**Descriptive statistics - 1**

type of data - quantitative (continuous, discrete), categorical (nominal, ordinal)

measures of centre - mean, median, mode

notation

random variable (a place holder to hold some data, say X)

**Descriptive statistics - 2**

measures of spread (best represented by boxplot)- range, inter quartile range(IQR), standard deviation and variance

5 munber summary of data - min, max, Q1, Q2(median), Q3

IQR = Q3 - Q1

range = max - min

measures of shape

normal distribution - mean == median == mode

left skewed distribution - mean < median < mode

right skewed distribution - mode < median < mean

measures of outliers

**Admissions case study**

simpson's paradox - use of statistics incorrectly

**Probability**

In probability we make predictions of future events based on models or causes that we assume || we are predicting data

In statistics we analyze data from past events to infer what those models or causes could be || we use data to predict

Non dependent events (non conditional): The outcome of a events donot depends on previous events, like if we toss coin twice, 2nd toss

is not dependent on 1st toss.

Probability of single event : P(A) = 1 - P(not A)

Probability of composite events : If P(Head) = 0.5, and we flip coin twice then probability of getting 2 heads is , P(H) \* P(H) = 0.5 \* 0.5 = .25

**Binomial distribution (used for independent events)**

The Binomial Distribution helps us determine the probability of a string of independent 'coin flip like events'.

The probability mass function associated with the binomial distribution is of the following form:

P(X=x) = ( n! / x!(n-x)! ) \* p^x \* (1-p) ^ (n-x)

where n is the number of events, x is the number of "successes", and p is the probability of "success"

EG: We flip a loaded coin 12 times (P(H) = 0.8) and we need to find probability of getting 9 heads.

( 12! / 9! \* 3 ! ) \* (0.8^9) \* (0.2^3) = 0.236

​​

**Conditional probability**

**Bayes rule**

**Python probability practice**

**Normal distribution**

**Sampling distribution and the central limit theorem**

Statistic - Any numerical summary calculated from sample.

Parameter - Any numerical summary calculated from population.

Inference - Drawing conclusions regarding a parameter based on our statistics is known as inference.

Sampling distribution: Suppose we have data of 21 students about who drink coffee and who don't. Now if we look at distribution of proportion of

students who drink coffee and who don't across all sample of size 5, it's call sampling distribution. Sample size could be

anything.

Sampling distributions is the distribution of a statistic.

Notation

2 theorems for sampling distributions - law of large numbers and central limit theorem.

The Law of Large Numbers states that as a sample size increases, the sample mean will get closer to the population mean.

In general, if our statistic is a "good" estimate of a parameter, it will approach our parameter with larger sample sizes.

The Central Limit Theorem states that with large enough sample sizes our sample mean will follow a normal distribution,

but it turns out this is true for more than just the sample mean.

Central limit theorem applies for:

- means

- proportion

- difference in means

- difference in proportion

**Bootstrapping**

Sampling with replacement.

EX: using np.random.choice with replace=True (default)

Ex : In dice rolls values could be (1,2,3,4,5,6), in case of non bootstrap sample size could be max 6 (all values repeating once only),

but if we use bootstrap sampling, we could get a sample size of even 100 where any value could repeat any number of time but they

will be only in range of 1-6, we could even have 1 repeating 100 times in corner cases.

Very useful in machine learning algorithms like random forest etc.

Bootstrapping focuses on concept of requiring no more data needed to gain a better understanding of the parameter.

Bootstrapping is a technique where we sample from a group with replacement.

By bootstrapping and then calculating repeated values of our statistics, we can gain an understanding of the sampling distribution of our statistics.

**Confidence intervals**

Zero is the null value of the parameter (in this case the difference in means).

If a 95% confidence interval includes the null value, then there is no statistically meaningful or

statistically significant difference between the groups.

Important : If confidence interval contains 0 it means its not statistically significant and H0 is not rejected.

**Hypothesis testing**

Ho - Null hypothesis is the condition we believe to be true even before we collect any data.

Mathematically null hypothesis is commonly a statement of 2 groups being equal or of an effect

being zero.

Null and alternative hypothesis should be competing and non overlapping hypothesis.

Holds some sort of equal sign (== , <= , >=)

H1 - Alternative hypothesis is the statement of what you want or what you want to prove to be true.

Holds != , < , > signs.

Type of error.

actual: guilty innocent

jury

guilty Type 1

innocent Type 2

Type 1 error

- If a person is innocent but we made them guilty.

- An error where alternative hypothesis is chosen but null hypothesis is actually true.

- You should set up your null and alternative hypothesis, so that the worse of your errors is the type I error.

- They are denoted by the symbol α.

- The definition of a type I error is: Deciding the alternative (H​1​​ ) is true, when actually (H​0​) is true.

- Type I errors are often called false positives.

Type 2 error

- If a person is guilty but we made them innocent.

- An error where null hypothesis is chosen but alternative hypothesis is actually true.

- They are denoted by the symbol β.

- The definition of a type II error is: Deciding the null (H​0) is true, when actually (H​1​​) is true.

- Type II errors are often called false negatives.

Common hypothesis test:

Common hypothesis tests include:

Testing a population mean (One sample t-test).

Testing the difference in means (Two sample t-test)

Testing the difference before and after some treatment on an the same individual (Paired t-test)

Testing a population proportion (One sample z-test)

Testing the difference between population proportions (Two sample z-test)

P-value

When you perform a hypothesis test in statistics, a p-value helps you determine the significance of your results.

The p-value is a number between 0 and 1 and interpreted in the following way: A small p-value (typically ≤ 0.05)

indicates strong evidence against the null hypothesis, so you reject the null hypothesis.

P value is the conditional probability of the data, given null hypothesis is true.

Large P value suggests we should not move away from null hypothesis.

Small P-value means its statistically significant and we should reject the null hypothesis.

Example 1: Which of two marketing campaign will drive more traffic to website.

H0 : Camp new = Camp old

H1 : Camp new != Camp old

Example 2: If a new campaign is better than old campaign.

H0 : Camp new <= camp old || camp new - camp old <= 0

H1 : Camp new > camp old || camp new - camp old > 0

Confidence interval could be used for 2 tailed hypothesis testing.

Say Ho : mean = 25 and H1 : mean != 25. alpha = 0.10

Since 90% CI means alpha level of 0.10

Assume at 90% confidence we got a range (10, 30), Now since mean = 25 falls in the interval

so we will not reject the null.

Basically, in 1-sided or 2-sided test if H0 lies in CI range then we fail to reject the null.

**A/B Test Case study**

User flow from start to end in a website is called user/customer funnel.

Click through rate = number of unique visitors clicked on a page / number of unique visitors visiting the page

**Regression**

Simple linear regression - get relationship between 2 quantitative variables.

Whatever we want to predict for becomes the Y axis.

The variable in Y axis is called Response or Dependent variable.

The variable in X axis is called Explanatory or Independent variable.

Correlation coefficients - Correlation coefficients provide a measure of the strength and direction of a linear relationship.

-tive value means -tive relationship and +tive means +tive.

1 - 0.7 -> strong relationship

0.3 - 0.7 -> moderate relationship

0 - 0.3 -> weak relationship

Line - A line is commonly identified by an intercept and a slope.

The intercept is defined as the predicted value of the response when the x-variable is zero.

The slope is defined as the predicted change in the response for every one unit increase in the x-variable.

We notate the line in linear regression in the following way:

​y^ = b0 ​+ b​1​​x​1

​​

where

y^ is the predicted value of the response from the line.

b0 is the intercept.

b1 is the slope.

x1 is the explanatory variable.

y is an actual response value for a data point in our dataset (not a prediction from our line).

Fitting A Regression Line - The main algorithm used to find the best fit line is called the least-squares algorithm.

R squared (square of correlation coefficient) . Closer this value is to 1 the better our model fits and the closer the value to 0,

the worst the model fits.

R squared is defined as amount of variability in the response (Y) that can be explained by your explanatory variable or model.

Say your r-squared is .67 of linear regression between price(y) and area(x) of your house, you are predicting price.

It means 67 % of variability in price can be explained by area and rest 33 % could be explained by other characteristic.

In linear regression by default for calculating p value it takes 2 tailed hypothesis test. So if p value is less than alpha then its statistically significant

that the value is important as it supports the alternate hypothesis in which its not equal to 0.

**Multiple linear regression**

Dummy variable for adding categorical values:

The way that we add categorical variables into our multiple linear regression models is by using dummy variables.

The most common way dummy variables are added is through 1, 0 encoding. In this encoding method, you create a new

column for each level of a category (in this case A, B, or C).

When we add these dummy variables to our multiple linear regression models, we always drop one of the columns.

The column you drop is called the baseline. The coefficients you obtain from the output of your multiple linear

regression models are then an indication of how the encoded levels compare to the baseline level (the dropped level).

Identify multicollinearity - we would like to have some relationship between X and Y variables, but if there is relationship between

2 X variables then this is bad for our prediction. We could identify multicollinearity by : Scatter plot matrix and VIFs

VIFs (variance inflation factors) - should be less than 10. If present there will be at least 2 variables having VIFs greater than 10. if we remove one and check r-squared.

It will be equal as earlier with both variables used to predict in model. It means both variables are not required for prediction only

1 (any one of them) is sufficient.

**Logistic regression:**

Confusion matrix – similar to error types matrix

Accuracy - number of correct labels / number of rows

Precision and recall – the probability of an algorithm to correctly identify something.

Say you have a picture of Mahatma Gandhi and various other people. You are showing them to various people for prediction. You actually showed picture of mahatma Gandhi 50 times, people predicted Mahatma Gandhi on any random picture 40 times, out of which 35 times they predicted correctly, rest 5 times they predicted mahatma Gandhi on someone else’s picture.

**Recall = 35 / 50 = 0.7 (no of correct prediction/ no of event)**

**Precision = 35 / 40 = 0.85 (no of correct prediction/ no of total prediction)**

**True positive = 35**

**False positive = 40 – 35 = 5 (type 1) (no of times Mahatma Gandhi not shown but predicted)**

**False negative = 50 – 35 = 15 (type 2) (no of times Mahatma Gandhi shown but not identified)**

True positive, false positive and false negative.

