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.

   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?

   Answer: a) Central Limit Theorem b) Central Mean Theorem

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

   Answer: b) Modeling bounded count data

4. Point out the correct statement.

   Answer: d) All of the mentioned

5. ______ random variables are used to model rates.

   Answer c) Poisson

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

   Answer b) False

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

   Answer :   b) Hypothesis

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

   Answer: a) 0

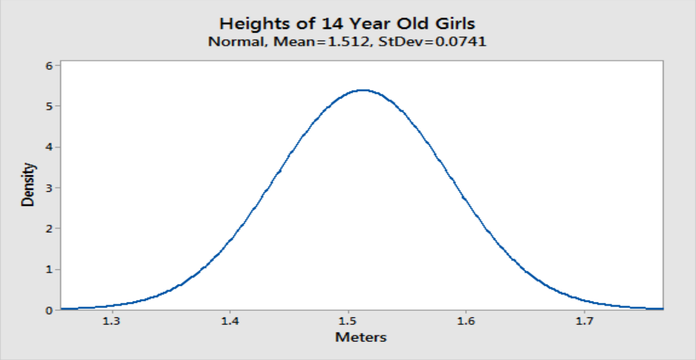
9. Which of the following statement is incorrect with respect to outliers?

   Answer: 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: The normal distribution, also known as the Gaussian distribution, is the most important probability distribution in statistics for independent, random variables. 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. Example:

Height data are normally distributed. The distribution in this example fits real data that I collected from 14-year-old girls during a study.

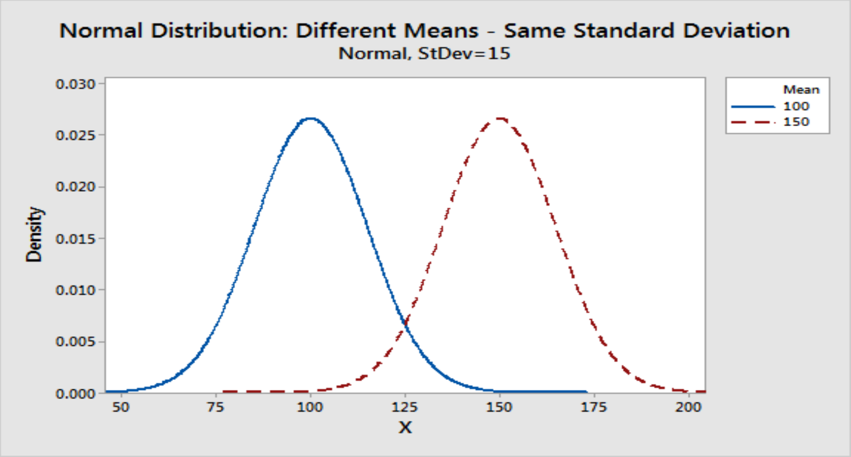


Parameters of the Normal Distribution

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.

Mean

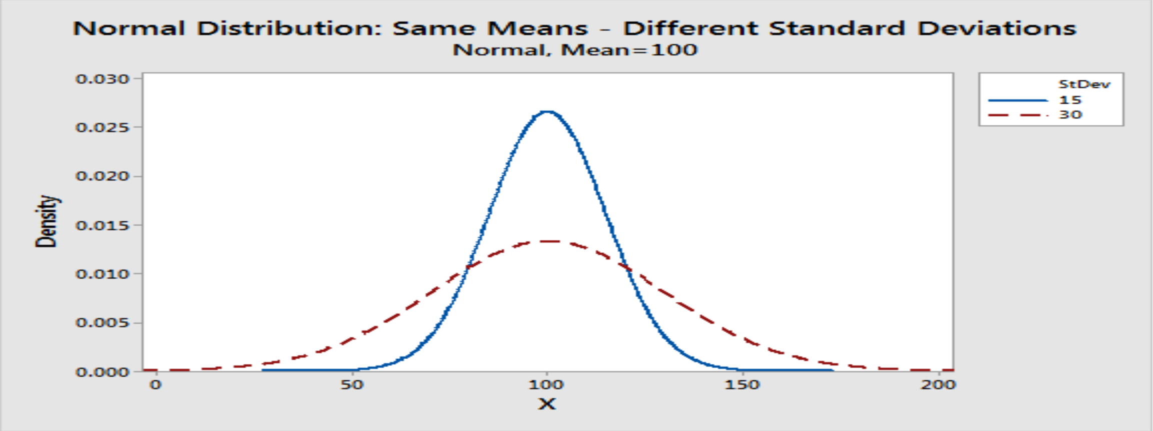
The mean is the central tendency of the normal distribution. It defines the location of the peak for the bell curve. Most values cluster around the mean. On a



graph, changing the mean shifts the entire curve left or right on the X-axis.

Standard deviation

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.



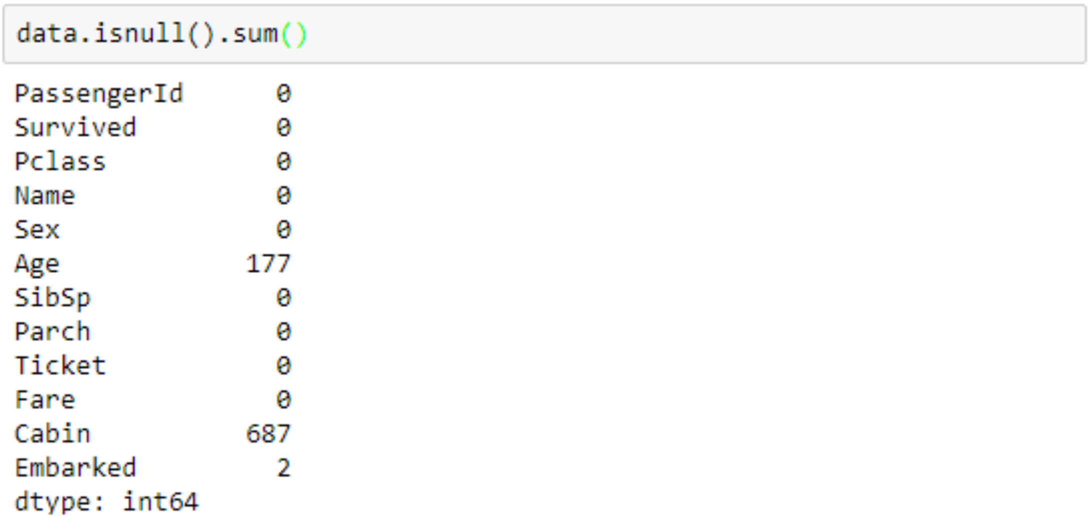
On a graph, changing the standard deviation either tightens or spreads out the width of the distribution along the X-axis. Larger standard deviations produce wider distributions.

When you have narrow distributions, the probabilities are higher that values won’t fall far from the mean. As you increase the spread of the bell curve, the likelihood that observations will be further away from the mean also increases.

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

Answer: there are five Ways To Handle Missing Values In Machine Learning Datasets

. To understand various methods we will be working on the Titanic dataset:

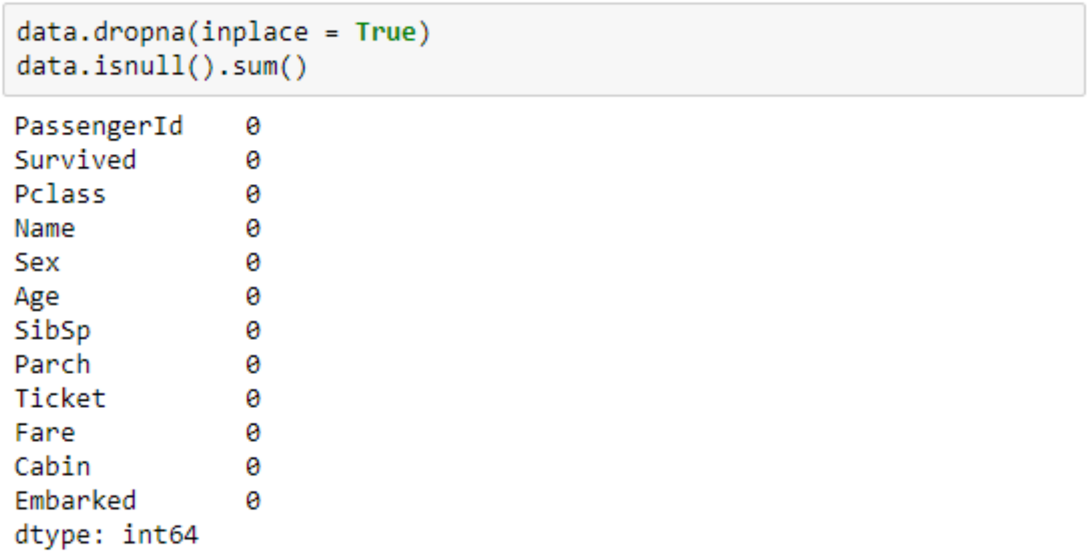


```
data.isnull().sum()
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

1. Deleting Rows

This method commonly used to handle the null values. Here, we either delete a particular row if it has a null value for a particular feature and a particular column if it has more than 70-75% of missing values.



```
data.dropna(inplace = True)
data.isnull().sum()
```

```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Cabin          0
Embarked       0
dtype: int64
```

Pros:

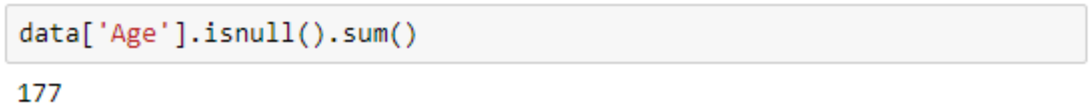
- Complete removal of data with missing values results in robust and highly accurate model
- Deleting a particular row or a column with no specific information is better, since it does not have a high weightage

Cons:

- Loss of information and data
- Works poorly if the percentage of missing values is high (say 30%), compared to the whole dataset

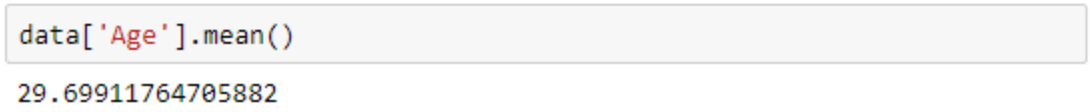
2. Replacing With Mean/Median/Mode

This strategy can be applied on a feature which has numeric data like the age of a person or the ticket fare. We can calculate the mean, median or mode of the feature and replace it with the missing values. This method is also called as leaking the data while training. Another way is to approximate it with the deviation of neighbouring values. This works better if the data is linear.



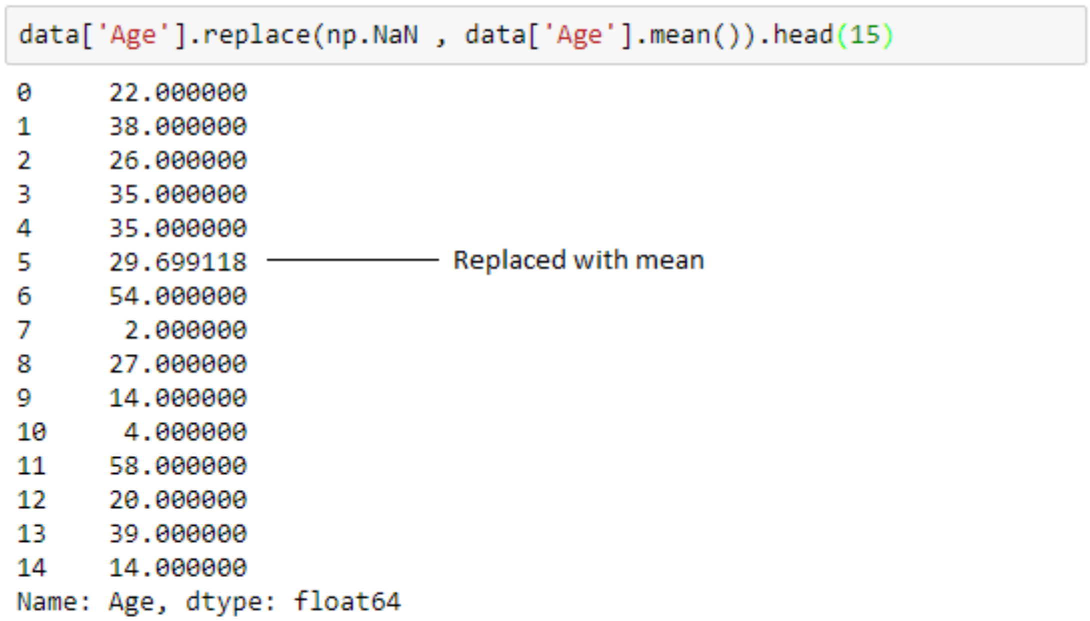
```
data['Age'].isnull().sum()
```

```
177
```



```
data['Age'].mean()
```

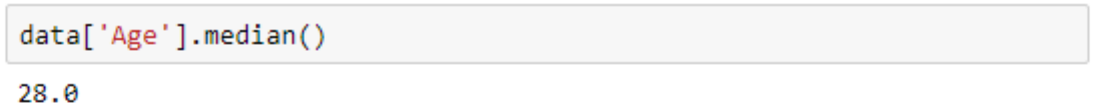
```
29.69911764705882
```



```
data['Age'].replace(np.NaN , data['Age'].mean()).head(15)
```

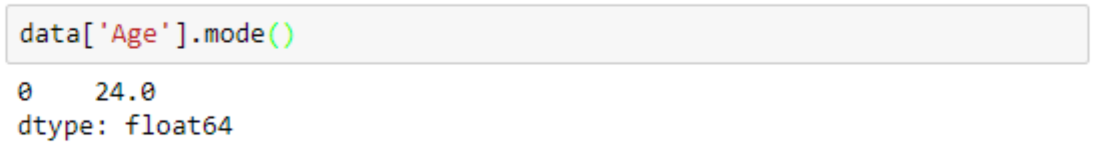
```
0     22.000000
1     38.000000
2     26.000000
3     35.000000
4     35.000000
5     29.699118
6     54.000000
7      2.000000
8     27.000000
9     14.000000
10     4.000000
11    58.000000
12    20.000000
13    39.000000
14    14.000000
Name: Age, dtype: float64
```

To replace it with median and mode we can use the following to calculate the same:



```
data['Age'].median()
```

```
28.0
```



```
data['Age'].mode()
```

```
0    24.0
dtype: float64
```

Pros:

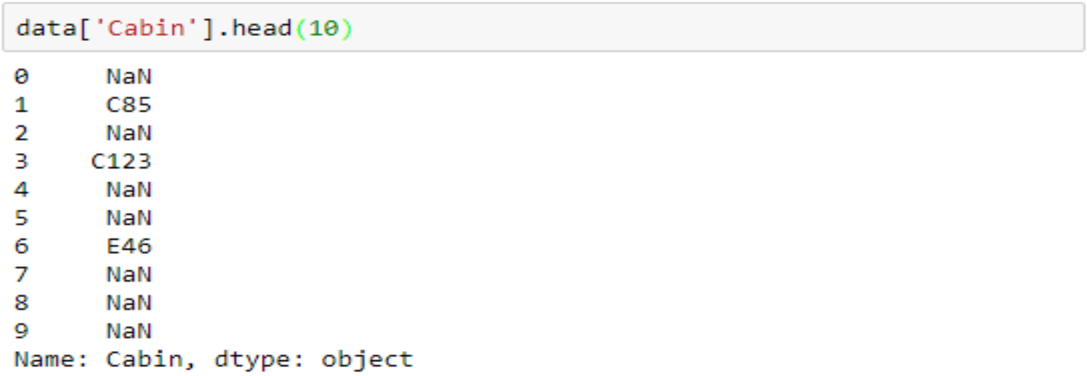
- This is a better approach when the data size is small
- It can prevent data loss which results in removal of the rows and columns

Cons:

- Imputing the approximations add variance and bias
- Works poorly compared to other multiple-imputations method

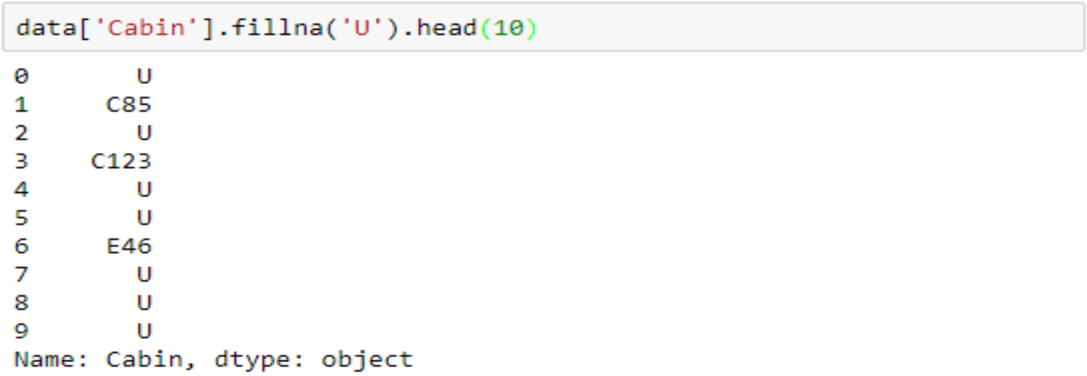
3. Assigning An Unique Category

A categorical feature will have a definite number of possibilities, such as gender, for example. Since they have a definite number of classes, we can assign another class for the missing values.



```
data['Cabin'].head(10)
```

```
0     NaN
1     C85
2     NaN
3    C123
4     NaN
5     NaN
6     E46
7     NaN
8     NaN
9     NaN
Name: Cabin, dtype: object
```



```
data['Cabin'].fillna('U').head(10)
```

```
0       U
1     C85
2       U
3    C123
4       U
5       U
6     E46
7       U
8       U
9       U
Name: Cabin, dtype: object
```

Pros:

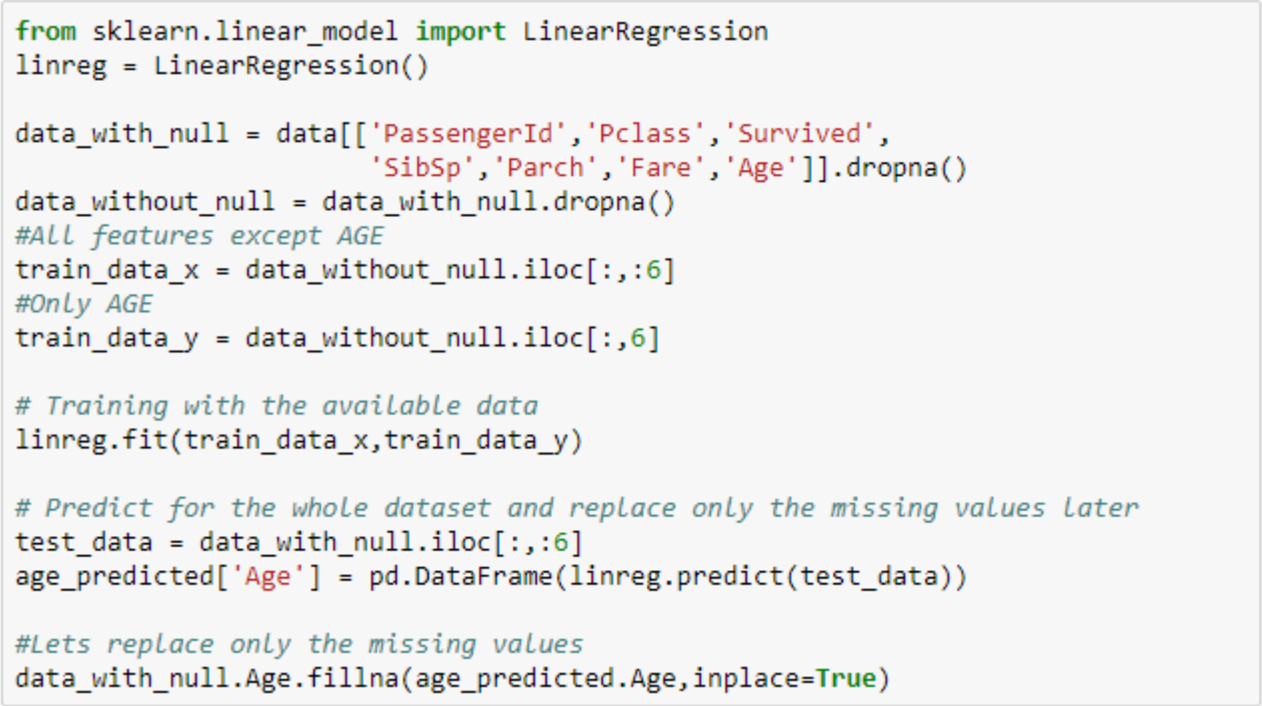
- Less possibilities with one extra category, resulting in low variance after one hot encoding — since it is categorical
- Negates the loss of data by adding an unique category

Cons:

- Adds less variance
- Adds another feature to the model while encoding, which may result in poor performance

4. Predicting The Missing Values

Using the features which do not have missing values, we can predict the nulls with the help of a machine learning algorithm. This method may result in better accuracy, unless a missing value is expected to have a very high variance. We will be using linear regression to replace the nulls in the feature ‘age’, using other available features.



```
from sklearn.linear_model import LinearRegression
linreg = LinearRegression()

data_with_null = data[['PassengerId','Pclass','Survived',
                       'SibSp','Parch','Fare','Age']].dropna()
data_without_null = data_with_null.dropna()
#All features except AGE
train_data_x = data_without_null.iloc[:,:6]
#Only AGE
train_data_y = data_without_null.iloc[:,6]

# Training with the available data
linreg.fit(train_data_x,train_data_y)

# Predict for the whole dataset and replace only the missing values later
test_data = data_with_null.iloc[:,:6]
age_predicted['Age'] = pd.DataFrame(linreg.predict(test_data))

#Lets replace only the missing values
data_with_null.Age.fillna(age_predicted.Age,inplace=True)
```

Pros:

- Imputing the missing variable is an improvement as long as the bias from the same is smaller than the omitted variable bias
- Yields unbiased estimates of the model parameters

Cons:

- Bias also arises when an incomplete conditioning set is used for a categorical variable
- Considered only as a proxy for the true values

5. Using Algorithms Which Support Missing Values

KNN is a machine learning algorithm which works on the principle of distance measure. This algorithm can be used when there are nulls present in the dataset. While the algorithm is applied, KNN considers the missing values by taking the majority of the K nearest values. In this particular dataset, taking into account the person's age, sex, class etc, we will assume that people having same data for the above mentioned features will have the same kind of fare.

Unfortunately, the SciKit Learn library for the K – Nearest Neighbour algorithm in Python does not support the presence of the missing values.

Another algorithm which can be used here is RandomForest. This model produces a robust result because it works well on non-linear and the categorical data. It adapts to the data structure taking into consideration of the high variance or the bias, producing better results on large datasets.

Pros:

- Does not require creation of a predictive model for each attribute with missing data in the dataset
- Correlation of the data is neglected

Cons:

- Is a very time consuming process and it can be critical in data mining where large databases are being extracted
- Choice of distance functions can be Euclidean, Manhattan etc. which is do not yield a robust result

12. What is A/B testing?

Answer: A/B testing is one of the most popular controlled experiments used to optimize web marketing strategies. It allows decision makers to choose the best design for a website by looking at the analytics results obtained with two possible alternatives A and B.

two alternative designs: A and B. Visitors of a website are randomly served with one of the two. Then, data about their activity is collected by web analytics. Given this data, one can apply statistical tests to determine whether one of the two designs has better efficacy.

Now, different kinds of metrics can be used to measure a website efficacy. With discrete metrics, also called binomial metrics, only the two values 0 and 1 are possible. The following are examples of popular discrete metrics.

- Click-through rate — if a user is shown an advertisement, do they click on it?
- Conversion rate — if a user is shown an advertisement, do they convert into customers?
- Bounce rate — if a user is visits a website, is the following visited page on the same website?

With continuous metrics, also called non-binomial metrics,, the metric may take continuous values that are not limited to a set two discrete states. The following are examples of popular continuous metrics.

- Average revenue per user — how much revenue does a user generate in a month?
- Average session duration — for how long does a user stay on a website in a session?
- Average order value — what is the total value of the order of a user?

13. Is mean imputation of missing data acceptable practice?

Answer : yes it is acceptable practice. mean imputation (also called mean substitution) really ought to be a last resort. mean imputation is the replacement of a missing observation with the mean of the non-missing observations for that variable.

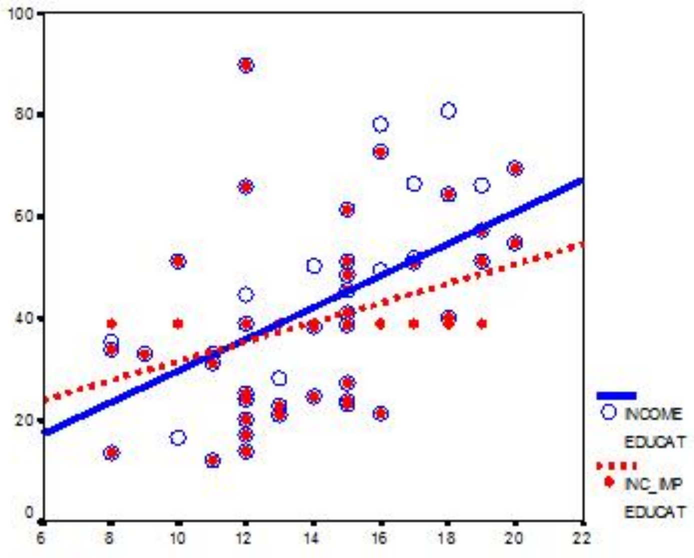
#1: Mean imputation does not preserve the relationships among variables.

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.

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.

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.

#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.

14.   What is linear regression in statistics?

Answer: he 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.
t 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.
regression analysis predicts trends and future values. The regression analysis can be used to get point estimates.

Types of Linear Regression

Simple linear regression
1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)

Multiple linear regression
1 dependent variable (interval or ratio) , 2+ independent variables (interval or ratio or dichotomous)

Logistic regression
1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)

Ordinal regression
1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)

Multinomial regression
1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)

Discriminant analysis
1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio)

When selecting the model for the analysis, an important consideration is model fitting. Adding independent variables to a linear regression model will always increase the explained variance of the model (typically expressed as R^2). However, overfitting can occur by adding too many variables to the model, which reduces model generalizability. Occam’s razor describes the problem extremely well – a simple model is usually preferable to a more complex model. Statistically, if a model includes a large number of variables, some of the variables will be statistically significant due to chance alone.

15. What are the various branches of statistics?

Answer: 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

FLIP ROBO