Writing Assignment 3

Probability & Statistics

Author

Paul Beggs
BeggsPA@Hendrix.edu

Instructor

Prof. Lars Seme, M.S.

Due

November 21, 2025

1 Introduction

It can often be the case that when we collect empirical data, it can be challenging to distinguish between known theoretical distributions based on summary statistics alone. Thus, the focus of this paper is to investigate three distinct datasets—**Data1**, **Data2**, **Data3**—where each has $n = 200$ samples, and each sample is generated by one of these methods:

1. $X \sim 15 + Gamma(6.25, 0.8)$
2. $Y \sim 20 + \left(\sqrt{4/3}\right) t(3)$
3. $Z \sim 20 + N(0, 4)$.

2 Data Analysis

A quick check to determine where the data came from is to find the mean and variance of each theoretical distribution, and compare it with the sample mean and sample variance from the empirical ones. First, let us investigate the theoretical distributions:

1. We know that X is gamma-distributed and shifted by 15, so its mean and variance formulas are $\mu = 15 + \alpha\theta$ and $\sigma^2 = \alpha\theta^2$. Substituting α for 6.25, and θ for 0.8, we see that
$$\mu = 15 + (6.25 \cdot 0.8) = \boxed{20} \quad \text{and} \quad \sigma^2 = (0.64)(6.25) = \boxed{4}.$$
2. For Y, since it is t-distributed, its mean is usually 0, and its variance is found by solving $\frac{r}{r-2}$ for $r > 2$. However, it is shifted by 20 and scaled by a factor of $\sqrt{4/3}$, so we find its mean and variance to be:
$$\mu = 0 + 20 = \boxed{20} \quad \text{and} \quad \sigma^2 = \frac{3}{3-2} \cdot \left(\sqrt{4/3}\right)^2 = \boxed{4}.$$
3. Then for Z, we find the mean and variance from the distribution parameters, but we must also shift the mean by 20:
$$\mu = 0 + 20 = \boxed{20} \quad \text{and} \quad \sigma^2 = \boxed{4}.$$

The empirical means and sample variances can be seen in Table 1. These results leave us in a rather awkward situation: all of our theoretical distributions have the same mean and variance, and (perhaps unsurprisingly) the empirical distributions also have similar means with each other, and with the theoretical distributions. Thus, we must tackle this investigation from another angle.

	Min.	Q_1	Median	Q_3	Max.	Mean	S. Var.
Data1	9.325	19.101	20.095	20.979	26.490	19.947	3.908
Data2	16.179	18.486	19.777	21.163	27.306	19.978	3.962
Data3	14.057	18.665	20.093	21.462	25.354	20.125	4.086

Table 1: Dataset Statistics

3 Box Plots & Histograms

For **Data1**, we see that its histogram follows a bell-shaped curve (Figure 1 (a)), and can see numerous outliers in its box plot (Figure 1 (b)). Then, for **Data2**'s histogram (Figure 2 (a)), we see that is also roughly bell shaped, but is right skewed. **Data3**'s histogram is also bell-shaped (Figure 3 (a)), with most of its data being falling within the interquartile range with only 1 outlier, as seen in its box plot (Figure 3 (b)). From these observations, we can make the draw the following connections:

- **Gamma Distribution:** We know that the gamma distribution follows a shape similar to that of **Data2**'s, where there is a right skew with a bell-shape. We also know that gamma's sample space is non-negative, so we know that it must start at 15 (due to Y's shifting). This leaves us with the clear choice that **Data2** must be gamma-distributed, as it is the only distribution that starts above 15 (as per Table 1).
- **Normal & T Distributions:** Like the gamma distribution, these both are bell-shaped, but the normal curve has the property where 95% of the data falls within two standard deviations of the mean. Compare that to **Data1**'s distribution that has two outliers that are over five standard deviations from the mean. This tells us that the **Data1** cannot be normally distributed.

Thus, through a process of elimination, we have strong evidence that **Data1** is t-distributed, **Data2** is gamma-distributed, and **Data3** is normally distributed. Now, to ensure our choices are valid, we will now investigate q-q plots.

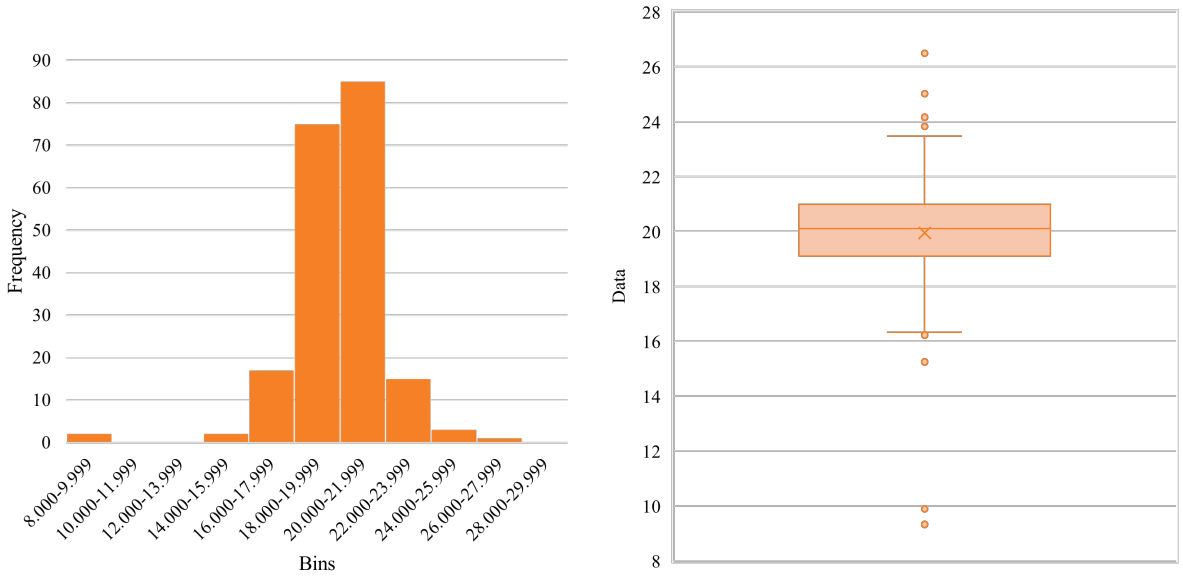
4 Q-Q Plots

Q-q plots are measurements that take a data values at specific percentiles, and compares them to theoretical counterparts. For example, the smallest value in **Data1**'s set is ≈ 9.326 with percentile ≈ 0.005. So, we can take that percentile and calculate the corresponding value in one of our distributions and compare values. It is important to note that if the empirical distribution's value and the theoretical distribution's value are equal for every percentile, then that shows that both distributions are equivalent. In q-q plots, that relationship is modeled as $y = x$, and can be seen as a straight line. Now, in our q-q plots (Figure 4), we see that **Data1** and the t distribution plot has a straight line (albeit with a few outliers); **Data2** and the gamma distribution plot is almost a near perfect straight line; and **Data3** and the normal curve plot show the straightest line for all datasets vs. the normal distribution.

5 Conclusion

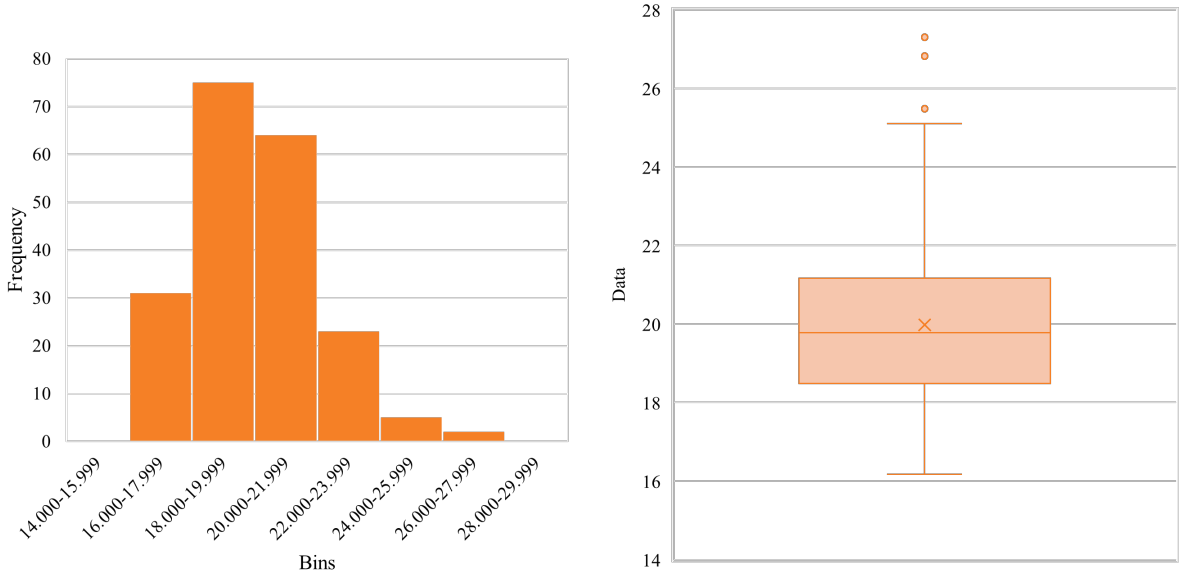
Through our analysis, we have shown that when you cannot rely on comparing means and variances of empirical distributions to theoretical distributions, you still have other tools that you can rely on. Specifically, we examined histograms and box plots and gained strong evidence that **Data1** was t-distributed because it contained outliers that made it so that it could not be gamma-distributed, nor could it be normally distributed because it had outliers that were more than 5 standard deviations from the mean. Then, we also saw that the only theoretical distribution that could be mapped to **Data2** was gamma because the other two empirical distributions started below 15, which gamma cannot because its support is non-negative. Finally, that left **Data3** to be mapped to the normal distribution. Finally, we looked at q-q plots that directly supported our conclusions from the histograms and box plots. Thus, through this investigation, we have definitively shown that **Data1** is t-distributed, **Data2** is gamma-distributed, and **Data3** is normally distributed.

Appendix A: Histograms & Box Plots



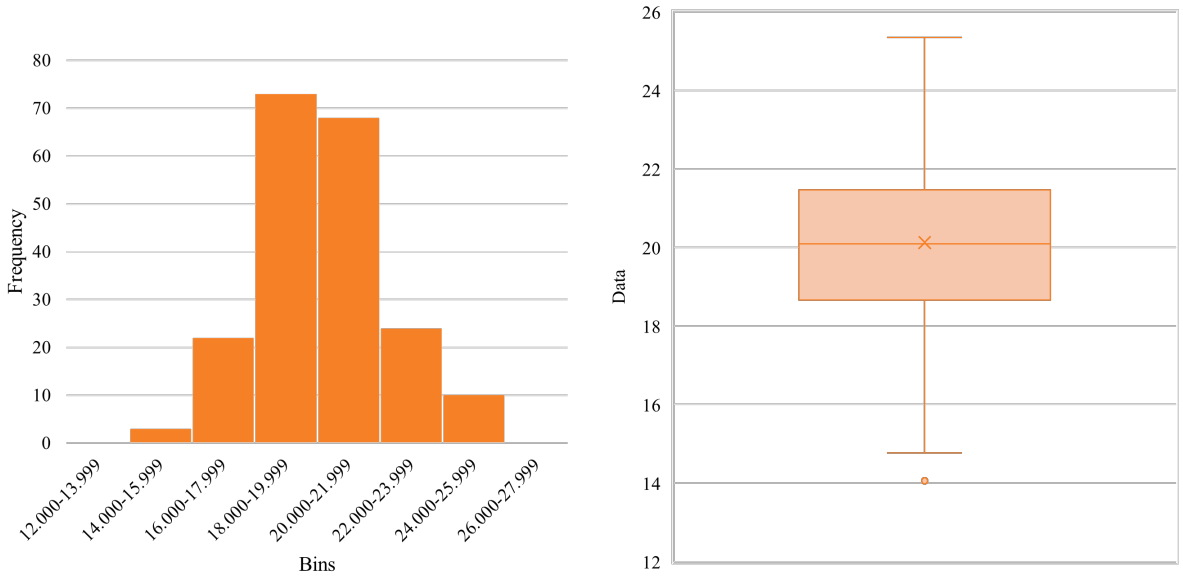
(a) Histogram (b) Box Plot

Figure 1: **Data1**'s Histogram & Box Plot



(a) Histogram (b) Box Plot

Figure 2: **Data2**'s Histogram & Box Plot



(a) Histogram (b) Box Plot

Figure 3: **Data3**'s Histogram & Box Plot

Appendix B: Q-Q Plots

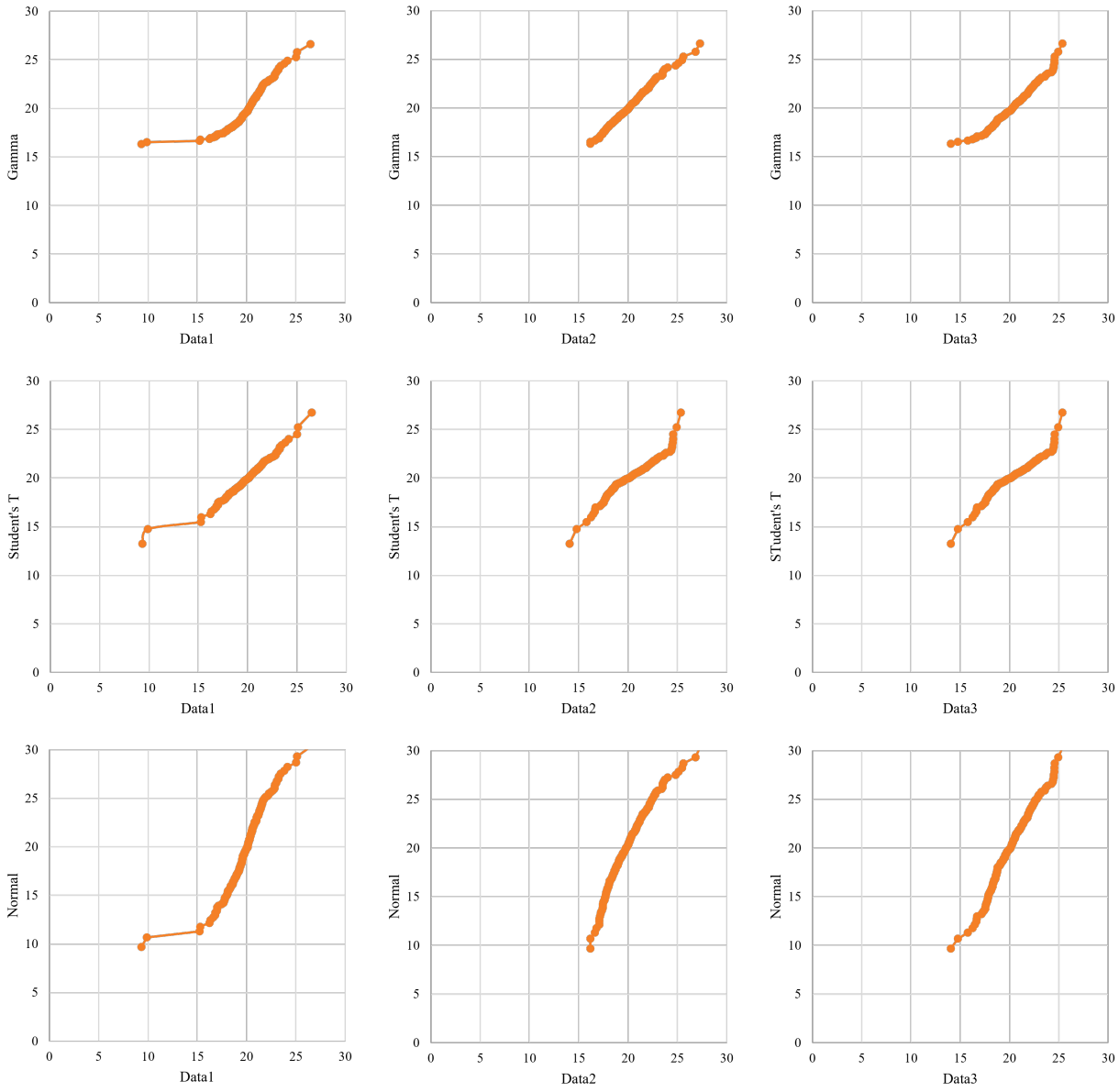


Figure 4: All Q-Q Plots; All plots are shown on the same data range for consistency.