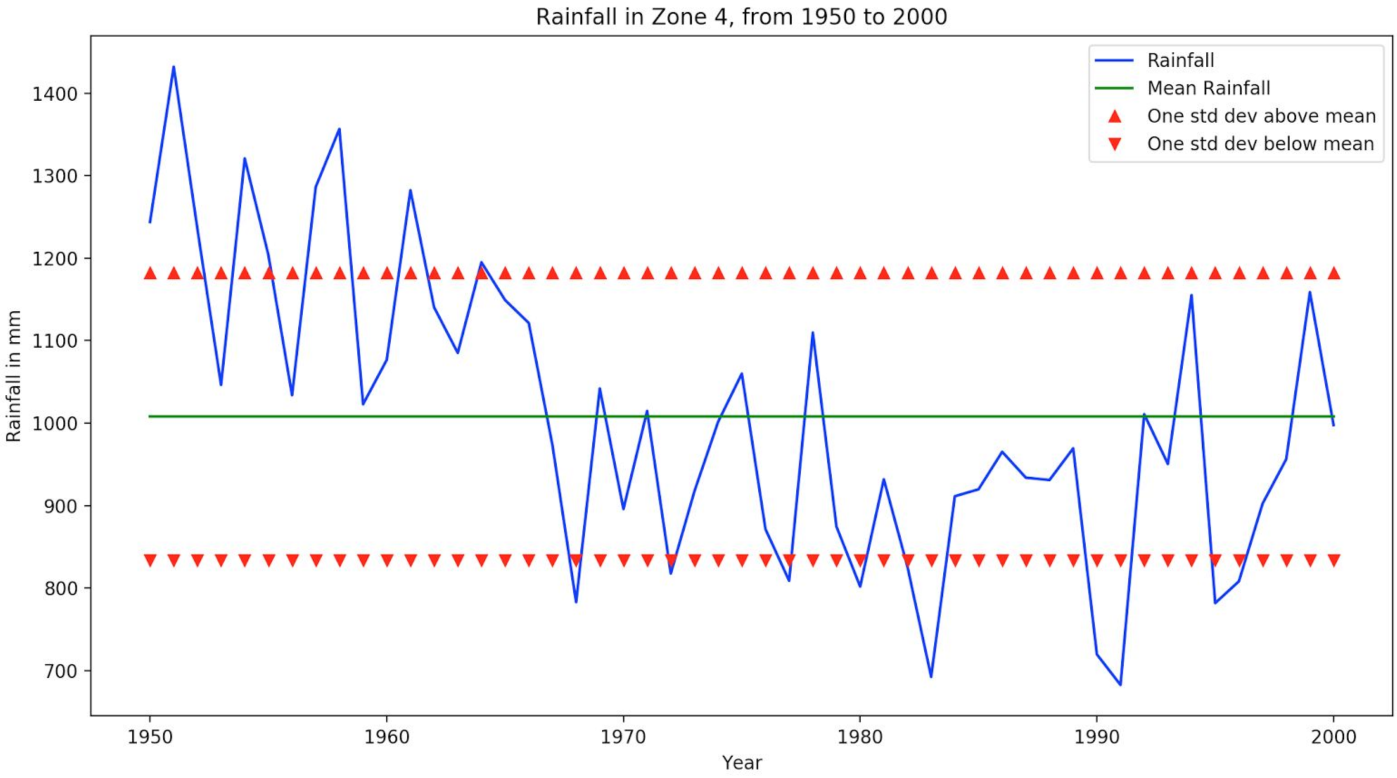

Figure: Rainfall in mm for Zone 4, displaying mean rainfall and one standard deviation from the mean

	Number of years when rainfall was within 1 standard deviation of the mean	**Percentage of years when rainfall was within 1 standard deviation of the mean**
Zone 1 (NW)	31	60.78 %
Zone 2 (NE)	31	60.78 %
Zone 3 (SW)	39	76.47 %
Zone 4 (SE)	32	62.75 %

Table: Years when rainfall was within one standard deviation of the mean

	Quartile 1 (25th Percentile)	**Quartile 2 (50th Percentile / Median)**	**Quartile 3 (75th Percentile)**	**Interquartile Range (Q3 - Q1)**
Zone 1 (NW)	396.11	487.13	592.56	196.45
Zone 2 (NE)	418.88	526.35	640.43	221.55
Zone 3 (SW)	961.30	1106.70	1319.85	358.55
Zone 4 (SE)	899.02	997.70	1130.85	231.83

Table: Quartiles and interquartile range of rainfall in mm

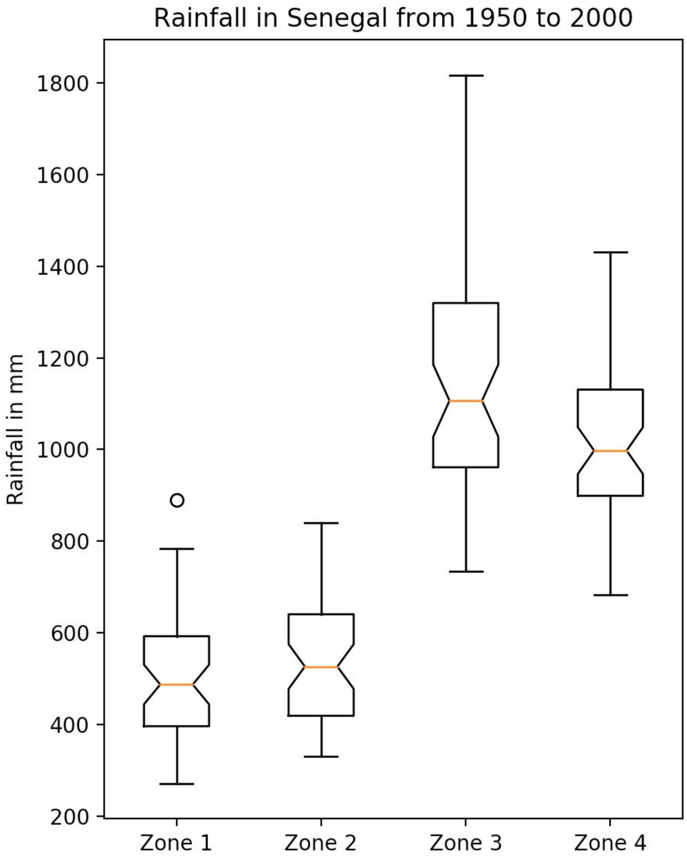

Figure: Boxplot illustrating range, median, and quartiles of rainfall in mm