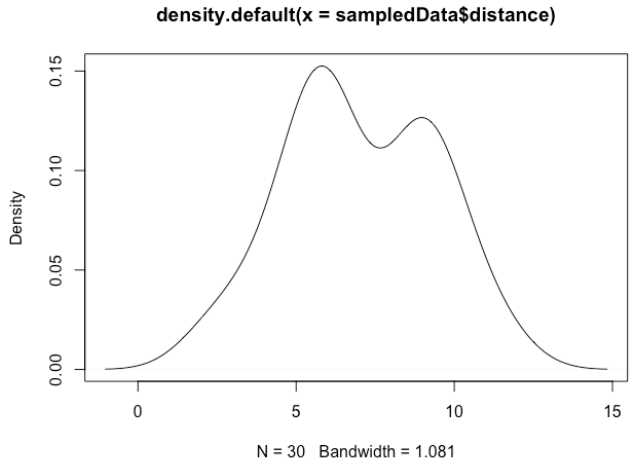
1. **Sampling process**

   The data is sampled from 62 sets of data , which represent the distances I walk on the days I come to the college, collected from apple fitness software to 30 sets of data. The sampling process was performed using stratified sampling to represent the first semester, when most of the activities were online, and the second semester, when all the activities are back in person, equally. The process used in sampling from each subset of data, semester 1 and semester 2, is simple random sample using R command to random 15 row index numbers, no replacement, from each subset of data and bind the data from sampled index together as in the attached "plotData.R".

2. **Data Description**



Population data:	Sampled data:
Mean = 7.3935	Mean = 7.0233
SD = 2.6369	SD = 2.3705
Median = 6.8	Median = 6.45
	Skewness = -0.03452
	Kurtosis = 2.1968

The data is negatively skewed which means that more values are concentrated to the right of the distribution. The kurtosis of approximately 2.2 indicates that the distribution has well behaved tails as the value is close to 3 (kurtosis of the standard normal distribution). High standard deviation, in relation to mean, shows that the datapoints are far from the centre which mean that the distribution has a big variability.

3. **Distribution to use for hypothesis testing and confidence intervals**

   Z-distribution

   If the population standard deviation is known, Z-distribution will give a more accurate results at the sample size (sample size = 30). However, if population standard deviation is unknown, T-distribution would be more appropriate as it allows a bigger variety of values to take place at the tails.

So, in this case, where the population SD is known (population SD = 2.6369) and the sample size is not big enough to make a difference between Z-distribution and T-distribution, Z-distribution would give more accurate result.

4. **Estimation**

95% confidence interval

$$x = \bar{x} \pm CV(\frac{\sigma}{\sqrt{n}})$$

$$x = 7.0233 \pm 1.96(\frac{2.6369}{\sqrt{30}})$$

$$95\%\ confidence\ interval\ is\ [6.0797, 7.9669]$$

97.5% confidence interval

$$x = \bar{x} \pm CV\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$x = 7.0233\ \pm 2.245(\frac{2.6369}{\sqrt{30}})$$

$$97.5\%\ confidence\ interval\ is\ [5.9425, 8.1041]$$

5. **Hypothesis Testing**

μ_0 = 6.45, α = 0.025, H0: The mean of the population is equal to 6.45.

$$Z = \frac{\bar{x} - \mu_0}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{7.0233 - 6.45}{\frac{2.6369}{\sqrt{30}}} = 1.19\ (rounded\ to\ three\ significant\ figures)$$

$$p - value\ =\ 0.12$$

At the significance alpha of 0.025, critical value for z-distribution is 2.245. So, the null hypothesis is not rejected, there is not enough evidence to reject, which can be interpreted by z test statistic falling between the confidence intervals or p-value being higher than α.

6. **Sample Size**

From $x = \mu \pm CV(\frac{\sigma}{\sqrt{n}})$, $margin\ of\ error = CV \times \frac{\sigma}{\sqrt{n}}$.

As the CI is 99%, CV for 0.995 is used which, in z-distribution, is 2.58.

For the margin of error to be less than $\frac{m}{2} = \frac{1.8872}{2} = 0.9436$.

$$CV \times \frac{\sigma}{\sqrt{n}} = d = \frac{\frac{m}{2}}{2}$$

$$2.58 \times \frac{2.6369}{\sqrt{n}} \leq \frac{0.9436}{2}$$

$$n \geq 208\ (rounded\ up\ to\ integer\ as\ sample\ size\ cannot\ be\ fraction)$$

Using the formular above, margin of error for $n \geq 208$ will be lower or equal to 0.9434 which is lower than $\frac{m}{2}$ which is 0.9436. This confirms the result we get that for the sample size of at least 208, 99% CI will not exceed $\frac{m}{2}$. While for $n \leq 207$, the margin of error will be 0.9457 which is higher than $\frac{m}{2}$. So, the calculation is accurate. However, it might not be reliable as it relies on an assumption that the sampling is simple random sampling while the actual process is not.