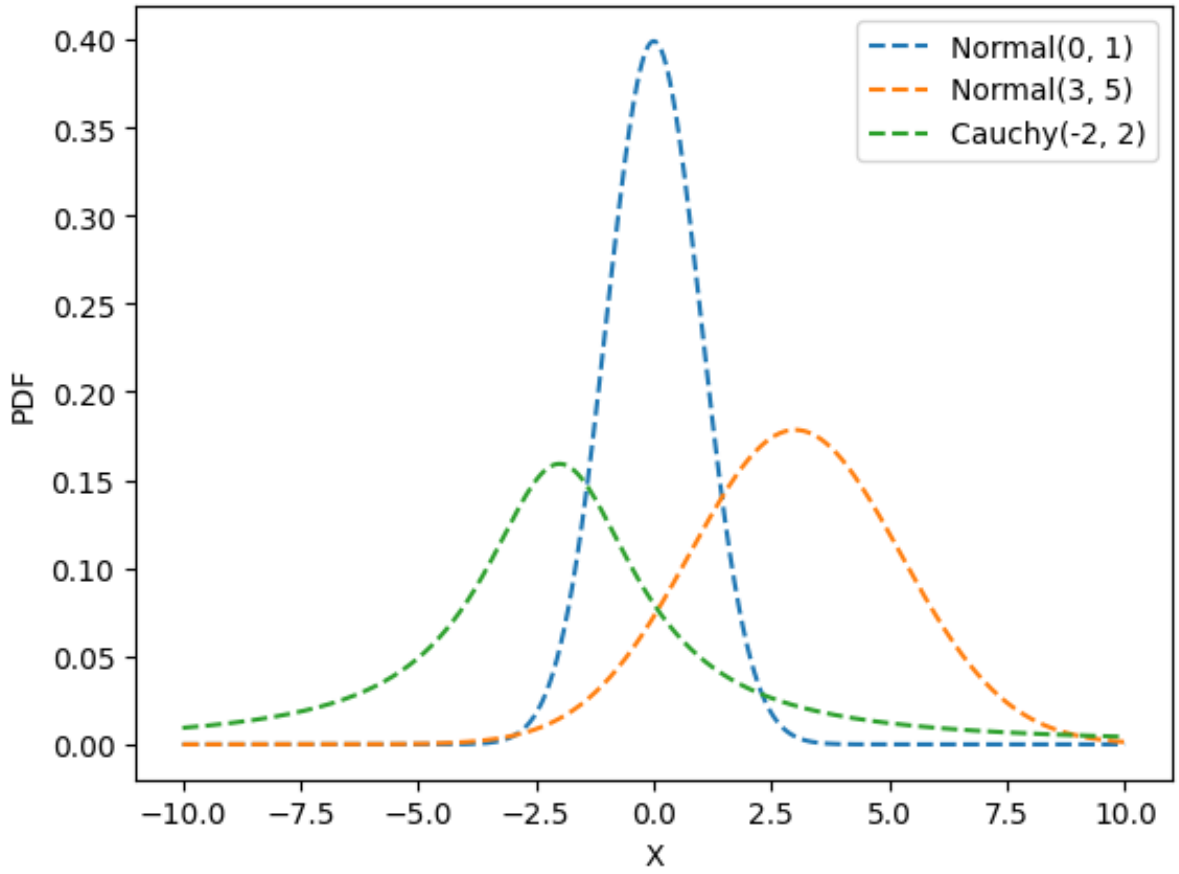
Statistics Exercise 1

Madhukara S Holla

25th September 2023

Question 1.a



Key Observation

Normal Distribution - $\mathcal{N}(0, 1)$

- Has a narrow curve due to low variance, resulting in a high probability density around the mean (0).
- It has shorter tails - samples drawn from this distribution will be closer to the mean.
- Does not have long tails - chances of drawing extreme values are low.

Normal Distribution - $\mathcal{N}(3, 5)$

In comparison with $\mathcal{N}(0, 1)$,

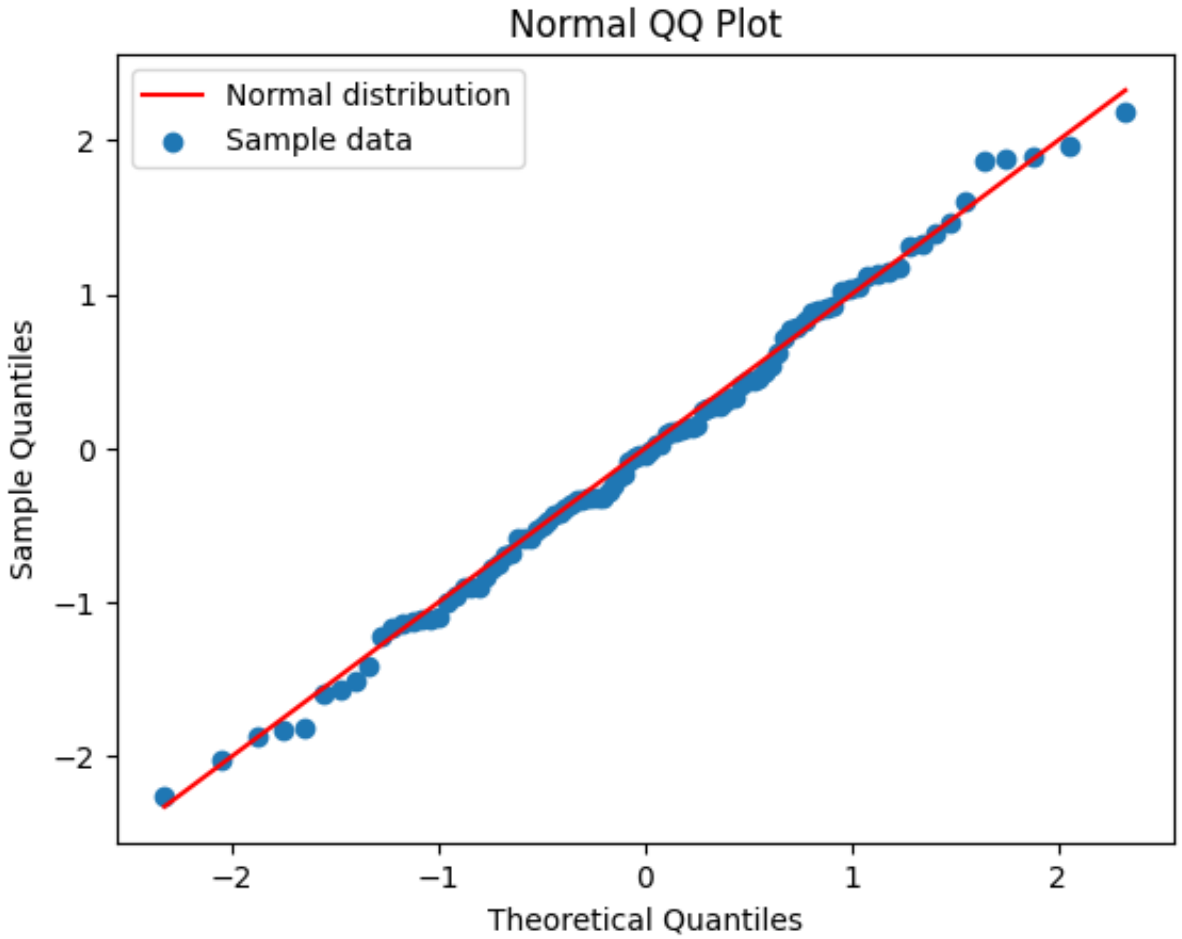
- Has a wider curve due to high variance, resulting in a lower probability density around the mean (3).
- It has longer tails - a significant number of samples drawn from this distribution can be away from the mean (due to high variance).
- Tails are slightly longer than $\mathcal{N}(0, 1)$, but chances of drawing extreme values are still low.

Cauchy Distribution - $Cauchy(-2, 2)$

- Has a narrower curve when compared to $\mathcal{N}(3, 5)$ and has longer tails.
- This distribution is not symmetric and does not have a mean or variance.
- Chances of drawing extreme values are higher when compared to Normal distribution.

Question 1.c

QQ plot for samples from $\mathcal{N}(0, 1)$

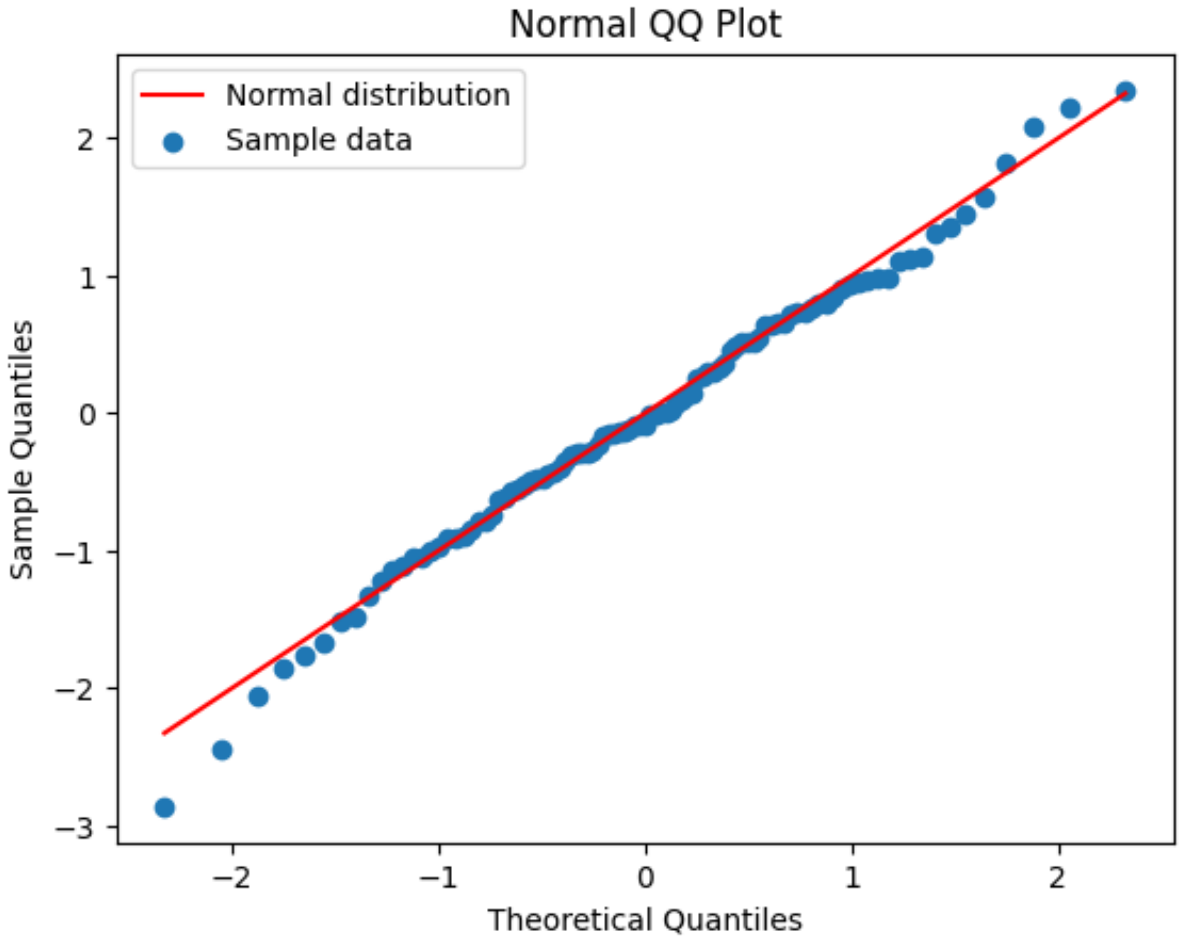


From the QQ plot,

- We can observe that the samples drawn from $\mathcal{N}(0, 1)$ closely align with the theoretical quantiles of normal distribution.
- They form a straight line, indicating that the samples are normally distributed.

Question 1.d

QQ plot for samples from $\mathcal{N}(3, 5)$

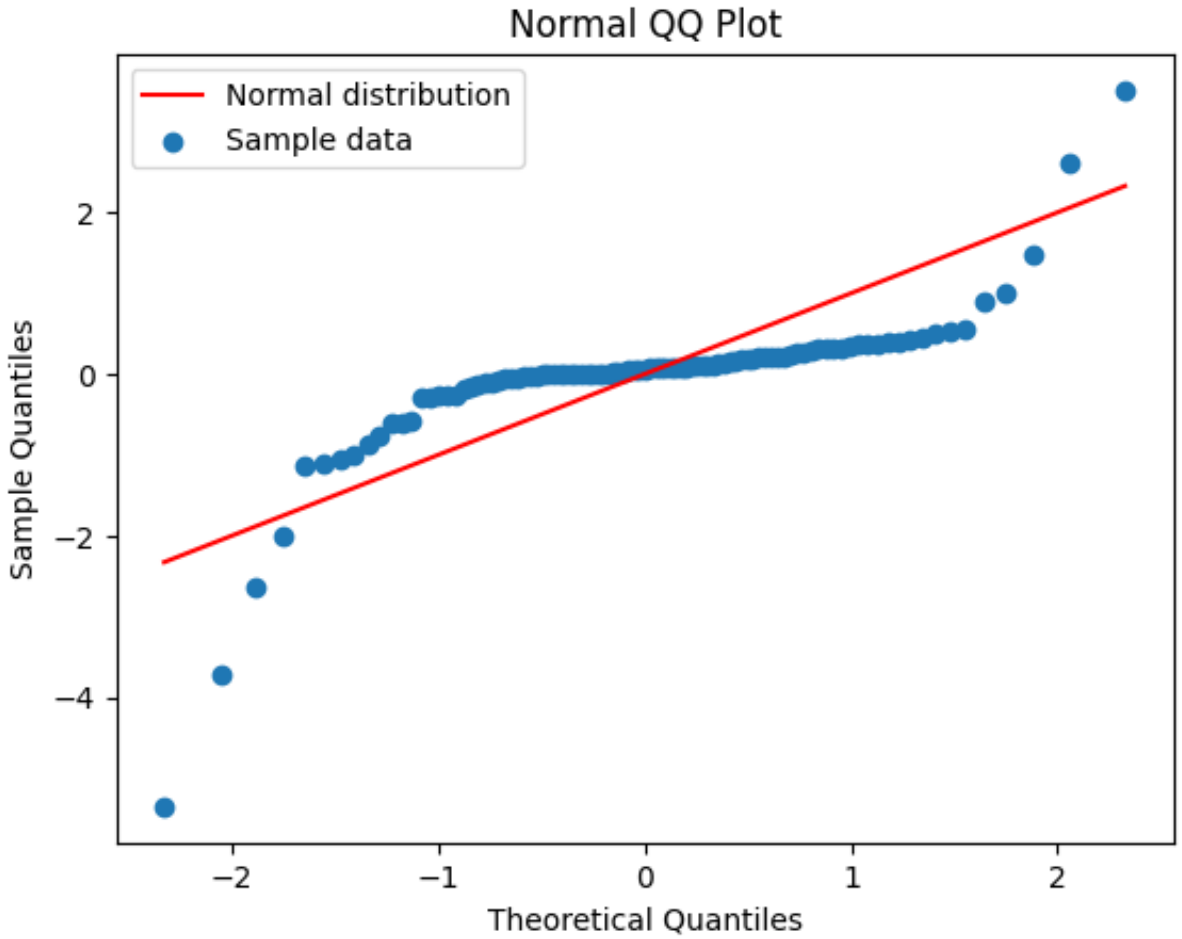


From the QQ plot,

- We can observe that the samples drawn from $\mathcal{N}(3, 5)$ closely align with the theoretical quantiles of normal distribution.
- We can see a few points away from the straight line in the plot. This is because of the high variance in the distribution.
- Ignoring a few outliers, the samples are closer to a straight line, indicating that the samples are normally distributed.

Question 1.e

QQ plot for samples from $Cauchy(-2, 2)$



From the QQ plot,

- We can observe that the samples drawn from $Cauchy(-2, 2)$ do not form a straight line - indicating that the samples are not normally distributed.
- In addition, we can see multiple outliers in the plot - indicating a longer tail.

Question 2.a

Given data: $x_1 - \bar{x}, x_2 - \bar{x}, x_3 - \bar{x} \ldots x_9 - \bar{x}$ and a mean $\bar{x}$ of 10 data points.

Mean is defined as:

$$\bar{x} = (x_1 + x_2 + x_3 \ldots + x_9 + x_{10})/10$$

Multiply both sides by 10

$$10.\bar{x} = (x_1 + x_2 + x_3 \ldots + x_9 + x_{10})$$

Subtract $9.\bar{x}$ from both sides

$$10.\bar{x} - 9.\bar{x} = (x_1 + x_2 + x_3 \ldots + x_9 + x_{10}) - 9.\bar{x}$$

Rearrange terms

$$\bar{x} = (x_1 - \bar{x} + x_2 - \bar{x} + x_3 - \bar{x} \ldots + x_9 - \bar{x}) + x_{10}$$

$$x_{10} = (x_1 - \bar{x} + x_2 - \bar{x} + x_3 - \bar{x} \ldots + x_9 - \bar{x}) - \bar{x}$$

By substituting the given data points and mean, we get the value of x_{10}.

Question 2.b

Sample data points: $x_1, x_2, x_3 \ldots x_9, x_{10}$.

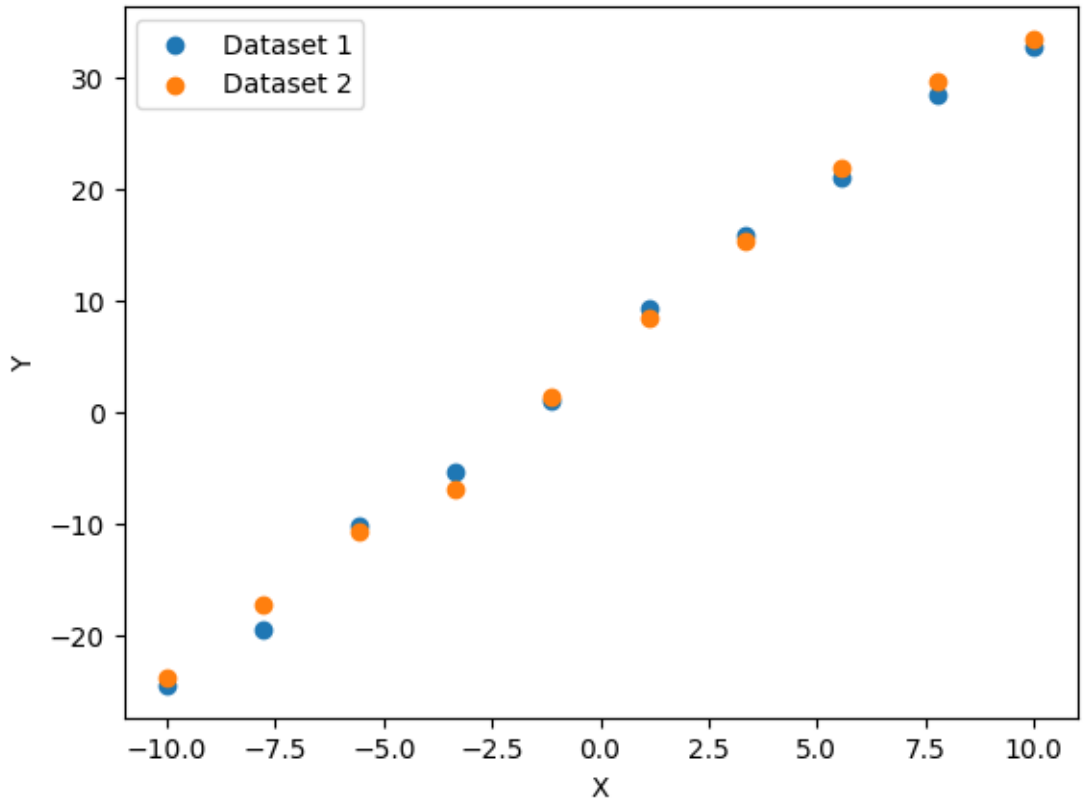
$$s^2 = \frac{\sum_{i=1}^{i=n}(x_i - \bar{x})^2}{n-1}$$

$\bar{x}$ is the mean of the sample data points and not the population mean.

In this case of sample variance, degrees of freedom is n-1 because we use the sample mean which is a calculated measure from the sample data. For n = 10, degrees of freedom is 9.

Given $n-1$ data points and a mean of n data points, we can easily calculate the missing data point. Which means only $n-1$ data points can vary, hence we lose one degree of freedom.

4.a



The plot above shows two datasets sampled from the same distribution $Y = 5 + 3.X + \varepsilon$. $\varepsilon \sim \mathcal{N}(0, 1)$

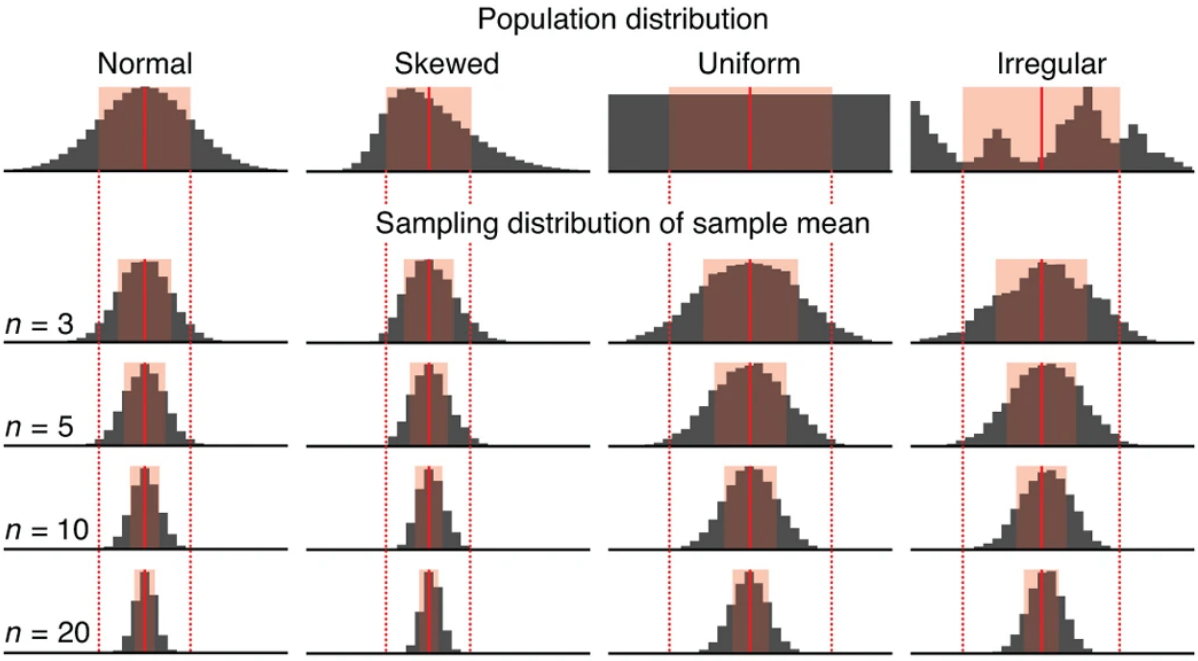
- Both datasets are similar and have a linear relationship.
- They have a similar slope and intercept.
- There is a small difference between each data point of either dataset due to the error term ε.
- The value of error term ε is different for each dataset as they are drawn independently from the same distribution.

4.b

Values obtained for dataset 1: $slope = 2.942$ and $intercept = 4.856$
Values obtained for dataset 2: $slope = 2.951$ and $intercept = 5.121$

- The outputs are slightly different between the two datasets due to the random noise ε in generating the data.
- The value of error term is different for each dataset as they are drawn independently.
- This randomness causes the estimated slope to deviate slightly from the true value of 3 in each dataset.

5.a



- From the figure above, we can see that the distribution of sample mean is approximately normal regardless of the population distribution.
- Between sample sizes of 3 and 5, we can see that the sample distribution is approximately normal (somewhat retains the shape of the population distribution).
- As the sample size increases, the sample distribution becomes more normal and the shape of the population distribution is lost.
- This is because the sample mean is an unbiased estimator of the population mean and the Central Limit Theorem states that the distribution of sample means is approximately normal regardless of the population distribution.
- Additionally, we can see that the high variance of the sample distribution decreases drastically as the sample size increases.
- This is because the more data we sample, the denominator of the sample variance increases, but the numerator does not increase by as much (because the sample mean remains closer to the population mean).