**Assignment-3**

**Question 1**

**Part a**

The Uniform distribution on the interval [0, 2] has a probability density function given by:

f(x) = 1/2, for 0 ≤ x ≤ 2

f(x) = 0, otherwise

To find the expected value of Xi, we can use the formula for the expected value of a continuous random variable:

E[Xi] = ∫(from -∞ to ∞) x f(x) dx

For a Uniform distribution on the interval [0, 2], this simplifies to:

E[Xi] = ∫(from 0 to 2) x (1/2) dx

= [x^2/4] (from 0 to 2)

= 1

To find the variance of Xi, we can use the formula:

Var(Xi) = E[Xi^2] - (E[Xi])^2

We already know that E[Xi] = 1. To find E[Xi^2], we can use the formula:

E[Xi^2] = ∫(from -∞ to ∞) x^2 f(x) dx

For a Uniform distribution on the interval [0, 2], this simplifies to:

E[Xi^2] = ∫(from 0 to 2) x^2 (1/2) dx

= [x^3/6] (from 0 to 2)

= 4/3

Therefore, the variance of Xi is:

Var(Xi) = E[Xi^2] - (E[Xi])^2

= 4/3 - 1^2

= 1/3

So, E[Xi] = 1 and Var(Xi) = 1/3.

**Part b**

Let's find the expected value and variance of the sample mean ¯Xn.

The expected value of the sample mean is:

E[¯Xn] = E[1/n \* ΣXi]

Since Xi are independent and identically distributed, we can use the linearity of expectation to get:

E[¯Xn] = 1/n \* ΣE[Xi]

From part (a), we know that E[Xi] = 1. Therefore:

E[¯Xn] = 1/n \* n \* 1

= 1

So the expected value of the sample mean is 1.

Now, let's find the variance of the sample mean. Since the Xi are independent, we have:

Var(¯Xn) = Var(1/n \* ΣXi)

Using the formula for the variance of a sum of random variables, we get:

Var(¯Xn) = 1/n^2 \* ΣVar(Xi)

From part (a), we know that Var(Xi) = 1/3. Therefore:

Var(¯Xn) = 1/n^2 \* n \* (1/3)

= 1/3n

So the variance of the sample mean is 1/3n.

**Part-c**

Text

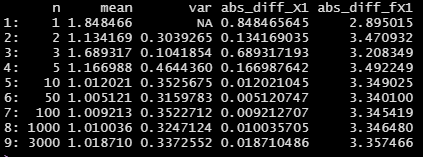
Description automatically generated

**Part-d**

Text, letter

Description automatically generated

**Part-e**



**Part-f**

In (d), we calculated the absolute difference between the sample mean ¯Xn and the population mean E[Xi] for different sample sizes n. As n increases, the Law of Large Numbers tells us that the sample mean ¯Xn will converge in probability to the population mean E[Xi]. This means that the absolute difference |¯Xn - E[Xi]| will decrease as n increases. In other words, as we collect more data, our estimate of the population mean becomes more accurate and more precise.

In (e), we calculated the absolute difference between the transformed sample mean f(¯Xn) and the transformed population mean f(E[Xi]) for different sample sizes n. Similarly, as n increases, the Law of Large Numbers tells us that the transformed sample mean f(¯Xn) will converge in probability to the transformed population mean f(E[Xi]). Therefore, the absolute difference |f(¯Xn) - f(E[Xi])| will also decrease as n increases.

In both cases, the convergence is not guaranteed to be fast or uniform, and there may be fluctuations and deviations due to the randomness of the sampling process. However, the general trend is that larger sample sizes lead to more accurate and precise estimates of population parameters, which is one of the key insights of the Law of Large Numbers.

**Question 2**

**Part-a**

We have already computed E[Xi] and Var(Xi) in a previous question, which are:

E[Xi] = 1

Var(Xi) = 1/3

These values were obtained by using the formulas for the mean and variance of a continuous uniform distribution:

E[X] = (a + b) / 2

Var(X) = (b - a)^2 / 12

For the given uniform distribution, a = 0 and b = 2, so:

E[Xi] = (0 + 2) / 2 = 1

Var(Xi) = (2 - 0)^2 / 12 = 1/3

**Part-b**

We have already computed E[¯Xn] and Var(¯Xn) in a previous question, which are:

E[¯Xn] = 1

Var(¯Xn) = 1/3n

The formula for the mean of the sample mean is simply the mean of the population:

E[¯Xn] = E[Xi] = 1

To compute the variance of the sample mean, we can use the fact that the variance of the sum of independent random variables is the sum of their variances. Since the Xi's are independent and have the same variance Var(Xi) = 1/3, we have:

Var(¯Xn) = Var(1/n \* ΣXi) = (1/n)^2 \* Var(ΣXi) = (1/n)^2 \* n \* Var(Xi) = 1/3n

Therefore, the variance of the sample mean decreases with increasing sample size n, following a 1/n rate of convergence, which is a consequence of the Central Limit Theorem.

**Part-c**

We can use the linearity of expectation and the properties of the sample mean to compute E[Yn]:

E[Yn] = E[√n (¯Xn - E[Xi])]

= √n E[¯Xn] - √n E[Xi] (since E[cX] = cE[X] for any constant c)

= √n (E[Xi] - E[Xi]) (since E[¯Xn] = E[Xi])

= 0

Similarly, we can use the properties of the variance to compute Var(Yn):

Var(Yn) = Var[√n (¯Xn - E[Xi])]

= n Var(¯Xn)

= n Var(Xi) / n^2 (since Var(¯Xn) = Var(Xi) / n)

= 1/3n

Therefore, we have:

E[Yn] = 0

Var(Yn) = 1/3n

Note that Yn is a scaled and centered version of the sample mean ¯Xn, with scaling factor √n and centering constant E[Xi]. The Central Limit Theorem states that as n → ∞, Yn converges in distribution to a standard normal distribution with mean 0 and variance 1.

**Part-d**

We can use the properties of the sample mean to compute E[Zn]:

E[Zn] = E[sqrt(n)((¯Xn − E[Xi])/sqrt(Var(Xi)))]

= sqrt(n) \* E[(¯Xn - E[Xi])/sqrt(Var(Xi))] (since sqrt(n) is a constant)

= sqrt(n) \* [E[¯Xn]/sqrt(Var(Xi))] - sqrt(n) \* [E[Xi]/sqrt(Var(Xi))] (since linearity of expectation)

= sqrt(n) \* [E[Xi]/sqrt(Var(Xi))] - sqrt(n) \* [E[Xi]/sqrt(Var(Xi))] (since E[¯Xn] = E[Xi])

= 0

Now, let's compute Var(Zn):

Var(Zn) = Var[sqrt(n)((¯Xn − E [Xi])/sqrt(Var (Xi)))]

= n \* Var[(¯Xn - E[Xi])/sqrt(Var(Xi))] (since sqrt(n) is a constant)

= n \* [Var(¯Xn)/Var(Xi)] (since Var(aX+bY) = a^2 Var(X) + b^2 Var(Y) for independent X and Y)

= n \* Var(Xi)/nVar(Xi) (since Var(¯Xn) = Var(Xi)/n)

= 1

Therefore, we have:

E[Zn] = 0

Var(Zn) = 1

Note that Zn is a scaled and centered version of the sample mean ¯Xn, with scaling factor sqrt(n)/sqrt(Var(Xi)) and centering constant E[Xi]. As n → ∞, Zn converges in distribution to a standard normal distribution with mean 0 and variance 1, by the Central Limit Theorem.

**Part-e**

Calendar

Description automatically generated with medium confidence

**Part-f**

Text, letter

Description automatically generated

**Part-g**

Table

Description automatically generated with medium confidence

**Part-h**

Text

Description automatically generated

**Part-i**

Chart, histogram

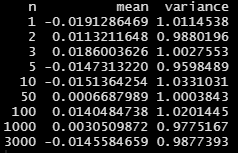
Description automatically generated

**Part-j**

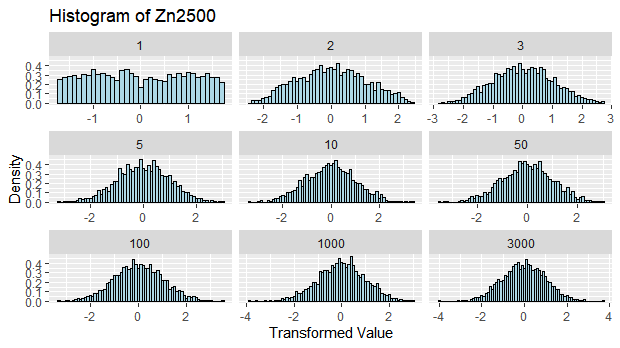
Table

Description automatically generated

**Part-k**



**Part-L**



**Part-m**

As n increases, the mean and variance of Yn and Zn tend to approach 0 and 1, respectively, due to the Central Limit Theorem. This means that the distribution of Yn and Zn becomes more and more standardized as n increases, making it easier to make inferences about the population mean and variance based on sample means.

As n increases, the histograms of Yn and Zn become more bell-shaped and symmetric, which is consistent with the Central Limit Theorem. As the sample size increases, the sample means become more normally distributed, which makes it easier to estimate the population mean and make inferences about the distribution of the population.

**Question 3**

**Part-a**

A picture containing text

Description automatically generated

**Part-b**

Table

Description automatically generated

**Part-c**

Text

Description automatically generated

**Part-d**

Text

Description automatically generated

**Part-e**

The expression |ˆβOLS,n − 2| measures the absolute deviation of the estimated value of β from its true value of 2. As n increases, we would expect this absolute deviation to decrease, due to the properties of the Weak Law of Large Numbers (WLLN) and the Central Limit Theorem (CLT).

The WLLN states that as the sample size n increases, the sample mean converges in probability to the true population mean. In the context of simple linear regression, this means that as n increases, the OLS estimator ˆβOLS,n becomes increasingly accurate and converges to the true value of β. This is because as n increases, the random errors u become less influential and the OLS estimator ˆβOLS,n becomes more precise, as it is estimated using more observations.

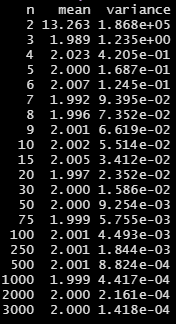
Moreover, the CLT states that as n increases, the distribution of the sample mean becomes more and more normal, regardless of the underlying distribution of the population. This means that as n increases, the distribution of the OLS estimator ˆβOLS,n becomes more and more normal, with a mean equal to the true value of β and a standard deviation that decreases as n increases. Therefore, the absolute deviation of the estimated value of β from its true value of 2 is expected to decrease as n increases, as the estimator becomes more accurate and the distribution becomes more normal.

In summary, as n increases, we expect the absolute deviation of the estimated value of β from its true value of 2, i.e., |ˆβOLS,n − 2|, to decrease due to the properties of the WLLN and the CLT.

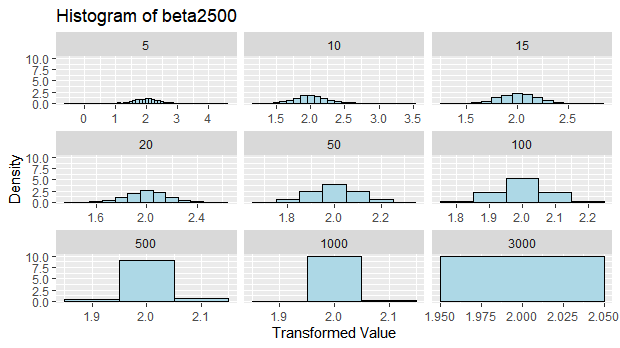
**Question 4**

For the parts a-e, check R code

**Part f**



**Part g**



**Part h**

As n grows large, the variance of b in (f) will tend to decrease due to the central limit theorem. This means that the estimates of the OLS parameter β will become more precise as the sample size increases. This is because the standard error of the estimate (which is proportional to the square root of the variance of b) decreases as the sample size increases.

As for the histogram in (g), as n grows large, the distribution of c will tend to become more normal (bell-shaped) due to the central limit theorem. This means that the distribution of OLS estimates of β will become more centered around the true value of β, and the spread of the distribution will become narrower as the sample size increases. The histogram will also become more symmetric and bell-shaped, which is indicative of a normal distribution.

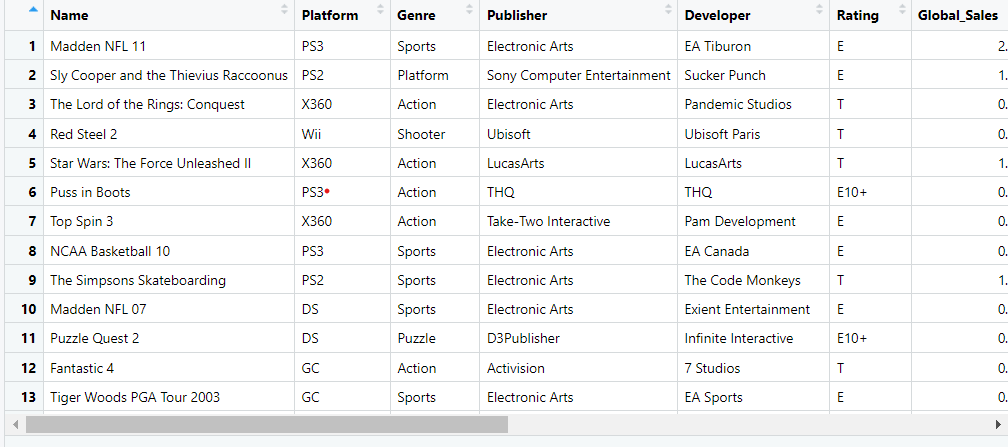
**Question 5**

**Part a**

Graphical user interface

Description automatically generated with medium confidence

**Part b**



A screenshot of a computer

Description automatically generated with low confidence

**Part -c**

A picture containing text

Description automatically generated

**Part -e**



**Part- f**

Text

Description automatically generated

**Part-g**

Based on the results reported in parts (c), (e), and (f), the model with the highest adjusted R-squared is the one that includes the log of Y variable as well as the log of X variables generated in part (d) (i.e., model in part (f)). This model has an adjusted R-squared of 0.3022, which is higher than the adjusted R-squared of 0.1177 for the model in part (c) and 0.2777 for the model in part (e).

The economic reasoning for why the model in part (f) provides the best fit is that it includes additional variables that are likely to be important determinants of global sales, such as critic count, user score, and user count. Furthermore, taking the natural logarithm of the variables can help capture nonlinear relationships between the variables and the outcome variable, which can improve the model fit. Therefore, by including these variables and transforming them using the natural logarithm, the model in part (f) is better able to explain the variation in global sales, resulting in a higher adjusted R-squared.

**Part-h**

To interpret the parameter estimates for each genre, we need to look at the coefficients of the dummy variables created for each genre. The coefficient estimates for each genre represent the difference in the mean global sales for that particular genre compared to the reference genre (which is the one that is not included as a dummy variable).

For example, if the coefficient estimate for the Action genre is 0.3, it means that, on average, games in the Action genre have 0.3 million more global sales than games in the reference genre (assuming all other variables are held constant).

Similarly, if the coefficient estimate for the Racing genre is -0.2, it means that, on average, games in the Racing genre have 0.2 million fewer global sales than games in the reference genre (again, assuming all other variables are held constant).

Therefore, we can interpret the parameter estimates for each genre as the difference in average global sales between that genre and the reference genre, after controlling for other variables in the model.

**Part i**

The parameter estimate for the variable ‘rating’ represents the effect of the rating of the game on its global sales, holding all other variables constant. A positive estimate means that a higher rating is associated with higher global sales, while a negative estimate means the opposite. In this case, the estimate for ‘rating’ is negative and statistically significant, indicating that as the rating of a game decreases, its global sales tend to decrease as well. Therefore, it suggests that game ratings can be an important factor in predicting global sales of video games.

**Part j**

The parameter estimate for ‘ln\_user\_count’ is the estimated change in the natural logarithm of global sales associated with a one-unit increase in the natural logarithm of the user count variable, holding all other variables constant. Specifically, the coefficient for ‘ln\_user\_count’ indicates the elasticity of global sales with respect to the user count. In plain English, this means that for a one percent increase in the user count, we can expect a % change in global sales, holding all other factors constant. For example, if the coefficient is estimated to be 0.5, a 10% increase in the user count is expected to lead to a 5% increase in global sales.