DISTRIBUTION

QUESTION-1: Simulate 30 rolls with =RANDBETWEEN(1,6). What is the probability of rolling a 3 exactly 5 times? (Hint: Use BINOM.DIST)

-> To find the probability of rolling a 3 exactly 5 times in 30 rolls, where each roll has a probability
p=$\frac{1}{6}$, Use the binomial distribution:

=BINOM.DIST(5, 30, 1/6, FALSE)

Result

The probability is approximately:

0.1921 (about **19.21%**)

This means there is about a **19% chance** of getting exactly five 3s in 30 rolls.

QUESTION-2: Generate 100 values in Excel using the continuous uniform distribution RAND() and plot a histogram. Describe the shape of the distribution.

-> If you generate 100 values in Excel using the continuous uniform distribution with:
=RAND()
and then plot a histogram, here's what you should expect:
Shape of the Distribution
A histogram of 100 RAND() values will generally show:

- Roughly equal frequency across all bins
- A flat, rectangular shape (because the continuous uniform distribution gives every value between 0 and 1 the same probability)
- Some random variation—with only 100 samples, it won't be perfectly flat, but it should still look fairly even overall

In Words

The distribution is approximately uniform — every interval between 0 and 1 is equally likely, so the histogram looks flat rather than peaked.

If you'd like, I can also generate a sample dataset and show you the histogram using Python.

DISTRIBUTION

QUESTION-3: A dataset has a mean of 50 and a standard deviation of 5. What percentage of values lie between 45 and 55 if the data follows a normal distribution?

-> For a normal distribution:

- Mean μ=50
- Standard deviation σ=5
- Range: 45 to 55

Compute the z-scores:

$$Z = \frac{45-50}{5} = -1$$

$$Z = \frac{55-50}{5} = 1$$

The percentage of values within ±1 standard deviation of the mean in a normal distribution is:
≈ 68%
So about 68% of the values lie between 45 and 55.

QUESTION-4: What is the concept of standardization (z-score), and why is it important in data analysis? Explain the formula and how standardization transforms a dataset.

-> Standardization, also known as computing a z-score, is the process of transforming data so that it has:

- Mean = 0
- Standard deviation = 1

This allows values to be interpreted relative to the distribution they came from.

What is a Z-Score?

A z-score tells you how many standard deviations a value is from the mean.

- A positive z-score → the value is above the mean
- A negative z-score → the value is below the mean
- A z-score of 0 → the value equals the mean

Formula for a $Z = \frac{x-\mu}{\sigma}$

Where: xxx = the original value

μ\muμ = mean of the dataset

σ\sigmaσ = standard deviation

DISTRIBUTION

Why Standardization Is Important

1. Makes variables comparable

Different variables often use different scales (e.g., height in cm vs. weight in kg).
Standardization puts them on the same scale, allowing fair comparison and analysis.

2. Essential for many machine-learning algorithms

Models like:

- K-means clustering
- Logistic regression
- SVM
- PCA
- K-nearest neighbors

use distance or variance; these models perform poorly when features have very different scales.

3. Identifies outliers

Values with $|z| > 3$ are often considered outliers.

4. Simplifies probability calculations

In statistics, once data is standardized, you can use the standard normal distribution table to compute probabilities.

How Standardization Transforms a Dataset

Standardizing each value using the z-score formula results in:

- New mean = 0
- New standard deviation = 1

But the shape of the distribution stays the same (it doesn't change skewness or kurtosis).

Example transformation

Original values:
50, 60, 70 (mean = 60)

Z-scores:

DISTRIBUTION

$\frac{50-60}{100} = -1$, $\frac{60-60}{100} = 0$, $\frac{70-60}{100} = 1$

Summary

Standardization is a key data preprocessing step that:

- Converts values into z-scores
- Centers the distribution at 0
- Scales it to unit variance
- Makes variables comparable
- Helps models converge faster and perform better

If you'd like, I can also show an example using a real dataset or generate one for you.

QUESTION-5: What is Kurtosis and their type?

-> Kurtosis is a statistical measure that describes the “tailedness” or shape of the distribution’s tails compared to a normal distribution.
It helps indicate whether a dataset has more or fewer extreme values (outliers).

Definition of Kurtosis: Kurtosis measures how heavily the tails of a distribution differ from the tails of a normal distribution.

Mathematically, kurtosis is based on the fourth moment of the distribution:

$$\text{Kurtosis} = \frac{\sum(x-\mu)4}{n\sigma 4}$$

Most software reports excess kurtosis:
Excess Kurtosis=Kurtosis−3\text{Excess Kurtosis} = \text{Kurtosis} - 3Excess Kurtosis=Kurtosis−3

- A normal distribution has kurtosis = 3
- And excess kurtosis = 0

Types of Kurtosis

1. Mesokurtic (Normal Kurtosis → K = 3)

- Excess kurtosis = 0
- Tails and peaks are similar to the normal distribution.
- Example: Normal distribution

Interpretation:
Balanced tails (not too heavy, not too light).

DISTRIBUTION

2. Leptokurtic (High Kurtosis → K > 3)

- Excess kurtosis > 0
- Heavier tails than normal
- Sharper peak
- More extreme values (outliers)

Interpretation:
Data are more concentrated near the mean but have more extreme deviations in the tails.

Example distributions:

- t-distribution with low degrees of freedom
- Financial returns

3. Platykurtic (Low Kurtosis → K < 3)

- Excess kurtosis < 0
- Lighter tails
- Flatter peak
- Fewer extreme values

Interpretation:
Data are more evenly spread out with fewer outliers.

Example distributions:

- Uniform distribution
- Some bounded datasets

Summary Table

Type of Kurtosis	Kurtosis (K)	Excess Kurtosis	Shape Characteristics
Mesokurtic	3	0	Normal peak, normal tails
Leptokurtic	> 3	> 0	High peak, heavy tails, more outliers
Platykurtic	< 3	< 0	Flat peak, light tails, fewer outliers

DISTRIBUTION

QUESTION-6: Explain why the uniform distribution is a good model for the outcome of rolling a fair die.

-> A uniform distribution is a good model for the outcome of rolling a fair die because every outcome is equally likely.

Why It Fits the Uniform Distribution

1. All outcomes have equal probability

A fair die has 6 faces:

1, 2, 3, 4, 5, 6

Each face has the same chance of landing face-up:

$$P(1)=P(2)=P(3)=P(4)=P(5)=P(6)=\frac{1}{6}$$

This matches the definition of a discrete uniform distribution, where every value in the set has identical probability

2. No outcome is favored

A fair die is designed so:

- Each face has the same weight
- The shape is symmetrical
- The rolling surface is not biased

Therefore, the die does not systematically favor any number over the others.

3. Theoretical and experimental results align

If you roll a fair die many times:

- The observed frequencies of each face become close to each other
- This supports the assumption of uniformity (Law of Large Numbers)

4. Simplicity and correctness in probability modeling

Using a uniform distribution:

- Makes calculations straightforward
- Matches the physical nature of a fair die
- Enables accurate probability predictions

DISTRIBUTION

Conclusion: The uniform distribution is a good model for rolling a fair die because each outcome is equally likely, making the probabilities constant across all possible results.

QUESTION-7: Use Excel to compute the probability of getting at least 8 successes in 15 trials with success probability 0.5

-> To compute the probability of getting at least 8 successes in 15 trials with p = 0.5 in Excel, you use the BINOM.DIST function.

Excel Formula

We want:

$P(X \geq 8) = P(8) + P(9) + \cdots + P(15)$

Excel can compute this in one step using the cumulative distribution:

=1 - BINOM.DIST(7, 15, 0.5, TRUE)

Explanation:

Explanation:

- BINOM.DIST(7,15,0.5,TRUE) gives $P(X \leq 7)$
- Subtracting from 1 gives $P(X \geq 8)$

Result

If you plug this formula into Excel, you get approximately:

0.5000

So the probability of getting at least 8 successes in 15 trials with p = 0.5 is about 50%.

If you want, I can also show you how to compute this using Python or list the individual binomial probabilities.

DISTRIBUTION

QUESTION-8: How does log transformation help in stabilizing variance and making data more normally distributed?

-> Log transformation is one of the most common and effective techniques in data analysis for dealing with skewed data, unequal variance, and non-normal distributions. Here's how and why it works:

1. Stabilizes Variance (Reduces Heteroscedasticity)

In many datasets, variance increases with the mean.
Examples: income, population, reaction times, biological measurements.

This results in a "fanning out" pattern when you plot the data.

How log transformation helps

The log function compresses large values more strongly than small values:

- Small numbers change very little
- Large numbers shrink significantly

This reduces the spread of high values, making the variance more uniform (homoscedastic).

Example

Before log:
10, 50, 100, 300 → widely spread

After log:
log(10), log(50), log(100), log(300) → spread reduced

2. Reduces Right Skewness

Many real-world variables follow a positively skewed (right-tailed) distribution.

Why right skew happens

- Some values can be extremely large
- Most values are clustered near the lower end

This violates assumptions of tools like regression, ANOVA, and parametric tests.

How log transformation helps

- Logs shrink large values more than small ones
- This brings extreme values closer to the rest
- The distribution becomes more symmetric

DISTRIBUTION

3. Makes Data More Normally Distributed

Because log transformation compresses the long right tail:

- Peak becomes more centered
- Tails become shorter
- Distribution looks more bell-shaped

Many statistical tests require approximate normality.
Log-transforming skewed data often makes them more suitable for:

- Linear regression
- t-tests
- ANOVA
- Machine learning algorithms

4. Converts Multiplicative Relationships to Additive

If a process grows *multiplicatively*, like:

$y=a \cdot bx$
taking logs gives: $\log(y)=\log(a)+x\log(b)$
This linearizes exponential relationships, making modeling easier.

5. Helps Handle Outliers

Right-skewed data often contain very large outliers.
The log function dampens the impact of these extremes, reducing their influence on:

- Means
- Regression coefficients
- Statistical tests