Explain Goodness of fit Test? in detail

The goodness of fit test is a statistical method used to assess how well a sample data set aligns with a specific probability distribution or theoretical model. This test helps in determining whether the observed data fits the expected distribution within a certain margin of error. Goodness of fit tests are commonly used in various fields such as economics, biology, social sciences, and quality control to validate assumptions and make reliable conclusions based on data analysis.

Here is a more detailed explanation of the goodness of fit test:

1. **Formulating Hypotheses**:
   - Null Hypothesis (H0): This hypothesis assumes that the sample data follows a specified distribution or model.
   - Alternative Hypothesis (Ha): This hypothesis states that the sample data does not fit the specified distribution.
2. **Selecting a Test Statistic**:
   - The choice of test statistic depends on the type of data and the distribution being tested. Common test statistics include the chi-square test, Kolmogorov-Smirnov test, Anderson-Darling test, etc.
3. **Collecting Data**:
   - A sample data set is collected or observed, and the expected distribution or model is specified based on theoretical considerations.
4. **Calculating the Test Statistic**:
   - The test statistic is computed using the sample data and the specified theoretical distribution or model. This calculated statistic quantifies the difference between the observed data and the expected distribution.
5. **Determining the Significance Level**:
   - Based on the test statistic and the degrees of freedom, a p-value is calculated. The p-value represents the probability of obtaining results as extreme as the ones observed, assuming that the null hypothesis is true.
6. **Making a Decision**:
   - If the p-value is less than a chosen significance level (commonly 0.05), the null hypothesis is rejected. This suggests that there is significant evidence to conclude that the observed data does not fit the specified distribution.

Goodness of fit tests are valuable tools for researchers and analysts to validate statistical models, ensure data quality, and make informed decisions based on the analysis of data sets. By determining how well the collected data fits the expected distribution, researchers can draw meaningful conclusions and make accurate predictions.

Give one example...

Outlier:

The Five Number Summary

The five numbers that help describe the **center, spread and shape of data are:**

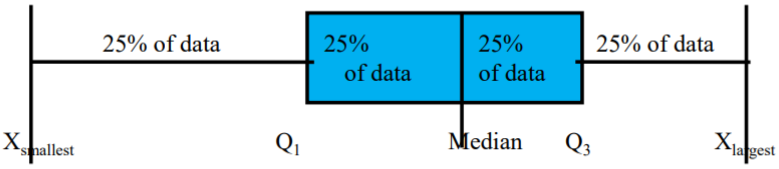
- $X_{smallest}$
- First Quartile (Q_1)
- Median (Q_2)
- Third Quartile (Q_3)
- $X_{largest}$

Five Number Summary and The Boxplot

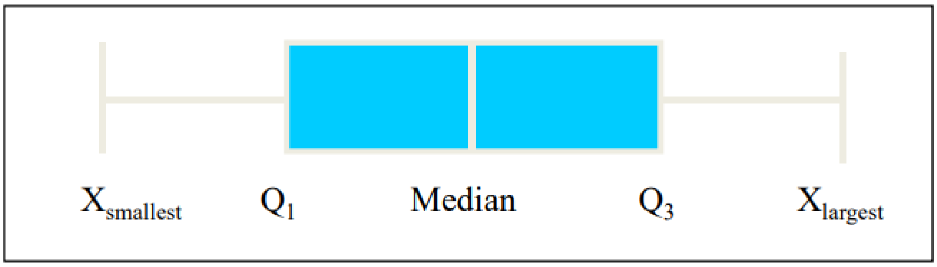
- The Boxplot: A Graphical display of the data based on the five-number summary:

$X_{smallest}$ -- Q_1 -- Median -- Q_3 -- $X_{largest}$

Example:



- If data **are symmetric around** the median then the box and central line are **centered between the endpoints**



- A Boxplot can be shown in either a **vertical or horizontal** orientation

Arrange value in ascending order.

Locating Quartiles

Find a quartile by determining the value in the appropriate **position** in the ranked data, where

First quartile position: $Q1 = (n+1)/4$ ranked value

Second quartile position: $Q2 = (n+1)/2$ ranked value

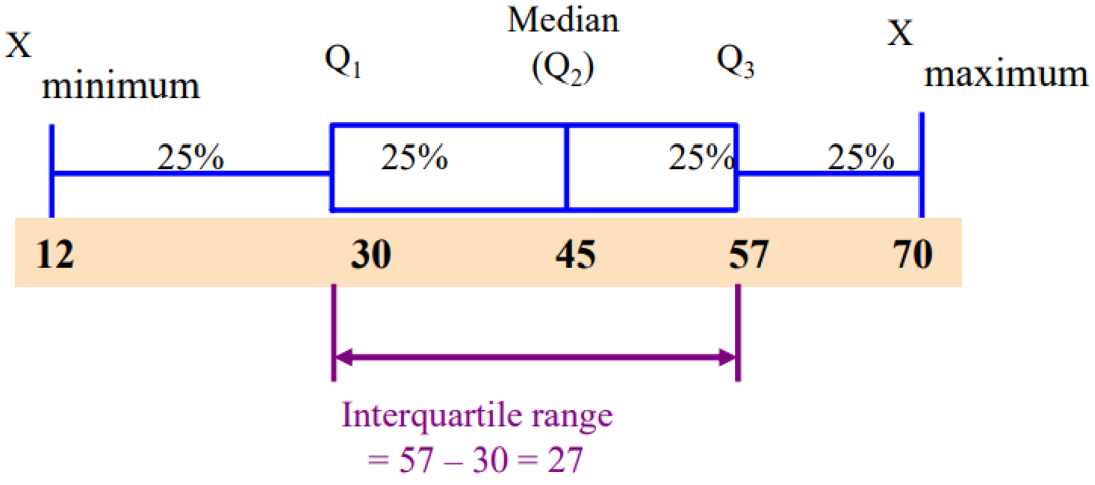
Third quartile position: $Q3 = 3(n+1)/4$ ranked value

- where n is the number of observed values

Quartile Measures: The Interquartile Range (IQR)

- The IQR is $Q_3 – Q_1$ and measures the spread in the **middle 50% of** the data
- The IQR is also called the **midspread** because it covers the middle 50% of the data
- The IQR is ***a measure of variability that is not influenced by* *outliers*** or extreme values
- Measures like Q_1, Q_3, and IQR that are ***not influenced by outliers are called* *resistant measures***

Example:



Multiplication Law of Probability

The Multiplication Law of Probability is used to determine the probability of the intersection of two events, that is, the probability that both events occur. The formula depends on whether the events are independent or dependent.

Independent Events

Two events AA and BB are independent if the occurrence of one does not affect the occurrence of the other. The multiplication law for independent events is given by:

$$P(A\cap B)=P(A)\cdot P(B)P(A\cap B)=P(A)\cdot P(B)$$

Example:

Suppose you flip a coin and roll a die. Let event AA be getting a head on the coin, and event BB be rolling a 4 on the die. These events are independent.

- Probability of getting a head, $P(A)=12P(A)=21$
- Probability of rolling a 4, $P(B)=16P(B)=61$

Using the multiplication law for independent events:

$$P(A\cap B)=P(A)\cdot P(B)=12\cdot 16=112P(A\cap B)=P(A)\cdot P(B)=21\cdot 61=121$$

Dependent Events

Two events AA and BB are dependent if the occurrence of one affects the probability of the occurrence of the other. The multiplication law for dependent events is given by:

$$P(A∩B)=P(A)\cdot P(B|A)P(A∩B)=P(A)\cdot P(B|A)$$

where $P(B|A)P(B|A)$ is the conditional probability of BB given AA.

Example:

Suppose you draw two cards from a deck without replacement. Let event AA be drawing an Ace first, and event BB be drawing a King second.

- Probability of drawing an Ace first, $P(A)=452P(A)=524$
- Probability of drawing a King second given an Ace was drawn first, $P(B|A)=451P(B|A)=514$

Using the multiplication law for dependent events:

$$P(A∩B)=P(A)\cdot P(B|A)=452\cdot 451=162652=4663P(A∩B)=P(A)\cdot P(B|A)=524\cdot 514=265216=6634$$

Bayes' Theorem

Bayes' Theorem is a way to find the probability of an event given the probability of another related event. It relates the conditional and marginal probabilities of random events. The formula is:

$$P(A|B)=P(B|A)\cdot P(A)P(B)P(A|B)=P(B)P(B|A)\cdot P(A)$$

where:

- $P(A|B)P(A|B)$ is the probability of event AA given event BB.
- $P(B|A)P(B|A)$ is the probability of event BB given event AA.
- $P(A)P(A)$ and $P(B)P(B)$ are the probabilities of events AA and BB respectively.

Example:

Suppose there is a test for a disease that is 99% accurate (both for true positives and true negatives). The disease prevalence in the population is 1%. We want to find the probability that a person has the disease given that they tested positive.

- Let A be the event "person has the disease".
- Let B be the event "person tests positive".

Given:

- $P(A)=0.01$ (prevalence of the disease)
- $P(B|A)=0.99$ (probability of testing positive if diseased)
- $P(B|\neg A)=0.01$ (probability of testing positive if not diseased)

We need to find $P(A|B)$. First, we find $P(B)$:

$$P(B)=P(B|A)\cdot P(A)+P(B|\neg A)\cdot P(\neg A)$$

$$P(B)=0.99\cdot 0.01+0.01\cdot 0.99=0.0099+0.0099=0.0198$$

Now, using Bayes' Theorem:

$$P(A|B)=\frac{P(B|A)\cdot P(A)}{P(B)}=\frac{0.99\cdot 0.01}{0.0198}=\frac{0.0099}{0.0198}=0.5$$

So, the probability that a person has the disease given that they tested positive is 50%. This example illustrates how even a highly accurate test can yield a lower than expected probability of actually having the disease, due to the low prevalence of the disease.

---

1 card is drawn at random from the pack of 52 cards. (i) Find the Probability that it is an honor card. (ii) It is a face card.

- ➔There are 52 cards in a deck.
- There are 4 suits: hearts, diamonds, clubs, and spades.
- Each suit has 13 cards: Ace, 2, 3, ..., 10, Jack, Queen, King.
- Face cards are Jack, Queen, and King. Each suit has 3 face cards, so there are $3\times 4=12$ face cards in total.
-

- ### (i) Probability that it is an honor card

In many contexts, honor cards are defined as the Ace, King, Queen, Jack, and 10. However, if we define honor cards as just the face cards (Jack, Queen, and King), we proceed accordingly.

Number of honor cards:

For face cards, there are 12 in total (3 in each of the 4 suits). If honor cards include 10s, Aces, Kings, Queens, and Jacks, each suit has 5 honor cards (10, J, Q, K, A), making a total of $5\times4=205\times4=20$ honor cards.

Probability calculation:

If honor cards are defined as just face cards (Jack, Queen, King):

- Number of honor cards = 12
- Total number of cards = 52

P(honor card)=Number of honor cardsTotal number of cards=1252=313P(honor card)=Total number of cardsNumber of honor cards=5212=133

If honor cards include 10s, Aces, Kings, Queens, and Jacks:

- Number of honor cards = 20
- Total number of cards = 52

P(honor card)=2052=513P(honor card)=5220=135

(ii) Probability that it is a face card

Number of face cards:

- There are 3 face cards (Jack, Queen, King) in each suit.
- Number of face cards = $3\times4=123\times4=12$

Probability calculation:

- Total number of cards = 52

P(face card)=Number of face cardsTotal number of cards=1252=313P(face card)=Total number of cardsNumber of face cards=5212=133

Frequency Distribution of soft drink purchases:

Soft Drink	Frequency
Coca-Cola	19
Diet Coke	8
Dr. Pepper	5
Pepsi	13
Sprite	5

total	**50**

Calculate the relative frequency for Coca-Cola and Diet Coke. Compute Mean, variance & standard deviation for given soft drink data.

The formula for relative frequency is:

Relative Frequency=Frequency of the soft drinkTotal frequencyRelative Frequency=Total frequencyFrequency of the soft drink

Given data:

- Frequency of Coca-Cola = 19
- Frequency of Diet Coke = 8
- Total frequency (total purchases) = 50

Relative Frequency for Coca-Cola

Relative Frequency of Coca-Cola=19/50=0.38

Relative Frequency for Diet Coke

Relative Frequency of Diet Coke=8/50=0.16

Calculate mean, variance and standard deviation by yourself.