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)
Answer: Each roll of a fair die has probability
p=P(rolling a 3)=1/6p

We want the probability of getting exactly 5 threes in 30 rolls, which follows a binomial distribution:

$$P(X=5)=(30/5)\ (1/6)^{5}\ (56)^{25}$$

Using Excel (as hinted):
=BINOM.DIST(5, 30, 1/6, FALSE)
=0.192

Result:

$$P(X=5)\approx 0.19P$$

Probability ≈ 0.19 (or 19%)

So, there is about a **19% chance** that a 3 appears **exactly 5 times 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.

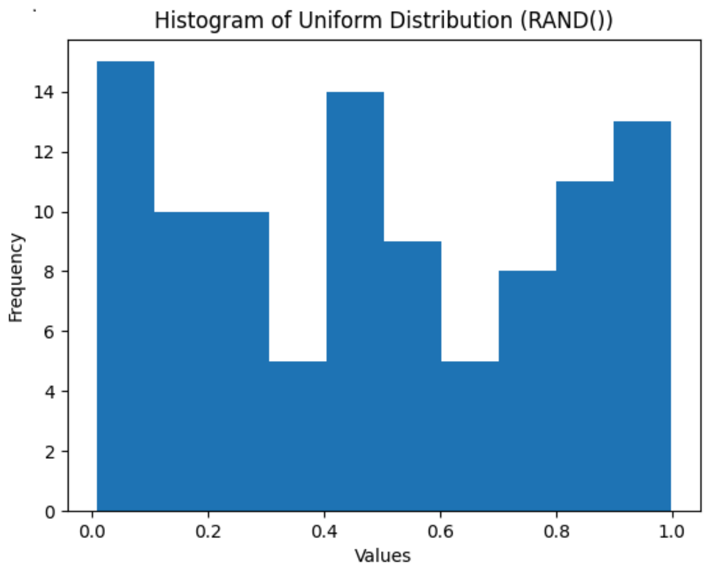
Answer: **Step 1: Generate 100 Uniform(0,1) values**

1. Open Excel.
2. In cell **A1**, type:
   `=RAND ()`
3. Press **Enter**.
4. Click on cell A1 and drag the fill handle down to **A100**.

This generates 100 random values from a **continuous uniform distribution on [0, 1]**.

Distribution

Step 2: Create a histogram

5. 

Step 3: Describe the shape of the distribution

- The histogram should appear **approximately flat (rectangular)**.
- The bars should have **roughly equal heights** across the interval from 0 to 1.
- Small irregularities are normal due to **random sampling** and the relatively small sample size (100 values).

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?

Answer: Since the data follows a **normal distribution**:

- Mean (μ) = 50
- Standard deviation (σ) = 5
- Range 45 to 55 = **μ ± 1σ**

According to the **Empirical Rule (68–95–99.7 rule)**:

- About **68%** of the values lie within **1 standard deviation** of the mean.

Answer: Approximately 68% of the values lie between **45 and 55**.

Distribution

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.

Answer: <u>Concept of Standardization (Z-score)</u>

Standardization is a statistical technique used to convert data values into a common scale called z-scores. A z-score tells us how many standard deviations a data point is away from the mean.

Z-score Formula

$$z = (X - \mu)/\sigma$$

Where:

- (X) = original data value
- (μ) = mean of the dataset
- (σ) = standard deviation of the dataset
- (z) = standardized value (z-score)

How Standardization Transforms a Dataset

After standardization:

- The **mean becomes 0**
- The **standard deviation becomes 1**
- Original values are converted into **unit-free** z-scores

Example: If a value has a z-score of +2, it means the value is 2 standard deviations above the mean.

Why Standardization Is Important in Data Analysis

1. **Comparison Across Different Scales**
   Allows comparison of variables measured in different units (e.g., height vs salary).
2. **Identifying Outliers**
   Very high or low z-scores indicate potential outliers.

Distribution

3. **Improves Model Performance**
   Many machine learning algorithms (e.g., k-means, SVM, linear regression) perform better with standardized data.
4. **Probability & Normal Distribution Analysis**
   Helps in finding probabilities using the **standard normal distribution table**.
5. **Feature Scaling**
   Prevents variables with larger scales from dominating analysis.

Question 5: What is Kurtosis and their type?

Answer: Kurtosis is a statistical measure that describes the shape of a data distribution, specifically how peaked or flat it is compared to a normal distribution. It also indicates the weight of the tails (presence of extreme values).

Types of Kurtosis →

There are **three main types of kurtosis**:

Type	Shape	Tails	Kurtosis Value
Mesokurtic	Normal	Moderate	≈ 3
Leptokurtic	Sharp peak	Heavy	> 3
Platykurtic	Flat	Light	< 3

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

Answer: The **uniform distribution** is a good model for the outcome of rolling a **fair die** because **all possible outcomes are equally likely**.

Explanation

When you roll a fair six-sided die, the possible outcomes are:
{1, 2, 3, 4, 5, 6}

Distribution

For a die to be *fair*, each number has the **same probability** of occurring:
P(1) = P(2) = P(3) = P(4) = P(5) = P(6) =1/6

This exactly matches the definition of a **uniform distribution**, where:

- Every outcome in the sample space has **equal probability**
- No outcome is favored over another

Key Reasons

1. **Equal Likelihood**
   Each face of a fair die has the same chance of appearing.
2. **Discrete and Finite Outcomes**
   The die has a fixed number of outcomes (6), which suits a **discrete uniform distribution**.
3. **No Bias or Weighting**
   A fair die is designed so that shape, weight, and balance do not influence results.

Conclusion

Because all six outcomes are equally probable and independent, the **discrete uniform distribution** is the most appropriate and accurate model for rolling a fair die.

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

Answer: To compute this in **Excel**, we use the **binomial distribution**.

Given:

- Number of trials (n) = **15**
- Probability of success (p) = **0.5**
- We want **at least 8 successes** → ($P(X \geq 8)$)

Excel Formula

In Excel, use:

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

Distribution

Explanation:

- `BINOM.DIST(7,15,0.5,TRUE)` calculates
  $P(X \le 7)$
- Since
  $P(X \ge 8) = 1 - P(X \le 7)$
  we subtract from 1.

Result (Approximate)

The probability is:

$$P(X \ge 8) \approx 0.5$$

Alternative Method (Direct Sum)

You could also calculate:

=BINOM.DIST(8,15,0.5,FALSE)

+ BINOM.DIST(9,15,0.5,FALSE)

+ ...

+ BINOM.DIST(15,15,0.5,FALSE)

But the **first method is faster and recommended**.

Final Answer:

Using Excel, the probability of getting **at least 8 successes in 15 trials** with **p = 0.5** is approximately **0.5**.

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

Answer: Log transformation is a common data-preprocessing technique used to **reduce skewness**, **stabilize variance**, and make data **closer to a normal distribution**, which is important for many statistical methods.

1. What is Log Transformation?

In log transformation, each data value (X) is replaced by:

Distribution

Y=log(X)(or log(X+c) if zeros exist)

Common logs used:

- **Natural log (ln)**
- **Log base 10**
- **Log base 2**

2. How Log Transformation Stabilizes Variance

Problem: Heteroscedasticity

In many datasets, variance **increases with the mean** (e.g., income, sales, population).

How Log Helps

- Log **compresses large values** more than small values
- Reduces the spread of high-value observations
- Makes variance more constant across different levels of data

Example:
Raw data: 10, 100, 1000
Log data: 1, 2, 3

Large gaps become smaller → **variance stabilizes**

3. How Log Transformation Makes Data More Normal

Problem: Right-Skewed Data

Many real-world datasets are **positively skewed** (long right tail).

How Log Helps

- Pulls in extreme high values
- Reduces right skewness
- Makes the distribution more **symmetric and bell-shaped**

Example:
Income or salary data often become more normally distributed after log transformation.

4. Additional Benefits

Distribution

1. **Reduces effect of outliers**
   Extreme values have less influence after log transformation.
2. **Improves model assumptions**
   Helps meet assumptions of:
   - Linear regression
   - ANOVA
   - Correlation analysis
3. **Linearizes relationships**
   Turns exponential relationships into linear ones.

5. When to Use Log Transformation

✅ Data is **right-skewed**
✅ Variance increases with mean
✅ Values are **positive**
❌ Not suitable for zero or negative values (unless adjusted)

Conclusion:

Log transformation compresses large values, stabilizes variance, reduces skewness, and helps data better approximate a normal distribution—making statistical analysis more reliable and accurate.