# STATISTICS WORKSHEET-1 (*SOLUTIONS*)

*Q1 to Q9 have only one correct answer. Choose the correct option to answer your question*.

**1. Bernoulli random variables take (only) the values 1 and 0**.

a) True
b) False

**Solution: (A) True**

**2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?**

a) Central Limit Theorem
b) Central Mean Theorem
c) Centroid Limit Theorem
d) All of the mentioned

**Solution: (A) Central Limit Theorem**

**3. Which of the following is incorrect with respect to use of Poisson distribution?**

a) Modeling event/time data
b) Modeling bounded count data
c) Modeling contingency tables
d) All of the mentioned

**Solution: (B) Modeling bounded count data**

**4. Point out the correct statement.**

a) The exponent of a normally distributed random variables follows what is called the log- normal distribution
b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent
c) The square of a standard normal random variable follows what is called chi-squared distribution
d) All of the mentioned

**Solution: (D) All of the mentioned**

**5. ______ random variables are used to model rates.**

a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned

**Solution: (C) Poisson**

**6. Usually replacing the standard error by its estimated value does change the CLT.**

a) True
b) False

**Solution: (B) False**

**7. Which of the following testing is concerned with making decisions using data?**

a) Probability
b) Hypothesis
c) Causal
d) None of the mentioned

**Solution: (B) Hypothesis**

**8. Normalized data are centered at______and have units equal to standard deviations of the original data.**

a) 0
b) 5
c) 1
d) 10

**Solution: (A) 0**

**9. Which of the following statement is incorrect with respect to outliers?**

a) Outliers can have varying degrees of influence
b) Outliers can be the result of spurious or real processes
c) Outliers cannot conform to the regression relationship
d) None of the mentioned

**Solution: (C) Outliers cannot conform to the regression relationship**

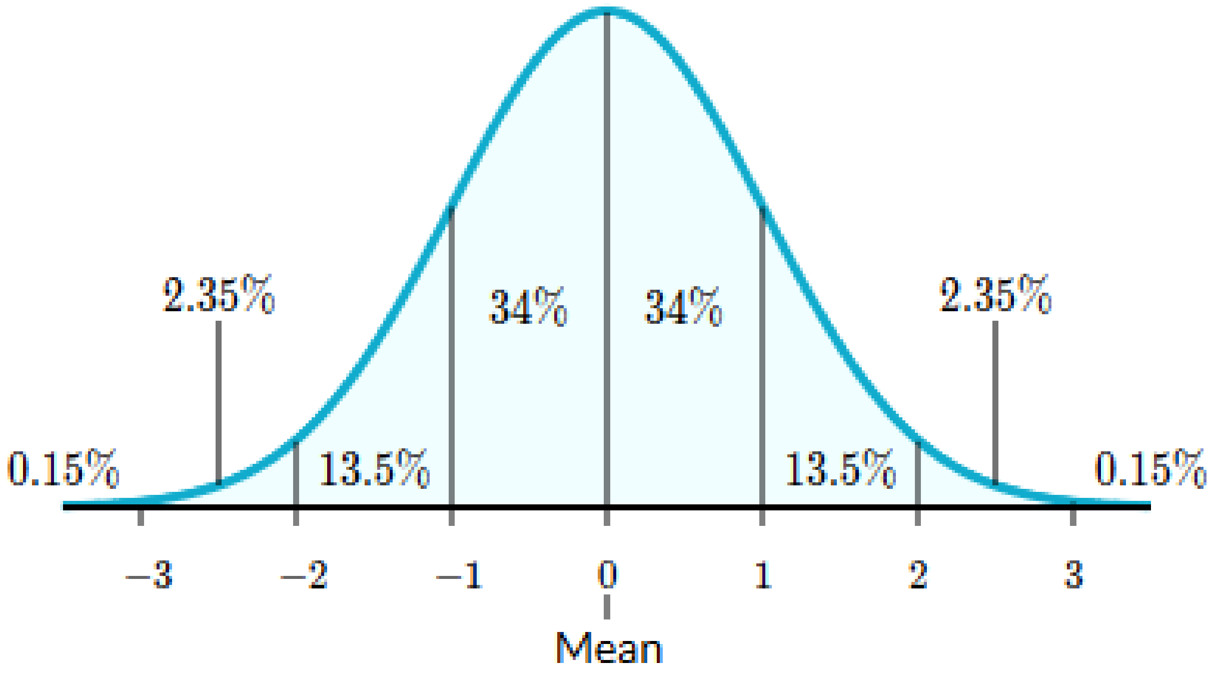
*Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.*

**10. What do you understand by the term Normal Distribution?**

**Answer:**

**Normal Distribution:**

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.



Normal distributions have the following features:

- symmetric bell shape.
- mean and median are equal; both located at the center of the distribution
- ≈68% of the data falls within 1 standard deviation of the mean
- ≈95% of the data falls within 2 standard deviations of the mean
- ≈99.7% of the data falls within 3 standard deviations of the mean

**11. How do you handle missing data? What imputation techniques do you recommend?**

**Answer:**

**Missing data:**

In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.

**Handling Missing data:**

- Method 1 is deleting rows or columns. We usually use this method when it comes to empty cells.
- Method 2 is replacing the missing data with aggregated values.
- Method 3 is creating an unknown category.
- Method 4 is predicting missing values.

**Imputation techniques:**

A common technique is to use the **mean** or **median** of the non-missing observations. This can be useful in cases where the number of missing observations is low. However, for large number of missing values, using mean or median can result in loss of variation in data and it is better to use imputations.

## 12. What is A/B testing?

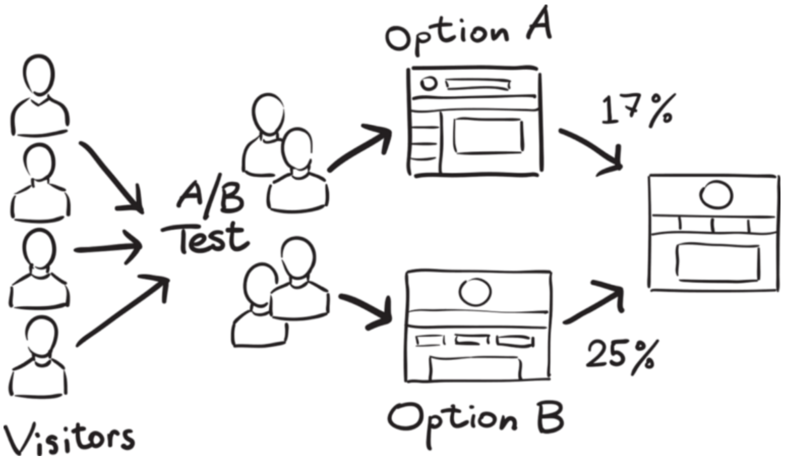
**Answer:**

**A/B testing:**

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

For instance, let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.



It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The **population** refers to all the customers buying your product, while the **sample** refers to the number of customers that participated in the test.

## 13. Is mean imputation of missing data acceptable practice?

**Answer:**

**Mean Imputation:**

The process of replacing null values in a data collection with the data's mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

## 14. What is linear regression in statistics?

**Answer:**

**Linear Regression:**

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

**Simple Linear Regression:**

Simple linear regression is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line. Both variables should be quantitative.

The formula for a simple linear regression is:

$$y = \beta_0 + \beta_1 X + \varepsilon$$

- **y** is the predicted value of the dependent variable (**y**) for any given value of the independent variable (**x**).
- **$B_0$** is the **intercept**, the predicted value of **y** when the **x** is 0.
- **$B_1$** is the regression coefficient – how much we expect **y** to change as **x** increases.
- **x** is the independent variable ( the variable we expect is influencing **y**).
- **e** is the **error** of the estimate, or how much variation there is in our estimate of the regression coefficient.

Linear regression finds the line of best fit line through your data by searching for the regression coefficient ($B_1$) that minimizes the total error (e) of the model

## 15. What are the various branches of statistics?

**Answer:**

**Branches of statistics:**

There are two main branches of statistics:
(1) Descriptive statistics and
(2) Inferential statistics

### Descriptive statistics

The branch of statistics that involves the collecting, organization, summarization, visualizing and analyzing data.

*Descriptive Statistics classified as:*

**(1) Measure of Central Tendency**
- (a) Mean
- (b) Median
- (c) Mode

**(2) Measure of Variation (Dispersion)**
- (a) Range
- (b) Standard Deviation
- (c) Variance

**(3) Measure of Frequency**
- (a) Count
- (b) Percent
- (c) Frequency

**(4) Measure of Position**
- (a) Percentile Ranks
- (b) Quartile Ranks

### Inferential statistics

The branch of statistics that involves using a sample to draw conclusions about the population.

*Inferential Statistics classified as:*

**(1) Hypothesis development**
- (a) Experimental
- (b) Statistical
    - (i) Null
    - (ii) Alternative

**(2) Statistical tests**
- (a) Parametric
- (b) Non-Parametric