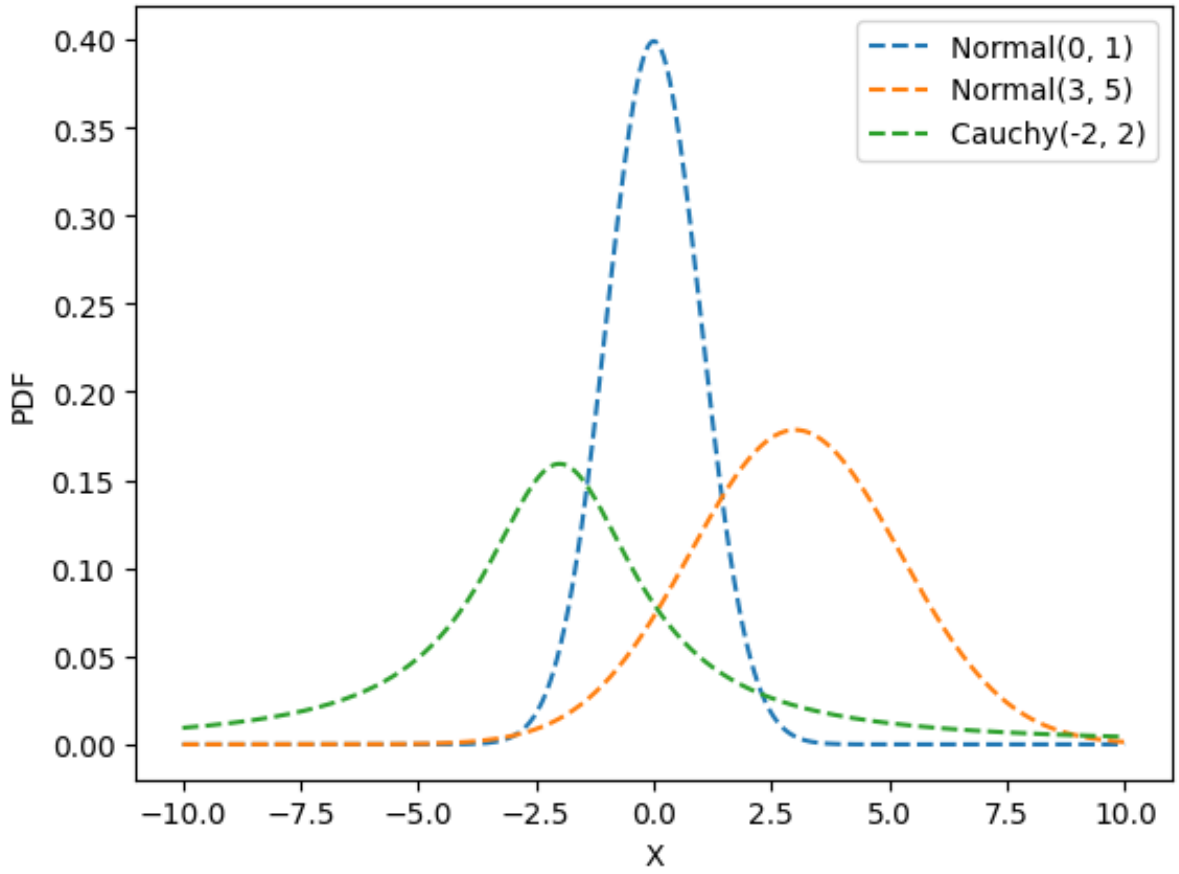
Statistics Exercise 1

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 $\mathcal{N}(0, 1)$,

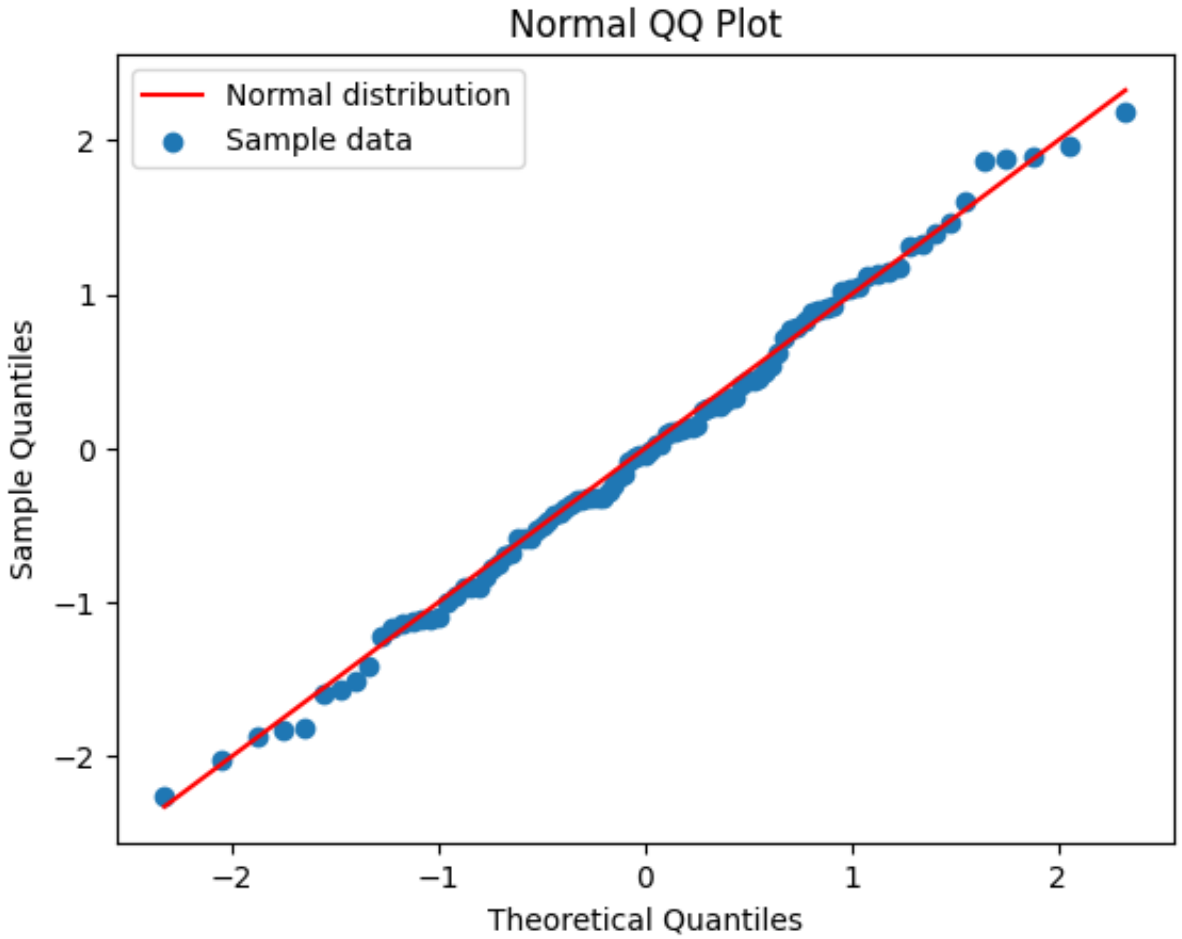
- Has a wider curve due to high variance, resulting in a lower probability density around the mean (3).
- It has longer tails - a significant number of samples drawn from this distribution can be away from the mean (due to high variance).
- Tails are slightly longer than $\mathcal{N}(0, 1)$, but chances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

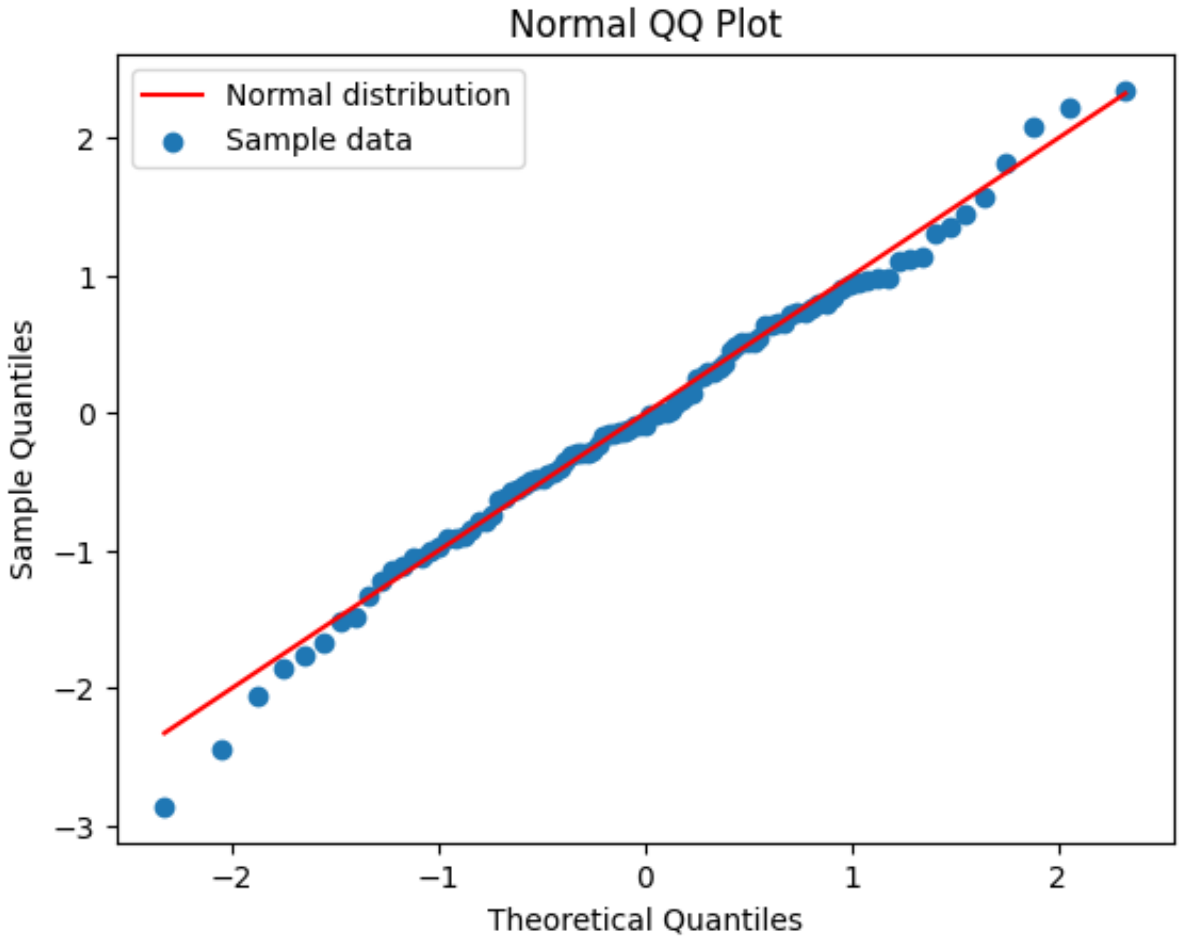


From the QQ plot,

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

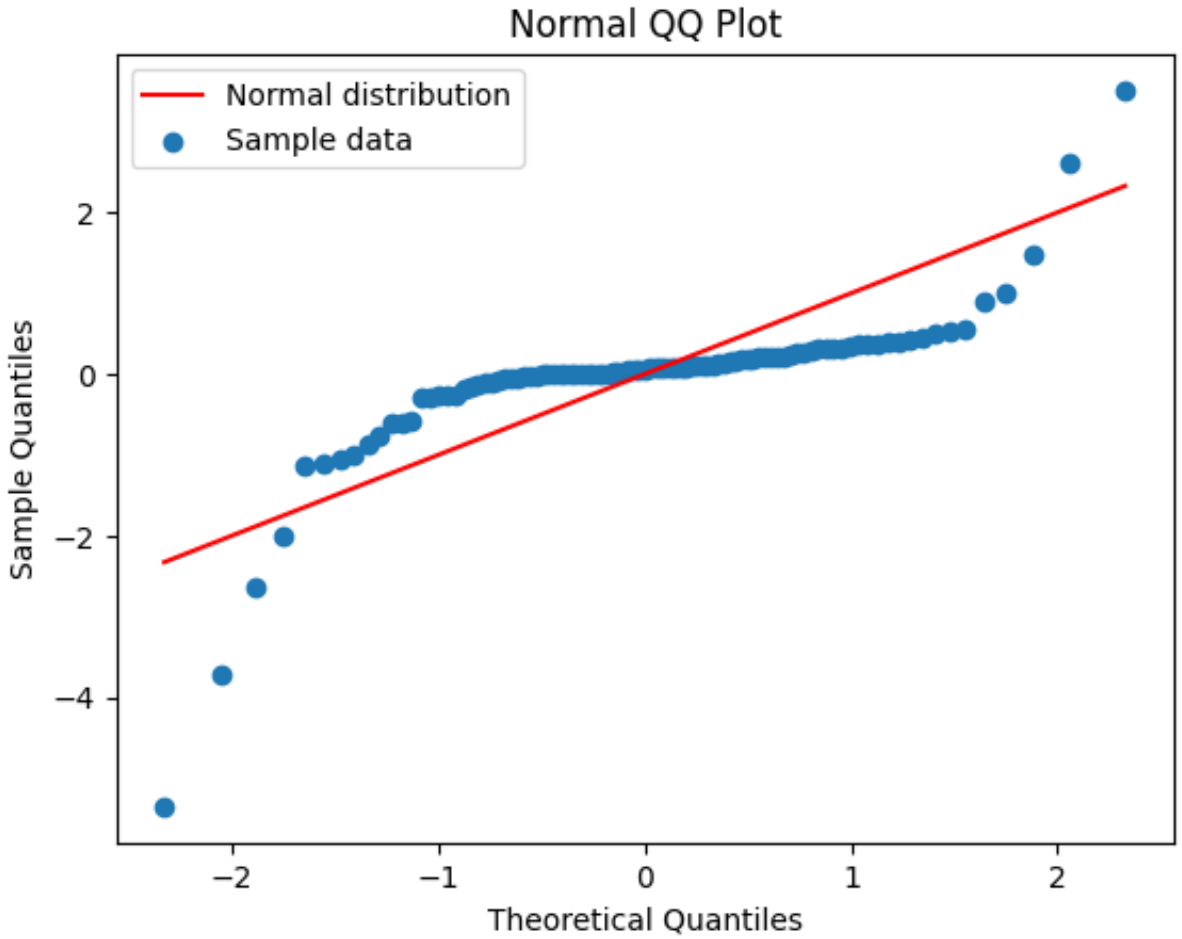


From the QQ plot,

- 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: $x_1, x_2, x_3 \ldots x_9, x_{10}$.

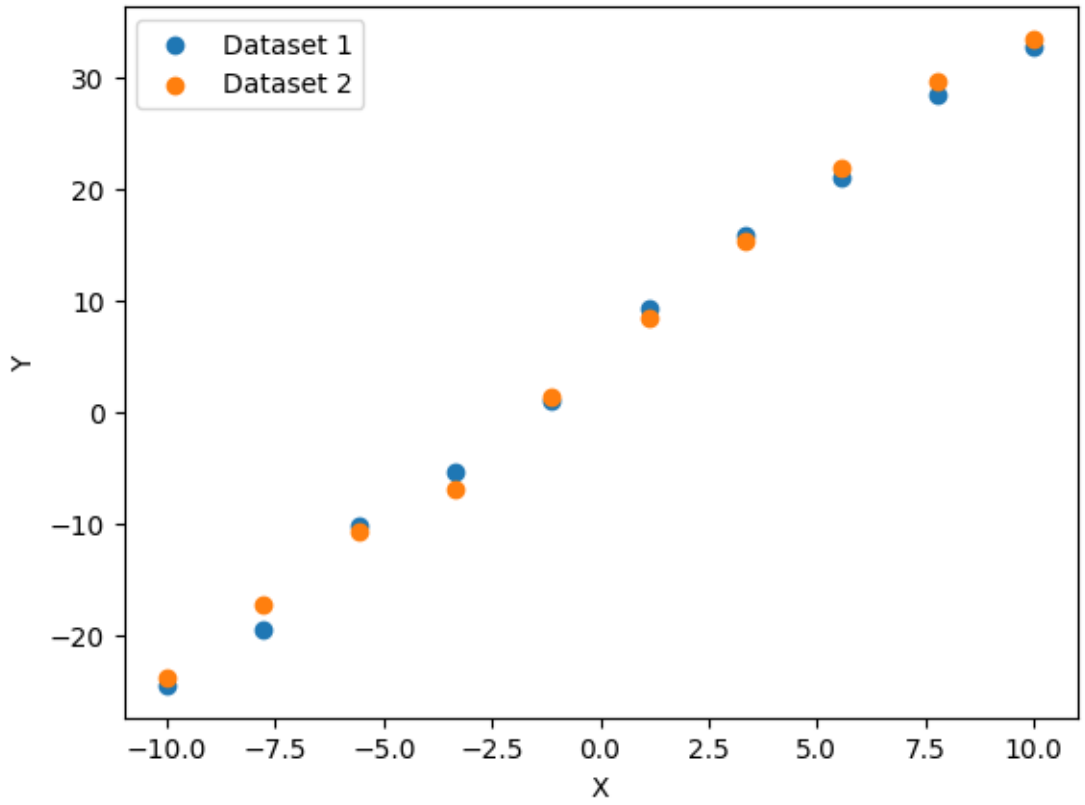
$$s^2 = \frac{\sum_{i=1}^{i=n}(x_i - \bar{x})^2}{n-1}$$

$\bar{x}$ 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 $Y = 5 + 3.X + \varepsilon$. $\varepsilon \sim \mathcal{N}(0, 1)$

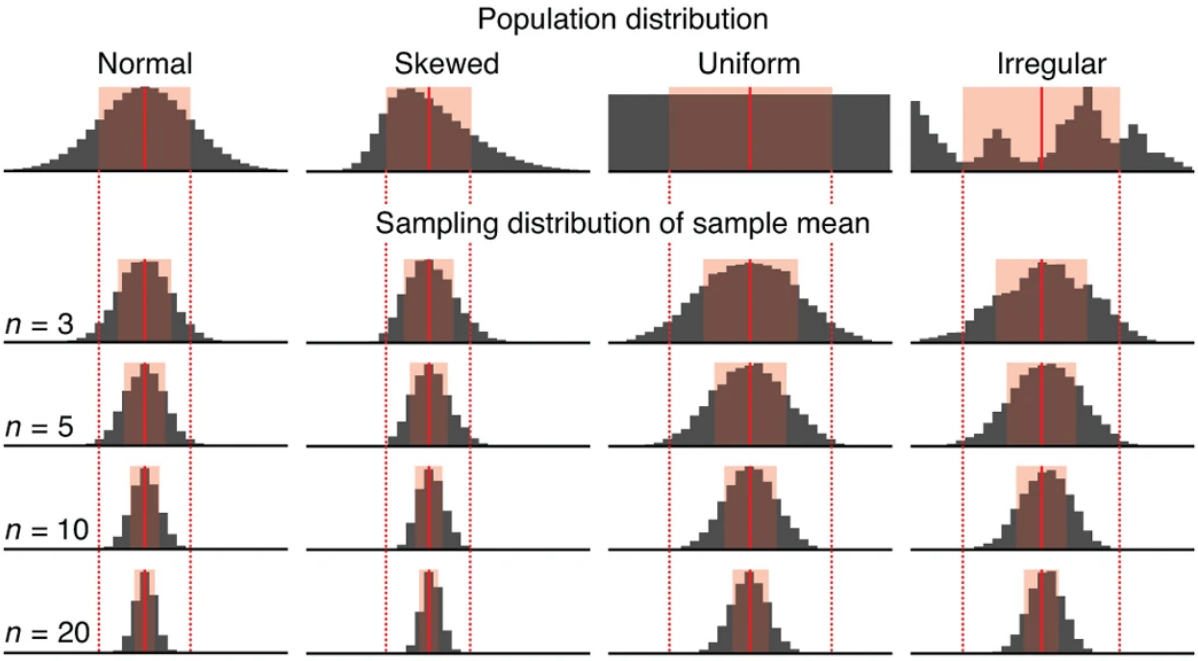
- Both datasets are similar and have a linear relationship.
- They have a similar slope and intercept.
- There is a small difference between each data point of either dataset due to the error term ε.
- The value of error term ε is different for each point as they are drawn independently from the same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

4.b

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

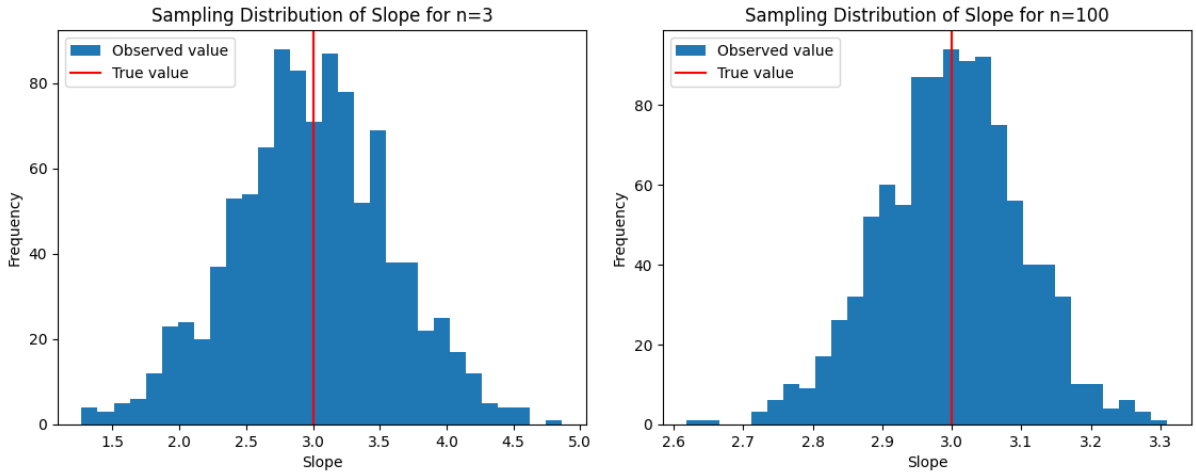
- The outputs are slightly different between the two datasets due to the random noise ε in generating the data.
- The value of error term is different for each dataset as they are drawn independently.
- This randomness causes the estimated slope to deviate slightly from the true value of 3 in each dataset.

5.a



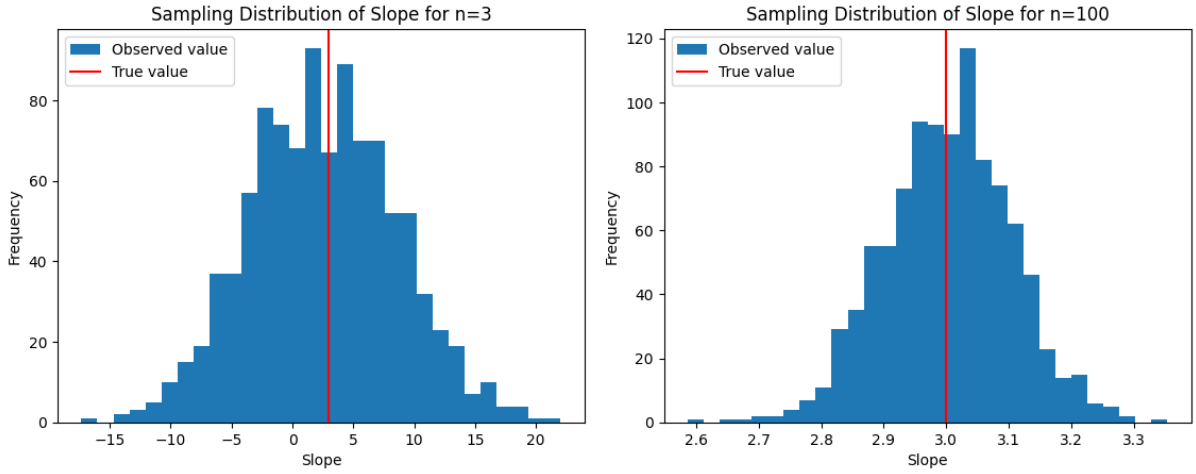
- From the figure above, we can see that the distribution of sample mean is approximately normal regardless of the population distribution.
- Between sample sizes of 3 and 5, we can see that the sample distribution is approximately normal (somewhat retains the shape of the population distribution).
- 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).

5.e



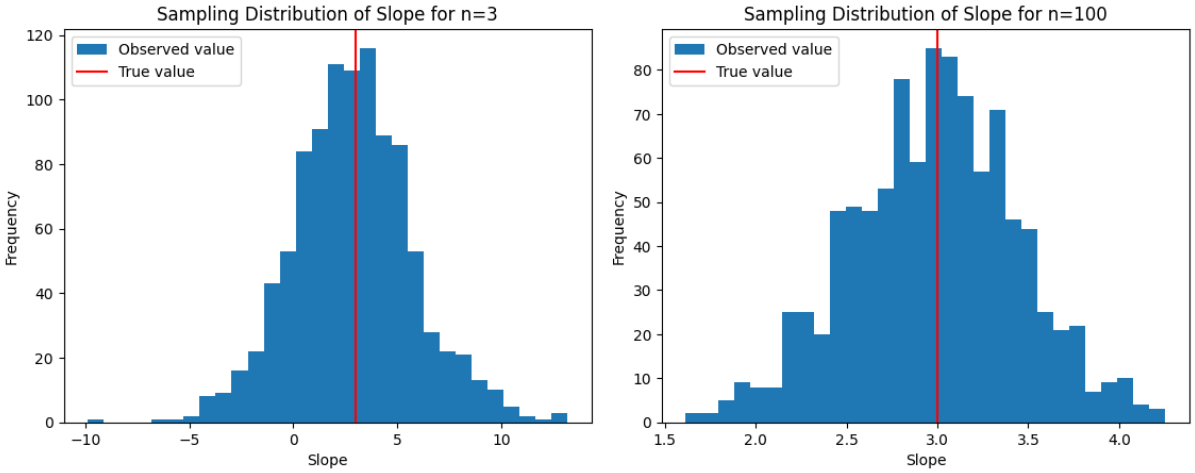
Plot from data sampled with $n = 3$ vs $n = 100$ using normal error term $\varepsilon \sim \mathcal{N}(0, 1)$.

- We can see that the distribution with $n = 100$ is much narrower than the distribution with $n = 3$.
- We see a higher variance in the mean with $n = 3$ because denominator is a smaller number (smaller sample size).
- As the sample size increases, the denominator increases and the variance decreases (because the numerator does not grow by as much).



Plot from data sampled with $n = 3$ vs $n = 100$ using normal error term $\varepsilon \sim \mathcal{N}(0, 100)$.

- We can see that the distribution with $n = 3$ and $\varepsilon \sim \mathcal{N}(0, 100)$ has a wider distribution compared to $\varepsilon \sim \mathcal{N}(0, 1)$ for same data size.
- This is because of the huge variance in the error term.
- But as the sample size increases to 100, we can see that the variance has decreased drastically and is similar to the distribution with $\varepsilon \sim \mathcal{N}(0, 1)$ for the same sample size.
- Even with a high variance in the error term, the mean of 100 numbers is closer to the population mean and the variance decreases as the sample size increases.



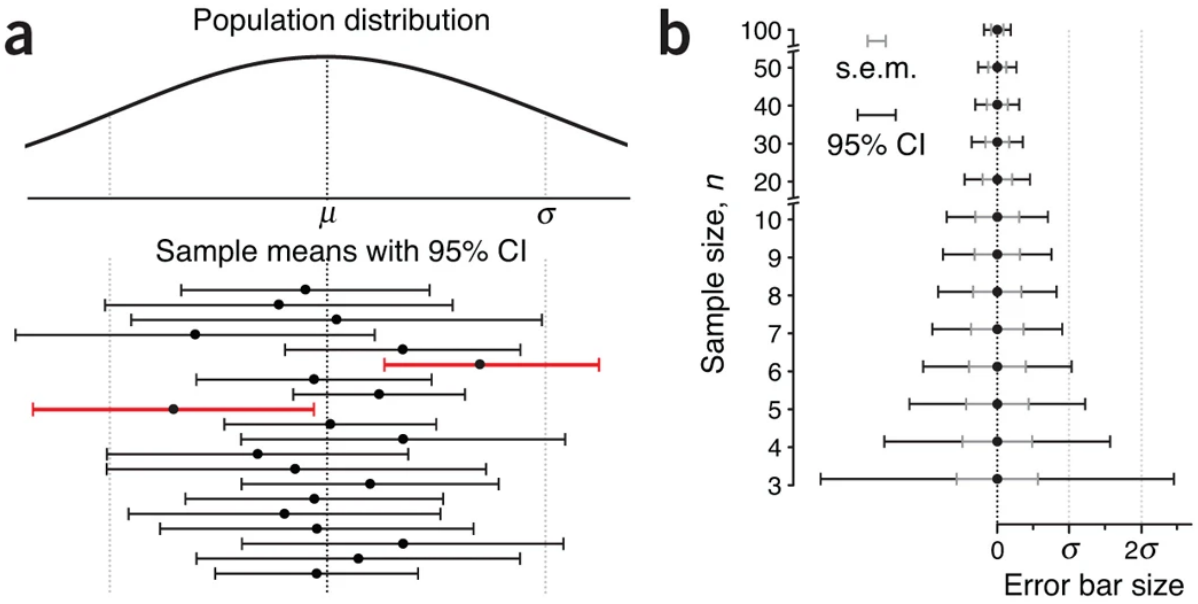
Plot from data sampled with $n = 3$ vs $n = 100$ using Chi Squared error term $\varepsilon \sim \chi^2(10)$.

- We can see that the distribution with $n = 100$ is narrower compared to the distribution with $n = 3$.
- Both of these distributions are approximately normal despite the population distribution being Chi Squared.
- This is because of the Central Limit Theorem which states that the distribution of sample means is approximately normal regardless of the population distribution.

Conclusion

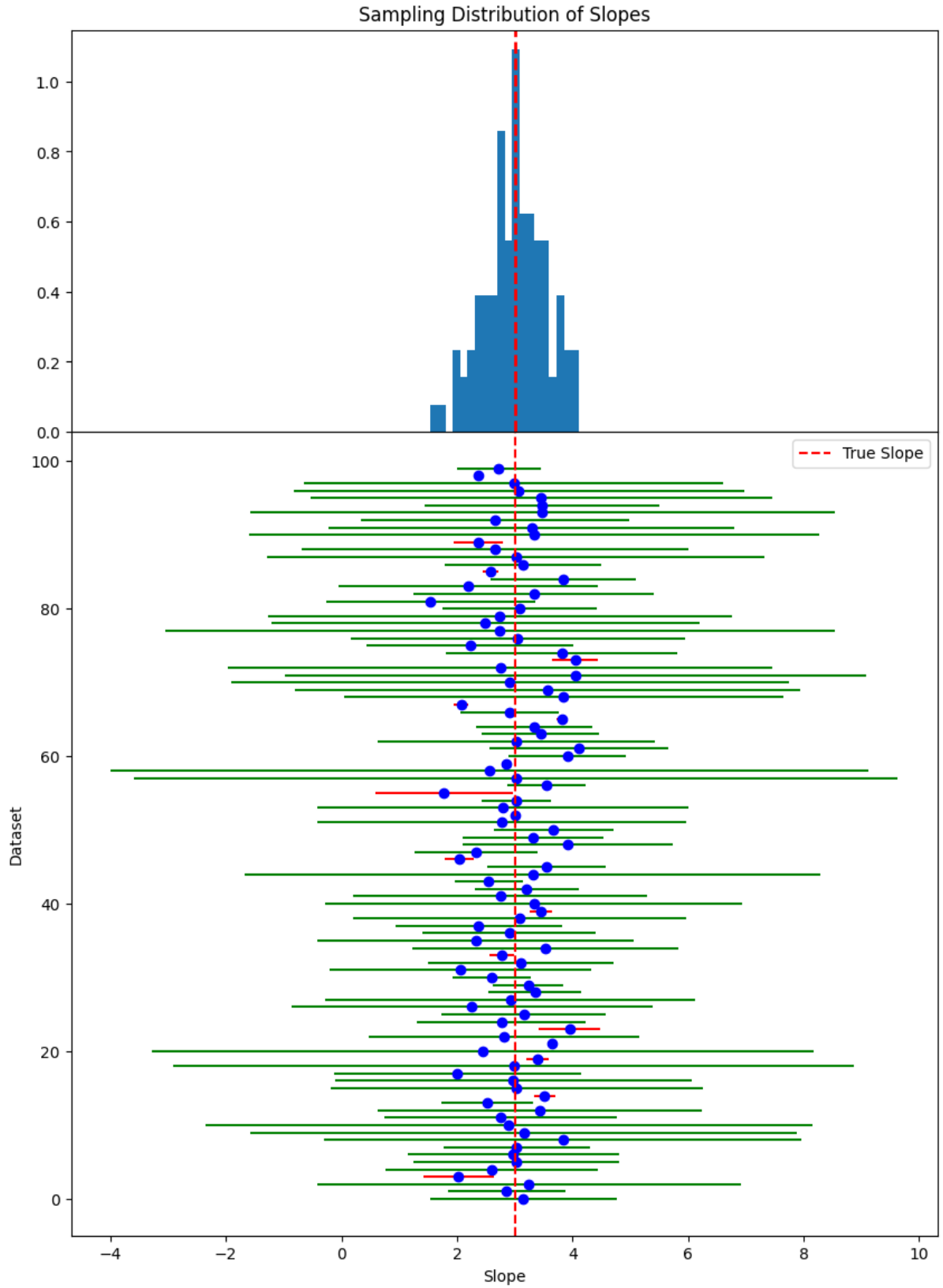
- In summary, the distribution with $n = 3$ and $\varepsilon \sim \mathcal{N}(0, 100)$ has the widest distribution due to a high variance in the error term and a small sample size.
- As the sample size increases, the variance decreases and the distribution is approximately normal.

6.a



- In the figure above (a), we can see a plot of means from 20 datasets sampled from the same distribution, along with their confidence interval.
- Each dot on the line represents the mean of one of the 20 datasets and the horizontal line around each mean represents the 95% confidence interval for that sample mean.
- A confidence interval of 95% means on an average 95% of the sample means will capture the population mean in their confidence interval.
- In the sample above, we can see that only 2 out of 20 intervals fail to capture the population mean (as expected).

6.b

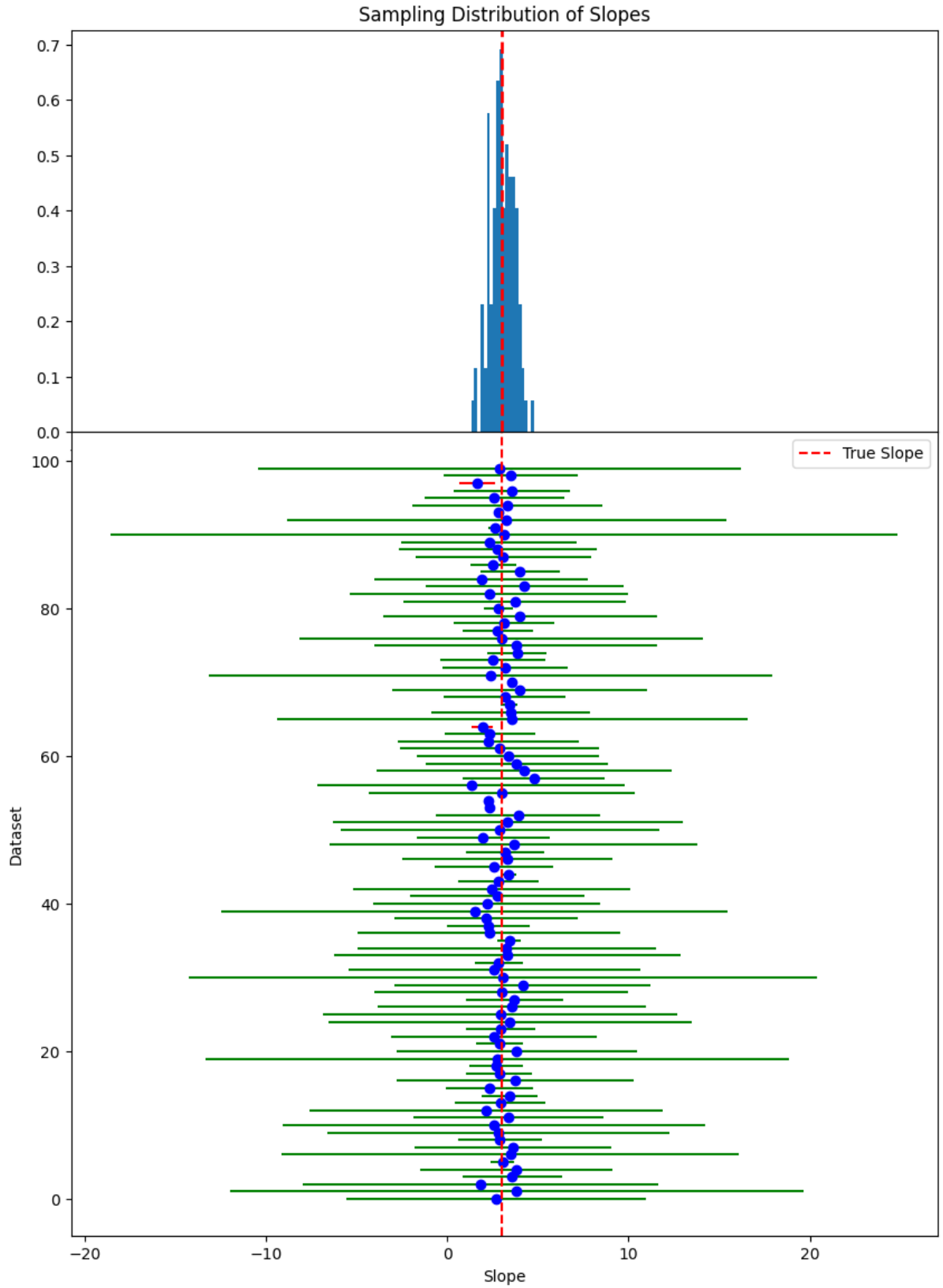


Plot for data sampled with n = 3 and confidence interval of 95%.

- From the plot above, we can see that not all the confidence intervals capture the true slope.

- This is expected, as we set the confidence interval to 95%.
- This means that on an average, 5% of the intervals will not capture the true slope.
- By setting a higher confidence interval, we can reduce the number of intervals that fail to capture the true slope - but this will result in wider confidence intervals.

6.c

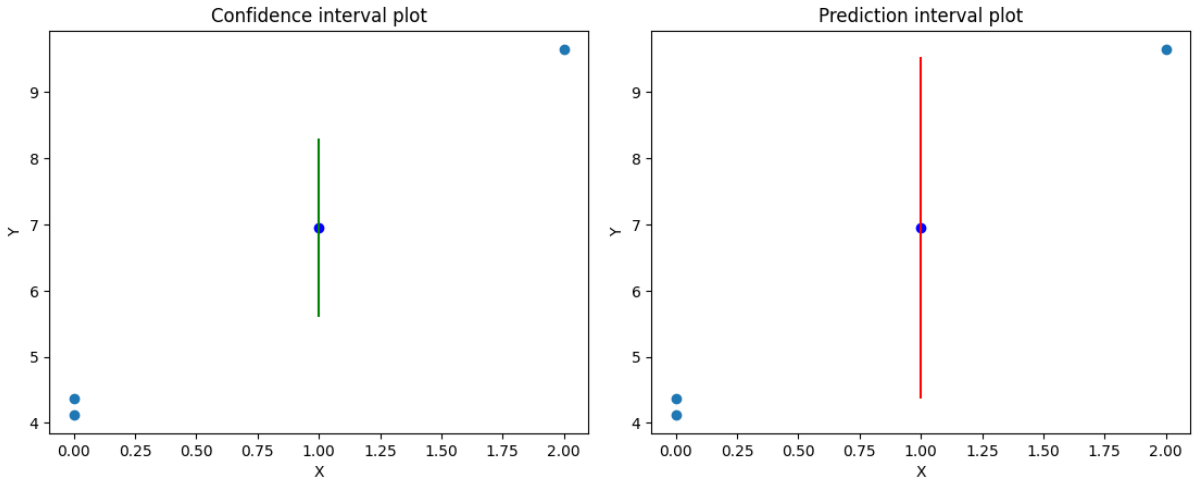


Plot for data sampled with n = 3 and confidence interval of 99%.

- From the plot we can see that the confidence intervals are wider than the previous plot. (Ranging from -20 to 20 as opposed to -5 to 10 in the previous plot).

- This is because we set a higher confidence interval of 99%. This means on an average, only 1% of the intervals will not capture the true slope.
- By setting a higher confidence interval, we reduced the number of intervals that do not capture the slope, but the intervals are huge and do not provide much information.

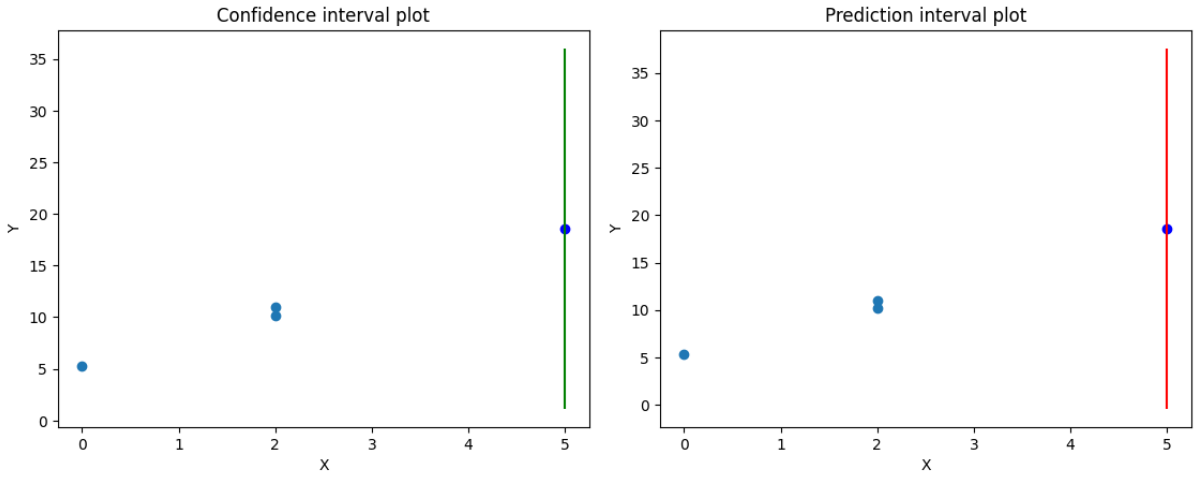
6.d



The plot above shows the 95% confidence interval and prediction interval for the mean $E\{Y_h|X_h = 1\}$.
As we can see, the prediction interval is wider than the confidence interval.

- The confidence interval estimates the range within which the population mean (average) of the response variable (Y) is likely to fall with a certain level of confidence.
- The width of the confidence interval is primarily determined by the sample size and the variability of the data (residuals). As sample size increases, the width of the confidence interval decreases.
- The prediction interval for individual observations (PI) estimates the range within which an individual new observation (Yh) is likely to fall, given a specific value of X.
- Prediction interval accounts for both the uncertainty in estimating the mean (as in the CI) and the variability of individual data points around the mean. Hence it has a wider range.

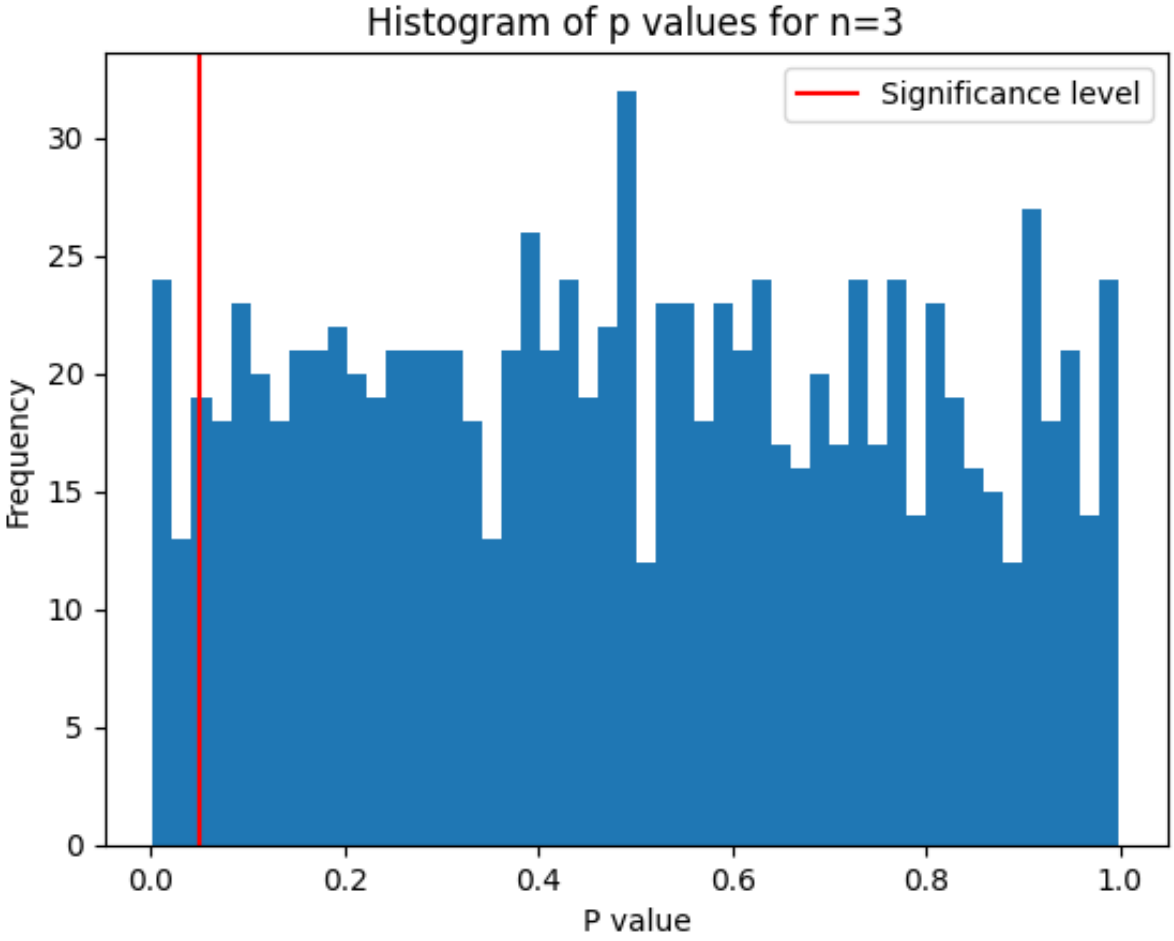
6.e



The plot above shows the 95% confidence interval and prediction interval for the mean $E\{Y_h|X_h = 5\}$.

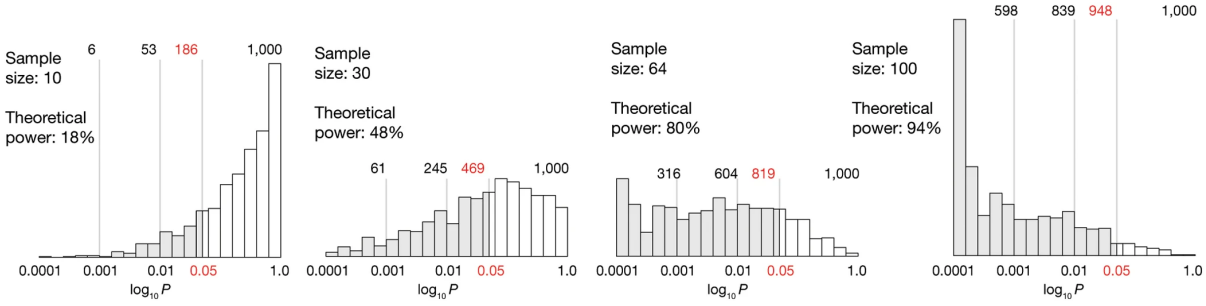
- The confidence interval and prediction interval for $X_h = 5$ are wider than the intervals for $X_h = 1$.
- When we extrapolate to new X values far from the observed range, the uncertainty in estimating the mean and prediction intervals increases.
- This is because model accounts for the variability in mean and individual observations - which becomes significant as we extrapolate.

7.a



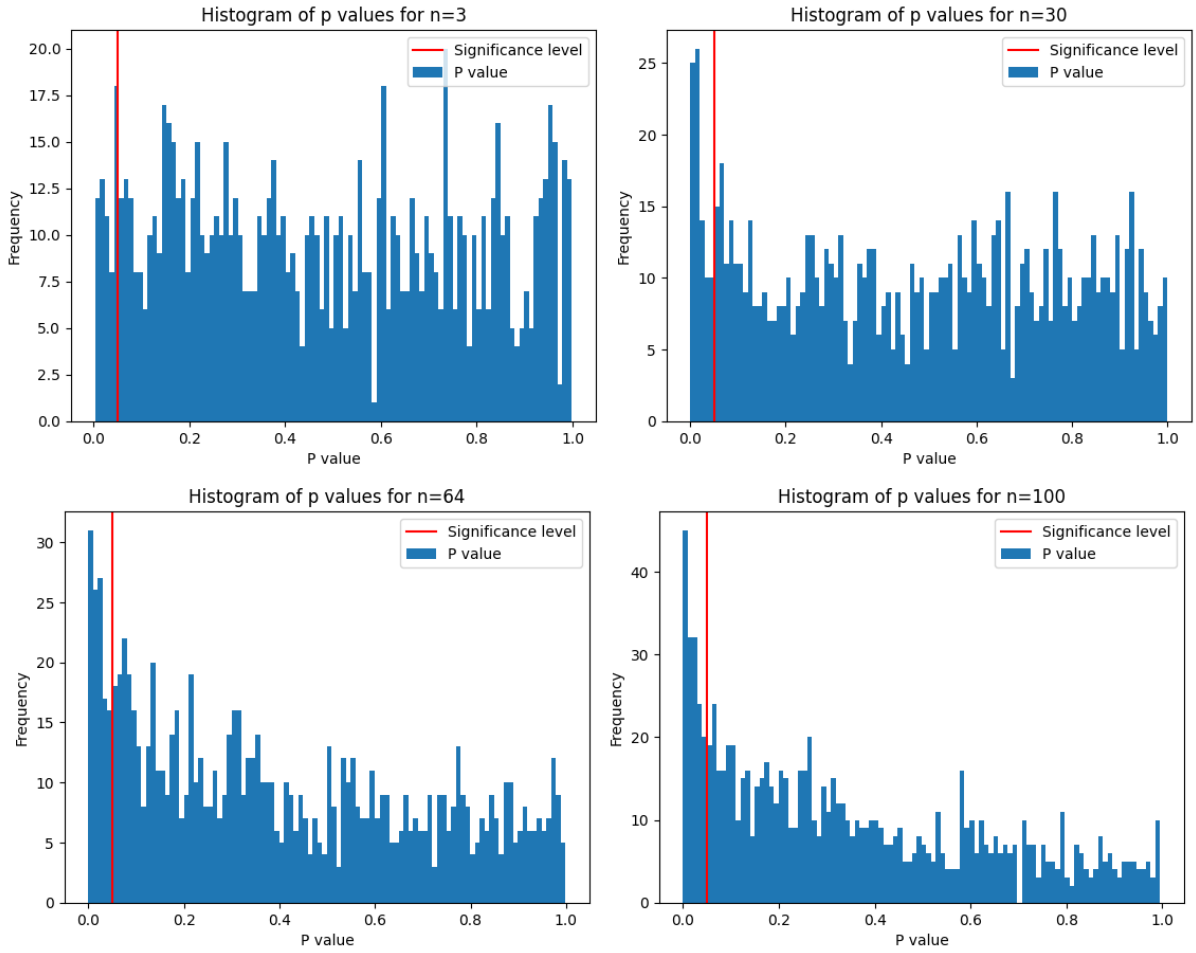
- In the plot above, we can see a histogram of p-values for null hypothesis $H_0 : \beta_1 = 0$.
- The data has been sampled from a distribution with a slope of $\beta_1 = 0$.
- Since the true slope is zero H_0 is true, we see a uniform distribution of p-values between 0 and 1.
- This is because under the null hypothesis, there is no true effect of X on Y, and any observed associations are due to random sampling variability.

7.b



- The plot above shows the importance of sample size in hypothesis testing and p-values of small sample sizes can be misleading.
- For the sample size of 10, The histogram of p-values shows that most p-values are skewed towards 1.0, indicating that the test often fails to reject the null hypothesis (no significant difference between the samples).
- As sample size increases, we can see that the histogram is more uniform indicating that larger samples are more likely to detect true differences and produce lower P-values
- As we reach a sample size of 100, we can see that H_0 is rejected. This suggests that with a larger sample size, the test not only detects differences but also provides highly significant results.
- In practical terms, this means that larger sample sizes increase the power of your statistical tests and improve your ability to detect true effects or differences when they exist.

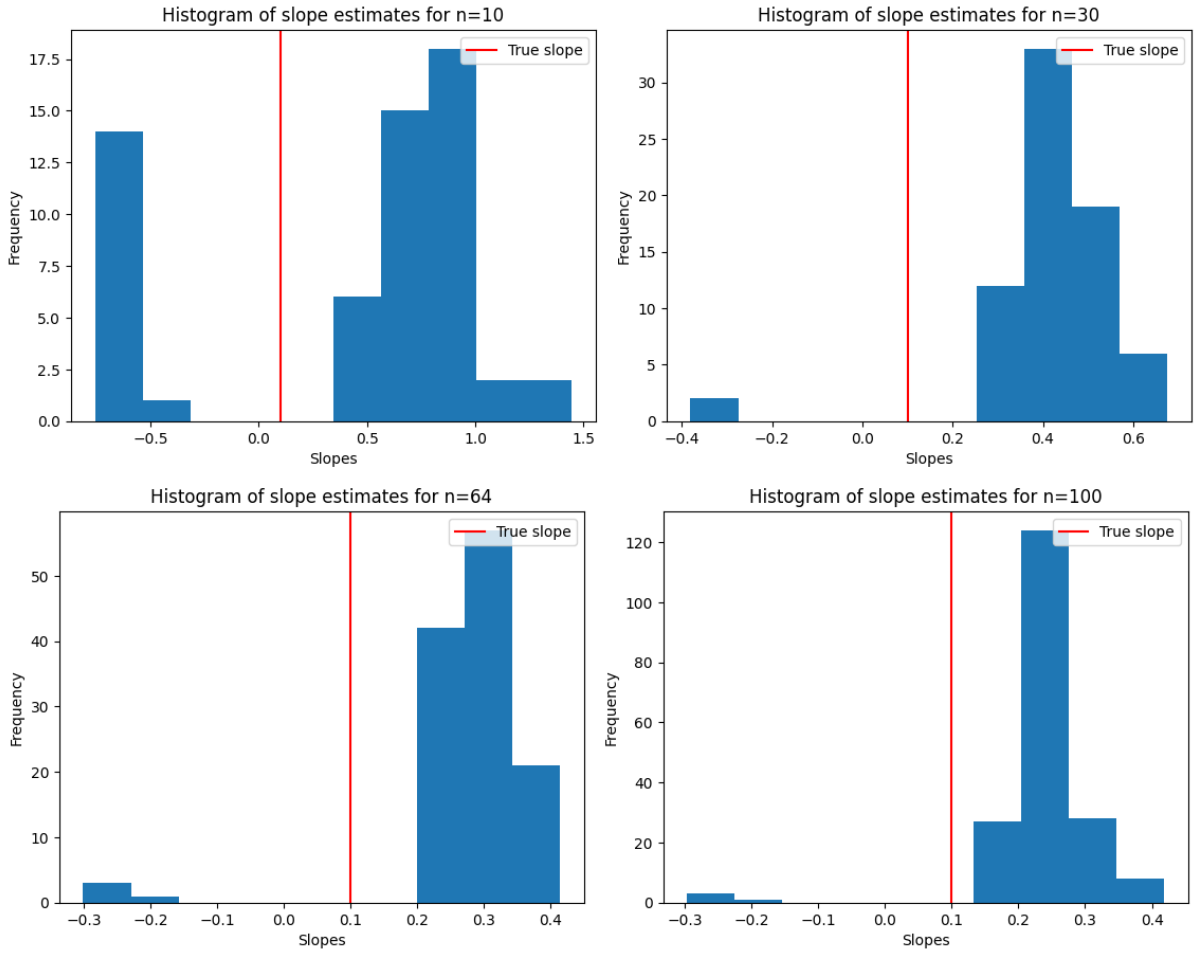
7.c



Above is a plot of p-values for 1000 studies with varying sample sizes, sampled from the same distribution $5 + 0.1.X + \varepsilon$. Where $\varepsilon \sim \mathcal{N}(0, 1)$.

- From the plots above we can see that p-values of hypothesis tests do not provide accurate information for small sample sizes. Even though the true slope is 0.1, the hypothesis test does not seem to reject the null hypothesis.
- As a result, if we were to reproduce a study with a small sample size, we would get a different p-value and a different conclusion.
- By increasing the sample sizes, we can get more accurate p-values that lead to consistent conclusions.
- From the plots above, we can deduce that having a sample size of 30 or more will lead to consistent conclusions over multiple studies.

7.d



Above is a plot of estimated slopes for 1000 studies with varying sample sizes, sampled from the same distribution $5 + 0.1.X + \varepsilon$. Where $\varepsilon \sim \mathcal{N}(0, 1)$.
Estimated slopes have been plotted only for significant hypothesis tests. $(p{-}value < 0.05)$

- From the plots above, we can see the estimated slopes are not closer to the true slope. Since the true slope is much smaller than the variance of the random term, we cannot estimate the population slope accurately.
- Hence, it is not ideal to plot only the significant hypothesis tests as they do not provide accurate estimation of the population parameters.