

STATISTICS WORKSHEET-1

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
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
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
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**
5. _____ random variables are used to model rates.
a) Empirical
b) Binomial
c) **Poisson**
d) All of the mentioned
6. 10. Usually replacing the standard error by its estimated value does change the CLT.
a) True
b) **False**
7. 1. Which of the following testing is concerned with making decisions using data?
a) **Probability**
b) Hypothesis
c) Causal
d) None of the mentioned
8. 4. Normalized data are centered at _____ and have units equal to standard deviations of the original data.
a) **0**
b) 5
c) 1
d) 10
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

Q10 and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

Ans: The normal distribution, also known as the Gaussian distribution, is the most important probability distribution in statistics for independent, random variables. Most people recognize its familiar bell-shaped curve in statistical reports.

The normal distribution is a continuous probability distribution that is symmetrical around its mean, most of the observations cluster around the central peak, and the probabilities for values further away from the mean taper off equally in both directions. Extreme values in both tails of the distribution are similarly unlikely. While the normal distribution is symmetrical, not all symmetrical distributions are normal. For example, the Student's t, Cauchy, and logistic distributions are symmetric.

As with any probability distribution, the normal distribution describes how the values of a variable are distributed. It is the most important probability distribution in statistics because it accurately describes the distribution of values for many natural phenomena. Characteristics that are the sum of many independent processes frequently follow normal distributions. For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution.

Parameters of the Normal Distribution

As with any probability distribution, the parameters for the normal distribution define its shape and probabilities entirely. The normal distribution has two parameters, the mean and standard deviation. The Gaussian distribution does not have just one form. Instead, the shape changes based on the parameter values, as shown in the graphs below.

Mean μ : The mean is the central tendency of the normal distribution. It defines the location of the peak for the bell curve. Most values cluster around the mean. On a graph, changing the mean shifts the entire curve left or right on the X-axis. Statisticians denote the population mean using μ (mu).

Standard deviation σ : The standard deviation is a measure of variability. It defines the width of the normal distribution. The standard deviation determines how far away from the mean the values tend to fall. It represents the typical distance between the observations and the average. Statisticians denote the population standard deviation using σ (sigma). On a graph, changing the standard deviation either tightens or spreads out the width of the distribution along the X-axis. Larger standard deviations produce wider distributions.

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

Missing data can be dealt with in a variety of ways. I believe the most common reaction is to ignore it. Choosing to make no decision, on the other hand, indicates that your statistical programme will make the decision for you.

Your application will remove things in a listwise sequence most of the time. Depending on why and how much data is gone, listwise deletion may or may not be a good idea.

Another common strategy among those who pay attention is imputation. Imputation is the process of substituting an estimate for missing values and analysing the entire data set as if the imputed values were the true observed values.

And how would you choose that estimate? The following are some of the most prevalent methods:

Mean imputation

Calculate the mean of the observed values for that variable for all non-missing people. It has the advantage of maintaining the same mean and sample size, but it also has a slew of drawbacks. Almost all of the methods described below are superior to mean imputation.

Substitution

Assume the value from a new person who was not included in the sample. To put it another way, pick a new subject and employ their worth instead.

Hot deck imputation

A value picked at random from a sample member who has comparable values on other variables. To put it another way, select all the sample participants who are comparable on other factors, then choose one of their missing variable values at random.

One benefit is that you are limited to just feasible values. In other words, if age is only allowed to be between 5 and 10 in your research, you will always obtain a value between 5 and 10. Another factor is the random element, which introduces some variation. For exact standard errors, this is crucial.

Cold deck imputation

A value picked deliberately from an individual with similar values on other variables. In most aspects, this is comparable to Hot Deck, but without the random variance. As an example, under the same experimental condition and block, you can always select the third individual.

Regression imputation

The result of regressing the missing variable on other factors to get a predicted value. As a result, instead of utilising the mean, you're relying on the anticipated value, which is influenced by other factors. This keeps the associations between the variables in the imputation model, but not the variability around the anticipated values.

Stochastic regression imputation

The predicted value of a regression plus a random residual value. This has all of the benefits of regression imputation plus the random component's benefits. The majority of multiple imputation is based on stochastic regression imputation.

Interpolation and extrapolation

An estimate based on other observations made by the same person. It generally only works with data that is collected over time. Proceed with caution, though. For a variable like height in children—one that cannot be reduced through time—interpolation would make more sense. Extrapolation entails estimating beyond the data's true range, which necessitates making more assumptions than is necessary.

Single or Multiple Imputation

- Single and multiple imputation are the two forms of imputation. When people say imputation, they usually mean single.
-

- The term "single" refers to the fact that you only use one of the seven methods to estimate the missing number outlined above.
- It's popular since it's simple to understand and generates a sample with the same number of observations as the complete data set.
- When listwise deletion eliminates a considerable amount of the data set, single imputation appears to be a tempting option. It does, however, have certain restrictions.
- Unless the data is Missing Completely at Random, certain imputation processes, such as means, correlations, and regression coefficients, result in skewed parameter estimations. The bias is frequently worse than with listwise deletion, which is most software's default.
- The level of the bias is determined by a number of factors, including the imputation technique, the missing data mechanism, the fraction of missing data, and the information in the data set.

Furthermore, standard errors are underestimated by all single imputation approaches. Because the imputed observations are estimates, their values have a random error associated with them. However, your programme is unaware of this when you enter that estimate as a data point. As a result, it ignores the additional source of error, resulting in too-small standard errors and p-values.

And, while imputation is straightforward in theory, it is difficult to master in reality. As a result, it isn't perfect, although it may suffice in some circumstances.

As a result of multiple imputation, numerous estimates are generated. In multiple imputation, two of the approaches indicated above—hot deck and stochastic regression—work as the imputation method.

The multiple estimates varied significantly because these two approaches contain a random component. This reintroduces some variance that your program can account for in order to provide reliable standard error estimates for your model.

12. What is A/B testing?

A/B Testing, also known as split testing, is a method that is used quite extensively in the field of data science. Its aim is to enable data scientists to make accurate, evidence-based decisions based on the insights derived from testing two different variables.

A/B testing in data science is a powerful tool used to assess the efficiency or efficacy of any proposed change against the existing design. It involves the comparison of two groups, which are virtually identical except for one variation that differentiates the groups.

This direct comparative study allows data scientists to actively experiment and measure their results, making statistically knowledgeable decisions.

Without A/B testing, data scientists would essentially be making blind guesses about what might work. By conducting A/B tests, they can eliminate the guesswork and gain valuable insights into the impact of their changes.

13. Is mean imputation of missing data acceptable practice?

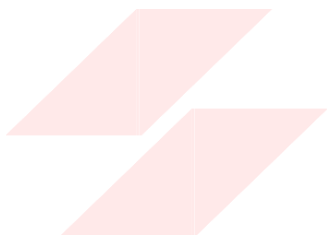
Bad practice in general . If just estimating means: mean imputation preserves the mean of the observed data . Leads to an underestimate of the standard deviation. Distorts relationships between variables by “pulling” estimates of the correlation toward zero

14. What is linear regression in statistics?

linear regression is a statistical model which estimates the linear relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).

15. What are the various branches of statistics?

There are three real branches of statistics: data collection, descriptive statistics and inferential statistics



FLIP ROBO