**STATISTICS WORKSHEET-1**

Q1) A – True

Q2) A – Central limit theorem

Q3) B – Modeling bounded count data

Q4) D – All of the mentioned

Q5) C – Poisson

Q6) B – False

Q7) B – Hypothesis

Q8) A – 0

Q9) C – Outliers can confirm to the regression relationship.

Q10) Normal Distribution

The normal distribution is the most widely known and used of all distributions. Because the normal distribution approximates many natural phenomena so well, it has developed into a standard of reference for many probability problems.



Many things actually are normally distributed, or very close to it. For example, height and intelligence are approximately normally distributed; measurement errors also often have a normal distribution. The normal distribution is easy to work with mathematically. In many practical cases, the methods developed using normal theory work quite well even when the distribution is not normal. There is a very strong connection between the size of a sample N and the extent to which a sampling distribution approaches the normal form. Many sampling distributions based on large N can be approximated by the normal distribution even though the population distribution itself is definitely not normal.

Standard Normal Distributed Curve:

As you might suspect from the formula for the normal density function, it would be difficult and tedious to do the calculus every time we had a new set of parameters for μ and σ. So instead, we usually work with the standardized normal distribution, where μ = 0 and σ = 1, i.e., N (0,1). That is, rather than directly solve a problem involving a normally distributed variable X with mean μ and standard deviation σ, an indirect approach is used. We first convert the problem into an equivalent one dealing with a normal variable measured in standardized deviation units, called a standardized normal variable. To do this, if X ∼ N(μ,𝜎2), then 𝑍=𝑋−𝜇𝜎~ 𝑁(0,1)

If necessary, we can then convert back to the original units of measurement. To do this, simply note that, if we take the formula for Z, multiply both sides by σ, and then add μ to both sides, we get 𝑋=𝑍𝜎+𝜇.

The interpretation of Z values is straightforward. Since σ = 1, if Z = 2, the corresponding X value is exactly 2 standard deviations above the mean. If Z = -1, the corresponding X value is one standard deviation below the mean. If Z = 0, X = the mean, i.e., μ. It is very important to understand how the standardized normal distribution works, so we will spend some time here going over it. Recall that, for a random variable X, (𝑥)=𝑃(𝑋≤𝑥).

Q11)

**Imputation using Statistics:**

It can be “Mean” or “Median” or “Most Frequent”.

“Mean” will replace missing values using the mean in each column. It is preferred if data is numeric and not skewed.

“Median” will replace missing values using the median in each column. It is preferred if data is numeric and skewed.

“Most frequent” will replace missing values using the most frequent in each column. It is preferred if data is a string (object) or numeric.

Q12)

A/B testing in its simplest sense is an experiment on two variants to see which performs better based on a given metric. Typically, two consumer groups are exposed to two different versions of the same thing to see if there is a significant difference in metrics like sessions, click-through rate, and/or conversions.

Using the visual above as an example, we could randomly split our customer base into two groups, a control group and a variant group. Then, we can expose our variant group with a red website banner and see if we get a significant increase in conversions. It’s important to note that all other variables need to be held constant when performing an A/B test.

Getting more technical, A/B testing is a form of statistical and two-sample hypothesis testing. **Statistical hypothesis testing**is a method in which a sample dataset is compared against the population data. **Two-sample hypothesis testing**is a method in determining whether the differences between the two samples are statistically significant or not.

Q13)

 There are three problems with using mean-imputed variables in statistical analyses:

* Mean imputation reduces the variance of the imputed variables.
* Mean imputation shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval.
* Mean imputation does not preserve relationships between variables such as correlations.

Q14)

Linear regression is a basic and commonly used type of predictive analysis.  The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable?  (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?  These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.  The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b\*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

First, the regression might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable.  Typical questions are what is the strength of relationship between dose and effect, sales and marketing spending, or age and income.

Second, it can be used to forecast effects or impact of changes.  That is, the regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables.  A typical question is, “how much additional sales income do I get for each additional $1000 spent on marketing?”

Third, regression analysis predicts trends and future values.  The regression analysis can be used to get point estimates.  A typical question is, “what will the price of gold be in 6 months?”

Q15)

**The Branches of Statistics**

Two branches, *descriptive statistics* and *inferential statistics*, comprise the field of statistics.

**Descriptive Statistics**

**CONCEPT** The branch of statistics that focuses on collecting, summarizing, and presenting a set of data.

**EXAMPLES** The average age of citizens who voted for the winning candidate in the last presidential election, the average length of all books about statistics, the variation in the weight of 100 boxes of cereal selected from a factory's production line.

**INTERPRETATION** You are most likely to be familiar with this branch of statistics, because many examples arise in everyday life. Descriptive statistics forms the basis for analysis and discussion in such diverse fields as securities trading, the social sciences, government, the health sciences, and professional sports. A general familiarity and widespread availability of descriptive methods in many calculating devices and business software can often make using this branch of statistics seem deceptively easy. (Chapters 2 and 3 warn you of the common pitfalls of using descriptive methods.)

**Inferential Statistics**

**CONCEPT** The branch of statistics that analyzes sample data to draw conclusions about a population.

**EXAMPLE** A survey that sampled 2,001 full-or part-time workers ages 50 to 70, conducted by the American Association of Retired Persons (*AARP*), discovered that 70% of those polled planned to work past the traditional mid-60s retirement age. By using methods discussed in Section 6.4, this statistic could be used to draw conclusions about the population of all workers ages 50 to 70.

**INTERPRETATION** When you use inferential statistics, you start with a hypothesis and look to see whether the data are consistent with that hypothesis. Inferential statistical methods can be easily misapplied or misconstrued, and many inferential methods require the use of a calculator or computer.