**What is the difference between inferential statistics and descriptive statistics?**

Descriptive statistics – provides exact and accurate information.

Inferential statistics – provides information of a sample and we need to inferential statistics to reach to a conclusion about the population.

**What is the difference between population and sample in inferential statistics?**

From the population we take a sample. We cannot work on the population either due to computational costs or due to availability of all data points for the population.

From the sample we calculate the statistics

From the sample statistics we conclude about the population

**What are descriptive statistics?**

Descriptive statistic is used to describe the data (data properties)

5-number summary is the most commonly used descriptive statistics

**Most common characteristics used in descriptive statistics?**

* Center – middle of the data. Mean / Median / Mode are the most commonly used as measures.
  + Mean – average of all the numbers
  + Median – the number in the middle
  + Mode – the number that occurs the most. The disadvantage of using Mode is that there may be more than one mode.
* Spread – How the data is dispersed. Range / IQR / Standard Deviation / Variance are the most commonly used as measures.
  + Range = Max – Min
  + Inter Quartile Range (IQR) = Q3 – Q1
  + Standard Deviation (σ) = √(∑(x-µ)2 / n)
  + Variance = σ2
* Shape – the shape of the data can be symmetric or skewed
  + Symmetric – the part of the distribution that is on the left side of the median is same as the part of the distribution that is on the right side of the median
  + Left skewed – the left tail is longer than the right side
  + Right skewed – the right tail is longer than the left side
* Outlier – An outlier is an abnormal value
  + Keep the outlier based on judgement
  + Remove the outlier based on judgement

**What is quantitative data and qualitative data?**

Quantitative data is also known as numeric data

Qualitative data is also known as categorical data

**How to calculate range and interquartile range?**

IQR = Q3 – Q1

Where, Q3 is the third quartile (75 percentile)

Where, Q1 is the first quartile (25 percentile)

**What is the benefit of using box plot?**

Shows the summary pictorially:

Low extreme (minimum)

Lower quartile (Q1)

Median

Upper quartile (Q3)

Upper extreme (maximum)

**What is the meaning of standard deviation?**

It represents how far are the data points from the mean

(σ) = √(∑(x-µ)2 / n)

Variance is the square of standard deviation

**What is left skewed distribution and right skewed distribution?**

* Left skewed
  + The left tail is longer than the right side
  + Mean < median < mode
* Right skewed
  + The right tail is longer than the right side
  + Mode < median < mean

**What does symmetric distribution mean?**

The part of the distribution that is on the left side of the median is same as the part of the distribution that is on the right side of the median

Few examples are – uniform distribution, binomial distribution, normal distribution

**What is the relationship between mean and median in normal distribution?**

In the normal distribution mean is equal to median

**What does it mean by bell curve distribution and Gaussian distribution?**

Normal distribution is called bell curve distribution / Gaussian distribution

It is called bell curve because it has the shape of a bell

It is called Gaussian distribution as it is named after Carl Gauss

**How to convert normal distribution to standard normal distribution?**

Standardized normal distribution has mean = 0 and standard deviation = 1

To convert normal distribution to standard normal distribution we can use the formula

X (standardized) = (x-µ) / σ

**What is an outlier?**

An outlier is an abnormal value (It is at an abnormal distance from rest of the data points).

**Mention one method to find outliers?**

Boxplot

Widely used – Any data point that lies outside the 1.5 \* IQR

Lower bound = Q1 – (1.5 \* IQR)

Upper bound = Q3 + (1.5 \* IQR)

**What can I do with outlier?**

* Remove outlier
  + When we know the data-point is wrong (negative age of a person)
  + When we have lots of data
  + We should provide two analyses. One with outliers and another without outliers.
* Keep outlier
  + When there are lot of outliers (skewed data)
  + When results are critical
  + When outliers have meaning (fraud data)

Part 2 – Advance Statistics and Hypothesis Testing

**What is the difference between population parameters and sample statistics?**

* Population parameters are:
  + Mean = µ
  + Standard deviation = σ
* Sample statistics are:
  + Mean = x (bar)
  + Standard deviation = s

**Why we need sample statistics?**

Population parameters are usually unknown hence we need sample statistics.

**How to find the mean length of all fishes in the sea?**

Define the confidence level (most common is 95%)

Take a sample of fishes from the sea (to get better results the number of fishes > 30)

Calculate the mean length and standard deviation of the lengths

Calculate t-statistics

Get the confidence interval in which the mean length of all the fishes should be.

**What are the effects of the width of confidence interval?**

* Confidence interval is used for decision making
* As the confidence level increases the width of the confidence interval also increases
* As the width of the confidence interval increases, we tend to get useless information also.
  + Useless information – wide CI
  + High risk – narrow CI

**What is the difference between 95% confidence level and 99% confidence level?**

The confidence interval increases as me move from 95% confidence level to 99% confidence level

**What do you mean by degree of freedom?**

DF is defined as the number of options we have

DF is used with t-distribution and not with Z-distribution

For a series, DF = n-1 (where n is the number of observations in the series)

**What do you think if DF is more than 30?**

As DF increases the t-distribution reaches closer to the normal distribution

At low DF, we have fat tails

If DF > 30, then t-distribution is as good as normal distribution

**When to use t distribution and when to use z distribution?**

* The following conditions must be satisfied to use Z-distribution
  + Do we know the population standard deviation?
  + Is the sample size > 30?
  + CI = x (bar) – Z\*σ/√n to x (bar) + Z\*σ/√n
* Else we should use t-distribution
  + CI = x (bar) – t\*s/√n to x (bar) + t\*s/√n

**What is p-value in hypothesis testing?**

* If the p-value is more than then critical value, then we fail to reject the H0
  + If p-value = 0.015 (critical value = 0.05) – strong evidence
  + If p-value = 0.055 (critical value = 0.05) – weak evidence
* If the p-value is less than the critical value, then we reject the H0
  + If p-value = 0.055 (critical value = 0.05) – weak evidence
  + If p-value = 0.005 (critical value = 0.05) – strong evidence

**How to calculate p-value using manual method?**

Find H0 and H1

Find n, x(bar) and s

Find DF for t-distribution

Find the type of distribution – t or z distribution

Find t or z value (using the look-up table)

Compute the p-value to critical value