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

ANSWER:-(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

ANSWER:- (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

ANSWER:- (A) Modeling event/time 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 5. ______ random variables are used to model rates. a) Empirical b) Binomial
c) Poisson

d) All of the mentioned

ANSWER:-(D) All of the mentioned

5. ______ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

ANSWER:-(B) Binomial

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

a) True

b) False

ANSWER:- (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

ANSWER:- (B) HYPOTHESIS

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

ANSWER:- (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

ANSWER:- (A) Outliers can have varying degrees of influence

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

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

Normal Distribution in Statistics

By Jim Frost 181 Comments

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.

In this blog post, learn how to use the normal distribution, about its parameters, the Empirical Rule, and how to calculate Z-scores to standardize your data and find probabilities.

Example of Normally Distributed Data: Heights

Height data are normally distributed. The distribution in this example fits real data that I collected from 14-year-old girls during a study. The graph below displays the probability distribution function for this normal distribution. Learn more about Probability Density Functions.

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

Introduction

If you are aiming for a job as a data scientist, you must know how to handle the problem of missing values, which is quite common in many real-life datasets. Incomplete data can bias the results of the machine learning models and/or reduce the accuracy of the model. This article describes missing data, how it is represented, and the different reasons data values get missed. Along with the different categories of missing data, it also details out different ways of handling missing values with dataset examples.

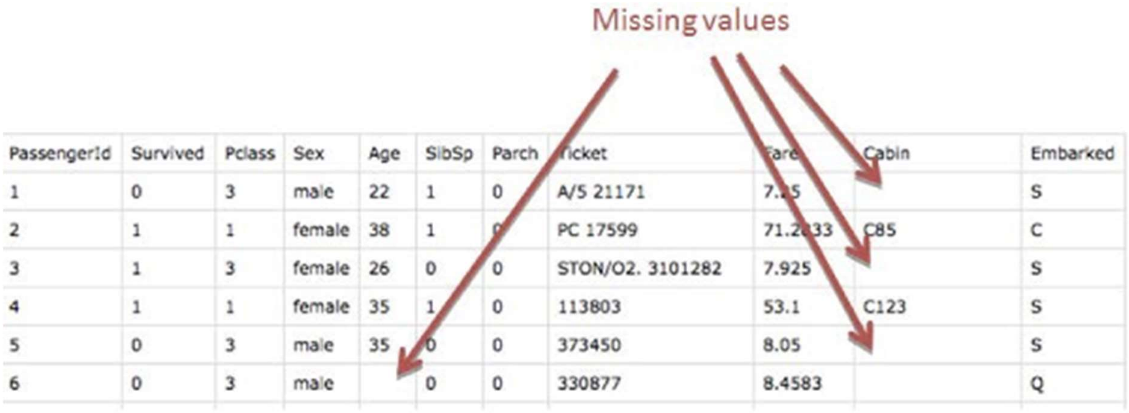
Learning Objectives

- In this tutorial, we will learn about missing values and the benefits of missing data analysis in data science.
- You will learn about the different types of missing data and how to handle them correctly.
- You will also learn about the most widely used imputation methods to handle incomplete data.

Table of Contents

What Is a Missing Value?

Missing data is defined as the values or data that is not stored (or not present) for some variable/s in the given dataset. Below is a sample of the missing data from the Titanic dataset. You can see the columns 'Age' and 'Cabin' have some missing values.


Missing values

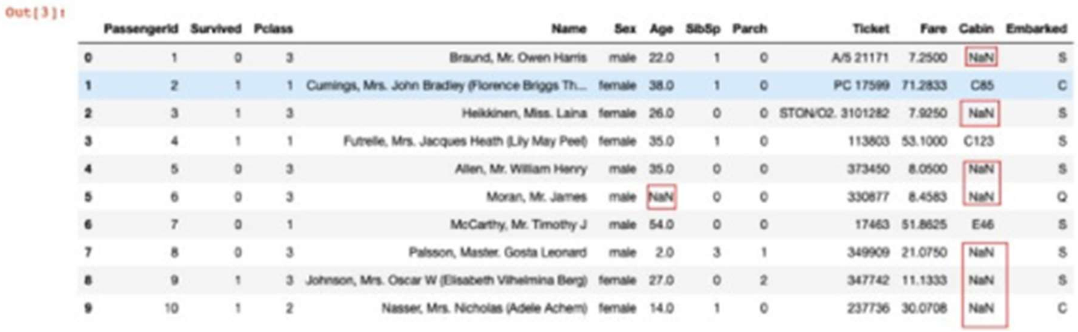
PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

Source: analyticsindiamag

How Is a Missing Value Represented in a Dataset?

In the dataset, the blank shows the missing values.

In Pandas, usually, missing values are represented by **NaN**. It stands for **Not a Number**.



Out[3]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C

13.What is A/B testing?

A/B testing (also known as **bucket testing**, **split-run testing**, or **split testing**) is a user experience research methodology.[1] A/B tests consist of a randomized experiment that usually involves two variants (A and B),[2][3][4] although the concept can be also extended to multiple variants of the same variable. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics. A/B testing is a way to compare multiple versions of a single variable, for example by testing a subject's response to variant A against variant B, and determining which of the variants is more effective.[5]

Overview[edit]

"A/B testing" is a shorthand for a simple randomized controlled experiment, in which a number of samples (e.g. A and B) of a single vector-variable are compared.[1] These values are similar except for one variation which might affect a user's behavior. A/B tests are widely considered the simplest form of controlled experiment, especially when they only involve two variants. However, by adding more variants to the test, its complexity grows.[6]

A/B tests are useful for understanding user engagement and satisfaction of online features like a new feature or product.[7] Large social media sites like LinkedIn, Facebook, and Instagram use A/B testing to make user experiences more successful and as a way to streamline their services.[7]

Today, A/B tests are being used also for conducting complex experiments on subjects such as network effects when users are offline, how online services affect user actions, and how users influence one another.[7] A/B testing is used by data engineers, marketers, designers, software engineers, and entrepreneurs, among others.[8] Many positions rely on the data from A/B tests, as they allow companies to understand growth, increase revenue, and optimize customer satisfaction.[8]

Version A might be used at present (thus forming the control group), while version B is modified in some respect vs. A (thus forming the treatment group). For instance, on an e-commerce website the purchase funnel is typically a good candidate for A/B testing, since even marginal-decreases in drop-off rates can represent a significant gain in sales. Significant improvements can be sometimes seen through testing elements like copy text, layouts, images and colors,[9] but not always. In these tests, users only see one of two versions, since the goal is to discover which of the two versions is preferable.[10]

Multivariate testing or multinomial testing is similar to A/B testing, but may test more than two versions at the same time or use more controls. Simple A/B tests are not valid for observational, quasi-experimental or other non-experimental situations - commonplace with survey data, offline data, and other, more complex phenomena.

A/B testing is claimed by some to be a change in philosophy and business-strategy in certain niches, though the approach is identical to a between-subjects design, which is commonly used in a variety of research traditions.[11][12][13] A/B testing as a philosophy of web development brings the field into line with a broader movement toward evidence-based practice. The benefits of A/B testing are considered to be that it can be performed continuously on almost anything, especially since most marketing automation software now typically comes with the ability to run A/B tests on an ongoing basis.

Common test statistics[edit]

"Two-sample hypothesis tests" are appropriate for comparing the two samples where the samples are divided by the two control cases in the experiment. Z-tests are appropriate for comparing means under stringent conditions regarding normality and a known standard deviation. Student's t-tests are appropriate for comparing means under relaxed conditions when less is assumed. Welch's t test assumes the least and is therefore the most commonly used test in a two-sample hypothesis test where the mean of a metric is to be optimized. While the mean of the variable to be optimized is the most common choice of estimator, others are regularly used.

13.Is mean imputation of missing data acceptable practice?

Mean imputation: So simple. And yet, so dangerous.

Perhaps that’s a bit dramatic, but mean imputation (also called mean substitution) really ought to be a last resort.

It’s a popular solution to missing data, despite its drawbacks. Mainly because it’s easy. It can be really painful to lose a large part of the sample you so carefully collected, only to have little power.

But that doesn’t make it a good solution, and it may not help you find relationships with strong parameter estimates. Even if they exist in the population.

On the other hand, there are many alternatives to mean imputation that provide much more accurate estimates and standard errors, so there really is no excuse to use it.

This post is the first explaining the many reasons not to use mean imputation (and to be fair, its advantages).

First, a definition: mean imputation is the replacement of a missing observation with the mean of the non-missing observations for that variable.

Problem #1: Mean imputation does not preserve the relationships among variables.

True, imputing the mean preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the mean remains unbiased. That's a good thing.

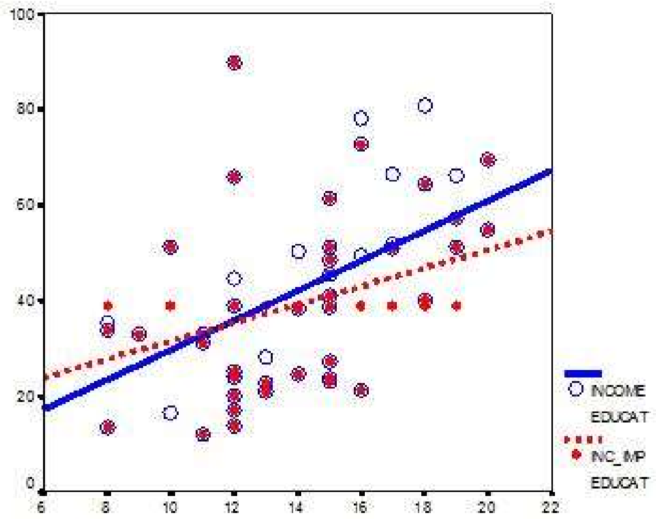
Plus, by imputing the mean, you are able to keep your sample size up to the full sample size. That's good too.

This is the original logic involved in mean imputation.

If all you are doing is estimating means (which is rarely the point of research studies), and if the data are missing completely at random, mean imputation will not bias your parameter estimate.

It *will* still bias your standard error, but I will get to that in another post.

Since most research studies are interested in the relationship among variables, mean imputation is not a good solution. The following graph illustrates this well:



This graph illustrates hypothetical data between X=years of education and Y=annual income in thousands with n=50. The blue circles are the original data, and the solid blue line indicates the best fit regression line for the full data set. The correlation between X and Y is r = .53.

I then randomly deleted 12 observations of income (Y) and substituted the mean. The red dots are the mean-imputed data.

Blue circles with red dots inside them represent non-missing data. Empty Blue circles represent the missing data. If you look across the graph at Y = 39, you will see a row of red dots without blue circles. These represent the imputed values.

The dotted red line is the new best fit regression line with the imputed data. As you can see, it is less steep than the original line. Adding in those red dots pulled it down.

The new correlation is r = .39. That's a lot smaller that .53.

The real relationship is quite underestimated.

Of course, in a real data set, you wouldn't notice so easily the bias you're introducing. This is one of those situations where in trying to solve the lowered sample size, you create a bigger problem.

One note: if X were missing instead of Y, mean substitution would artificially *inflate* the correlation.

In other words, you'll think there is a stronger relationship than there really is. That's not good either. It's not reproducible and you don't want to be overstating real results.

This solution that is so good at preserving unbiased estimates for the mean isn't so good for unbiased estimates of relationships.

Problem #2: Mean Imputation Leads to An Underestimate of Standard Errors

A second reason is applies to any type of single imputation. Any statistic that uses the imputed data will have a standard error that's too low.

In other words, yes, you get the same mean from mean-imputed data that you would have gotten without the imputations. And yes, there are circumstances where that mean is unbiased. Even so, the standard error of that mean will be too small.

Because the imputations are themselves estimates, there is some error associated with them. But your statistical software doesn't know that. It treats it as real data.

Ultimately, because your standard errors are too low, so are your p-values. Now you're making Type I errors without realizing it.

That's not good.

14.What is linear regression in statistics?

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).
Generate predictions more easily

You can perform linear regression in Microsoft Excel or use statistical software packages such as IBM SPSS® Statistics that greatly simplify the process of using linear-regression equations, linear-regression models and linear-regression formula. SPSS Statistics can be leveraged in techniques such as simple linear regression and multiple linear regression.

You can perform the linear regression method in a variety of programs and environments, including:

- R linear regression
- MATLAB linear regression
- Sklearn linear regression
- Linear regression Python
- Excel linear regression

Why linear regression is important

Linear-regression models are relatively simple and provide an easy-to-interpret mathematical formula that can generate predictions. Linear regression can be applied to various areas in business and academic study.

You’ll find that linear regression is used in everything from biological, behavioral, environmental and social sciences to business. Linear-regression models have become a proven way to scientifically and reliably predict the future. Because linear regression is a long-established statistical procedure, the properties of linear-regression models are well understood and can be trained very quickly.
A proven way to scientifically and reliably predict the future

Business and organizational leaders can make better decisions by using linear regression techniques. Organizations collect masses of data, and linear regression helps them use that data to better manage reality — instead of relying on experience and intuition. You can take large amounts of raw data and transform it into actionable information.

You can also use linear regression to provide better insights by uncovering patterns and relationships that your business colleagues might have previously seen and thought they already understood. For example, performing an analysis of sales and purchase data can help you uncover specific purchasing patterns on particular days or at certain times. Insights gathered from regression analysis can help business leaders anticipate times when their company's products will be in high demand.

15. What are the various branches of statistics?

Descriptive Statistics

Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment.
Different areas of study require different kinds of analysis using descriptive statistics. For example, a physicist studying turbulence in the laboratory needs the average quantities that vary over small intervals of time. The nature of this problem requires that physical quantities be averaged from a host of data collected through the experiment.

Inferential Statistics

Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.
Most predictions of the future and generalizations about a population by studying a smaller sample come under the purview of inferential statistics. Most social sciences experiments deal with studying a small sample population that helps determine how the population in general behaves. By designing the right experiment, the researcher is able to draw conclusions relevant to his study.
While drawing conclusions, one needs to be very careful so as not to draw the wrong or biased conclusions. Even though this appears like a science, there are ways in which one can manipulate studies and results through various means. For example, data dredging is increasingly becoming a problem as computers hold loads of information and it is easy, either intentionally or unintentionally, to use the wrong inferential methods.
Both descriptive and inferential statistics go hand in hand and one cannot exist without the other. Good scientific methodology needs to be followed in both these steps of statistical analysis and both these branches of statistics are equally important for a researcher.