n models with changing data
samples for the abundant class

ANSWERSHEET 1 (MACHINE LEARNING)

Ans 1 (B) In hierarchical clustering you don't need to assign number of clusters in beginning.

Ans 2 (A) max_depth

Ans 3 (C) RandomUnderSampler

Ans 4 (C) 1 and 3

Ans 5(D) 1-3-2

Ans 6 (B) Support Vector Machines

Ans 7 (**C)** CART can only create binary trees (a maximum of two children for a node), and CHAID can create multiway trees (more than two children for a node)

Ans 8 (A) Ridge will lead to some of the coefficients to be very close to 0.

(D) Lasso will cause some of the coefficients to become 0.

Ans 9 (C) Use ridge regularization

(D) use Lasso regularization

Ans 10 (A) Overfitting

(D) Outliers

ANSWERS 11 TO ANSWERS 15

Ans 11

n models with changing data
samples for the abundant class

When the number of categories in the dataset is quite large. One Hot Encoding should be avoided in this case as it can lead to high memory consumption.

When the categorical features present in the dataset are ordinal i.e., for the data being like Junior, Senior, Executive, Owner., then we should be avoided one hot encoding.

We can use Binary encoding technique can be used in such a case because in binary encoding the categorical feature is first converted into numerical using an ordinal encoder. Then numbers are transformed in the binary number. After that binary value is split into different columns. Binary encoding works really well when there are a higher number of categories.

Ans 12

1. ***Use the right evaluation metrics***

Applying evaluation metrics for model using imbalanced data can be dangerous. Imagine our training data is the one illustrated in graph above. If accuracy is used to measure the goodness of a model, a model which classifies all testing samples into “0” will have an excellent accuracy (99.8%), but obviously, this model won’t provide any valuable information for us.

In this case, other alternative evaluation metrics can be applied such as:

- Recall/Sensitivity: how many relevant instances are selected.
- F1 score: harmonic mean of precision and recall.
- Precision/Specificity: how many selected instances are relevant.

n models with changing data
samples for the abundant class

- MCC: correlation coefficient between the observed and predicted binary classifications.
- AUC: relation between true-positive rate and false positive rate.

2. Resample the training set

Two approaches to make a balanced dataset out of imbalanced dataset out of an imbalanced.

1. Under-sampling
2. Over-sampling

i) **Under-sampling**

Under-sampling balances the dataset by reducing the size of the abundant class.. This method is used when quantity of dataset is sufficient. By randomly selecting an equal number of samples in the abundant class, a balanced new dataset can be retrieved for modelling.

ii) **Over-sampling**

Over-sampling is used when quantity of data is insufficient. It tries to balance the dataset by increasing the size of rare samples. New rare samples are generated by using e.g. repetition, SMOTE (Synthetic Minority Over-sampling Technique).

3. Use K-fold Cross-Validation In The Right Way

Over-sampling takes observed rare samples and applies SMOTE to generate new random data based on distribution function. If cross validation is

n models with changing data
samples for the abundant class

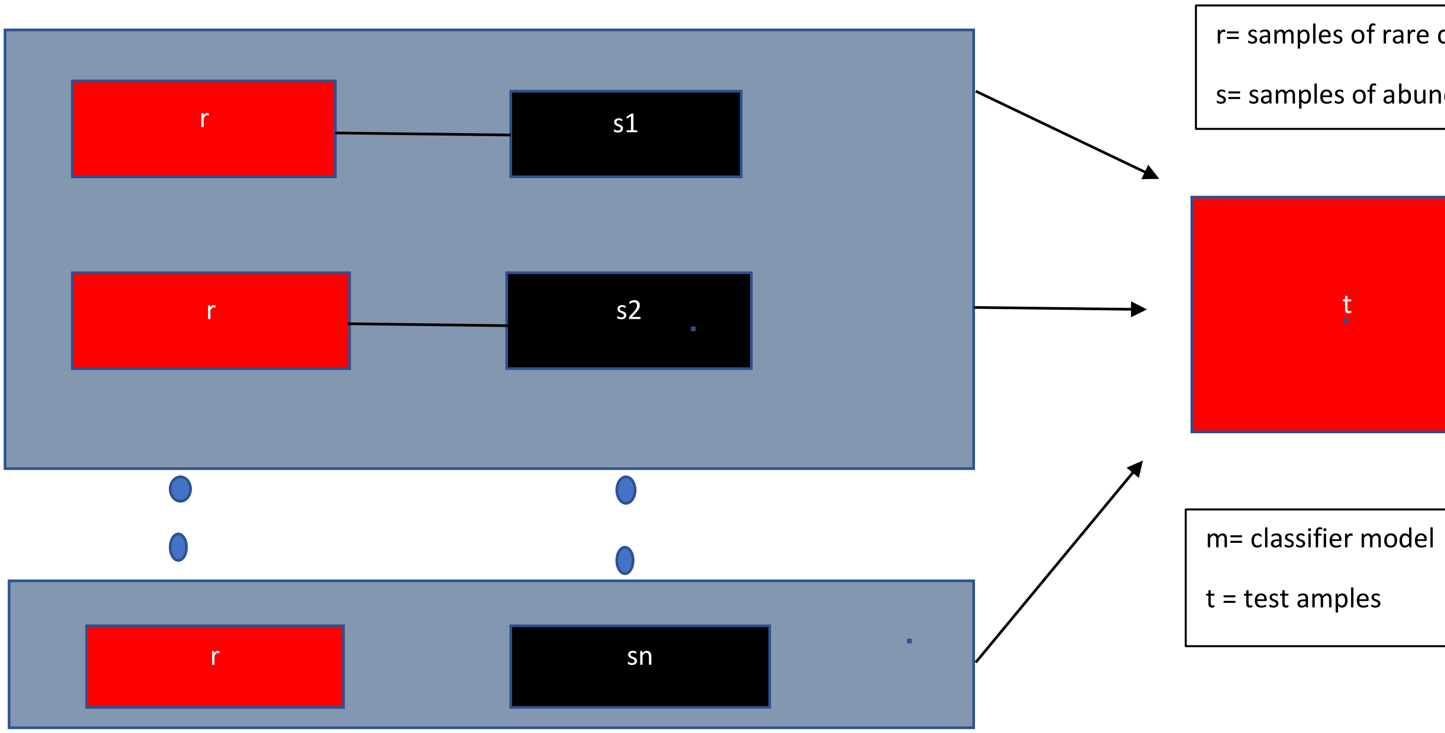
applied after over-sampling basically what we are doing is overfitting our model to a specific artificial SMOTE result That is why cross-validation should always be done before over-sampling the data, just as how feature selection should be implemented. Only by resampling the data repeatedly, randomness can be introduced into the dataset to make sure that there won't be an overfitting problem.

4. ***Ensemble Different Resample Dataset***

The easiest way to successfully generalize a model is by using more data. The problem is that out-of-the-box classifiers like logistic regression or random forest tend to generalize by discarding the rare class. One easy best practice is building n models that use all the samples of the rare class and n-differing samples of the abundant class.

Suppose we want to ensemble 10 models, we would keep e.g. the 500 cases of the rare class and randomly sample 5000 cases of the abundant class. Then you just split the 5000 cases in 10 chunks and train 10 different models.

n models with changing data samples for the abundant class



This approach is simple and perfectly horizontally scalable if you have a lot of data. Ensemble models also tend to generalize better, which makes this approach easy to handle.

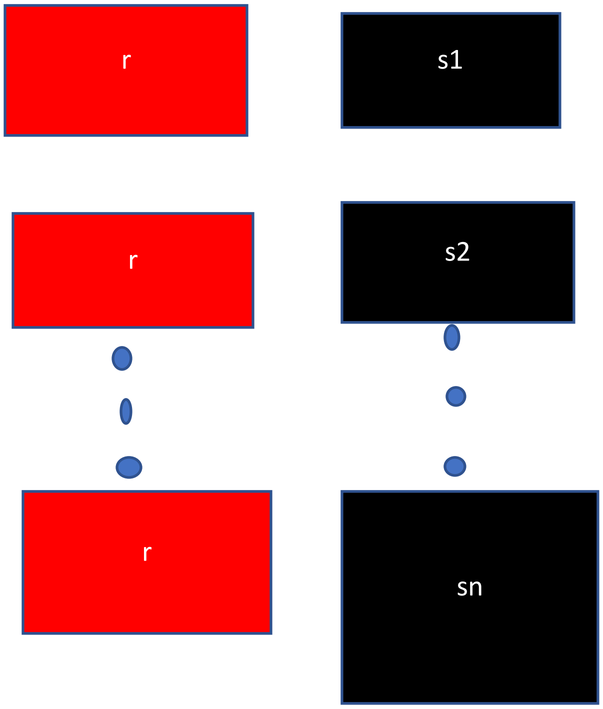
5. ***Resemble With Different Ratios***

The previous approach can be fine-tuned by playing with the ratio between the rare and the abundant class. The best ratio heavily depends on the data and the models that are used.

But instead of training all models with the same ratio in the ensemble, it is worth trying to ensemble different ratios So if 10 models are trained, it might make sense to have a model that has a ratio of 1:1 (rare:abundant) and another one with 1:3, or even 2:1. Depending on the model used this can influence the weight that one class gets.

n models with changing data
samples for the abundant class

n models with changing ratios between rare and abundant class.



6. ***Cluster The Abundant Class***

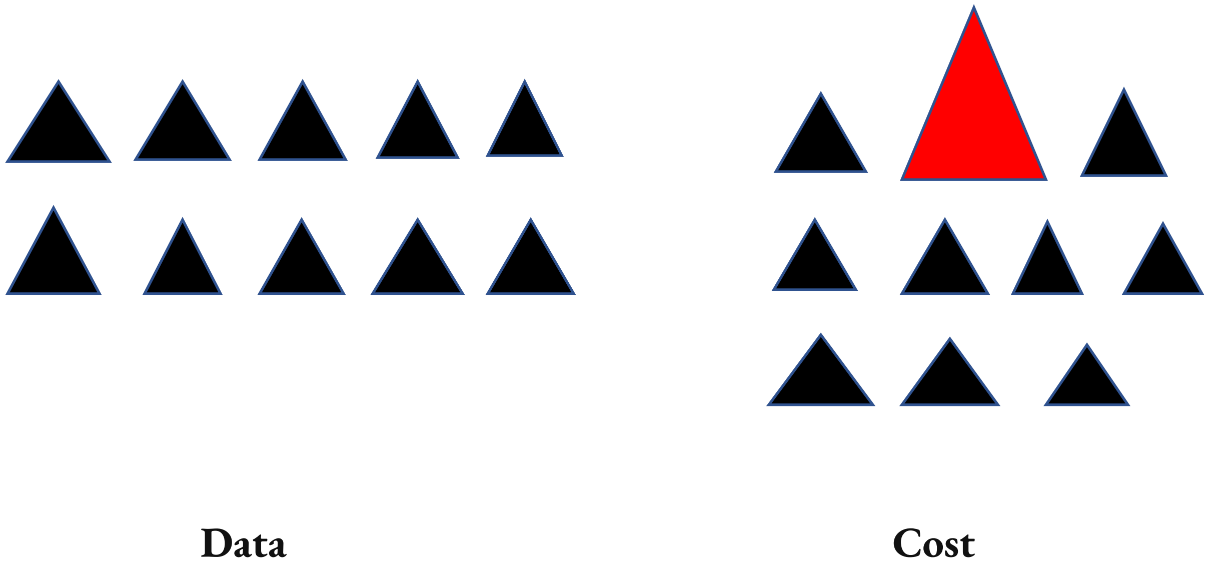
Sergey on Quora suggests clustering the abundant class in r groups, with r being the number of cases in r. For each group, only the medoid (centre of cluster) is kept. The model is then trained with the rare class and the medoids

7. ***Design Our Model***

The famous XGBoost is already a good starting point if the classes are not skewed too much, because it internally takes care that the bags it trains on are not imbalanced. But then again,

n models with changing data
samples for the abundant class

it is possible to design many models that naturally generalize in favour of the rare class. For example, tweaking an SVM to penalize wrong classifications of the rare class by the same ratio that this class is underrepresented.



<u>Ans 13</u>

Synthetic Minority Oversampling Technique (SMOTE) is a statistical technique for increasing the number of cases your dataset in a balanced way. The component works by generating new instances from existing minority cases that you supply as input.

ADASYN is a improved version of Smote. What it does is same as SMOTE just with a minor improvement. After creating those sample it adds a random small values to the points thus making it more realistic.

n models with changing data
samples for the abundant class

The key difference between ADASYN and SMOTE is that the former uses a density distribution, as a criterion to automatically decide the number of synthetic samples that must be generated for each minority sample by adaptively changing the weights of the different minority samples to compensate for the skewed ..

Ans 14

GridSearchCV is a technique for finding the optimal parameter values from a given set of parameters in a grid.

No, GridSearch CV is not preferable to use in case of large datasets because you do grid search on all of your data, the error on your test set will be biased low, and when we go to apply your model to new data, the error could be much higher (and likely will, except for the effects of randomness).

we should only use grid search on the training data *after* doing the train/test split, if we want to use the performance of the model on the test set as a metric for how your model will perform when it really does see new data.

Ans 15

There are six evaluation metrics used to evaluate a regression model.

1. ***Mean Absolute Error***

MAE is a very simple metric which calculates the absolute difference between actual and predicted values.

n models with changing data
samples for the abundant class

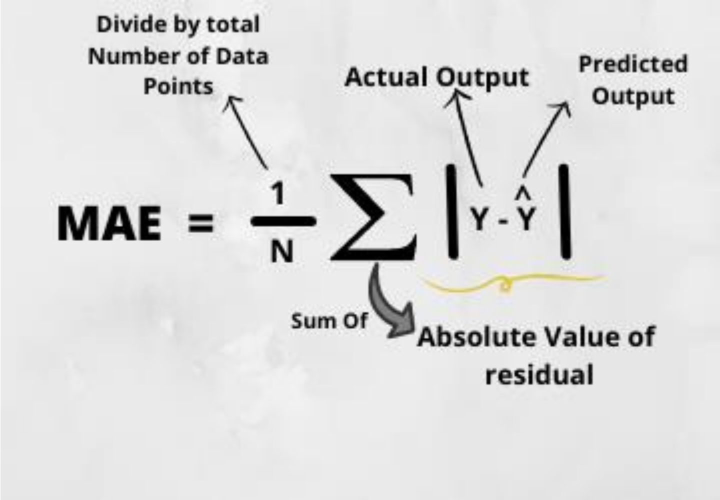
let’s take an example you have input data and output data and use Linear Regression, which draws a best-fit line.

We have to find the MAE of your model which is basically a mistake made by the model known as an error. Now find the difference between the actual value and predicted value that is an absolute error but we have to find the mean absolute of the complete dataset.

sum all the errors and divide them by a total number of observations And this is MAE. And we aim to get a minimum MAE because this is a loss.

The mean absolute error between your expected and predicted values can be calculated using the mean absolute error() function from the scikit-learn library.

The function takes a one-dimensional array or list of expected values and predicted values and returns the mean absolute error value.


Divide by total Number of Data Points
Actual Output
Predicted Output
MAE = $\frac{1}{N}\sum|Y - \hat{Y}|$
Sum Of
Absolute Value of residual

Advantages of MAE

- The MAE you get is in the same unit as the output variable.
- It is most Robust to outliers.

n models with changing data
samples for the abundant class

Disadvantages of MAE

- The graph of MAE is not differentiable so we have to apply various optimizers like Gradient descent which can be differentiable.

2) ***Mean Squared Error(MSE)***

(MSE) squared error states that finding the squared difference between actual and predicted value.

It represents the squared distance between actual and predicted values. we perform squared to avoid the cancellation of negative terms and it is the benefit of MSE.

between actual and
predicted

Advantages of MSE

The graph of MSE is differentiable, so we can easily use it as a loss function.

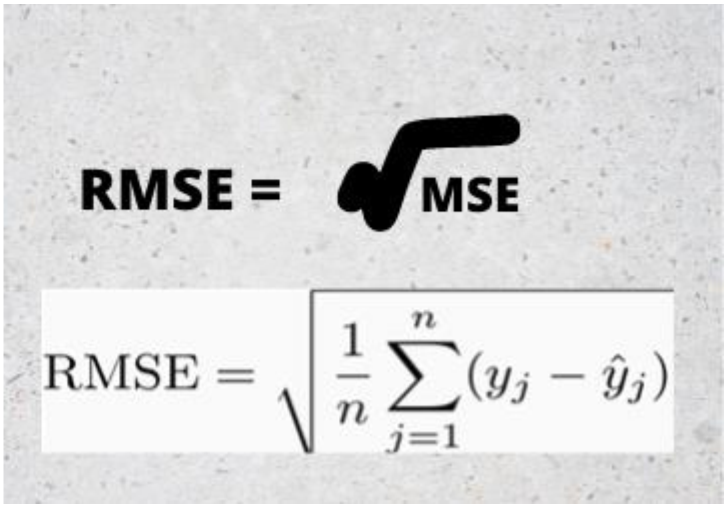
Disadvantages of MSE

- If we have outliers in the dataset then it penalizes the outliers most and the calculated MSE is bigger.
- The value we get after calculating MSE is a squared unit of output. for example, the output variable is in meter(m) then after calculating MSE the output we get is in meter squared.

n models with changing data
samples for the abundant class

3. ***Root Mean Squared Error(RMSE)***

It is a simple square root of mean squared error.



$$\text{RMSE} = \sqrt{\text{MSE}}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)}$$

Advantages of RMSE

- The output value you get is in the same unit which makes interpretation of loss easy.

Disadvantages of RMSE

- It is not that robust to outliers as compared to MAE.

4) ***Root Mean Squared Log Error(RMSLE)***

The output will vary on a large scale to control this situation of RMSE we take the log of calculated RMSE error and resultant we get as RMSLE.

For RMSLE we have to use the NumPy log function over RMSE.

n models with changing data
samples for the abundant class

5) ***R Squared (R2)***

R2 score is a metric that tells the performance of our model, not the loss in an absolute sense that how many wells did our model perform.

MAE and MSE depend on the context as we have seen whereas the R2 score is independent of context.

R2 squared calculates how must regression line is better than a mean line. Hence, R2 squared is also known as Coefficient of Determination .

6) ***Adjusted R Squared***

The disadvantage of the R2 score is while adding new features in data the R2 score starts increasing or remains constant but it never decreases because It assumes that while adding more data variance of data increases. But the problem is when we add an irrelevant feature in the dataset then at that time R2 sometimes starts increasing which is incorrect.

To control this situation Adjusted R Squared .

R_a^2 = adjusted R^2

.

Now as K increases by adding some features so the denominator will decrease, n-1 will remain constant. R2 score will remain constant or will increase slightly so the complete answer will increase and when we subtract this from one then

n models with changing data
samples for the abundant class

the resultant score will decrease. so this is the case when we add an irrelevant feature in the dataset.

And if we add a relevant feature then the R2 score will increase and 1-R2 will decrease heavily and the denominator will also decrease so the complete term decreases, and on subtracting from one the score increases.

ANSWERSHEET-1 (PYTHON)

Ans 1 (C) %

Ans 2 (D) 0.67

Ans 3 (C) 24

Ans 4 (D) 0

Ans 5 (C) 0

Ans 6 (C) The finally block will be executed no matter if the try block raises an error or not.

Ans 7 (A) It is used to raise an exception.

Ans 8 (C) In defining a generator

ANSWERS 9 TO ANSWERS 10

Ans 9 (A)_abc

(B) 1abc

(C) abc2

Ans 10 (A) yield

(B) raise

Ans11 =A python program to find the factorial of a number.

In [2]:

```python
# Python 3 program to find
# factorial of given number

def fact(n):
    return 1 if (n==1 or n==0) else n * fact(n - 1);

num = 5

print("Factorial of",num,"is",)

fact(num)
```

```
Factorial of 5 is
```

Out[2]:

```
120
```

Ans 12= A python program to find whether a number is prime or composite.

In [17]:

```python
num = 11
# If given number is greater than 1
if num > 1:
    # Iterate from 2 to n / 2
    for i in range(2, int(num/2)+1):
        # If num is divisible by any number between
        # 2 and n / 2, it is not prime
        if (num % i) == 0:
            print(num, "is not a prime number")
            break
    else:
        print(num, "is a prime number")
else:
    print(num, "is a composite number")
```

```
11 is a prime number
```

Ans 13= A python program to check whether a given string is palindrome or not.

In [18]:

```python
# function to check string is
# palindrome or not
def isPalindrome(str):

    # Run loop from 0 to len/2
    for i in range(0, int(len(str)/2)):
        if str[i] != str[len(str)-i-1]:
            return False
    return True

# main function
s = "malayalam"
ans = isPalindrome(s)

if (ans):
    print("Yes")
else:
    print("No")
```

```
Yes
```

Ans 14= A Python program to get the third side of right-angled triangle from two given sides.

In [30]:

```python
import math
a = 6
b = 8
c = math.sqrt(a ** 2 + b ** 2)
print(c)
```

```
10.0
```

Ans 15= A python program to print the frequency of each of the

characters present in a given string.

In [31]:

```
str = "YOLO LIFE"

# create dictionary to store key value pair
dict = {}

for i in str:
    # if i already appears as key in dict, increment the count
    if i in dict:
        dict[i] += 1

    # else i appears for the first time, add to dict
    else:
        dict[i] = 1

# printing result
print(dict)
```

```
{'Y': 1, 'O': 2, 'L': 2, ' ': 1, 'I': 1, 'F': 1, 'E': 1}
```

In [ ]:

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

WORKSHEET –1 (STATISTICS)

Ans 1 (B) The probability of failing to reject H0 when H1 is true.

Ans 2 (B) Null hypothesis.

Ans 3 (D) Type I error.

Ans 4 (B) The t distribution with n - 1 degrees of freedom

Ans 5 (C) Rejecting Ho when it is false.

Ans 6 (D) A two-tailed test.

Ans 7 (B) The probability of committing a Type I error.

Ans 8 (A) The probability of committing a Type II error.

Ans 9

Ans 10 (C) The level of significance.

Ans 11 (A) Level of significance.

Ans 12 (B) The t-ratio.

Ans 13

Analyse in variance (AVONA) in SPSS must have a dependent variable which should be metric. ANOVA in SPSS must have one or ore independent variables, which should be categorical in nature. In ANOVA in SPSS, categorical dependent variables are called factors. A particular combination of factor levels, or categories, is called treatment.

Analyse in variance ANOVA in SPSS , is used for examining the differences in the mean values of dependent variable associated with the effect of the controlled independent variables after taking influence of the uncontrolled independent variables. ANOVA in SPSS is used as the test of means for two or more populations.

Ans 14

ASSUMPTIONS OF ANOVA

1. The experimental errors of our data are normally distributed.

2. Equal variance between treatments.

3. Independence of samples.

ASSUMPTION 1 - . The experimental errors of our data are

“If I was to repeat my sample repeatedly and calculate the means, those

means would be normally distributed.”

Testing for Normality - Shapro Wilks Test

Tests the hypotheses:

$HO: distribution\ of\ residuals = normal\ distribution$

$Ha: distribution\ of\ residuals \neq normal\ distributi$on

Non- significant p-value = Normal distribution

ASSUMPTION 1 # The experimental errors of our data are

1) **Shaprio-Wilks normality test** — if our data is mainly unique values.
2) **Spiegelhalter's T' normality test** – powerful non-normality is due to kurtosis, but bad if skewness is responsible.

3) **D'Agostino-Pearson normality test –** if you have lots of repeated values.
4) **Lilliefors normality test** – mean and variance are unknown.

Central Limit Theorem:

"Sample means tend to cluster around the central population value."

<u>For large N</u>:

The assumption for Normality can be relaxed ANOVA not really compromised if data is non-normal.

<u>Assumption of Normality is important when:</u>

1) Very small N.
2) Small effect size.
3) Highly non-normal.

Assumptions 2# - Equal variance between treatments.

<u>Bartlett Test</u>

Tests the hypotheses: Ho: $varianceA$ = $varianceB$

Ha: $varianceA$ ≠ $variance$B

<u>NORMAL distribution</u>: equal number of points along observed

- EQUAL variances: equal spread on either side of the mean predicted value=0

- Good to go!

<u>NON-NORMAL distribution</u>: unequal number of points along observed

- EQUAL variances: equal spread on either side of the meanpredicted value=0

- Optional to fix.

NORMAL/NON NORMAL: look at histogram or test.

- UNEQUAL variances: cone shape – away from or towards zero.

- This needs to be fixed for ANOVA.(transformation)

OUTLIERS: points that deviate from the majority of data points.

- This needs to be fixed for ANOVA (transformation or removal).

ASSUMPTION 3# Independence of samples.

"Your samples have to come from a randomized or randomly sampled design."

Pseudoreplication

A particular combination of experimental design (or sampling) and statistical analysis which is inappropriate for testing the hypothesis of interest.

Occurs when a number of observations or the number of data points are treated inappropriately as independent replicates.

Observations may not be independent if:

(1) repeated measurements are taken on the same subject

(2) observations are correlated in time

(3) observations are correlated in space.

Pseudoreplication

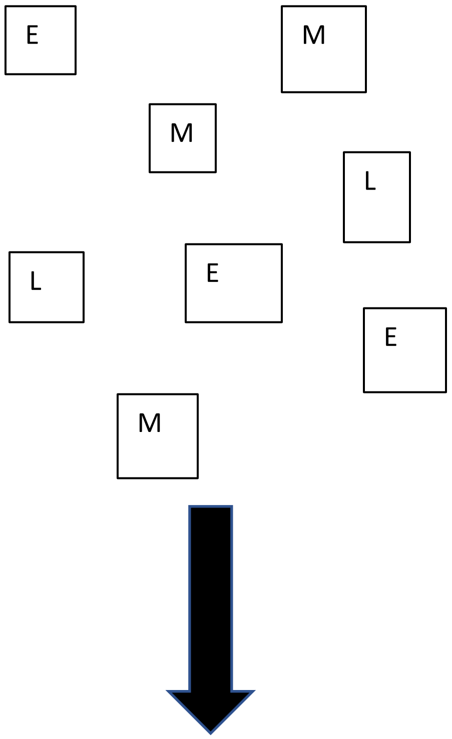
# Difficult to measure the variance between (signal) and the variance within (noise).

# Pseudoreplicates because they are not spatially independent.

# Could measure alternative variables (treatements, covariates) – but often hard to do.

# Cannot prove there are no confounding factors

- Environmental gradient
- Topographical gradient

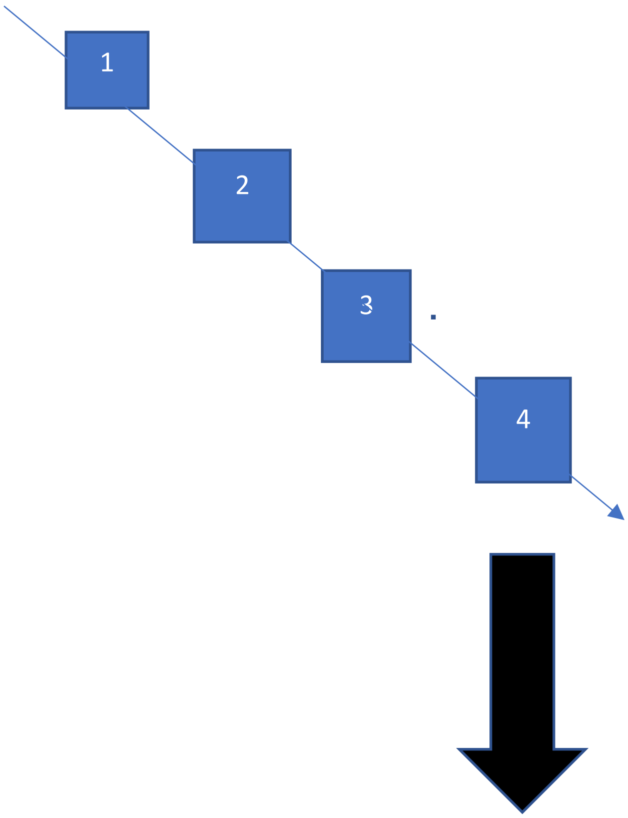


Plot ID	Stage Scale	Sp1	Sp2	Sp3
1	E			

2	**M**			
3	**L**			
4	**M**			
5…	**L**			

# Pseudoreplication – experiments with transects

A



PLOT ID	**TRANSECT**	**PLOT 1**	**PLOT 2**	**PLOT 3**	**PLOT 4**	**ALL PLOTS**
1	**A**					
2	**B**					
3	**C**					
4	**D**					
5…	**E…**					

ASSUMPTION 3# Independence of samples.

Systematic arrangements

A	**C**	**B**
B	**A**	**C**
C	**B**	**A**

Poor practice

@ Distinct pattern in how treatments are laid out.

@ “More random than randomized”.

@ If your treatments effect one another – the individual treatment effects could be masked or overinflated.

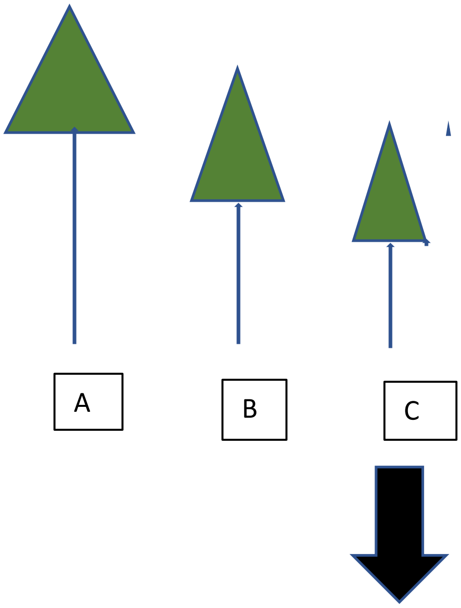
Randomized

A	**B**	**A**
B	**C**	**B**
C	**C**	**A**

@ Good practice

@ No distinct pattern in how treatments are laid out.

@ If your treatments effect is strong enough it will emerge as significant despite the leaching issue.

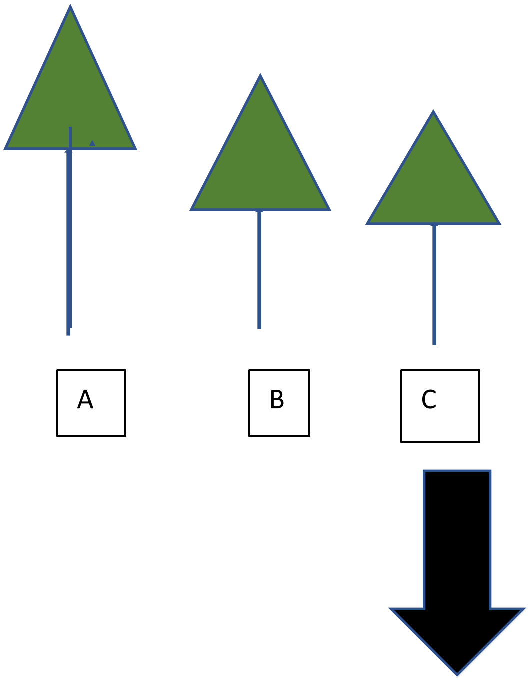


ID	VARIETY	YEAR	HT
1	A	1	17
2	A	2	18
3	A	3	19
4	B	1	12
5	B	2	14
6	B	3	13
7	C	1	7
8	C	2	8
9	C	3	9

@ ANOVA assume each row of data you enter is an independent observation.

@ So if we run a simple ANOVA to determine the effect of VARIETY on HT we would me misinforming the analysis.

<u>Temporal Independence</u>



ID	VARIETY	YEAR	HT1	HT2	HT3
1	A	1	17	18	19
2	B	2	12	13	14
3	C	3	7	8	9

1. We need multiple (independent) trees for each VARIETY to correctly answer this question

2. We would put HT in separate columns.

Ans 15

The difference between One-way Anova and Two-way Anova are: -

<u>One Way Anova</u>:

- One Independent Variable
- One way ANOVA is a hypothesis test, used to test the equality of three of more population means simultaneously using variance.
- Three or more levels of one factor.
- Need not to be same in each group.

- Need to satisfy only two principles.

Two way Anova:

- Two Independent Variables.
- Two-way ANOVA is a statistical technique wherein, the interaction between factors influencing variables can be studied.
- Two Independent Variables.
- Effect of multiple levels of two factors.
- Need to be equal in each group.
- All three principles needs to be satisfied