



# Data Application Lab

## Answers for Probability and Statistics Quiz

### 1. What is the Central Limit Theorem? Explain it. Why is it important?

The CLT states that the arithmetic mean of a sufficiently large number of iterates of independent random variables will be approximately normally distributed regardless of the underlying distribution. i.e: the sampling distribution of the sample mean is normally distributed.

- Used in hypothesis testing
- Used for confidence intervals
- Random variables must be iid: independent and identically distributed
- Finite variance

### 2. What is Bayes' Theorem? How is it useful in a machine learning context?

Bayes' Theorem gives you the posterior probability of an event given what is known as prior knowledge. It's expressed as the true positive rate of a condition sample divided by the sum of the false positive rate of the population and the true positive rate of a condition.

### 3. What's the differences between the Poission Distribution and Normal Distribution?

- A Poisson distribution is discrete while a normal distribution is continuous
- A Poisson random variable is always  $\geq 0$ .
- When the mean of a Poisson distribution is large, it becomes similar to a normal distribution.

### 4. What is hypothesis testing?

A statistical hypothesis is an assumption about a population parameter. Hypothesis tests are used to test the validity of a claim that is made about a population. There are two types of statistical hypotheses:

**Null hypothesis:** The null hypothesis, denoted by  $H_0$ , is usually the hypothesis that sample observations result purely from chance.

**Alternative hypothesis:** The alternative hypothesis, denoted by  $H_1$ , is the hypotheses that sample observations are influenced by some non-random cause.

## 5. What is p-value?

In statistical hypothesis testing, the p-value or probability value is the probability for a given statistical model that, under the null hypothesis, of obtaining a result equal to or more extreme than what was actually observed. The hypothesis tests ultimately use a p-value to weigh the strength of the evidence (what the data are telling you about the population).

The p-value is a number between 0 and 1 and interpreted in the following way:

- A small p-value (typically  $\leq 0.05$ ) indicates strong evidence against the null hypothesis, so with a small p-value, you reject the null hypothesis.
- A large p-value ( $> 0.05$ ) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.
- P-values very close to the cutoff (0.05) are considered to be marginal.

## 6. The mean weight of a new product needs to be below 20lbs if the design can meet the new product design requirements. Ten sample products are manufactured for testing purpose. The weight of each product is determined. The sample weight data is:

15.6, 16.2, 22.5, 20.5, 16.4, 19.4, 16.6, 17.9, 12.7, 13.9

**Does the data supply sufficient evidence to conclude that this type of product meets the new standard? Assume we are willing to risk a Type I error with probability = 0.01.**

$$\bar{X} = 17.17, S = 2.98$$

$$H_0: \mu = 20 \text{ lbs}, H_1: \mu < 20 \text{ lbs}$$

$$t = \frac{17.17 - 20}{\frac{2.98}{\sqrt{10}}} = -3.00$$

From the t-table, we can tell that the type I error with probability = 0.01 is -2.821, which is large than -3.00, the probability to get this is less than 1%. So we can reject the null hypothesis.

We can say that  $\mu < 20$  lbs.

