Q1

(a) By properties of sums, we know that

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\overline{y}) &= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\overline{y} \\ &= \sum_{i=1}^{n} y_i - \overline{y}\sum_{i=1}^{n} 1 \\ &= n\overline{y} - \overline{y}n \\ &= 0\end{aligned}$$

Where in the last line for the first term, we used the definition of $\overline{y}$. In similar fashion, we have that

$$\begin{aligned}S_{yy} &= \sum_{i=1}^{n}(y_i-\overline{y})^2 \\ &= \sum_{i=1}^{n}(y_i^2 - 2y_i\overline{y} + \overline{y}^2) \\ &= \sum_{i=1}^{n} y_i^2 - \overline{y}\sum_{i=1}^{n} 2y_i + \overline{y}^2\sum_{i=1}^{n} 1 \\ &= \sum_{i=1}^{n} y_i^2 - 2n\overline{y}^2 + n\overline{y}^2 \\ &= \sum_{i=1}^{n} y_i^2 - n\overline{y}^2\end{aligned}$$

As desired.

(b) Let $n = 2$ and $M \in \mathbb{Z}$

(i) Let $y_1 = 1$ and $y_2 = -1$, then we have that

$$\begin{aligned}\overline{y} &= \frac{1}{2}(1 + (-1)) = 0 \\ S_{yy} &= (1)^2 + (-1)^2 - n \cdot 0 = 2\end{aligned}$$

as desired.

(ii) Let $y_1 = 1$ and $y_2 = 1$, then we have that

$$\begin{aligned}\overline{y} &= \frac{1}{2}(1 + 1) = 1 \\ S_{yy} &= (1)^2 + (1)^2 - 2 \cdot 1 = 0\end{aligned}$$

as desired.

(c) Let $y_1, \ldots, y_n$ represent the amount of food in kilograms from n purchases for a typical shopper, and let $z_1, \ldots, z_n$ represent these same purchases measured in pounds.

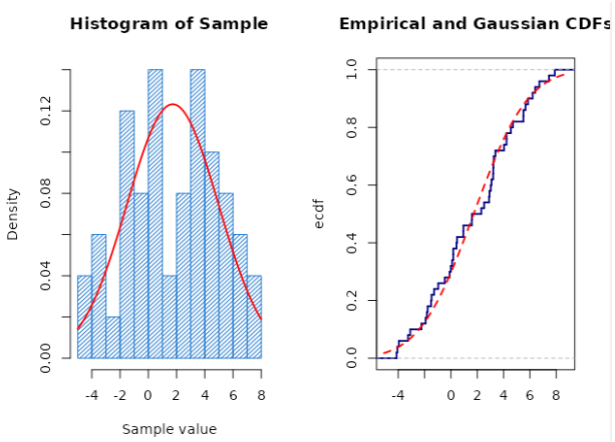
(i) $y_i = 0.454 z_i$ and $z_i = \dfrac{1}{0.454} y_i$

(ii) $\overline{y} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} y_i \implies \overline{z} = \dfrac{1}{(0.454)n} \sum_{i=1}^{n} y_i = \dfrac{1}{0.454} \overline{y}$. Given $\overline{y} = 32$, $\overline{z} = 70.485$

(iii) $S_{zz} = \displaystyle\sum_{i=1}^{n} z_i^2 - n\overline{z}^2 = \sum_{i=1}^{n} \left(\frac{1}{0.454} y_i \right)^2 - n \left(\frac{1}{0.454} \overline{y} \right)^2 = \left(\frac{1}{0.454} \right)^2 S_{yy}$. Given $S_{zz} = 20$, $S_{yy} = 4.122$

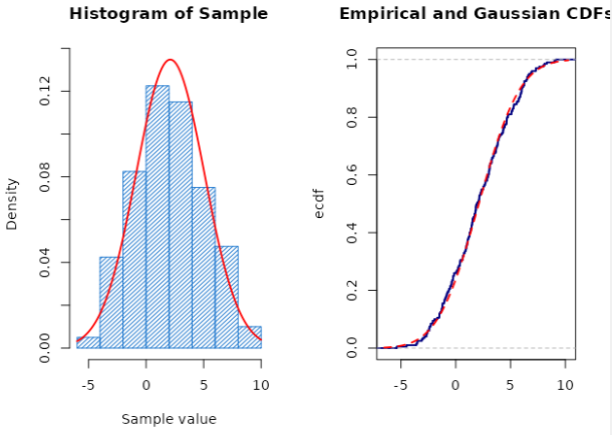
Q2

(a) $\mu = 2$, $\sigma = 3$, and sample size of 50



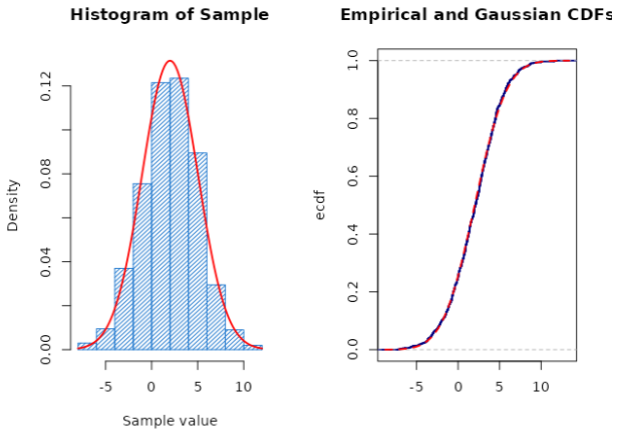
Statistic	Value	Statistic	Value	Statistic	Value
Mean	1.729	Variance	10.478	Skewness	-0.041
Median	1.952	IQR	5.007	Kurtosis	2.057

(b) $\mu = 2$, $\sigma = 3$, and sample size of 200



Statistic	Value	Statistic	Value	Statistic	Value
Mean	2.103	Variance	8.760	Skewness	0.030
Median	1.935	IQR	4.501	Kurtosis	2.419

(c) $\mu = 2$, $\sigma = 3$, and sample size of 1000



Statistic	Value	Statistic	Value	Statistic	Value
Mean	1.998	Variance	9.202	Skewness	-0.089
Median	2.112	IQR	4.070	Kurtosis	3.111

(d) As the sample size increases, all the numerical summaries approach their expected values. The mean started at 1.729 with a sample size of 50, but increases to 1.998 with the sample size of 1000. The same thing happens with σ; it starts at 10.478, but then climbs down to a more accurate 9.202 with the increases sample size. The skewness was always near zero for all the samples, but stayed relatively close to 0. The kutosis rapidly approaches 3 as the sample size increases, meaning it's becoming more and more like a standard normal distribution. In other words, as the sample size increases, all the numerical and graphical summaries approach their expected values corresponding to a normal distribution.

(e) When we decreases the number of bins to 5 or less than 5, we notice that the histogram is relatively peaked near the mean. The extremes of the histogram are also quite low compared to the bins near the mean. If we increase the number of bis above 20, we notice that the distribution flattens out, most of the are the same size, and the extremes are all the same size and are abundant in the tails.

Q3

(a) For the given information, an appropriate potential population is "2500 *active mobile players as of present*"

(b) Since the players are not interacted with by the observer (very specific data collected from the central server), and the fact that this was a smaller, finite, population of the entire whole player base, this is a survey study.

(c) skill_grade is an ordinal variate; it's a measure of the player's skill tier via a discrete "grading" system from A to F. time_overworld is a continuous variate (could be discrete is the we can only measure up to minutes, or second, or some finite time); theoretically we can find the time spent on the game to utmost accuracy, meaning it's a continuous variate. device_age is a discrete variate since it's rounded down; if let's say a device was a year and 2 months old, it would be rounded down to 1 (this means the only possible values this variable can take on are natural numbers; a discrete variate!)

Q4

(a) Since time_combat represents the time spent in player-vs-player combat (in hours), each of the numbers represent a the quartiles of the data (along side $y_{(0)}$ and $y_{(n)}$

$$\begin{aligned} y_{(0)} &= 0 \\ q(0.25) &= 2.94 \\ q(0.50) &= 7.09 \\ q(0.75) &= 14.44 \\ y_{(n)} &= 65.65 \end{aligned}$$

The range and IQR are simply

$$\begin{aligned} \textbf{range} &= y_{(6)} - y_{(1)} = 65.65 - 0 = 65.65 \\ \textbf{IQR} &= q(0.75) - q(0.25) = 14.44 - 2.94 = 11.5 \end{aligned}$$

(b) The sample mean is 10.121 while the sample median is 7.090, from this, we notice they differ by quite a bit. This tells us that there exists a an outlier, and that the data is positively skewed (there are extreme scores/outliers in the top 50% of the data), which means the graph of the distribution has more values to the left of the mean, and a very flat tail to the right of the mean. An exponential distribution also has positive skewness (actually one can show that a perfect exponential distribution has skewness of 2), and exhibits the same behaviour as describe above (tail becomes very flat on the right side of the graph).

(c) First, we know that the median of the exponential model is simply $\mathrm{m}[X] = \ln(2)/\lambda$ (where λ is the rate parameter), and the expectation value is $\mathrm{E}[X] = 1/\lambda$. In the model, it can be shown that $\mathrm{m}[X] < \mathrm{E}[X]$, which is also true of our numerical data. Second, another property of the exponential model is that the standard deviation is equal to the mean, and in our numerical data, this is quite close to being true (given a larger sample size, the gap between the mean and SD would shrink to eventually be the same). Third, if we examine the histogram of the time in combat, plotted alongside the Exp pdf, we notice that the pdf is an amazing approximation to the density of the histogram. In other words, the Exp pdf is a great approximation to the pdf of our distribution. Fourth, finally, we notice that the the tail at the right of our mean flattens our very fast, exponentially even! This is what we noticed above in part (b), where the data tells us our distribution is skewed positively, and the exponential is also skewed positively.

Q5

Consider the ordered set of data

$$y = 39, 41, 42, 46, 46, 46, 48, 49, 52, 53, 59$$

(a) The five number summary of y is the minimum ($y_{(1)}$), the first quartile ($q(0.25)$), the median ($q(0.50)$), the third quartile ($q(0.75)$) and the maximum ($y_{(n)}$)

(i) The minimum is simply the first data point in the ordered list of points, therefore $y_{(1)} = 39$

(ii) To find the first quartile we first need $m = (12)(0.25) = 3$, therefore $q(0.25) = y_{(3)} = 42$

(iii) To find the median we first need $m = (12)(0.50) = 6$, therefore $q(0.50) = y_{(6)} = 46$

(iv) To find the third quartile we first need $m = (12)(0.75) = 9$, therefore $q(0.50) = y_{(9)} = 52$

(v) Finally, the maximum is the last point in the ordered data set, therefore $y_{(11)} = 59$

(b) We have already calculated the first, second (median), and third quartiles in (a), therefore we can just plug these into the formula

$$\begin{aligned}\text{Skew}_{Bowley} &= \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} \\ &= \frac{52 + 42 - 2(46)}{52 - 42} \\ &= 0.2\end{aligned}$$

(c) First we find that the mean, rounded to 3 digits, is simply

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{11} y_i = \frac{39 + 41 + 42 + 46 + 46 + 46 + 48 + 49 + 52 + 53 + 59}{11} = 47.364$$

Thus, using our equations for the sample skewness from class (calculated on desmos for simplicity) is

$$\text{Skew} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2\right)^{3/2}} = \frac{\frac{1}{11}\sum_{i=1}^{11}(y_i - 47.364)^3}{\left(\frac{1}{11}\sum_{i=1}^{11}(y_i - 47.364)^2\right)^{3/2}} \approx 0.459$$

(d) The Bowley skewness gave us an answer near zero, meaning that the shape of the data set is almost symmetrical but since it is positive, it is positively skewed (just ever so slightly). The Bowley skewness tells us the difference in the quartiles, since in a symmetric distribution the first and third quartiles are at equal distances from the mean. This coincides with what we got in (c), where we found a positive, near zero, skewness; meaning the distribution is positively skewed.